Development Effectiveness

Impact-Evaluation Guidelines
Technical Notes
No. IDB-TN-333
November 2011

Evaluating the Impact of Science, Technology and Innovation Programs: a Methodological Toolkit

Gustavo Crespi
Alessandro Maffioli
Pierre Mohnen
Gonzalo Vázquez

This page has intentionally left in blank

Evaluating the Impact of Science, Technology and Innovation Programs: a Methodological Toolkit

Inter-American Development Bank

2011

http://www.iadb.org

The Inter-American Development Bank Technical Notes encompass a wide range of best practices, project evaluations, lessons learned, case studies, methodological notes, and other documents of a technical nature. The information and opinions presented in these publications are entirely those of the author(s), and no endorsement by the Inter-American Development Bank, its Board of Executive Directors, or the countries they represent is expressed or implied.

Gustavo Crespi. IDB-Competitiveness and Innovation Division (CTI). gcrespi@iadb.org

Alessandro Maffioli. IDB-Strategy Development Division. alessandrom@iadb.org

Evaluating the Impact of Science, Technology and Innovation Programs: a Methodological Toolkit

Abstract

Gustavo Crespi,[1] Alessandro Maffioli,[2] Pierre Mohnen,[3] Gonzalo Vázquez[4]

The purpose of this guideline is to provide ideas and technical advice on how to measure the effectiveness of Science, Technology and Innovation Programs (STIP). The paper addresses the specific challenges of evaluating STIP, from the assessment of the intervention logic to the choice of the most appropriate method to solve the attribution problem. Much attention is devoted to the topic of data, discussing pros and cons of different data sources, data quality issues, and strategies for data collection. The paper analyzes in detail the potential application of experimental and quasi-experimental methods to STIP. For each method, the paper highlights characteristics and assumptions, practical issues related to the implementation, and strengths and weakness specifically related to the application to STIP. Other specific issues related to the evaluation of STIP are also considered: the timing of effects, intensity of treatment, multiple treatments, impact heterogeneity, externalities, and general equilibrium effects. Concrete examples of rigorous evaluations of STIP support the discussion of the various topics throughout the guideline.

JEL Classification: H43, O12, O14, O22, O31, O32, O38

Keywords: Impact Evaluation, Science, Technology Innovation, Development Effectiveness, COLCIENCIAS

[1] Gustavo Crespi, IDB-Competitiveness and Innovation Division (CTI), gcrespi@iadb.org;
[2] Alessandro Maffioli, IDB-Strategy Development Division, alessandrom@iadb.org;
[3] Pierre Mohnen, UN-MERIT, p.mohnen@maastrichtuniversity.nl
[4] Gonzalo Vázquez, IDB-Strategy Development Division, gonzalovaz@iadb.org

Table of Contents

1. Introduction

Science, Technology and Innovation Programs (STIP) have become a key ingredient in the mix of Productive Development Policies (PDP) of most emerging economies. Although the general need for such interventions is rarely questioned, scholar and policy makers often debate the proper approaches and instruments for effectively pushing the technology frontier of an economy. For this reason, the need to produce rigorous evaluations of STIP has increasingly gained relevance for governments, multilateral organizations, and civil societies.

The set of guidelines contained in this paper aims at providing ideas and technical advice on how to measure the effectiveness of STIP. This objective implies a twofold limit to the scope of this publication. First, we cover only one key aspect of the evaluation process, i.e., the attribution of effects. This means that all the methods and techniques proposed here deal with the fundamental problem of identifying a causal relationship between policy intervention and observed changes in the target population. Other important elements of a comprehensive evaluation process–such as efficiency, relevance and institutional coherence–are not covered by this version of the guidelines. Second, we focus exclusively on quantitative approaches. For this reason, we mostly concentrate on the methodological literature based on counterfactual analysis that stems from the application of experimental and quasi-experimental methods to the evaluation of public policies.

We also restrict our analysis to a homogenous set of STIP. We concentrate mainly on technology and innovation programs, while devoting less attention to science promotion programs. Although at times we make reference to science programs, we do not systematically address the specific challenges of evaluating this kind of intervention. With this choice, not only we hope to offer clearer and more focus ideas and suggestions, but also we acknowledge that the complexity and idiosyncrasies of science programs require a specific and separate discussion.

The rest of the paper is structured as follow. Section two briefly reviews the economics of STIP. There is no other reasonable starting point for an evaluator than a thorough assessment of the main characteristics of the policy to be evaluated. For this reason, this section discusses the rationale underlying the intervention logic of STIP and analyzes the main features of their implementation processes.

Section three defines the general conceptual framework and introduces the main challenges that need to be tackled when evaluating STIP. In particular, we discuss the key evaluation questions to be addressed, identify the specific outcomes and indicators to be considered, and introduce the concept of counterfactual.

Section four is entirely devoted to the data topic. Impact evaluations require the use micro-data, which can be achieved by primary data collection or by gaining access to existing datasets. The section discusses the pros and cons of different data sources, data quality issues, and strategies for data collection. Particular attention is devoted to one of the most important source of information in this field, i.e., innovation surveys.

Section five reviews the evaluation methods that could be applied to measure the impact of STIP. We start from the highest methodological standard by discussing the randomized control trial (RCT) approach, and then move on to various quasi-experimental methods. For each method, we discuss key characteristics and main assumptions, practical issues related to the implementation, and strengths and weakness specifically related to the application to STIP. In doing so we refer to concrete examples of rigorous impact evaluations in this area. We conclude this section by briefly introducing structural models as a possible alternative approach. Structural models are particularly suitable for the evaluation of policies with universal coverage, which makes the use of experimental and quasi-experimental approaches particularly challenging.

Section six explores a series of key issues to be considered when evaluating STIP, including the timing of effects, intensity of treatment and multiple treatments, impact heterogeneity across subgroups of beneficiaries, and externalities and general equilibrium effects. We conclude with possible directions for future research.

2. Economics of Science, Technology and Innovation Programs

2.1 The Rationale for Science, Technology and Innovation Programs

The fundamental premise of STIP is that government intervention would be necessary if profit driven actors underperform with regards to the production and/or exchange of technological knowledge from a social welfare perspective (Steinmueller, 2010). Under this perspective the benchmark case for STIP policy emerges from the need to correct market failures caused by the "public good" nature of knowledge. Indeed since the seminal works by Nelson (1959) and Arrow (1962), scientific and technological knowledge has been regarded as a non-excludable and non-rival good. To the extent that private benefits associated with knowledge creation are not fully appropriable by innovators this will create a wedge between the private and social returns of knowledge investments, leading to a rate of investment in knowledge that will fall short from socially optimum levels. It is important to notice that the benchmark case for STIP applies not only to the levels of innovation efforts but also to the direction of these efforts. Certainly, the "public good" rationale of knowledge applies more strongly in the case of scientific rather than technological knowledge.[5] To the extent that the latter is more applied, predictable, and linked to firm specific assets, it is more likely that innovators will collect a larger share of the value of innovation to society; as a result, private sector investments in technological knowledge will be closer to the optimum social levels.[6] In summary, the benchmark case for STIP is a lack of appropriability of returns to innovation. The difference between private and social returns to knowledge investment affects not only the level of knowledge investment but also its composition, undermining private actors' willingness to invest in scientific research in particular. However, this by no means implies that firm investment in technological knowledge will be socially optimal; appropriability problems also exists with respect to technological knowledge, as the coverage offered by intellectual property rights protection is usually limited.

The productivity slowdown in the developed world during the 70s (Griliches, 1979) together with advances in the literature of the economics of information (Stiglitz and Weiss, 1981) and innovation systems (Freeman, 1987; Lundvall, 1992, and Nelson, 1993), suggested

[5] On the other hand, projects with a significant component of basic research are unlikely to produce results with commercial application in the short run. Although this may discourage private investments, the projects could still have a high social return because of the skills and knowledge produced during their development, apart from their final achievements.

[6] The applied nature of technological knowledge also made it more likely of being protected by intellectual property rights.

additional ailments that hinder innovation, dramatically enlarging the scope for STIPs. Indeed, the economics of information literature makes clear that the problem of asymmetric information in market transactions can affect firm innovation from two different perspectives.

From the perspective of investment theory, R&D has several peculiar characteristics, which differentiate it from ordinary investment (Hall and Lerner, 2009): First, innovation projects are riskier than physical investment projects. Consequently, external investors might require a higher risk premium for the financing of innovation activities. Second, because of the problem of appropriability, innovators are themselves reluctant to share information about their projects with potential outside investors. This asymmetric information problem hampers the financing of innovation. Providing convincing signals about the quality of an innovation project is costly (Bhattacharya and Ritter, 1983) and sometimes leads to market failures (due to the problems of adverse selection and moral hazard). Third, the difficulty of using intangible assets as collateral also leads to increased costs of external capital in the form of a risk premium. In summary, asymmetric information can lead to a wedge between the opportunity cost that private innovators are willing to bear for undertaking innovation investments and the capital cost that external investors are willing to charge to finance innovation projects. The result will be that privately profitable and socially beneficial innovation projects will not materialize due to the fact that financing costs are simply too high.

The second perspective –how asymmetric information affects knowledge production and dissemination– relates to the fact that private actors do not have “perfect information” about technology or production possibilities. In the same way, adverse selection and moral hazard problems also extend to the (imperfect) operation of technology markets. This claim is consistent with two empirical findings: (i) that there exist persistent differences in the technological performance between countries and so that catching-up is very far from being an automatic process consistent with the idea of knowledge as a global public good (Fagerberg and Verspagen, 2002), and (ii) that the process of technological diffusion, even within narrowly defined industries, is very sluggish leading to a persistent firm heterogeneity in productive performance (Nelson, 1981).

Finally, the innovation systems literature makes three important contributions to the field of STIP design. The first contribution is that innovation is the result of multiple inputs and far more than R&D. Certainly, successfully placing a new product in the market requires a full set of complementary investments such as acquisition of external knowledge, purchase of specialized machinery and equipment, and training and commercialization, among others (Smith, 2006). Market failures in the markets where each one of these investments are being decided could also undermine innovation.

The second contribution of the innovation systems literature to STIP design is that successful innovation is about far more than having access to financial resources for investments; it also depends on the quantity, quality, and composition of complementary assets such as human capital, technological infrastructure and knowledge. The market conditions under which complementary assets operate are also key. For example, despite the demand pull generated by the diffusion of the knowledge economy paradigm, the institutions and markets responsible for the supply of human capital normally show a remarkably slow response pace. As a result, producing necessary human capital requires far more than augmenting the funding of scholarships targeting Science, Technology, Engineering and Math (STEM) professions and technicians; targeted educational funding should be complemented by the development of incentives to make education and vocational training institutions more demand responsive towards STEM areas.

A third contribution of the innovation systems literature is recognizing that knowledge has a significant tacit component, and as such innovation is the result of feedback and interaction involving numerous actors. Although many of these interactions are market mediated, a large proportion of them are governed by non-market institutions. Because the efficiency of this process observed at the macro level depends on the behavior of individual actors and the institutions that govern their interaction, coordination problems might arise (Soete et al. 2010).

An illustrative example of these coordination problems could emerge is the setting of standards for producer-user interactions in the case of General Purpose Technologies (GPTs). GPTs are a set of technologies that spread out across different economic activities, driving innovations in affected sectors. Progress in the adopting sectors feeds back into GPT development, generating a process of sustainable growth (Bresnahan and Trajtemberg, 1995). It is also clear that the way in which GPTs contribute to growth is not only through the development of GPT intensive sectors per se (the supply side) but also, and even more critically, through the development of the complementary innovations that facilitate their wider adoption across the other sectors of the economy, which start to innovate as a consequence of this interaction (Trajtember, 2006). This requires solving coordination problems, which in turn requires paying attention to linkages among different actors as well as their absorptive capacities (Cohen and Levinthal, 1989).

The concept of absorptive capacities is a key ingredient of the new literature of innovation, in particular from the perspective of catching-up economies. Following Steinmueller (2010), absorptive capacity refers to fact that new knowledge might not be employable without substantial investments by the users of innovations in complementary human capital and learning; furthermore, it also implies that new knowledge might not be reproducible without the direct assistance of the original innovator. The result is higher costs of imitation and replication, and a shift in policy emphasis from the appropriability problem and the importance of IPRs towards the importance of coordination and diffusion programs.

To the extent that human interactions are governed by institutions, the innovation systems literature puts a strong emphasis on institutional governance and institutional change, i.e., institutional design arrangements that foster public-private interactions and at the same time minimize problems of moral hazard. Institutional change interventions refer also to arrangements that build linkages between the different actors involved in the innovation process (such as, universities, public research organizations, technology producers, consumers, etc.) either by defining new roles for existing institutions (such as allowing the patenting of university research in order to encourage technology transfer), or by creating clubs or consortiums that regulate interactions between the agents (Steinmueller, 2010). These sorts of arrangements may lead to an improved equilibrium by eliminating redundancy in innovation efforts or through the internalization of externalities. In this case, public intervention is often required to reduce transaction costs that may hamper the formation of joint ventures and to regulate the activities of different actors in order to achieve the desired balance between cooperation and competition.[7]

[7] The regulation may allow and encourage firms to coordinate their R&D investment during the first stage of a project (e.g., the basic research stage) and then force them to engage in Cournot or Bertrand-type competition in the second stage (e.g., prototype development). On this topic, see among others Martin and Scott (2000).

2.2 Implementation of STIP

The previous analysis offers different conceptual frameworks that justify the implementation of STIP based on the idea that profit-seeking agents will produce both a level and composition of knowledge which will fall short of socially optimal outcomes. These different conceptual frameworks have created roles for a menu of different STIP designs (including the corresponding programs at the implementation level). These different policy designs have also permeated over time to the LAC region where countries have evolved in terms of policy practice and institutional development at different rhythms and frequently in accordance with prevalent economic agendas and administrations. (Sagasti, 2011)

Without losing the perspective that the within region heterogeneity is large in terms of the complexity of the national programming frameworks and budgets (Navarro and Llisterri, 2010), it would help the analysis to provide a taxonomy to classify the different interventions. The above mentioned literature provides at least two different taxonomies: a taxonomy that focuses on the dichotomy between knowledge users and producers or the supply versus demand cut, and an alternative taxonomy focuses more on the nature of knowledge by focusing on the institutions in which knowledge is produced and their different incentive regimes. This leads to the dichotomy between science policies versus technology policies.

Both taxonomical approaches have advantages and costs. In the LAC tradition, the standard approach has been to use the supply versus demand split. However, supply policies have often been confounded with science policies. This confusion is because of the assumption that Latin-American firms are consumers of knowledge rather than producers. Table 1 presents a summary of the typical STIP (policies and programs) followed in the region clustered under four broad categories: (i) science policies, (ii) technology policies, (iii) coordination policies, and (iv) institutional change policies.

These policy approaches have been gradually implemented by the different countries over time. They are not substitutes for each other but rather sub-components of a broad policy strategy, and they can complement each other depending on the policy concerns and targets of individual countries. Policy instruments within each approach can be implemented through different institutional modalities and, naturally, in accordance with each country’s institutional and policy context. The methods presented in this guideline are related to the impact evaluation of interventions in the first three policy realms (science, technology and coordination policies); institutional strengthening programs are out of the scope of the current guidelines.

Table 1. Innovation Systems Support Policies. A Taxonomy for LAC

Policy Area	**Policy Designs**	**Programs (examples)**
Science Policy		
	Ex ante subsidies for scientific research	Scientific research funds (horizontal funding)
	Thematic Funding	Scientific research funds (directed funding)
	Advanced human capital supply	Scholarships for STE graduate and PG studies.
	Ex post subsidies for scientific research	Wage incentives based on research outputs
	Support of Scientific Infrastructure	Funding for complex and robust equipment
Technology Policy		
	Horizontal subsidies for pre-competitive innovation	Technology Development Funds (matching grants)
	Horizontal fiscal Incentives for pre-competitive innovation	Tax credits or deductions
	Signaling strategies	Contests, public procurement.
	Technology Diffusion	Adoption subsidies, information diffusion, and certification.
	Specialized human capital	Training programs
	Financial Measures	Venture Capital and Seed Capital programs, Technology Guarantee programs and conditional loans
Coordination Policy		
	Linkages Strategies	Cluster Programs, University-Industry collaborative research programs, incubators, Public Private Partnerships, etc
	Quasi-public goods designs	Sectorial Funds, Technology Consortiums
	International Technology Transfer	Attraction of Technology Based FDI, labor mobility programs
Institutional Strengthening Policy		
	Innovation System Governance	Innovation Councils, Public Private Partnerships, Research Councils, Innovation Agencies, Metrology and Standardization System, intellectual property organization, etc

	Complementary Institutions	Bridging institutions (e.g. University Technology Transfer Offices), Export Promotion Agency, FDI Attraction Agency, Training Institutions
	Regulation	Higher education regulation, environmental regulation, bio-ethical regulation, competition regulation, trade regulation and financial regulation, etc.

Source: Authors' elaboration based on Steinmueller (2010)

An alternative way of understanding STIP is by making use of the framework developed by IDB (2008) and used in several IDB studies to assess productive development policies in LAC. This approach identifies two dimensions that are relevant to classify STIP. The first dimension is the scope of policies. Policies could be sector focused or transversal to the extent to which they do not aim at benefiting any particular sector but rather the economy as a whole. Following this taxonomy, we will refer to transversal policies as horizontal policies while we will call interventions with a sector-specific approach as vertical policies. The second dimension relates to the type of intervention: programs could generate market interventions (to the extent to which they seek to alter the behavior of economic agents through affecting market prices through subsidies, tax incentives, tariffs, etc.); or interventions that target the generation of public goods that contribute to higher innovation in particular by private sector. These two dimensions define a 2x2 matrix that allows re-classifying the whole set of STIPs in four quadrants (an example of this, in Table 2.

Table 2. Scope versus Type in LAC STIP Policies

		Type	
		Horizontal	**Vertical**
Scope	**Public good**	Protection of Intellectual Property Education and training (including Higher Education)	Standards and metrology Technological consortia Technological extension (INTAs, EMBRAPA)
	Market	R&D subsidies (matching grants) Fiscal Credits for R&D	Sectorial funds (software law) Public procurement

Table 2 provides a non-exhaustive list of examples of STIPs classified along these four dimensions. For example, the typical case of a public good horizontal type of policy is an intervention that contributes to the business environment in which innovation related investments decisions are taken by the private actors; a textbook example of this could be the strengthening of intellectual property rights (IPR). It is obvious that a properly designed IPR system is a public good whose provision is a State responsibility that a priori it is not designed to benefit any particular economic sector. A horizontal market intervention is a policy that aims at reducing the private cost of implementing some activities that are believed to generate wide economy spillovers, such, for instance the establishment of an R&D matching-grants scheme or the implementation of an R&D tax credit program. A vertical public good type of intervention (in the top right hand side corner of the table) consists in the provision of a key input which is a public good for a particular sector but that cannot be easily deployed in other sectors of the economy. An example is the implementation of a quality control and certification program, which promotes the diffusion of green technologies in some particular sector. Finally, a vertical market intervention policy includes programs that not only alter market prices of key inputs (such as labor or capital costs) but focuses only on some particular sectors. An example is the introduction of a matching-grants or tax incentive program with a particular sector focus, such as tax cuts that promote the software industry or some particular technology (such as biotechnology or nanotechnology). This is the case of the GPT whose development and deployment could have impacts across the economy at large, particularly through the development of network externalities and the generation of a pool of highly qualified human resources (Bresnahan and Trajtember, 1995).

The relevance of this approach is that in each quadrant the public policy considerations for the assessment of these policies are different. One important consideration across the different quadrants is the extent to which the different public policies located in them might give rise to rent-seeking and regulatory capture by the private and public actors. These are precisely the types of problems that have led to the poor reputation of industrial policy (sometimes also extended to innovation policies) in the past. In general, it is expected that rent seeking and regulatory capture will be more severe in the case of market interventions (such as subsidies or tax credits) than in the context of provision of public goods. At the same time, rent-seeking and regulatory capture are likely to be more relevant in the case of vertical than in horizontal interventions. IDB (2008) is particularly relevant for this toolkit in the sense that it provides a road map that should be follow for the evaluation of STIP programs. In other words, evaluators of STIP programs should be pay particular attention first to those interventions in the bottom-right quadrants, followed by those interventions in the bottom left quadrant.

The actual implementation of these programs in the LAC region also reflects changes in the policy approach. Beginning in the late 80s and early 90s the policy focus shifted from budget transfers to higher education, public research organization and, firms towards horizontal policies where resources allocation is driven by the demand (from researchers, firms or both). The assumption was that more horizontal and neutral policies towards funding would help align researcher incentives more closely with the needs of the productive sector, while at the same time the introduction of competition would lead to higher quality and productivity. The funding structure of the sector changed dramatically: budget transfers to innovation systems actors in the areas of science policy were significantly reduced (a factor that also led to an important reduction in the funding of research infrastructure) while those allocated to technology and coordination policies were increased. Moreover, in the area of science policy, allocation mechanisms also changed. Allocation mechanisms moved from budget transfers to institutions towards the direct support of scientists based on the quality of the research being done and the ability to build linkages with the production sector and adhere to broadly defined national priorities.

These policy changes were mainly implemented through the establishment of competitive funding programs operated by either newly created institutions (such as innovation agencies) or by augmenting the scope of already existent institutions (such as national research councils). Programs normally placed emphasis on two different types of funds: (i) *Science Funds*, which mainly finance scientific research projects, targeting higher education institutions and public research organizations, even though they are also open to private sector researchers, and (ii) *Technology Funds*, which mainly finance technology development and business innovation projects, focusing on firms and fostering linkages among firms and research institutions.

Although the implementation of these two types of instruments varies significantly in terms of targeting, operational mechanisms, and administration rules, in all the cases competition among project proposals from the potential beneficiaries plays a central role in determining resource allocation. In order to somehow control for moral hazard problems, the typical financial instrument operated by these funds is the matching grant. The subsidy never covers all costs of financed projects. How this co-funding is actually implemented depends on the type of project.[8] Indeed, in the case of innovation projects the share of total cost paid by the subsidy increases when the project implies the participation of more than one beneficiary since it is expected that projects involving more than one agent might generate relatively higher spillovers. Alternatively, sometimes there is an increase in the subsidy component when a project's beneficiary is a small firm under the assumption that the intensity of market failures faced by small and medium-sized enterprises (SMEs) is higher.

The kind of expenditures that are eligible for a subsidy could vary significantly. However, the typical matching grant program normally pays for variable research costs (researchers' salaries, research inputs and the costs of outsourced R&D), although in some cases a fraction of the capital cost of laboratory and testing equipment is also included.[9] The operation of the co-funding mechanism is normally implemented through the ex-post reimbursement of the

[8] Co-funding normally varies between 30% to 80% of the project costs.

[9] Some programs also include among eligible costs those expenses related with either the acquisition of intellectual property rights such as the purchase of a license or patent/trademarks application costs.

approved expenditures that qualify for the subsidy.[10] In the case of scientific research projects, co-funding incentives operate through a different mechanism: the subsidy normally covers input and research support costs but typically include little to no support for the compensation of the principal researcher.

The typical science or technology fund in Latin America follows a standard production process: (i) the administration agency launches a call for proposals for research or innovation projects following certain guidelines with regards to eligible expenditures and the criteria for project selection,[11] (ii) together with the technical background of the research team, applicants are also asked to submit a technical proposal and oftentimes a cost-benefit analysis of the project, (iii) the proposal is then evaluated by two or more technical evaluators (normally external to the administration agency), and (iv) proposals and the evaluation forms are sent to an adjudication committee (with representatives from the academic, public and private sectors) where the final decisions are made. Normally a minimum cut-off point–representing a minimum level of quality requested to the proposals–is set, and all project proposals that do not reach the cut-off point are eliminated. The remaining proposals are ranked according to the merit score obtained in the technical evaluation and allocations are done until all the pre-selected proposals have been funded or until available resources for the competition are disbursed (normally it is the second case).

The matching-grants programs in Latin America have also followed a clear evolution over time. After beginning with a purely horizontal approach they have gradually moved towards a more targeted approach (or, according to the Table 2, they have moved from the bottom left quadrant towards the top right quadrants). There are two main rationales for this: (i) to avoid stretching limited resources too thinly and enabling the level of funding to reach the critical mass needed to have impacts, and (ii) that the policy learning achieved during the horizontal phased allowed the policy markets to learn about main market failures and other constraints that firms

[10] This condition does not make the matching-grant approach very suitable for the promotion of entrepreneurship. Given that subsidies are paid ex post against receipts, if it is the case that the (new) entrepreneur is credit constrained, this type of funding may be of little help. Some designs are trying to correct for this through the inclusion of partial advanced funding provisions for new firms, but even in this case this advanced cash needs to be covered with guarantees.

[11] Typical there are several competitions per year and some agencies also operate an "open window system" where firms could apply at any time. In this case projects are still evaluated from a technical point of view and, sometimes, from a cost-benefit point of view; and also the project, if it passes the cut-off score, still needs to go the adjudication board for approval. The main difference with the previous system is that in this case there is no call for proposals, and so firms that file their proposals earlier are more likely to be funded.

face in order to innovate, leading them to the development of more tailored STIP. Obviously, impact evaluation has been an important policy tool for this.

Another interesting development has occurred in the context of the co-existence of matching-grants with subsidized or conditional loan programs. At the beginning of the 90s several countries experimented also with these sorts of loans. Under this approach loans could be partially or even totally forgiven on the basis of three criteria: the success or the failure of the project, the nature of the beneficiary, and the level of project technological risk. However, the co-existence of the loans with similarly oriented matching-grants programs led to competition between both interventions and a very little interest among firms for conditional loans. As a result, conditional loans schemes were phased out over time and the overall system of direct transfers was simplified. More recently there has been a re-emergence of subsidized loans in some countries, now with a clearer focus on funding the adoption of innovative technologies by firms (in particular, technologies embodied in machinery and equipment). It is important to take into consideration that the rationale for this is to some extent different from what is normally used for the support of R&D expenditures. In the case of the adoption of embodied technology, the subsidized loan argument is normally based on the potential spillovers that a technology generates to the rest of the sector or the economy. In other words, subsidized loans are used to address an asymmetric information problem, and once the benefits of a new technology have been demonstrated, the subsidy should stop. Obviously, actual implementation of the scheme requires fine-tuning by the implementation agency of what should or should not be considered an innovative technology. Despite these problems, a nice feature of subsidized loans is that subsidies are small and they do not reduce the capital cost below the opportunity costs of the firm internal funding. As such, subsidized loans can be a very powerful tool for self-selection of potential innovators that face liquidity constraints rather than plain rent-seekers.

The role played by the agency in charge of managing science or technology funds is critical in the successful implementation of STIP programs. Matching-grants schemes are quite “active” policy tools (even within the domain of horizontal demand driven programs). The actual implementation of the scheme depends on the decision making institutional capacities, indeed, although firms self-select to apply for the grants, final selection is done by the implementing agency. Following Hall and Maffioli (2008), the role of this agency is at least three-fold. First, the agency acts as a screener, conveying the technical knowledge that the financial markets lack or are not willing to develop. This role should reduce asymmetry of information between the financial sector and the innovative firms. Second, it has the ability and authority to monitor firm investment. This function could diminish the risk of financing firms diverting resources to other uses. Third, the agency tries to select those projects that have the highest social returns or where the wedge between private and social returns is the highest. Under these assumptions, the agency should be selecting projects that unlikely to be financed otherwise.

The fulfillment of these roles requires the building of strong institutional capacity. For this purpose, two additional conditions need to be met: first, the overall process needs to be predictable and policy experimentation needs to be supported by well rooted monitoring and evaluation systems; second, properly skilled human resources –both internal to the agency and external in the users and examiners– must be available. Unless both conditions are met, a matching grants scheme will be unlikely to succeed (Toivanen, 2009).

The successful administration of a matching grant scheme requires strong institutional capacity on the part of the implementing agency and, to a certain extent, of the beneficiaries. When these capacities are not available the outcome might be very high administration and compliance costs. In countries that are starting policy experimentation in this area, the hurdles to apply for benefits could be particularly high, the speed at which applications can be processed may be too slow, and opportunity costs of applying almost forbidding, especially for SMEs and new entrepreneurs. The existence of these barriers to participation has led some countries to follow an alternative policy design: the implementation of R&D tax incentives.

Developed countries have increasingly adopted R&D tax incentives (OECD, 2010). Tax incentives operate through different approaches, including tax credits, enhanced allowances, and accelerated depreciation of intangible investments. Tax credits allow a direct deduction from tax owed, while enhanced allowances and accelerated depreciation represent a deduction (above the normal deduction rate of 100%) from the taxable income of the company. The main difference between the two mechanisms is that the former directly reduces the tax liability, while in the latter the reduction of the tax liability depends on effective tax rates. Tax incentives normally apply to corporate income tax, though some countries have experimented with other approaches such as reductions in tariffs for imports of research machinery and equipment, deductions in the value added tax, and discounts in the social security and employers' contributions to payroll taxes of researchers' salaries.

The implementation of R&D tax incentives requires one to consider the following issues: (i) the definition of a target group (the tax incentives can be made available to all firms or the support can be made more generous for SMEs or some specific sectors), (ii) the regulatory labeling of R&D activities (countries normally apply some variation of the OECD's Frascati Manual definitions), (iii) the qualification of those activities eligible for the tax incentive (these might be salaries of R&D personnel, R&D expenditures –salaries plus research inputs costs- and capital R&D expenditures), and (iv) a decision on whether the scheme will follow a volume base (deductions based on the total amount of previous qualified expenditures) or incremental (base of an growth of R&D expenditures in which case it is necessary to define the base amount upon which the growth will be calculated) (Van Pottelsberghe et.al, 2009).

In recent years, R&D tax incentives schemes have become both more generous and an important component of stimulus packages. The list of types of eligible expenditures has also become longer, growing to include expenditures on intangibles, such as intellectual property rights and licensing in and out– as well as adoption of certain types of technologies, in particular green technologies. Countries such as the USA and Canada have even extended R&D tax incentives to the sub-national level. The portions of tax codes relevant to R&D are increasingly being used to create incentives for innovation (EU, 2009).

From an evaluation point of view, it is worth emphasizing the main differences between matching grant subsidies and tax incentives. Matching grants are a type of direct support for business R&D which is linked to specific projects. They modify the firms' marginal cost of capital schedule and they may raise the private marginal rate of return of the innovation investment. Because they are project based, matching grants allow public agencies to target projects with perceived high marginal social rates of return (such as stimulating the formation of collaborative technological consortiums). The government has full control of the fiscal cost of the scheme as the fiscal resources for transfers to the private sector are normally included in the state budgets. The main problems with the matching-grants are that they need important institutional capacities in the executing agency and when these capacities are not present the efficiency of the whole operation dramatically decreases.

On the other hand, R&D tax incentives are based on firm-level–rather than project-level–R&D activities. This scheme allows firms to get support for the whole portfolio of R&D activities without having to submit a project proposal for each one of them, dramatically reducing firms' compliance costs as well as agency administration costs.

However, two important caveats need to be considered. First, the actual impact of R&D tax incentives on the marginal cost of capital of R&D activities depends on the general fiscal environment as fiscal incentives are less effective in a country with a low rate of corporate income tax. In fact, this is one of the main reasons on why the empirical literature has normally found that R&D tax incentives are less effective as a stimulus for R&D in the SMEs rather than in large companies (Harris, et.al 2009). SMEs simplified tax treatments –as the ones that abound in LAC– normally imply that corporate tax rates are smaller in the case of SMEs. Effects of tax incentives in the case of SMEs are likely to be different also because they strongly depends on the firm's tax position and on its ability to make profits, which is unlikely for young SMEs just entering the market. Therefore, R&D tax incentives might not be very useful as a stimulus for "self-discovery". This can be ameliorated through the inclusion in the scheme of carry-over provisions that allow unused portions of the credit to be carried forward to the next fiscal years. In some developed countries carry-forward provisions are combined with direct cash refunds, in which case the R&D tax incentives actually become R&D grants (e.g., see the cases of France and the Netherlands in Criscuolo, 2009).

Second, the fiscal costs of R&D tax incentives could be substantial and, to the extent that carry-forward provisions exist, the rejection incentives of the administrative agencies could be relatively more relaxed given that these costs will be absorbed by future administrations. In developed countries the fiscal costs of tax credits in terms of forgone revenues have systematically increase over the last twenty years with values in the range between 0.06% of the GDP in the case of the United Kingdom up to 0.29% of the GPD in the case of France (OECD, 2010). For a country that spends about 2% of the GDP in R&D, of which 60% correspond to the private sector, R&D tax incentives may represent a significant part of the overall effort.

Several LAC countries have established R&D tax incentives during the last 15 years (Argentina, Colombia, Brazil, Mexico–recently discontinued–and lately Chile and Uruguay). The typical LAC R&D tax incentive presents some important differences with the standard approach in developed countries. For instance, sometimes indirect taxes are also included among the deductions (such as the value added tax or some import taxes). The most important difference is that the implementation of the scheme in LACs is mostly project based: in order to qualify for tax incentive firms are usually asked to submit a project proposal to the innovation agency which will review whether the project qualifies as an R&D. The agency will then recommend the approval of the eligible expenditures to the tax authorities who will issue an tax credit certificate. The rationale for this approach is to have tighter control of the fiscal costs of the scheme. The major problems are that many of the administration and compliance costs of matching grants reappear, but without the benefits of the matching grants system as the investment decision remains completely under the firm control. That is, if a firm submits a project proposal that under the law qualifies as an R&D project, the agency is obliged to issue the R&D tax credit certificate, even when social returns of the project are low and not very different from the private ones.

From an evaluation point of view R&D tax incentives also pose important challenges. Firstly, for the purpose of policy assessments, firms cannot be legally excluded from a tax incentive to which they are entitled. This removes the possibility of evaluating R&D tax credits by constructing a control group using randomization techniques. Even the implementation of quasi-experimental techniques might be difficult when all the qualifying firms (firms that do R&D for example) receive the incentive. For this reason, one of the favorite approaches for the impact evaluation of R&D tax credits schemes resides in the utilization of structural modeling techniques (Hall and Van Reenen, 2001, and OECD, 2010).

3. Conceptual Framework and Main Challenges in an STIP Impact Evaluation

3.1The Rationale for the Evaluation of STIP

STIP can be justified by the presence of various market, coordination, and systemic failures. An effective response to these failures requires government ability to both design and deliver proper policy solutions. In reality, governments face informational constraints that may be at least as severe as those of firms. Firms, R&D projects, and innovations are highly heterogeneous, and a high degree of heterogeneity means that a policy that is optimal in the sense of Pareto efficiency for one firm or project may not be so for another. This need to tailor policies to circumstances puts administrative agencies under severe informational stress (Toivanen, 2009). In summary, although there might be a strong case to be made in favor of STIP, actual implementation could easily have unintended consequences, e.g., public support could lead to both crowding in or crowding out of private funding.

The study of the relationship between the different sources of funds is important in order to identify complementary or substitution effects. That is, it is not possible to guarantee that a given increase in government funds to R&D is going to be linearly transmitted to the firm's innovation budget: the presence of a crowding in (or crowding out) phenomenon in relation to private sources must be taken into account. In certain situations a positive correlation between private and public resources for R&D might be expected; for instance, public resources can be used to finance fixed capital costs (such as the building of laboratories), allowing the firm to price research services at variable cost. Public resources can be used to fund the riskier component of the research project (i.e., the basic research) and then private funds can be used to complete the development phase of the project. However, the relationship might also lead to a substitution effect. This would occur when the type of project being funded by the innovation agencies was very similar to the types of projects funded by firms, or, more indirectly, when the distortions introduced by the tax system to collect the additional resources for funding public R&D reduce the private returns to investment in general, and in particular to R&D.

Firms normally have a portfolio composed of several innovation projects. From a firm's point of view, receiving a subsidy under a matching-grants system may turn an unprofitable project into a profitable one. Alternatively, a matching grant subsidy may speed up the completion of a project already underway. If subsidies involve setting up or upgrading research facilities, this will reduce the fixed costs of other current and future research projects, increasing their probability of being completed or undertaken as a result. Also, the learning and know-how gained from the project being supported can spill over to other current and future research projects, thereby enhancing their prospects for success. In all of these ways, a matching grant subsidy may stimulate current and future research projects. If these hypotheses are true, we can expect a crowding in effect of public financing on private innovation investment.

However, it is also common sense to expect that some research projects can be carried out without government funding. There are several external sources of funds for research proposals including public, foreign or multilateral institutions (such as the IDB, WB, etc.), the financial sector and charities. The possibility of substitution can be increased by administrators who are often under pressure to avoid the appearance of 'wasting' public funds and who may tend to fund projects with a higher success probability and with clearly identifiable results (projects that are likely to have a range of alternative sources of funds). These are projects that could have been financed by other sources of funds, suggesting that the public funds can in fact be superfluous.

Another reason for substitution is that a project enhanced by a subsidy could have an effect on the price of inelastically supplied research time (no one can work more than 24 hours a day). If the subsidy turns an unattractive project into an attractive one, but there are human capital constraints, the researcher or the team may decide to discontinue what previously was an attractive project (which might have been funded by other sources of funds). The commitment to undertake the subsidised project may crowd out other non-subsidised projects (and their accompanying resources). Thus, we cannot simply extrapolate that a given increase in public funds to the firm innovation budget will lead immediately to a proportional increase in project execution. We need to investigate the effect on other potential sources of funds.

We can position this analysis clearly in term of an accounting framework (for details, see Lach, 2000). Let us assume that the size of R&D projects is fixed. The only decision the firm makes is whether or not to undertake the project. Each firm has n potential projects, each of size α_i, i=1,...,n. Then the total R&D budget of a firm is given by:

$$R^0 = \sum_{i=1}^{n} \alpha_i \chi_i^o \quad (1)$$

where χ_i^0 is a binary variable indicating whether the project i is undertaken or not if a subsidy is not received. Assume that the firm applies for a subsidy only for the n_{th} project. Then the R&D budget with innovation agency support is:

$$R^1 = \sum_{i=1}^{n} \alpha_i \chi_i^1 \ (2)$$

Note that receiving the subsidy can change the decision to undertake any of the first (n-1) projects and that the subsidised project (project n) is always implemented, so $\chi_n^1 = 1$, because of the contractual agreement. Then the increase in the firm R&D budget as a consequence of receiving a public subsidy is:

$$\Delta = R^1 - R^0 = \sum_{i=1}^{n-1} \alpha_i (\chi_i^1 - \chi_i^0) + \alpha_n (1 - \chi_n^0) \quad (3)$$

In order to evaluate the potential impact of the public subsidy, suppose first that the subsidy does not change the decision about the unsupported projects, $\chi_i^1 - \chi_i^0 = 0$ for i=1,....,n-1. Then,

$$\Delta = \begin{cases} \alpha_n & if \quad \chi_n^0 = 0 \\ 0 & if \quad \chi_n^0 = 1 \end{cases} \quad (4)$$

Clearly Δ is positive only when the subsidy causes the subsidised research project to be implemented and Δ is zero if the subsidised project would have been undertaken even in the absence of the subsidy. When the decision to implement the other non-subsidised projects can change as a result of receiving the subsidy, the consequences are not so clear. Without loss of generality, let us assume that only the decision about the $(n-1)_{th}$ project can change. We then have several possibilities:

$$\Delta = \begin{cases} (1)\ \alpha_{n-1} + \alpha_n \quad if \quad \chi_n^0 = 0 \quad \& \quad (\chi_{n-1}^1 - \chi_{n-1}^0) = 1 \\ (2)\ -\alpha_{n-1} + \alpha_n \quad if \quad \chi_n^0 = 0 \quad \& \quad (\chi_{n-1}^1 - \chi_{n-1}^0) = -1 \\ (3)\ \alpha_{n-1} \quad if \quad \chi_n^0 = 1 \quad \& \quad (\chi_{n-1}^1 - \chi_{n-1}^0) = 1 \\ (4)\ -\alpha_{n-1} \quad if \quad \chi_n^0 = 1 \quad \& \quad (\chi_{n-1}^1 - \chi_{n-1}^0) = -1 \end{cases} \qquad (5)$$

The gain in terms of the global research budget is positive when both projects are implemented as a result of receiving the subsidy, as in case (1) of equation (5). This is the best result: the subsidy makes attractive not only the subsidised project, but also a non-subsidised one and we have clear case of crowding in. This may happen when the subsidised project involves the setting up or upgrading of research facilities, thereby lowering the costs of other current (and non-subsidised) research projects. There may also be spillovers of learning and know-how gained from the subsidised project to other current (and future) research projects, thus increasing their attractiveness. On the other hand, the opposite effect may occur when severe time constraints reduce the attractiveness of the (n-1)$_{th}$ project. The firm in this case may find it more attractive to discontinue the non-subsidised project (case 2). Then the firm's total R&D budget may decrease or increase as a result of the subsidy depending on the relative size of each R&D project.

Cases (3) and (4) involve situations where the subsidised project would have been undertaken even without subsidy ($\chi_n^0 = 1$). In this respect, the subsidy is superfluous and does not contribute to the research budget at all. If, however, the funds released by the subsidy are used to implement an additional project which could not have otherwise been undertaken, then the subsidy effect becomes positive (case 3). The last case is when the (n-1)$_{th}$ project is closed down ($\chi_{n-1}^1 = 0$) as result of receiving the subsidy (case 4) and in this instance the firm's budget will unambiguously decrease.

The main conclusion of this analysis is that the impact of public subsidies on the other sources of funds is not an obvious one and that in order to determine the impact we have to look at more than the total R&D budget of the firm. This analysis is important for impact evaluation because it implies that the right unit of analysis for program evaluation is the beneficiary firm rather than the individual R&D project.

3.2 Outcomes of Interest

One of the first issues to be defined in an Impact Evaluation is when to measure the effects of the program, i.e., the outcomes of interest. In the spirit of the CDM model (Crépon, Duguet and Mairesse, 1998), a distinction can be made between innovation-input (short-term) indicators, innovation-output (medium-term) indicators and economic-performance (long-term) indicators. Innovation-input indicators are the indicators more directly affected by the intervention. For instance, for an R&D subsidy program for small firms, an innovation-input indicator is total expenditures in R&D. While the relationship between the subsidy and total expenditures in R&D seems almost tautological, this is not necessarily true. In fact, a large literature studies whether this type of subsidies generates a crowding out effect where public funding displaces private spending, leaving the level of total expenditures in R&D unchanged (see e.g. David, Hall and Toole, 1999; Duguet, 2004, González and Pazó, 2008). In other words, R&D and other innovation investments that are affected by the changes in the firm's marginal cost of capital of investments allow us to assess the extent to which STIP generate input *additionality*.

However, just assessing whether innovation efforts increase as a consequence of a subsidy is not enough for policy evaluation purposes. As is clear from our previous discussion, the whole portfolio of R&D projects held by the firm is usually affected. As a result, projects with different levels of productivity might be executed while others might be postponed or canceled. Therefore, assessing the outputs of innovation investments is also important. This would allow us assessing *output additionality*. Innovation outputs are variables where the concrete realization of innovation activities is observed. The oldest recorded measure of innovation is the patent, the earliest appearance of which dates back to 1474 in Venice (van Pottelsberghe and Guellec 2007). To the extent that the idea that is patented is sufficiently new (the invention step), a patent can be accorded to the applicant by a national or a supranational patent office such as the European Patent Office (EPO). Patent applications can in certain cases be a sufficiently useful measure of innovation output, although patents granted are a better measure of invention. It is known that the distribution of patent values is highly skewed: only a few patents are worth a lot (Giuri and Myriam, 2007) Therefore a more revealing measure is the citations-based patent count, where patents that receive subsequently many forward citations in other patent applications are given more weight. Patents are easily accessed electronically at a modest cost, but the linking of patent and other firm data is a time-consuming job. In some parts of the world such linked data have already been set up. The drawbacks of patents as a success indicator are: (i) patents can be applied for reasons other than the appropriation of R&D efforts, such as cross-patenting, signaling, and financing; (ii) in many sectors few firms patent and rather favor other means of appropriation over patents (like secrecy, complexity or being the first on the market); and (iii) patents, again partly for strategic reasons, may remain on the shelves, or not lead to successful products in the market.

For scientific research funding programs, it is common to use bibliometric data such as the number of publications in scientific and/or in technical journals. Publications can be weighted by the number of citations in subsequent patent filings or publications, or by the rankings of the scientific journals. There are also well known programs on this regards. There may be a publication bias in some academic journals in the sense that some topics are more fashionable than others, some articles have a lower diffusion because of the language in which they are written, and it may take some time before a scientific article gets published. Ultimately one may consider that what really counts is not so much the effort in innovation or the actual output of that innovative effort, but its effect on economic performance. For example, in the case of business innovation programs, important output variables which measure the impacts of STI programs on output additionality are productivity growth, employment, wages, and exports to just cite a few.

The above mentioned indicators refer mainly to direct impacts of public support of innovation. There may also be positive or negative indirect effects. For example, for small firms being awarded with a matching grant might have a certification effect (as found by, e.g. Lerner 1999; Blanes and Busom, 2004). This lowers firms' cost of capital at the margin, when applying for external sources of financing. Grants act as a signal of "good quality" for firms and projects and reduce information asymmetry problems relaxing so the financial constraints. Other positive direct effects might be the attraction of R&D based FDI to a region or country, the inducement for a firm to start doing R&D, the building of linkages with between firms and universities to conduct joint R&D projects. However, very generous R&D subsidy schemes might also have negative indirect effects such as an increase in researcher salaries so that the real amount of knowledge that the R&D subsidy actually buys ends up being lower. Finally, key indirect effects are the knowledge and rent spillovers that beneficiaries could cause on the other actors of the innovation system. Indeed, treated firms could generate (positive) knowledge spillovers on non-participant firms (a fact that creates serious identification problems for impact evaluation) and (negative) rent spillovers in their competitors (the business stealing effect). They could also generate spillovers in other firms (either clients or suppliers) that are users of the technology when the price charge to the innovation does not cover the changes in quality of the novel inputs or outputs. When the final user is the consumer and when the innovation does not discriminate perfectly in monopolistic way, an innovation will lead to an increase in the consumer surplus. A complete impact evaluation should also require considering all these other indirect effects in order to generate enough information for a full social cost-benefit analysis of a public policy. The complexities underlying this task are enormous and so this is rarely done.[12] We will get back to the spillovers issues later on in this guide.

Table 3. Indicators for STIP Impact Evaluation

Direct Effects (on the beneficiary)	**Indicator**
Input Additionality	Total R&D expenditure as % of sales Total Innovation investment as % of sales Private R&D expenditure as % of sales Private Innovation expenditures as % of sales Start investing in R&D (0/1) Other private external funding % of R&D projects in collaboration with universities
Output Additionality	Patents applications and citations Publications and citations Innovative Sales as % of Sales Employment Revenues Labor Productivity Total Factor Productivity Fixed Capital Investment
Indirect Effects (on other actors)	
Input Additionality	
	Salaries of Researchers R&D of other firms

[12] An exception is Lokshin and Mohnen (2010), and Parson and Phillips (2007).

	Innovation expenditures by other firms
	% of other firms that start doing R&D
Output Additionality	Employment of other firms
	Revenues of other firms
	Labor or TFP of other firms
	Changes in Consumer Surplus
	Papers and Publications by universities collaborating with beneficiaries

3.3 Counterfactuals and the Fundamental Problem of Causal Inference

Even after carefully considering and selecting the relevant outcomes and indicators, evaluating the impacts of public programs is not a trivial task, especially when the goal of an evaluation is measuring a causal impact of program participation on outcomes of interest. In impact evaluation, the definition of causality is based on the concept of counterfactual. For instance, suppose a firm receives a subsidy for R&D investment, and suppose we observe the value of a given outcome of interest Y for that firm. The public subsidy is said to have a causal effect on Y if the outcome of the firm in the absence of subsidy, holding everything else equal, would have been different. In other words, the program or "treatment"[13] has a causal effect if the observed outcome when the firm receives a subsidy is different from the counterfactual outcome, i.e., the outcome that would have been observed if the firm had not received the subsidy. While this definition of causality is relatively simple and intuitive, it introduces a serious empirical complication: by definition, the counterfactual outcome is never observed. In other words, if a firm receives a subsidy, it is impossible to know with certainty how this firm would have done without it, or vice versa. The impossibility of observing a given unit both with and without receipt of treatment at any particular moment in time was labeled the *Fundamental Problem of Causal Inference* by the statistician Paul Holland (Holland, 1986).

These ideas can be formalized using the Rubin Causal Model (Holland, 1986), as follows.[14] Let Y_1 and Y_0 denote the potential outcomes for a unit in the presence and absence of treatment, respectively. The *observed outcome Y* for an individual will be Y_1 if the individual is treated and Y_0 otherwise. We will use the binary variable T to indicate the treatment status of the

[13] The words program, treatment and intervention will be used as synonyms in this document.

[14] The following paragraphs are taken from Heinrich, Maffioli and Vazquez (2010).

observed units, namely, T=1 for those who participate and T=0 for those who do not participate. Then we can write the observed outcome as:

$$\boldsymbol{Y = Y_0.(1 - T) + Y_1.T}$$

In this context, Y_0 is the counterfactual outcome for treated units, and Y_1 for untreated units. As explained in the previous paragraph, the impact of the program for unit *i* is defined as the difference between the two potential outcomes:

$$\boldsymbol{\delta_i = Y_{1i} - Y_{0i}}$$

which cannot be observed. In general, impact evaluations focus on estimating the average effect of treatment as opposed to individual treatment effects. In practice, several different average effects can be estimated. The first of these is the Average Treatment Effect (ATE), or the average impact of receiving treatment for the whole population:

$$\boldsymbol{ATE = E(\delta) = E(Y_1 - Y_0)}$$

The Average Treatment Effect on the Treated (ATT) is the average impact of receiving treatment for the treated population:

$$\boldsymbol{ATT = E(\delta|T = 1) = E(Y_1 - Y_0|T = 1)}$$

The Average Treatment Effect on the Untreated (ATU) is the impact that the program would have had on the population that did not participate in the program:

$$\boldsymbol{ATU = E(\delta|T = 0) = E(Y_1 - Y_0|T = 0)}$$

However, none of these parameters are observable. For instance, the ATT can be rewritten as:

$$\boldsymbol{ATT = E(Y_1|T = 1) - E(Y_0|T = 1)}$$

where the second term is not observable, since it measures the average outcome that the treated population would have obtained in absence of treatment. A feasible possibility is to exchange the second term by $\boldsymbol{E(Y_0|T = 0)}$, which is the average observed outcome for the untreated population. Then:

$$\boldsymbol{\Delta = E(Y_1|T = 1) - E(Y_0|T = 0)}$$
$$\boldsymbol{\Delta = E(Y_1|T = 1) - E(Y_0|T = 1) + E(Y_0|T = 1) - E(Y_0|T = 0)}$$
$$\boldsymbol{\Delta = ATT + SB}$$

where the last term is usually known as selection bias. This term captures the difference between the counterfactual for treated individuals and the observed outcome for the untreated individuals. Therefore, unless the selection bias is zero (which, as explained later, is unlikely in practice), econometric techniques need to be applied to correctly estimate the average impact of the program. The different methodologies will be described in Section 5.

3.4 Relevant Questions for an Impact Evaluation

The model presented in the previous subsection, which should always be the starting point for the design of an impact evaluation, may give a simplistic picture of the challenges an evaluator has to face and the amount of information that can be obtained. While parameters such as the ATT and ATE are of the highest importance for assessing the effectiveness of an intervention, a well-designed evaluation may provide additional information that is crucial for analyzing the effectiveness of a program and deriving precise conclusions that can contribute to the design of successful policies. In this subsection we briefly sketch some aspects that should always be considered in the design of an impact evaluation. Methodological aspects related to these issues will be discussed in Section 6.

3.4.1 Timing of the Effects

In general, it takes time for the effects of certain programs on innovation efforts, the innovation output of those efforts, or their effects on economic performance to be felt. The process of setting up a research program, finding the right people, financing the project, organizing the research and networking will likely generate adjustment lags in research projects.

The materialization of concrete outcomes requires a period of gestation after investment in research; it takes time to apply for a patent and get it approved, publish an article in a scientific journal, or launch a new product. These time lags may differ according to the chosen indicator of innovation output. For instance, it may take more time before the innovation output turns into higher profits or productivity. At the beginning of a new technology, there might be a very steep learning curve, or, in presence of network externalities, it is only when a new product is shared by many consumers that it becomes profitable. More generally, the impact of different programs may display very different patterns over time. An intervention may generate a one-shot increase in the outcome, may have strong impacts that fade out progressively with time; the impact of a program may only appear after a certain period, or may even generate an initial drop in the outcome that is later overshot by increases in subsequent years.

As a result, a proper consideration of the timing of program effects is crucial in an impact evaluation, and failing to account for these issues may lead to misleading conclusions and policy recommendations. A clear distinction should be made between short-run, medium-run, and long-run effects in order to properly evaluate the costs and benefits of a public program. For instance, considering only a short period of time after an intervention may result in underestimating impacts if program effects take several years to appear. On the other hand, evaluations focusing only on later periods may end up underestimating the costs if an adjustment process occurs in the first years.

3.42 Intensity of Treatment and Dosage Effects

While the Impact Evaluation literature usually analyzes the binary case of participation versus nonparticipation in a certain program, in practice it is generally the case that units may differ not only in their binary treatment status (participant versus non participant) but also in treatment intensity. For instance, firms may receive different amounts of public subsidy, and different research teams may be granted different levels of funding. This fact raises important issues to consider when designing an evaluation: the question of interest is not only whether participants perform better than nonparticipants, but also how different intensities of treatment may affect performance and whether it is possible to find an "optimal level" for the intervention (e.g. the amount of financing that maximizes the effect on firm performance).

3.4.3 Multiple Treatments

In contexts where multiple treatments are available, the evaluator may be interested not only in the individual effects of each treatment, but also in the effects generated by their interactions. In fact, it is not obvious that the effect of multiple programs will be additive; instead, it may be the case that combining different interventions has multiplicative effects, or that one treatment cancels out the effect of another. Therefore, the investigation of the joint effect of different types of interventions may be crucial for the design of effective programs.

3.4.4 Heterogeneity of Impact

In most relevant contexts, it may be hard to accept that a given intervention will have a constant effect, i.e., the same impact on all units under study. Two main types of impact heterogeneity may arise. One occurs when interventions have differential effects for different groups; for instance, research funding may have a higher impact for young researchers, and the effect of public credit may be stronger for firms that would otherwise face liquidity constraints. The second type of heterogeneity is related to the distribution of program effects; for instance, two programs may have the same average impact, but one of them may concentrate the effects on the lower part of the distribution (Frölich and Melly, 2009). Heckman, Smith and Clements (1997) list other parameters that may be of interest for an evaluator:

- The proportion of people taking the program who benefit from it
- The proportion of the total population that benefits from the program
- Selected quantiles of the impact distribution
- The distribution of gains at selected base state values

In these contexts, restricting the analysis to the average impact for the whole population (or treated population) may give an incomplete or at least imprecise assessment of the effect of a program. It is therefore of great interest to account for the possibility of impact heterogeneity to give a precise assessment of the effects of an intervention.

4. Data

Another factor that may condition the choice of how to evaluate the success of a policy intervention is the kind of data available. Here it is useful to review the different characteristics of the data.

4.1 Level of Aggregation

Data may be available at different levels of aggregation. Aggregate or macro data is usually measured at a wide level like geographical region or industry while micro data can be measured at the individual (firm, researcher, etc.) or project level. Macro data has the advantage of being comprehensive whereas micro data may be representative of sub-populations (i.e., cover the various industries, size classes, regions, etc.) but only cover part of the whole population of firms. As we shall see below, externalities are often present when it comes to STI effects and ought to be included in an evaluation. Aggregate data include the country/region internal externalities (i.e. those between firms of the same country/region) without a need to measure them explicitly. Externalities can also be captured with micro data but at the cost of a more complex evaluation design, depending on the choice of proximities or channels of spillover transmissions.

Furthermore, the finer the level of detail at which data are available, the larger will be the size of the dataset, and the more the effect of an intervention can be monitored at a micro level and be shown to possibly depend on characteristics of firms in the industry. On the other hand, at the micro level some variables are difficult to measure (e.g. input prices), whereas at a macro level price indices are constructed from national income accounts.

4.2 Data Sources

The second characteristic is the source of the data. Statistical offices are in charge of collecting data to describe and monitor the performance of an economy. Therefore, they conduct censuses and surveys on various aspects of the economic activity. Some interesting S&T statistics may come from particular government agencies (patents from patent offices, tax credits from the internal revenue service, R&D subsidies from ministries in charge of S&T, grant applications from granting agencies like the National Science Foundation, researchers' curriculum vitaes from universities or other research institutions, accounting data from publicly traded firms). We shall call the latter administrative data. The main problem with administrative data is that it is often confidential and hard to get. Although statistical offices are less in a position to argue that their data are not to be distributed, they are sometimes obliged not to share certain data, especially micro data, for reasons of confidentiality.

A related characteristic just alluded to is the public versus confidential nature of the data. Data concerning sectors or the whole economy are in principle publicly available, sometimes at a cost-recovering price for the statistical offices, whereas confidential data are only available to selected persons, often those working within the statistical office or then researchers sworn under oath not to disclose the information, and even then only for parts of the data (e.g. without knowing the identity or the address of the firm or with the constraint that the information can be physically accessed only within restricted centers connected to the statistical office). Confidential information can sometimes be made available to the public by adding noise to the data in such a way that the individual firm cannot be identified (e.g. the techniques of micro-aggregation, or replacing individual data with moving averages from a small number of similar firms).

Finally, the data may be extracted from a census of the frame population or a survey. Census data are supposed to cover all the firms of a frame population, e.g., firms with at least five employees that appear in the national registry of firms. Survey data are typically based on a stratified sample where the strata correspond to such things as regions, industries, size classes to make sure that all firms of these subdivisions are well represented in the finite subsample of the full population. R&D data are collected for all firms in a purposeful sample, i.e., firms that are believed of conducting R&D are contacted. Innovation survey data come from a stratified sample, where the coverage may differ according to the strata. For example, all firms with more than 250 employees are included in the sample, whereas for strata corresponding to smaller sizes only one half, one quarter or one tenth of the firms in the strata is contacted. The goal of this strategy is to include all the important firms and only a sample of the minor players. It is possible to blow up the sample to the level of the whole population by using the sampling weights. Some surveys are mandatory, and others are voluntary. This complicates the role of the analyst but can be controlled for if some information is available on the non-responding firms.

Innovation surveys (IS) can be key inputs for evaluating the impacts of STIP. Following the conceptual framework set by the Oslo Manual (or its Latin American adaptation: the Bogotá

Manual), all innovation surveys implemented thus far in Latin America have adopted a “subject” approach where the unit of analysis is the firm and its innovation behaviour as opposite to an “objective” approach where the unit of the analysis is an innovation output. In theory, this common conceptual framework could allow for harmonized basic definitions of key variables such as innovation outputs, R&D, impacts, linkages and obstacles. Indeed, a great advantage of the IS rests on the possibility of providing an integrated measurement of the innovation performance at the firm level rending information on inputs, outputs and also interactions and collaborations. Although implementation across the region is far of being homogeneous and sometimes large differences among the questionnaires and methodologies are observed, IS have become a key baseline against which to monitor the performance of different innovation systems at the national level (Crespi and Peirano, 2007). Furthermore, the relevance of IS for STI policy evaluation critically depends on the set of question modules or blocks that are included in the questionnaire. Although the final configuration of the questionnaires might vary across the different waves of the survey, in order to be useful for STIP impact evaluation at least the following modules should be present:

(i) *Module on Innovation Activities:* This module should include a set of questions on innovation investment decisions taken by the firm and the amount of resources spent in the different categories of investments. These categories normally are: R&D (both in-house and extramural), acquisition of external knowledge (such as licenses to use intellectual property or specialized services), acquisition of machinery and equipment (including computer hardware) in connection with innovations, investments in design, training and marketing. The consideration of all these categories is important to the extent that depending on the program to be evaluated either some or all of them might be considered as eligible expenditure for a subsidy.

(ii) *Human Resources:* Innovation surveys should also include a module on skills of the workforce at the firm level. This is important, in particular for developing countries where innovation activities are not fully formally in R&D labs or departments and they usually take place across many different departments of the firms. The traditional approach followed by IS in the region has been to ask only for the skills of the personnel working on R&D activities. This practice, however, should be discouraged, as it is the

complementary of skills across the whole plant, which ultimately affects the success of innovation.

(iii)Innovation Outputs: This module is important in order to assess the effectiveness of innovation inputs. It should cover at least the introduction of new products, new processes, new organizational methods and marketing systems. One particularly useful question in this context is to ask each firm for the fraction of sales that come from the selling of new products. Some IS in the region introduce a filter in the innovation outputs questions so that only those firms that have obtained a positive outcome are asked to answer for the innovation investments. This practice, however, should be discouraged as it not necessarily the truth that all firm-level investments in innovation result in successful outputs; many failures are to be expected.

(iv) Sources of Ideas for Innovation: An important module that should be present in all the IS because of its importance for understanding the transmission mechanism that led to a successful innovation is to ask for the sources of ideas for innovation. The standard practice has been to ask for internal (to the firm) sources such as R&D labs, departments, and external sources. These sources can be grouped as vertical external sources (clients and suppliers), horizontal (competitors), institutional (universities and public research organizations), and public (fairs, patents, etc). Asking for sources allows one to identify the main channels through which spillover effects might be in operation (see Crespi, et. al. (2008).

(v) Innovation Cooperation: Practitioners should also consider the inclusion of a module on cooperation in the standard IS questionnaire. This module normally asks for innovation activities done in collaboration with other innovation system actors such as other firms or universities. The inclusion of this module is also important for policy evaluation as many of STIP to be evaluated aim at either building or improving the linkages among firms or between firms and universities in order to foster innovation.

(vi) STIP: Is should also include a module in order to ascertain whether firms are aware of and have applied for or used STIP. While it is to be expected that final configuration of this module will vary substantially across countries, at least some degree of comparability could be obtained by making use of the taxonomy presented in Table 1. When a module

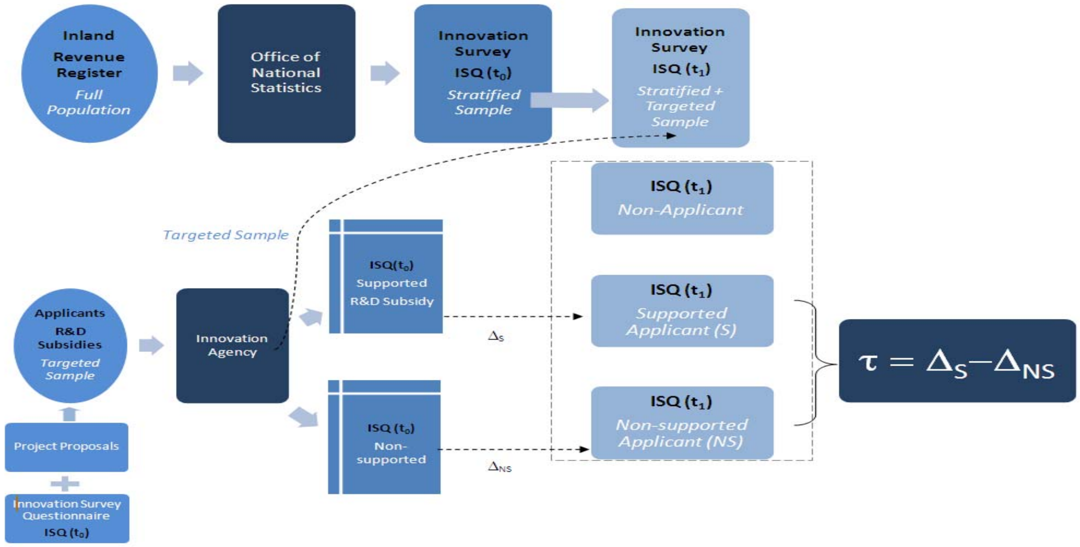
on STIP is included, it is important that it not only asks whether a firm has been using a given instrument, but that it also ask for how long the beneficiary has been using these instruments. This issue is important to achieve a proper identification of both treatment and control groups.

Despite this potential, in LAC Innovation Surveys still need to overcome at least three important constraints in order to make them fully useful for impact evaluation. First, so far as IS sampling frameworks have followed a cross-sectional nature, as we will see below in order to properly identify the causal effects of policy intervention we need to have rich baseline information. In terms of data collection, this implies moving towards a longitudinal or repeated cross-sectional approach. Some steps in this direction has been taken in Germany (Janz, et.al., 2001). In LAC, the case of the Chilean Enterprise Longitudinal Survey should be highlighted. Second, it is important to augment the sector (and regional) coverage of IS surveys. So far the majority of the IS in the LAC region have been implemented in the manufacturing sector. However, national accounts based evidence suggests that while natural resource-based sectors are highly dynamic (in terms of productivity growth), there is very little dynamism in the service sector. Natural resource-based sectors and the service sector are important in LAC, the former in terms of its contribution to competitiveness and the latter in terms of its importance for employment. However, neither of them is usually covered by standard national innovation surveys. A handful of countries have taken some steps in order to remediate this situation. Indeed, Uruguay, Brazil, Chile and Colombia have extended their IS include services. Lastly, in order to be fully effective for STIP impact evaluations, IS implementation in terms of data collection should be embedded within the policy implementation process itself. In other words, data collection should be an integral part of the STIP programs to be evaluated.

Box 1. Embedding Evaluation Process Data Collection into Program Implementation

A best practice approach for impact evaluation is to coordinate the processes of data collection normally done under the Offices of National Statistics with information emerging from policy implementation. A good example of this is being implemented by the Uruguay's National Innovation and Research Agency (ANII). In the case of Uruguay the Innovation Survey is collected regularly by the Office of National Statistics (INE) using the Inland Revenue business framework. The IS sampling framework is based on a stratified sample which is representative of the whole economy. On the other hand, whenever applicants submit proposals for funding to ANII they are asked to submit a project outline but also to fill the same questionnaire used by INE for the IS. After this, applicant information (and IS questionnaires) is sent to INE in order to be included in the future waves of the IS. This allows for both having a baseline with innovation information on applicants (both granted and unsupported firms) and follow-up information to be gathered by an independent party (INE). This produces information, which is rich enough to identify the impacts of policy. Figure 1 summarizes the whole process.

Figure 1. Embedding Innovation Surveys into the evaluation of Policy Impacts: The ANII Model



4.3 Frequency

The final characteristic is the frequency of the data. Some data are collected annually, like the R&D statistics or industrial surveys, some are collected bi-annually or every four years (like the innovation surveys in many countries), some irregularly or only once. Consequently for some variables we have only cross-sectional data, while for other variables repeated observations over time may be available.

Two main cases can be distinguished. First, repeated cross-sections are obtained when a survey is taken in successive years but considering different samples in each year. Thus, while the variables in the survey are measured in several years, the observed firms are different in each year. A usual way to work with repeated cross-sections is to aggregate the data by averaging firms in each year at some level like industry, city or other geographical and/or administrative areas. The main advantage of this survey structure over a single cross-section is the possibility to investigate dynamic effects, i.e. the time dimension of the responses to an S&T intervention, and also to control for some unobserved heterogeneity at least at the aggregate level, as will be explained in the next section.

Second, some surveys collect data from the same units over time, a data structure known as panel data. This can be considered the most informative type of survey from an Impact Evaluation point of view since it allows controlling for some types of selection biases at the individual level using fixed effects methods. This topic will be discussed in section 5. However, a disadvantage of panel data compared to repeated cross-sections is that the former may be subject to attrition bias, which occurs when firms are lost from the panel in a non-random fashion.

5. A Review of Evaluation Methods

This section provides a brief description of different methodologies usually applied in Impact Evaluation, focusing on intuition and main aspects of implementation. Technical details will be omitted as much as possible, and the reader interested in further exploring these techniques will be referred to the relevant literature.

The first subsection discusses Experimental Design, which is nowadays considered the "Gold Standard" in Impact Evaluation. Even in cases where a full experimental design is not feasible, it is usually a useful benchmark to which other methods can be compared. The second subsection describes several non-experimental and quasi-experimental techniques that can be applied when the experimental design is not feasible.

5.1 The "Gold Standard": Experimental Design

The experimental design, consisting of randomly dividing a representative sample into a treatment and a control group, is considered the superior design in the impact evaluation literature. This is because random assignment to treatment ensures that both on average observable and unobservable characteristics will be balanced between the treated and untreated units, making the two groups comparable and eliminating selection bias. As a result, the Fundamental Problem of Causal Inference can be overcome by using a randomly selected control group to estimate the counterfactual outcome of the treatment group.

Besides its undisputed usefulness in dealing with the missing counterfactual problem, experimental designs have other practical advantages. On the one hand, randomization allows estimation of the average impact of a program as a simple difference in means between the treated and control group without requiring the sophisticated econometric techniques needed in non-experimental contexts to account for different kind of biases. In addition, randomization can reduce data requirements relative to other non-experimental techniques, since estimating average program impacts with randomized assignment only requires information on post-treatment outcomes for the two groups and a handful of ex-ante characteristics to check that the randomization was successful. Of course, this is not to say that a rich dataset is unimportant for experimental evaluations. The more data are available, the more precise and informative the evaluation; for example, collecting data for many years after treatment may aid in investigating the long-term effects of a program, while rich pre-treatment data on outcomes and other observable factors can significantly improve the precision of estimated impacts, a key concern in studies using small samples. As such, the designer of an evaluation should always aim at collecting all the information that the budget permits.

While randomization is becoming a widespread approach to assess the impact of public policies in areas like Development and Labor Economics (see e.g. Banerjee and Duflo, 2008), randomization has not yet been applied to the evaluation of STIP. One possible explanation for this fact is that STIP are unlikely to meet the conditions in which random assignment is more feasible, namely, excess of demand. In general, randomized experiments to evaluate public interventions take advantage of the fact that high demand for these services and capacity constraints from the supply side generate excess demand. Under these conditions, random selection of the beneficiaries from a pool of eligible candidates is a clear and transparent way to ensure that every unit (individual, firm, etc.) has the same probability of participation.

However, an excess of demand is not necessary for application of an experimental design; randomization is compatible with treating all of the eligible population. For example, it is common to use randomization to divide the sample of eligible individuals into different groups and randomly assign the order in which they receive treatment, rather than whether or not they receive treatment. This allows one to use the groups treated later in time as comparison for the groups treated earlier. Nevertheless, certain program characteristics such as the type of project and the number of applicants may mean that randomization is not be politically or ethically feasible in some cases, but there is still a need to conduct an impact evaluation. Many non-experimental techniques have been developed to estimate the impact of public programs in lieu of random assignment. These techniques will be discussed in the next subsections.

5.2 Quasi and Non-experimental Methods

In the absence of random assignment, pre-program differences between participants and non-participants can generate biases that severely challenge the estimation of the program impacts. Of particular concern is selection bias, which may come from two sources. First, administrative bias (or program placement bias) occurs when program administrators select participants based on specific criteria that make them different from non-participants. Second, self-selection bias occur when individuals decide whether or not to participate based on some kind of cost-benefit analysis which again may result in significant differences between the pool of participants and non-participants. In practice it is very likely to have a combination of the two types of selection biases: in general, every public intervention has a target population (small firms, young researchers willing to study abroad, scientific projects with high potential success) and within the target population individuals or firms may decide whether to participate or not. As a consequence, a simple difference in means between treated and non-treated will not yield an accurate estimation of the program effect, since it will be contaminated by ex-ante differences between groups.

In this section, we briefly revise two main approaches can be used to address these issues. The first approach is to try to control for the factors that generate the selection bias; among these techniques, regression methods and propensity score matching explicitly control for observable variables differing between groups, while DD and fixed effects models use data before and after the program for the two groups to account for a certain type of unobserved heterogeneity. Second, the Instrumental Variables (IV) approach and Regression Discontinuity design (RD) both exploit particular features of assignment rules to try to replicate the experimental setting**.**

5.2.1 Regression Methods and Propensity Score Matching

As mentioned, in an experimental design the impact of the program can be estimated by the difference in means between groups, which is equivalent to running a linear regression of the outcome on a constant and a binary variable indicating the treatment status (treated/non-treated):

$$\boldsymbol{Y_i = \alpha + \beta T_i + \varepsilon_i}$$

In non-experimental settings, this technique is not enough to capture the parameter of interest. However, if all variables affecting both treatment status and the outcome variable are observable, it is straightforward to control for these variables by adding them to the linear regression:

$$\boldsymbol{Y_i = \alpha + \beta T_i + \gamma X_i + \varepsilon_i}$$

where $\boldsymbol{X_i}$ is a vector containing all the relevant control variables. If all relevant factors are accounted for, then it is a standard result that Ordinary Least Squares (OLS) yields a proper estimator of the treatment effect. This type of regression is easily implemented using standard statistical packages.

The key assumption here is that one can explicitly control for all the relevant variables, a statement usually known as *Conditional Independence Assumption* (CIA) or *Selection on Observables.* In the context of the CIA, Propensity Score Matching (PSM) can be seen as a somewhat more flexible technique to estimate the program impact relative to linear regression. The idea behind PSM is similar to that of the OLS estimator, namely, to control for observable differences between groups, but this time with fewer parametric assumptions. To understand how PSM works, suppose that treated and untreated individuals only differ in a single variable, *X*. Then the matching estimator assigns to each treated individual a comparison unit which has the most similar value of *X*. The treatment effect can be estimated in this case as an average of the differences between each treated unit and its nearest neighbor in terms of their values of *X*. When there are several factors differing between groups, however, the idea of closeness is not clearly defined, since individuals may be similar in certain aspects and dissimilar in others at the same time. To overcome this dimensionality problem, Rosenbaum and Rubin (1983) showed that if one knows all the relevant factors determining participation, then the matching procedure can be done based on the conditional probability of participation, or propensity score:

$$\boldsymbol{p(X) = P(T = 1|X)}$$

which represents the probability of participating in the program for a given value of the vector of characteristics X. The matching estimator is obtained in three steps:

1. Estimate the propensity score $\boldsymbol{p(X)}$ using a standard discrete-choice model (logit or probit);
2. Match each treated unit with one or several untreated units with similar propensity scores, and construct a common support by dropping from the sample units that do not have any comparable units with opposite treatment status;
3. Estimate the treatment effect on the treated as an average of the differences between each treated unit and its neighbor(s) in the common support.

In practice, several decisions need to be made to obtain a matching estimator. As to the first step, the substantive differences between logit and probit models are minor and the two options yield very similar results. In the second step, a choice among the different matching algorithms has to be made. Some of the possibilities include:

- Nearest neighbor matching, which consists on matching each treated unit to the non-treated individual with closest propensity score; this can also be extended to many nearest neighbors;
- Caliper and radius matching are similar techniques, but specify a "caliper" or maximum propensity score distance by which a match can be made;
- Kernel matching is a way of using all the information in the sample to construct a weighted average of non-treated units, where the weights are determined by the distance between the propensity scores.

A detailed discussion on these and other parameters to be chosen for PSM is beyond the scope of this document, and we refer the interested reader to Caliendo and Kopeinig (2005) and Heinrich et al. (2010). Crump et al. (2009) provide a technical discussion on how to construct the common support to ensure an appropriate overlap in the covariate distributions between groups.

It is worth mentioning that, while different matching algorithms exploit the information from the sample in different ways and thus may yield different estimated impacts, in practice the results of a study applying PSM should not depend critically on the matching procedure. As such, the sensitivity of estimations to different algorithms should always be studied to ensure that the results are robust and not just an artifact of model specification.

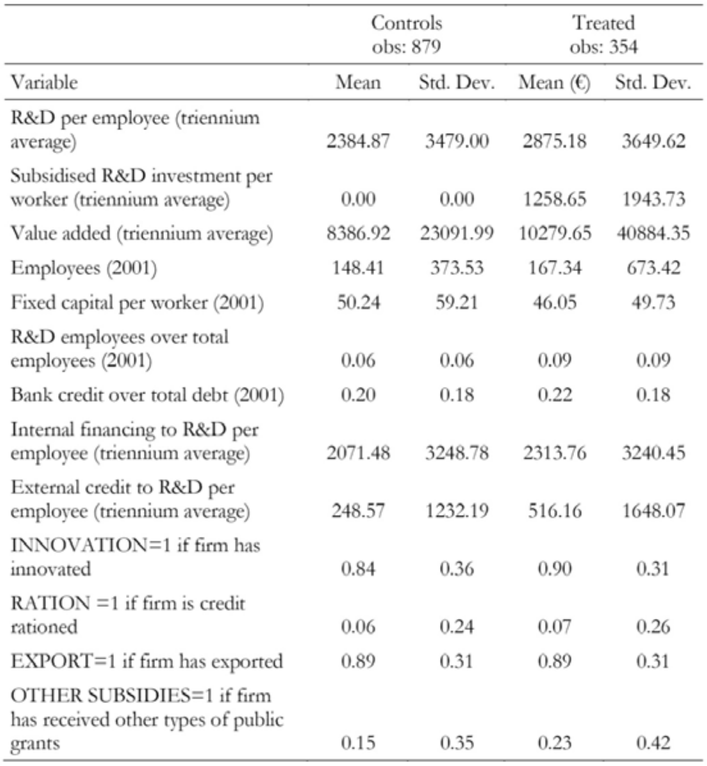
After running the matching procedure, it is important to provide some evidence showing that the treated and control groups are similar in the matched sample. One way to do this is to look at the differences in means or standardized differences (see Caliendo and Kopeinig, 2005) to show that the covariates are "balanced" between groups. It is also common to check that the distribution of the propensity scores is identical in both groups, either graphically or using statistical tests like the Kolmogorov-Smirnov test.

Currently, PSM appears to be a preferred approach for the evaluation of R&D public programs. Some examples from the literature are Czarnitzky, 2002; Almus and Czarnitzky, 2003; Duguet, 2004; Lööf and Heshmati, 2005; Bérubé and Mohnen, 2007; Carboni, 2008, and González and Pazó, 2008. Box 2 shows an application of PSM.

Box 2. Evaluating the Impact of R&D Subsidies Using PSM

Carboni (2008) analyzes the effect of public subsidies on private R&D spending using a firm-level dataset of manufacturing firms in Italy. Since R&D subsidies are not randomly assigned, large differences between subsidized and non-subsidized firms are expected, as confirmed by Table 4.

Table 4. Treated Versus Control Firms at Baseline

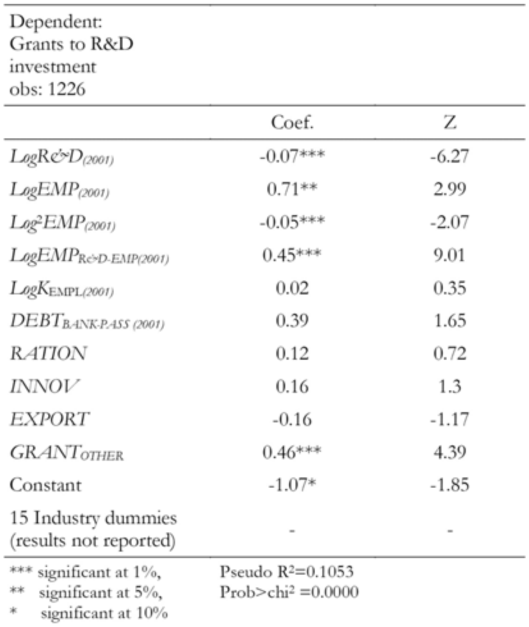


	Controls obs: 879		Treated obs: 354	
Variable	Mean	Std. Dev.	Mean (€)	Std. Dev.
R&D per employee (triennium average)	2384.87	3479.00	2875.18	3649.62
Subsidised R&D investment per worker (triennium average)	0.00	0.00	1258.65	1943.73
Value added (triennium average)	8386.92	23091.99	10279.65	40884.35
Employees (2001)	148.41	373.53	167.34	673.42
Fixed capital per worker (2001)	50.24	59.21	46.05	49.73
R&D employees over total employees (2001)	0.06	0.06	0.09	0.09
Bank credit over total debt (2001)	0.20	0.18	0.22	0.18
Internal financing to R&D per employee (triennium average)	2071.48	3248.78	2313.76	3240.45
External credit to R&D per employee (triennium average)	248.57	1232.19	516.16	1648.07
INNOVATION=1 if firm has innovated	0.84	0.36	0.90	0.31
RATION =1 if firm is credit rationed	0.06	0.24	0.07	0.26
EXPORT=1 if firm has exported	0.89	0.31	0.89	0.31
OTHER SUBSIDIES=1 if firm has received other types of public grants	0.15	0.35	0.23	0.42

Source: Carboni 2008

To reduce the bias that these differences in observed characteristics may generate, the author applies a PSM estimator. The probability of participation is estimated using covariates such as R&D spending, and the log number of employees at the beginning of the period, percentage of total workers who work in R&D, log capital intensity, the ratio of bank debt to total debt, dummies for innovation, credit rationing, exports, and other grants, as well as several quadratic terms, and industry dummies. The results of the model are shown in Table 5.

Table 5. Participation Model



Dependent: Grants to R&D investment obs: 1226		
	Coef.	Z
$LogR\&D_{(2001)}$	-0.07***	-6.27
$LogEMP_{(2001)}$	0.71**	2.99
$Log^2EMP_{(2001)}$	-0.05***	-2.07
$LogEMP_{R\&D\text{-}EMP(2001)}$	0.45***	9.01
$LogK_{EMPL(2001)}$	0.02	0.35
$DEBT_{BANK\text{-}PASS\ (2001)}$	0.39	1.65
RATION	0.12	0.72
INNOV	0.16	1.3
EXPORT	-0.16	-1.17
$GRANT_{OTHER}$	0.46***	4.39
Constant	-1.07*	-1.85
15 Industry dummies (results not reported)	-	-

*** significant at 1%,
** significant at 5%,
* significant at 10%

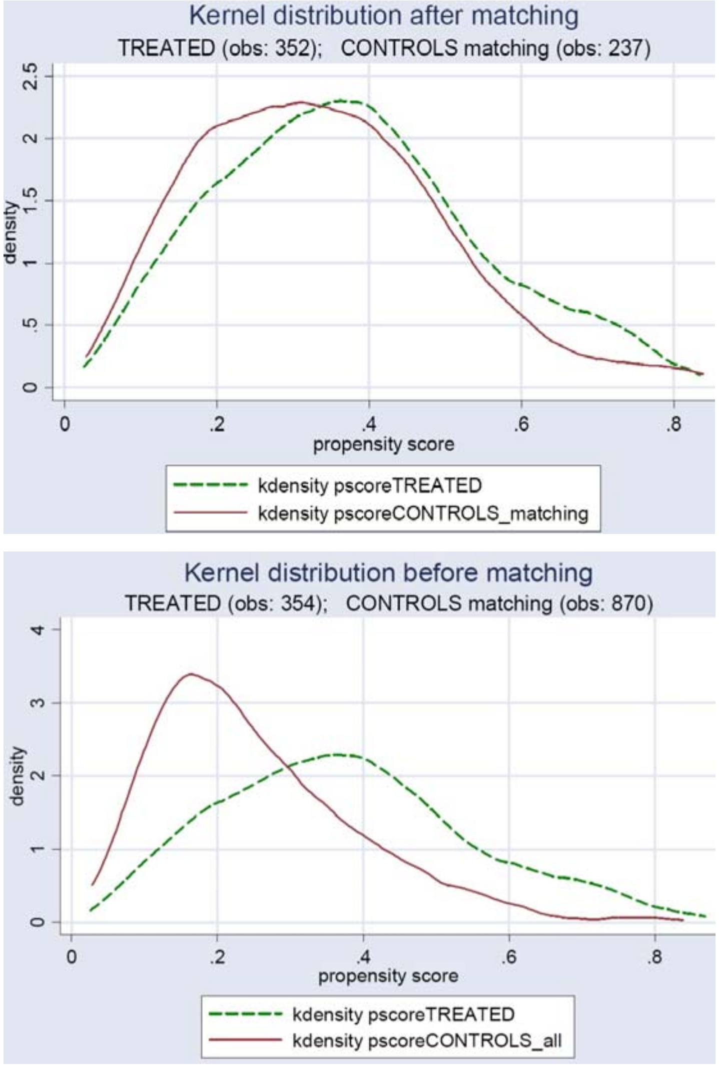
Pseudo R^2=0.1053
Prob>chi^2 =0.0000

Source: Carboni 2008

Some indicators to look at after running a participation model are the individual significance of the coefficients, the global significance of the model (chi-square statistic) and the pseudo R-squared. In general, significant coefficients and a low p-value of the chi-square test indicate that the variables included are relevant to explain participation, while a relatively high pseudo R-squared suggests a high predictive power of the model.

After estimating the conditional probability of participation, the author matched subsidized firms to non-subsidized firms with similar propensity scores using the nearest-neighbor algorithm. The author compares the distributions of the propensity scores before and after the matching to check that the matching is successful in making the groups more similar. (Figure 2)

Figure 2. P-score Distribution Before and After Matching

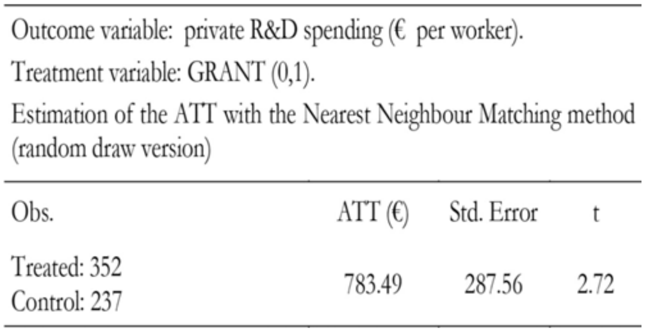


Source: Carboni 2008

The Figure 2 suggests that the samples are much more similar after matching. Note that the number of control units is significantly reduced (from 870 to 237), since the matching procedure drops the untreated observations that are not comparable to treated units. Moreover, two treated units that are outside the common support are dropped from the sample because they have no comparison unit.

Finally, the impact of R&D subsidies is estimated as the difference in averages on the matched sample (Table 6):

Table 6: PSM Results



Outcome variable: private R&D spending (€ per worker).
Treatment variable: GRANT (0,1).
Estimation of the ATT with the Nearest Neighbour Matching method (random draw version)

Obs.	ATT (€)	Std. Error	t
Treated: 352 Control: 237	783.49	287.56	2.72

Source: Carboni 2008

5.2.1.2 When are Regression and Matching Appropriate?

Although the literature appears to favor matching over simple regression methods, at least from a practical point of view the superiority of one technique over the other is not obvious. Matching can be viewed as a more flexible specification because of its non-parametric (or semi-parametric) nature, and as such it ensures that comparison units are obtained from the data instead of being constructed via extrapolation; it also emphasizes the need of for common support of observed characteristics to ensure the similarity of treated and comparison units. However, matching can also be a much more cumbersome procedure than OLS from a computational point of view; although it can now be implemented in several statistical packages, some issues related to the PSM method like the estimation of standard errors or finite sample properties are still being studied in the theoretical literature and can become a source of confusion for the practitioner (e.g., see Frölich, 2004 and Abadie and Imbens, 2008). Furthermore, matching is a “data-hungry” method that requires not only a very large set of covariates but also a large pool of potential controls to ensure that an appropriate match can be assigned to each treated unit.

It is common to compare the results from the two techniques, and large differences between linear regression and PSM should be interpreted with caution. In fact, leaving aside second-order issues of model specification, both regression and matching rest on the same assumption, i.e., the CIA. The bottom-line is that the discussion should not focus on whether to use one or the other, but whether the CIA is likely to hold. As any identification assumption, the CIA is not testable, and although some evidence can be provided to support it, its validity will have to be assessed for each specific context based on economic theory and a sound understanding of the program design and implementation.

In general, matching or regression methods may be appropriate when the researcher has a clear description of the selection process and has access to a rich set of information on the variables affecting participation. When the selection process is unclear, and especially when participation is likely to be driven by unobserved factors, matching and regression will perform poorly and alternative methods should be considered.

5.2.2 Difference-in-differences and Fixed Effects Models

Difference-in-differences (DD) models arose in the context of "natural experiments" used to evaluate the impact of aggregate-level policy changes on different outcomes. For instance, in one of the most cited papers using this technique, Card and Krueger (1994) exploit a law change in New Jersey to study the impact of minimum wages on employment, using Pennsylvania as a comparison group.

In the most simple two-period two-groups setting, the DD estimator can be understood as follows. Consider two cities, A and B, observed in two time periods, *t=0* and *t=1*. In the second period, a program to finance scientific research is implemented in city A. The idea of the DD method is simply to compare the changes in the outcome Y between cities, as illustrated in Figure 3.

Figure 3. DD Logic

The parameters $\boldsymbol{\Delta_A}$ and $\boldsymbol{\Delta_B}$ reflect the change in outcomes between the pre-treatment and post-treatment periods for each city. The objective is to estimate the impact of the intervention on the outcome of city A. The problem is that $\boldsymbol{\Delta_A}$ is not a good measure of this impact, since the outcome of city A may have changed because of other factors un related to the treatment, but this is unobservable. However, as long as one can argue that city B is a proper comparison, the change in its outcome, $\boldsymbol{\Delta_B}$, can be used to estimate this counterfactual. Then the impact of the program is calculated as the difference between the observed change in city A, $\boldsymbol{\Delta_A}$, and its estimated counterfactual, $\boldsymbol{\Delta_B}$.

In a regression framework, this is equivalent to using the following specification:

$$\boldsymbol{Y_{ist} = \alpha + \gamma CITY_s + \lambda YEAR_t + \beta CITY_s * YEAR_t + \varepsilon_{ist}}$$

where Y_{ist} is the outcome for the firm *i* in city *s* in year *t*, CITY is equal to one if the firm is located in city A, and *YEAR* is equal to one if *t=1*. In this setting, $\boldsymbol{\gamma}$ measures the difference in the average outcome between cities at the baseline *(t=0)*, $\boldsymbol{\lambda}$ reflects the time trend in outcomes that is not associated with the treatment (in this case, the trend of the outcome for the comparison group), and $\boldsymbol{\beta}$ captures the average effect of the treatment. This estimator is called DD because it can be obtained as a double difference:

$$\begin{aligned}\boldsymbol{\beta = \Delta_A - \Delta_B = [E(Y|CITY = A, YEAR = 1) - E(Y|CITY = A, YEAR = 0)]} \\ \boldsymbol{- [E(Y|CITY = B, YEAR = 1) - E(Y|CITY = B, YEAR = 0)]}\end{aligned}$$

This model can easily be extended to the case of multiple groups and time periods, and also to include control variables (see e.g. Imbens and Wooldridge, 2009). Moreover, the DD estimator can be adapted to the case where the treatment is assigned at the individual level; this allows to overcome one of the main limitations of regression and matching estimators, since it allows controlling for selection on unobservables as long as these unobserved factors are constant over time. Thus the DD model is an example of a fixed effects (FE) estimator, which assumes that any unobserved heterogeneity affecting outcomes and program participation is fixed over the recorded time horizon.

More precisely, the Fixed Effects (FE) model has the following structure:

$$\boldsymbol{Y_{it} = \alpha_i + \lambda_t + \beta T_{it} + \gamma X_{it} + \varepsilon_{it}}$$

where $\boldsymbol{\lambda_t}$ are time effects that are common to all units, $\boldsymbol{X_{it}}$ is a vector of observed control variables allowed to change over time and across individuals, $\boldsymbol{T_{it}}$ is the treatment indicator[15] and $\boldsymbol{\alpha_i}$ is the individual fixed effect that captures all observed and unobserved variables that are constant over time.

The main idea of fixed effects models is to eliminate the unobserved factors $\boldsymbol{\alpha_i}$ by exploiting the panel structure of the dataset. This could be done in different ways, for instance, running a linear regression on the differences in outcomes, $\boldsymbol{\Delta Y_{it} = Y_{it} - Y_{it-1}}$ (First-Difference estimator), by subtracting the individual averages over time (Within estimator) or by adding individual level dummies (Least Squares Dummy Variables estimator). These routines are easily implemented using standard statistical software (for example, the FE option of the xtreg command in Stata), and the technical details of the estimation procedures will not be covered in this chapter (the interested reader may refer to Wooldridge, 2002). As long as all the relevant[16] factors have been accounted for, the estimator for $\boldsymbol{\beta}$ captures the average impact of the program.

The identifying assumption of DD and the fixed effects model states that there are no unobserved time-varying factors affecting both the outcome and treatment status, which means that all unobserved relevant factors must be constant over time. This is a less restrictive assumption than the CIA, since it does not require the average outcomes of the treated and untreated groups to be equal pre-treatment, but allows them to differ by a constant magnitude. This weakening of assumptions comes at a cost, since panel data models require information on at least one pre-treatment and one post-treatment period, which can sometimes be difficult to obtain.

In other words, DD and FE models require that, in the absence of treatment, the two groups would have had the same trends. Although this assumption is not testable, its validity should always be carefully discussed. If data for several pre-treatment periods is available, a straightforward way to provide evidence to support this assumption is to show that trends were equal between groups before the program; the equality of pre-treatment trends suggests that the groups are indeed comparable and thus renders the identification assumption more plausible.[17]

[15] Usually, the treatment indicator is constructed as a dummy variable taking the value of one from the beginning of participation. However, other possibilities for constructing this variable will be discussed in Section 6.

[16] In this context, "relevant" means correlated with both the outcome and the treatment status.

[17] See Galiani et al. (2005) for a way of formally testing the equality of pre-treatment trends hypothesis.

5.2.2.1 Specification Checks

If, after controlling for a set of observable factors $\boldsymbol{X}$, the groups still have different pre-treatment trends, some adjustments will be needed to ensure that the control group is a proper comparison for the treated group.

One possibility, suggested by Angrist and Pischke (2009), is to add group-specific trends, allowing the treated and control groups to follow different trends. While this approach is more flexible, structure still needs to be placed on this specification. A linear trend is usually interacted with a group-specific intercept:

$$\boldsymbol{Y_{it} = \alpha_i + \lambda_t + \beta T_{it} + \gamma X_{it} + \varepsilon_{it}}$$

where $\boldsymbol{G_i}$ is equal to 1 if the observation belongs to the treatment group. This equation can only be estimated with three or more time periods.

Another feasible approach to try to correct for differences in ex-ante trends is to combine the DD estimator with PSM techniques. For the two-period, two-groups case, the Difference-in-differences Matching estimator, originally proposed by Heckman, Ichimura and Todd (1997), basically consists of applying the standard PSM estimator using the change in the outcome, $\boldsymbol{\Delta Y_{it} = Y_{it} - Y_{it-1}}$, instead of the outcome in levels. This technique aims at combining the advantages of both methods, since it allows one to control for unobserved heterogeneity fixed over time while reconstructing the counterfactual outcome using only the most similar observations from the pool of untreated units.

With several pre-treatment periods, lagged outcomes can be included in the propensity score in addition to other covariates to guarantee that the comparison units follow the same trajectories before the program. Then, the DD matching estimator described in the previous paragraph can be used to estimate the treatment effect, or one can run the previous equation in the matched sample.

One particular case in which trends between groups differ before the program occurs when individuals enter the program because of negative (or positive) transitory shocks in past outcomes. This fact was first noted by Ashenfelter (1978), who detected a sharp drop in the earnings of participants of a labor training program in the year previous to participation, a phenomenon currently known as Ashenfelter’s dip. In this case, including lagged outcomes $Y_{it\text{-}1}$, $Y_{it\text{-}2}$, etc. as control variables in the panel equation seems to be a natural choice to solve this problem. A precise description of this issues are beyond the technical level of this document, but it is worth mentioning that this strategy will introduce a bias and the estimation needs to be corrected using dynamic panel data models, which can be computationally cumbersome. For more information on this topic, see Bond (2002).

Some applications of DD and fixed effects models are provided by Chudnovsky et al. (2006) and Ubfal and Maffioli (2010), who analyze the impact of public funding on scientific research and research collaboration, and Görg and Strobl (2007), who study the effect of public subsidies on private R&D spending. See Box 3 for an example.

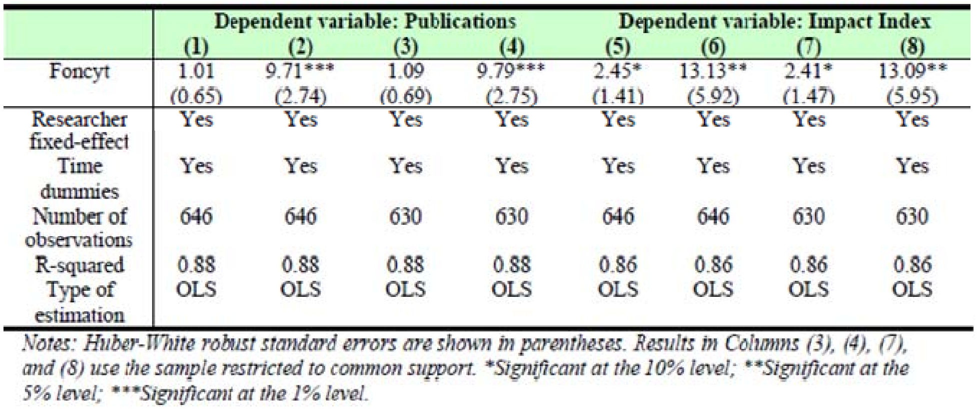
Box 3. Estimating the Impact of Public Funding on Scientific Research Using DD

Chudnovsky et al (2006) analyze the impact of a public funding program in Argentina, the FONCYT, on scientific productivity and research quality as measured by the number of publications and the impact factor of the journals where they were published. Using a panel dataset, the authors apply a DD estimator to account for time-constant unobservable factors that may affect participation and the outcome. They estimate the following equation:

$$\boldsymbol{Y_{it} = \beta D_{it} + \lambda X_{it} + \alpha_i + \mu_t + \varepsilon_{it}}$$

where $\boldsymbol{\beta}$ captures the average impact of the program. Moreover, to allow for heterogeneous effects of the treatment, the authors include interaction terms between participation and age, gender and having a doctorate. By combining the DD techniques with PSM, the estimations are restricted to a common support to ensure that observations in the sample are comparable. The results are shown in Table 7.

Table 7: DD and PSM Results



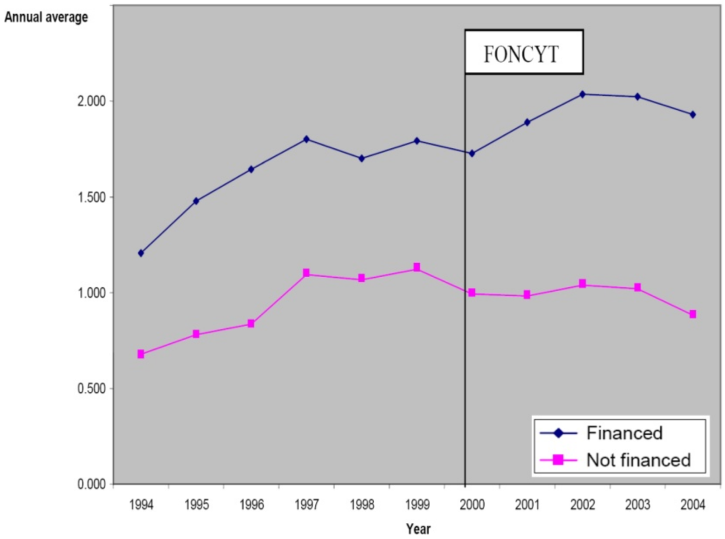
	Dependent variable: Publications				Dependent variable: Impact Index			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Foncyt	1.01 (0.65)	9.71*** (2.74)	1.09 (0.69)	9.79*** (2.75)	2.45* (1.41)	13.13** (5.92)	2.41* (1.47)	13.09** (5.95)
Researcher fixed-effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	646	646	630	630	646	646	630	630
R-squared	0.88	0.88	0.88	0.88	0.86	0.86	0.86	0.86
Type of estimation	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS

*Notes: Huber-White robust standard errors are shown in parentheses. Results in Columns (3), (4), (7), and (8) use the sample restricted to common support. *Significant at the 10% level; **Significant at the 5% level; ***Significant at the 1% level.*

Source: Chudnovsky et al (2006)

Finally, the validity of the common trends assumption is evaluated by observing the pre-treatment trend in the outcomes (in this case, number of publications). As mentioned, although the validity of the identification assumption cannot be proven, similar pre-treatment trends help to increase the credibility of the estimates. Figure 4 suggests that non-financed researchers in the sample have very similar trends before treatment and thus can be considered a good comparison group.

Figure 4. DD basic Assumption



Source: Chudnovsky et al (2006)

5.3.3 The Instrumental Variables Approach

The Instrumental Variables (IV) approach consists of exploiting particular features of the design and institutional setting of a program in order to find a source of exogenous variation that replicates as much as possible the conditions of a randomized trial. While the theoretical aspects of the IV method can be very complex, the intuition is straightforward; the idea is to find a variable that affects the probability of participation, but is not related to other variables affecting the outcome in any other way. In other words, an instrumental variable (or, simply, instrument) is a variable that affects the treatment status, but can be considered "as good as random". A famous example of instrumental variables is provided by Angrist and Krueger (1991), who noted that because of compulsory education laws, students born in the first quarter of the year could drop out of school earlier than other students and receive relatively less education as a result. Angrist and Krueger used quarter of birth as an instrument for educational attainment in their study of returns to education. Since there is no a priori reason to think that quarter of birth would affect earnings in any way other than its effect on educational attainment, quarter of birth is an exogenous source variation in education that acts as a natural experiment.

A good instrument must meet two conditions. The first one is usually known as the *relevance condition*, and states that an instrument must be correlated with the endogenous variable one wants to instrument, which in the case of impact evaluation is usually program participation. The stronger the relationship between the instrument and the probability of participation, the better the instrument, other things being equal.

The second condition, the *exogeneity* condition, states that an instrument has to be uncorrelated with other determinants of the outcome variable. As mentioned earlier, an instrument should replicate the conditions of random assignment. In the Angrist and Krueger example, quarter of birth is a good instrument because it is unlikely to be correlated with other factors like intelligence, motivation, innate ability, family background and other factors that are difficult to account for in a standard regression framework and could seriously compromise the estimation.

Under these two conditions, the IV estimator can be constructed in two stages (this is the reason why the IV estimator is also known as Two Stage Least Squares -2SLS- estimator): in the first stage, a linear regression of the treatment variable T on the instrument Z is run to capture the exogenous variation in T, this is, the part of the variability in T that is generated by the instrument and thus can be considered analogous to the variation one could obtain with random assignment. In the second stage, the predicted values from the first stage (i.e., the variation in T that is explained by Z) are plugged into a linear regression with the outcome as the dependent variable to obtain the IV estimator of the treatment effect. That is:

1. First stage: run $\boldsymbol{T_i = \pi_0 + \pi_1 Z + u_i}$ by OLS and obtain the predicted values $\boldsymbol{\widehat{T}_i = \widehat{\pi}_0 + \widehat{\pi}_1 Z}$
2. Second stage: run $\boldsymbol{Y_i = \beta_0 + \beta_1 \widehat{T}_i + \varepsilon_i}$ by OLS to obtain $\boldsymbol{\widehat{\beta}_{IV}}$

Several issues need to be clarified in this procedure. First, although these steps suggest that the 2SLS estimator can be obtained by running the two regressions "by hand", the standard errors obtained by doing this will be wrong.[18] Fortunately, statistical packages have routines that automatically make the necessary adjustments to provide valid inference (for example, the ivregress command in Stata). Second, other control variables X_i can safely be added to the model when needed, as long as they are included in both stages. Third, even when the instrumented variable is binary (as in this case) or discrete, the reader should never be tempted to replace the linear regression in the first stage with a non-linear model such as probit or logit. This is because only a linear first stage will result in predicted values of the treatment variable that are uncorrelated with unobserved determinants of the outcome (for further details, see Angrist and Pischke, 2009). Finally, if multiple instruments are available for T, they can be added to the first stage to increase efficiency (see, for instance, Wooldridge, 2002).

As to the two requirements for IV, the relevance condition is easily testable from the first stage. Thus, an IV-based evaluation should always start by analyzing the strength of the relationship between Z and T (this is done by looking at the $\boldsymbol{\pi_1}$ coefficient and testing whether this coefficient is significantly different from zero). However, the *exogeneity* assumption is not testable. Therefore, its validity cannot be assessed using statistical techniques, but should be based on a solid understanding of the program design, implementation and institutional setting, complemented with economic theory.

[18] This is so because "manual 2SLS" does not account for the additional variability introduced in the first stage and thus underestimates the standard errors.

5.2.3.1. Where to Find a Good Instrument?

While IV is one of the most powerful and credible tools for estimating the impact of the program, one of its main disadvantages is that good instruments are very hard to find. There is no formal theory or computational procedure indicating what variable makes a good instrument; instead, the search for an instrumental variable has to be based on a careful examination of the program design, target population and institutional context.

A brief look at the previous literature can provide some guidelines on this issue. As suggested by Jaffe (2002), variations in available budget resources are good candidates for instruments in cases like research funding. For instance, Wallsten (2000) uses this approach to evaluate the impact of the Small Business Innovation Research (SBIR) program. He notes that according to the SBIR legislation, each participating agency has to assign a fixed proportion of his total budget for SBIR grants; then, the number of grants that each agency can assign is directly linked to the total budget, since each agency must disburse its total SBIR budget each year. While the available budget for SBIR grants is clearly related to the probability of receiving a grant, the budget is not likely to be related to unobservable firm characteristics, and can therefore be used to construct an instrumental variable. A similar approach is adopted by Clausen (2008), who argues that budgets for R&D programs in Oslo are a result of national or supranational political processes and thus should not be correlated with firm characteristics.

Clausen (2008) also suggests that distance to the headquarter of the Norwegian Research Council and to the regional headquarters of the State Industrial and Development Fund are good instruments for participation, since being close to a headquarter positively affects the decision of participation but should not be related to other firm characteristics.

Binelli and Maffioli (2006) also use an IV approach to study the impact of the FONTAR, an innovation program in Argentina. The authors use the number of Technology Linkage Units (TLU), which are offices in charge of program promotion and advertisement and help firms to prepare project proposals, in each department to instrument program participation. The idea behind this approach is that the larger the number of TLUs in a certain department, the higher the probability of a firm participating in the program. As long as the number of TLUs is not associated to firm characteristics, it will be a valid instrument to evaluate the impact of the program.

Of course, the validity of an instrument depends on the specific context, and each particular case must be analyzed to determine whether the suggested instrument meets the two requirements. For instance, the distance to the closest headquarter may not be a valid instrument if firms consider this factor when choosing their locations. Similarly, the number of TLUs may not satisfy the exogeneity condition if TLUs are placed in departments where firms are more productive or have more innovative potential, since in this case the intended instrument would be correlated with unobserved firm characteristics that also affect their performance.

The bottom-line is that there are no general rules to find a valid instrument, and, although some basic guidelines can be provided, each particular case needs to be carefully analyzed to assess whether IV is an appropriate approach.

5.2.3.2 Some Pitfalls of the IV Estimator

Without going into technical details, it is important to mention some of the main problems that one can encounter when using an IV approach.

First, the finite sample properties of the IV estimator are not particularly appealing. In fact, it can be shown that, although consistent, the IV estimator is biased. This basically means that it can be misleading in small samples.

Second, very slight deviations from the *exogeneity* condition (i.e. a small correlation between Z and other unobserved factors determining Y) may generate serious problems of inconsistency, especially in the case of weak instruments (i.e. when the correlation between Z and T is small). In this case, instruments may perform much worse than standard OLS estimators. This reinforces the importance of carefully analyzing the validity of the identification assumptions when applying IV estimators.

Box 4. A Short Digression on Local Average Treatment Effects

At this point, the careful reader could suspect that using only part of the variation in the treatment status (namely, the exogenous variation generated by the instrument) could not possibly yield the same information on the treatment effect as random assignment. In other words, even if a good instrument ensures internal validity, there must be some loss for being in a second-best environment. In most cases, this will be true.

More precisely, two scenarios are possible. If the treatment effect is homogenous (i.e., equal for all individuals in the population), as long as the instrument is valid, in the sense of satisfying the two requirements described above, the variability generated by it will be enough to identify the average impact on the whole population being studied. However, when the treatment effect varies between individuals (treatment heterogeneity), which is likely to be the case in most evaluations, an IV only allows estimating the impact on the particular subpopulation of individuals whose behavior is affected by the treatment (the *compliers*). This is because the average program effect estimated by using the IV technique is identified by variation in program participation generated by the instrument. As a result, no information about impacts on individuals whose program participation decisions are not influenced by the instrument will be contained in the estimated average program impact. Individuals who make the same program participation decision regardless of the value of Z include those who always participate *(always-takers)* and those who never participate *(never-takers).* For this reason, when treatment effects are heterogeneous the IV approach estimates a Local Average Treatment Effect (LATE); the effect is local in the sense that it only described impacts on compliers.

Therefore, when using IV it is always crucial to carefully interpret the results, considering that the estimated impacts are only valid for a particular subpopulation which may or may not be the population of interest. However, this is not a necessarily limitation of the IV approach, but rather of the available data and of the design and implementation of the program being evaluated. In other words, it is a matter of clearly determining what can and cannot be learned from the impact of the program in the absence of random assignment. As Imbens (2009) clearly explains:

> "Reporting the local average treatment effect (…) is thus emphatically not motivated by a claim that [it] is the sole or primary effect of interest. Rather, it is motivated by a sober assessment that estimates for other subpopulations do not have the same internal validity, and by an attempt to clarify what can be learned from the data in the absence of identification of the population average effect. It is based on a realization that, because of heterogeneity in responses, instrumental variables estimates are a distinct second best to randomized experiments."

5.2.4 Regression Discontinuity Design

For some types of programs, particular features of the target population and the selection process generate conditions that, at some point, replicate the conditions of an experiment. In many cases, the selection process of a program needs to be based in somewhat arbitrary rules to define its objective population. This context can be exploited using a Regression Discontinuity (RD) design.

Suppose we want to measure the impact of research funding on the quality of funded projects. Let Y be some bibliometric measure (e.g., number of citations or number of publications) of an individual *i*. Also suppose that every candidate project is assigned a score base on its quality, measured by X, and projects with a score higher than a certain threshold k receive the funding. The main intuition of RD is that, since this cutoff is somewhat arbitrary, projects below and above but near it should be very similar; furthermore, some of them will receive the funding while some of them will not. Therefore, in this case projects near but above the threshold can be used to construct the counterfactual outcome of the projects near but below it. In other words, the arbitrary cutoff generates what Lee and Lemieux (2010) call a "local randomization" that allows estimating the impact of the program.

The main assumption of an RD design is that all the relevant covariates are continuous (i.e., do not "jump") at the cutoff to ensure that individuals from each side of the threshold can be considered similar. If, for example, the ability or motivation of a researcher takes a discrete jump at the cutoff, then individuals below the cutoff would not be a good comparison group for individuals above it. In more practical terms, the continuity assumption means that individuals do not have *precise* control of their values of X (Lee and Lemieux, 2010). Suppose that some researchers knew exactly how to design a project to receive the funding. Then these researchers would be able to control their values of X to ensure they receive the subsidy. Moreover, it is reasonable to think that these researchers may be more motivated, interested or able than the rest. In this case, researchers near but below the threshold will not be comparable to the ones above it.

Ideally, in an RD design one would compare observations within an interval $[\boldsymbol{k}-\boldsymbol{\varepsilon},\boldsymbol{k}+\boldsymbol{\varepsilon}]$, where $\boldsymbol{\varepsilon}$ is an arbitrarily small number. However, it is also true that the narrower the interval, the fewer observations in it. Therefore, in practice one will need to move away from the threshold to ensure a sufficiently large sample size, which could introduce bias since observations treated and untreated units far from the threshold are not necessarily comparable. In the previous example, for instance, it is reasonable to assume a positive relationship between *Y* and *X*: the higher the score, the better the project and thus the more likely it is to be published or cited. Thus, researchers with projects with very high scores will not be comparable to researchers with low scores, since these differences in scores may be determined by unobserved factors that affect the outcome like motivation or ability.

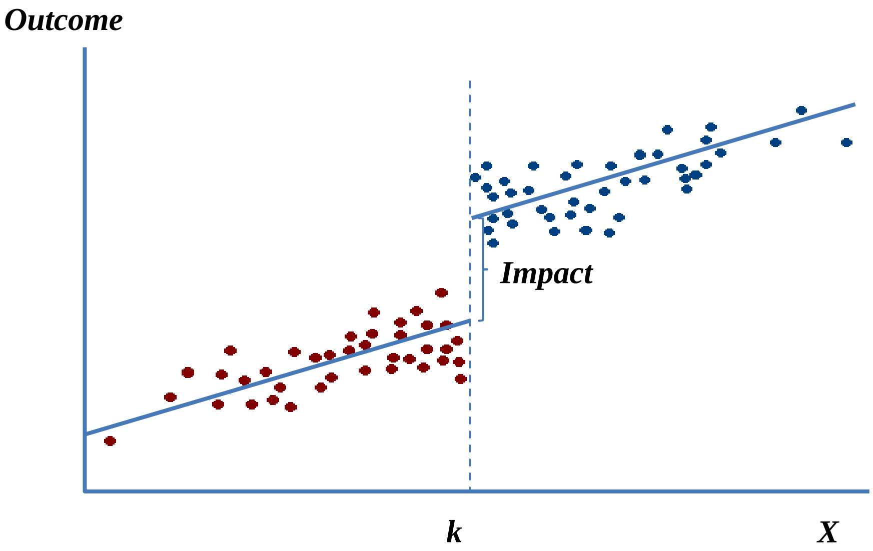
While the use of RD designs do not seem to be widespread in evaluations of STIP, some examples from the Impact Evaluation literature are Benavente et al., 2007; Jacob and Lefgren, 2007, and Bronchini and Iachini, 2011.

Two main strategies can be adopted in implementing a RD design. One is to make parametric assumptions on the relationship between the outcome and X (the running variable or forcing variable). The other one is to use nonparametric techniques to estimate the impact around a certain interval of the threshold. The next subsections briefly describe some implementation issues related to RD designs. More detailed discussions can be found in Imbens and Lemieux (2007) and Lee and Lemieux (2010).

5.2.4.1 Parametric RD

The simplest case of parametric RD is depicted in Figure 5, where a linear relationship between *Y* and *X* is assumed. This captures the idea that research projects with higher scores are more likely to be published or cited. However, if the score is higher than the cutoff, the project receives funding, which, in case of positive impact, will generate a jump in the outcome at the threshold. This discrete jump at the discontinuity is used to estimate the impact of the program.

Figure 5. Parametric RD



More formally, participation is determined based on the value of X according to the following rule:

$$\boldsymbol{T_i} = \begin{cases} \mathbf{1}\ \boldsymbol{if}\ \boldsymbol{X_i} \geq \boldsymbol{k} \\ \mathbf{0}\ \boldsymbol{if}\ \boldsymbol{X_i} < k \end{cases}$$

which means that unit *i* receives treatment if its value of *X* is larger than a threshold value k[19]. Then the treatment indicator can be defined as:

$$\boldsymbol{T_i} = \boldsymbol{I}(\boldsymbol{X_i} \geq \boldsymbol{k})$$

where *I(.)* is an indicator function. Then, the treatment effect can be estimated by running a regression of the outcome on the treatment indicator and the running variable:

$$\boldsymbol{Y_i} = \boldsymbol{\alpha} + \boldsymbol{\beta T_i} + \boldsymbol{\gamma X_i} + \boldsymbol{\varepsilon_i}$$

The treatment effect is captured by $\boldsymbol{\beta}$. Note that, in an RD context, there is no need to add additional covariates since, by construction, the treatment assignment is solely based on the value of *X*, although some covariates can be included to reduce the variance of the estimators.

The linearity assumption between Y and X can easily be relaxed by replacing the previous equation by:

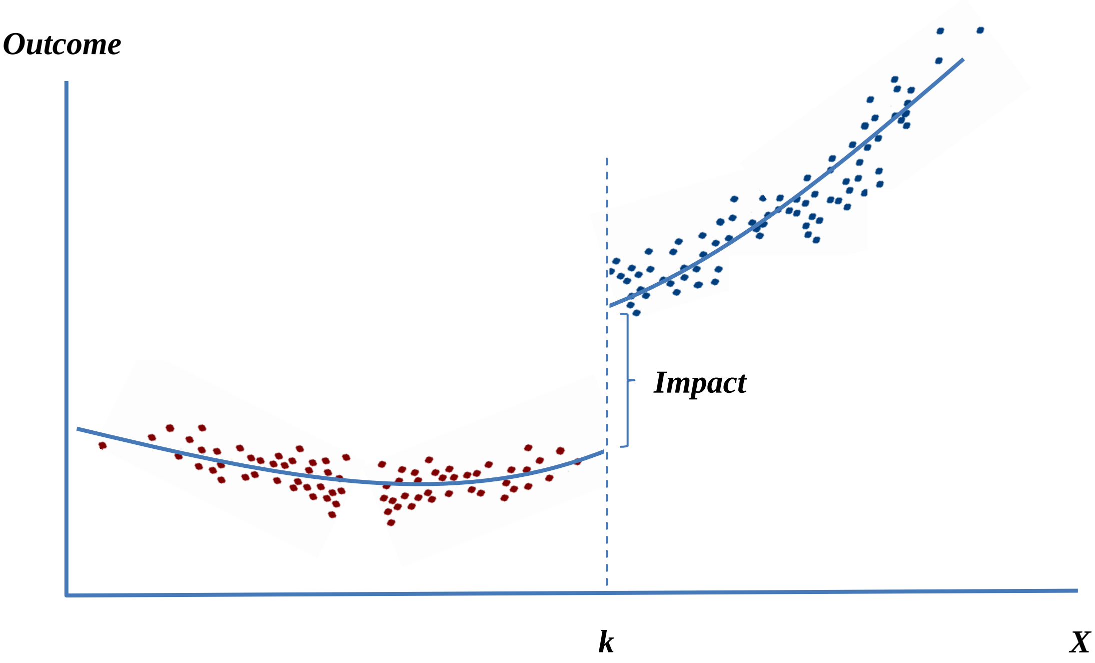
$$\boldsymbol{Y_i} = \boldsymbol{\alpha} + \boldsymbol{\beta T_i} + \boldsymbol{f}(\boldsymbol{X_i}) + \boldsymbol{\varepsilon_i}$$

[19] The methodology is analogous for the case where the treatment is assigned when $\boldsymbol{X} \leq \boldsymbol{k}$.

where *f(.)* is some flexible nonlinear function of *X*. In general, *f(.)* is modeled as a *q*-th degree polynomial, for instance:

$$\boldsymbol{Y_i = \alpha + \beta T_i + \gamma_1 X_i + \gamma_2 X_i^2 + \gamma_3 X_i^3 + \varepsilon_i}$$

Figure 6. Parametric RD without Linearity Assumption



Moreover, it is also possible to allow for different trends in each side of the cutoff by adding interaction terms between the treatment indicator and the covariate. In this case, it is common to subtract the value of the cutoff from the covariate to ensure that $\boldsymbol{\beta}$ captures the treatment effect at the threshold. Defining $\boldsymbol{\widetilde{X} = X - k}$, the regression in this case would look like:

$$\boldsymbol{Y_i = \alpha + \beta T_i + \gamma_1 \widetilde{X}_i + \gamma_2 \widetilde{X}_i^2 + \gamma_3 \widetilde{X}_i^3 + \delta_1 \widetilde{X}_i * T_i + \delta_2 \widetilde{X}_i^2 * T_i + \delta_3 \widetilde{X}_i^3 * T_i + \varepsilon_i}$$

In practice, an RD study would start with a simple specification for *f(.)* and progressively add higher order terms to check the robustness of the results. Moreover, statistical tests can be implemented to measure the goodness of fit of each specification (see e.g. Lee and Lemieux, 2010).

5.2.4.2 Nonparametric RD

Nonparametric specifications avoids making assumptions on the shape of the relationship between the outcome and the running variable. The main idea is to choose a small window around the threshold and then compare the average outcomes below and above the threshold inside the window.

Details on nonparametric specifications for RD designs are beyond the technical scope of this document, so this subsection will only sketch some implementation issues. The interested reader may refer to Imbens and Lemieux (2007) and Lee and Lemieux (2010) for further details.

Two main decisions have to be made to implement a nonparametric RD design: how to estimate the outcomes for treated and non-treated around the cutoff, and the bandwidth (i.e. the size of the window around the cutoff).

Related to the first issue, a straightforward solution is to estimate a simple average of the outcomes of the observations from each side of the cutoff. However, in general constant averages are biased when one needs to estimate values at the boundary of an interval (which is exactly the case for RD designs). Nonparametric techniques like local linear regression or local polynomial regression, which are implemented in standard statistical packages like Stata, can be used to reduce the bias.

As to the optimal bandwidth, the key issue to keep in mind is the classical bias-efficiency trade-off of nonparametric techniques: a narrower window will estimate the value with less bias but with less precision because of the smaller sample size. In an ideal situation one would start with a relatively small interval around the cutoff and progressively increase the size of the interval to see how different bandwidths affect the results. Hopefully, the results will be relatively stable. Recent advances in the literature suggest a procedure to choose the bandwidth that minimizes the Mean Squared Error (Imbens and Kalyanaraman, 2009), although details on these issues exceed the technical level of this guideline.

The bottom-line is that while nonparametric methods can be theoretically appealing since they avoid making parametric assumptions, they can be technically complex and computationally cumbersome. Thus practitioners should focus more on parametric methods. It is also worth mentioning, as suggested by Lee and Lemieux (2010), that nonparametric approaches should not be viewed as substitutes but as complements of parametric techniques, and checking the robustness of the results to different specifications help to significantly strengthen the credibility of the estimations.

5.2.4.3 Graphical Evidence

While the steps described in the previous paragraphs are necessary to estimate the magnitude of the impact and test its statistical significance, precise and clear graphical evidence is a very powerful tool to present the results and assess the robustness of an RD study.

According to Imbens and Lemieux (2003), three types of plots should be displayed. First, one must show the relationship between the running variable and the outcome; this helps to display graphically the discontinuity at the threshold (or the lack of it) and also to get a visual intuition of the functional form of the relationship between the two variables. Additionally, the predicted values from the model can be overlaid on a scatter plot of the raw data to get a more precise image of the estimated impact, as in figures 6 and 7.

Second, plots of different covariates against the running variable may help assess the validity of the identification assumption; if this assumption is true, then none of the covariates should present a discontinuity at the threshold. Of course this is not a test of the validity of the assumption since unobserved covariates cannot be analyzed, but it may help to strengthen its credibility. It is also important to plot the probability of receiving treatment against the running variable to show that there is a jump at the cutoff (in fact, under the assumptions made up to here, the probability of treatment can only take the values 0 or 1 depending on the side of the cutoff; this can be relaxed using Fuzzy designs, explained in the next subsection).

Finally, a plot of the distribution of the running variable may help determine the possibility of manipulation of this variable. For instance, if individuals could precisely manipulate the score, the distribution of the running variable would show a jump at the right of the threshold.

5.2.4.4 Fuzzy RD Design

In the RD design described in the previous paragraphs, the probability of receiving treatment is equal to zero on one side of the cutoff and one on the other. This design is known as Sharp Regression Discontinuity. However, in many cases, participation is not perfectly determined by the value of the running variable, but also by other factors. In these cases, being in one or the other side of the cutoff can be used as an instrumental variable, since it is clearly correlated with participation but assumed to be unrelated to other determinants of the outcome. This design is known as Fuzzy Regression Discontinuity. Thus, the variable $\boldsymbol{T_i = I(X_i \geq k)}$ (and possibly additional interaction terms) is used to instrument participation in the 2SLS context described in the previous section. Additional details on implementation issues of Fuzzy RD designs can be found in Imbens and Lemieux, 2007; Lee and Lemieux, 2010, and Angrist and Pischke (2009).

5.3. Structural Models

Conclusions on the effectiveness of public programs can also be derived from economic models. This is the main idea of Structural Models, which can be defined as "a collection of stylized mathematical descriptions of behavior and the environment which are combined to produce predictions about the effects of different choices, etc." (DiNardo and Lee, 2010). Thus, a structural model begins with the definition of a set of mathematical equations describing the behavior of relevant agents. For instance, a production functions and utility functions are usually employed to describe the firms' and consumers' behavior; these equations can be combined to define the market equilibrium. Finally, some parameters of the model can be changed to predict how these variations affect market outcomes.

As a specific example from the literature, Lokshin and Mohnen (2010) construct a structural model to evaluate the effectiveness of R&D tax credits in the Netherlands. The starting point of the model is a negative relationship between a firm's demand for R&D capital and its price (user cost), which is derived from the profit maximization problem of the firm. This implies that a tax deduction on R&D labor generates a reduction in user costs that triggers a partial adjustment process towards a higher level of R&D stock of the firms. The authors combine this theoretical model with firm-level data from the Netherlands to estimate the short and long-run elasticity of firm R&D capital formation to the user cost. Finally, this estimated elasticity is used in simulations to assess the effectiveness of a fiscal incentives program by comparing the additional R&D spurred by the fiscal incentives program to the cost to the government of supporting R&D with the tax scheme (Lokshin and Mohnen, 2010).

It is worth noting that in structural models the counterfactual is constructed from the underlying theoretical model, and thus can be used for ex-ante evaluation (since it does not require post treatment data) and also to evaluate programs when counterfactual outcomes cannot be constructed empirically (for instance, national-level programs, which is the case of many fiscal incentives programs). Further details on how to construct structural models can be found in Reiss and Wolak (2007). Harris et.al (2009) evaluate the impact of an R&D tax credit program implemented in Northern Ireland since 2000. The program covers expenditures on staffing costs, materials used in R&D (including computer software and energy), externally provided workers, and certain costs of subcontracted R&D. That is, the scheme does not cover capital expenditure associated with R&D on land, construction, production plants, and machinery. The current R&D tax credit for SMEs is set at 50% of qualifying revenues when calculating taxable profits, while in the case of large firms the incentive is set to 25%. For SMEs making losses, they can sacrifice the tax loss from R&D in exchange of a cash payment. In order to estimate the impacts of the fiscal incentives, Harris et al. (2009) derive an R&D stock demand from a CES production function that includes R&D as an additional factor input. By this way they manage to specific the following empirical model:

$$\boldsymbol{lnRD_{it} = \emptyset lnRD_{it-1} - \beta lnP_{it} + \propto lnY_{it} + \varepsilon_{it}}$$

where RD is the stock of R&D, Y is output and P is the user cost of R&D. This model allows for estimation of the own-price elasticity of R&D. The R&D user cost can be defined as:

$$\boldsymbol{P_{it} = \sum_{j=1}^{k} \omega_{it} \frac{1-\left(A_{c,ijt}+A_{d,ijt}\right)}{1-\tau_{it}}\left(r_{it}+\delta_{j}\right)}$$

where J refers to the assets being covered (different types of qualifying expenditures), $\boldsymbol{\omega_{ij}}$ refers to the relative amount spent on each asset, $\boldsymbol{\tau_{ij}}$ is the corporate tax on profits, $\boldsymbol{A_{c,ijt}}$ is the net present value of the tax credit, $\boldsymbol{A_{d,ijt}}$ is the net present value of the tax depreciation allowance, r the internal rate of return of the firm and $\boldsymbol{\delta_j}$ the economic depreciation rate. The authors estimate the above mentioned model for Northern Ireland by linking the R&D survey (BERD) with the manufacturing surveys (ARD). Given that R&D user costs in demand equation are endogenous, instrumental variables methods are used to estimate the own-price elasticity of R&D. Instruments are lagged values of the user costs and output. Additionally in order to control for adjustment costs, the R&D demand equation needs to be estimated using Dynamic Panel Data models.[20] The authors find that the estimated long run elasticity of R&D to user costs is -1.36 and that the introduction of the R&D tax credit scheme in Northern Ireland led to a fall in the user costs of 11%, this implies that the long run R&D stock should have risen by 15% due to the policy.[21]

[20] On the top of this, it is also necessary to correct for sample selection as not all firms carry-out R&D investments.

[21] Harris et.al. (2009) also explores the impacts of these increase on productivity. They found that impacts are very modest and that, at least for the case of Northern Ireland, in order to have an important impact on productivity, the R&D tax credit scheme should be substantially more generous than it is. However, increase the generosity of the incentive faces two problems: (i) it increases the fiscal costs in terms of forgone revenue, and (ii) it also the having enough availability of complementary assets, in particular human capital.

6. Dealing with Key Issues of an STIP Impact Evaluation

6.1 Accounting for the Timing of Effects

As mentioned, a proper assessment of the timing of effects is crucial for correctly identifying the costs and benefits of an intervention.

An example is provided by Chudnovsky et al (2006), who analyze the impact of the FONCYT program in Argentina on scientific productivity and research quality (see Box 3). Since the publication of a research paper takes time, observing a very short period of time after the subsidy is granted would lead to the estimation of a null effect. For this reason, the authors measure the number of publications in a four-year window after receiving the subsidy to take into consideration the lag between receiving the subsidy and the publication of the findings of the project.

The analysis of the dynamic impacts of a program requires large amounts of data, especially because a panel where the same units are followed over a period of time is needed, or at least a cross-section where different units entered the program in different periods (although the first case is certainly preferred).

If data of this type are available, dynamic effects can be estimated in a regression framework by changing the usual binary treatment indicator by a discrete variable indicating the number of years since participation (1, 2, 3 years, etc.). A much more flexible specification that allows for nonlinear marginal effects consists on adding dummy variables by year since participation. Then, a statistical test of whether all the coefficients are significantly different may help determine if the effect of the program is constant or varies with time.

6.2 Intensity of Treatment

Although much of the methodological discussion in the previous section focused on estimating the effect of a binary treatment variable, the extension of these techniques to the context of multivalued treatments is straightforward. For instance, one may be interested in estimating not only the effect of participation, but also the effect of participating multiple times (e.g. participating in a training course more than once), or the effect of different intensities (e.g. different amounts of financing).

In fact, although it was assumed in Section 2 that the treatment variable was binary, i.e. $\boldsymbol{T_i} \in \{\boldsymbol{0}, \boldsymbol{1}\}$, in most cases this assumption is not necessary for the validity of the methodologies, and other cases can be assessed by letting $\boldsymbol{T_i}$ to be take multiple values, either discrete (for example, $\boldsymbol{T_i} \in \{\boldsymbol{0}, \boldsymbol{1}, \boldsymbol{2}, \boldsymbol{3}\}$) or continuous.[22] Thus, linear regression, panel data, and IV methods are perfectly valid when the treatment variable is not binary, as long as the correct interpretation is given to the estimates. For instance, while in the binary case the coefficient $\boldsymbol{\beta}$ in a linear regression

$$\boldsymbol{Y_i = \alpha + \beta T_i + \gamma X_i + \varepsilon_i}$$

captures the average difference between treated and untreated units, in the continuous case the coefficient measures the impact of a unit increase in $\boldsymbol{T}$ on the average outcome (the impact of one additional year of training, or one additional dollar in funding). It is also possible to "decompose" a multivalued treatment variable in several dummies to allow for nonlinear effects, for example, a dummy variable for one year of treatment, a dummy variable for two years, and so on.

Box 5. Assessing the Long-run Effects of Technology Development Funds

Technology Development Funds (TDF) are a key component of the set of policy instruments that the different LAC governments have implemented so far in the region in order to support innovation. TDF are complex instruments that aim to correct the different market failures that harm innovation, in particular with regards to the firm-level disincentives generated by a lack of appropriability of innovation outcomes, lack of external funding due to information asymmetries in the financial markets, and lack of access to complementary inputs due to poor density of local innovation systems. Despite their popularity, further research is still needed to understand the long-run effect of technology development funds (TDFs). Past IDB evaluations have shown quite consistently that TDFs are effective at the level of research and development (R&D) input additionality (Hall and Maffioli, 2008). In particular, IDB studies have found that public funding does not crowd out private investment and in many cases has a positive

[22] For the particular case of PSM, the extension to continuous treatments (Generalized Propensiy Score) introduces some technical aspects that are beyond the scope of this document. The interested reader may refer to Hirano and Imbens (2004) and Heinrich, Maffioli and Vázquez (2010).

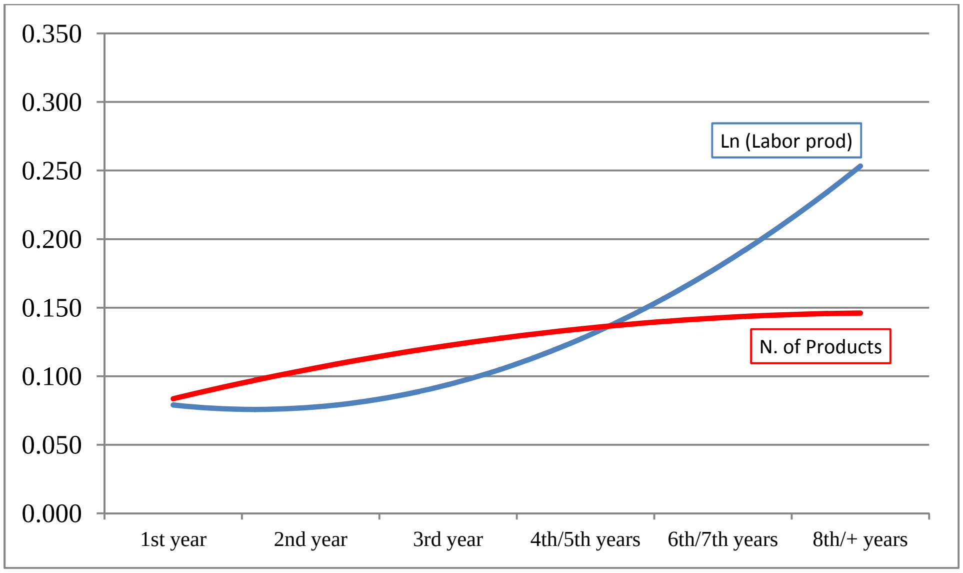
effect on the firm-level intensity of R&D. However, evidence from IDB studies regarding the impact that TDFs have on innovative outputs and firm performance is inconclusive overall.

To fill this gap, the Bank carried out a study on the long-run dynamic effects of a matching-grant program managed by the Colombian innovation agency, COLCIENCIAS. The study was designed to take advantage of a panel data set of sufficient length to detect the long-run effects of the program (Crespi et al. 2011). The study shows that COLCIENCIAS funding not only had a positive impact on firms' investment in innovation, but that also had a significant impact on their performance. Crespi et al. also provide evidence that these effects remain and sometimes increase over time. Of particular interest are program effects on productivity. Over the period 1995-2007 COLCIENCIAS funding had an average impact on introduction of new products and labor productivity of around 12% and 15% respectively, with these effects becoming more significant between three to five years after the firms started being treated (see Figure 7). These findings imply not only that beneficiary firms become more efficient, but that they grow more and gain a greater market share than the control group. The consequence is that economic resources are being reallocated towards more productive firms, which should positively impact economy-wide productivity.

The importance of these results is twofold: first they confirm that TDF are effective not only in promoting R&D investments, but also in boosting firm performance in the long run. Secondly they show that longer-term impact evaluations of such projects enable the detection of impacts on some of the most relevant variables of interest. This highlights the important of doing a careful follow-up of beneficiaries over a relatively long time period; such an approach can be done more efficiently by building the impact evaluation strategy within the policy design at the start of the program. This is precisely the approach taken by the US Congress for the evaluation of the Small Business Innovation Research (SBIR) program. The moment the program was approved in the early 1980s, Congress asked the Small Business Administration (SBA) to ensure that beneficiaries of the first three cohorts be followed up over the next decade (Lerner, 2002).

To obtain these results, the COLCIENCIAS evaluation relied on a unique data set. In particular, the study required matching the administrative records from COLCIENCIAS with the Annual Manufacturing Survey (EAM, from its Spanish name) and the Colombian National Innovation Survey (EDIT, from its Spanish name), both collected and managed by the Colombian Statistical Office DANE. In this way, it was eventually possible to estimate the effect of public funding on firm-level indictors over a 13-year period. The study adopted rigorous impact-evaluation techniques, most notably a specification that aims to remove biases due to firm-level fixed effects (observable and unobservable). In addition, to test the robustness of the results of this specification, the study combined the fixed-effect estimations with the matching of firms' characteristics at the baseline and included placebo tests to check for endogeneity through the estimation of anticipatory effects. One limitation of this approach is that data linking can only be done on already existing data registers. So, in this case, the use of the EAM somehow restricts the analysis to manufacturing firms and firms with more than 10 employees.

Figure 7. The Impacts of COLCIENCIAS on Productivity and Number of Products
(% differences with regards to the control group)



6.3 Multiple Treatments

As discussed in previous sections, when multiple programs are available an evaluator will be interested in estimating not only the individual impact of each alternative but also potential interaction effects between them. The most straightforward approach is to use a simple linear regression model (or panel data model) adding dummies for each treatment, and including interaction terms. For instance, for the case of two programs, $\boldsymbol{T_1}$ and $\boldsymbol{T_2}$:

$$\boldsymbol{Y_i = \alpha + \beta_1 T_{1i} + \beta_2 T_{2i} + \beta_3 T_{1i} * T_{2i} + \gamma X_i + \varepsilon_i}$$

In this context, the coefficients $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ capture the individual effect of each treatment, this is, the mean difference between units that participate in one program and the untreated units:

$$\boldsymbol{\beta_1 = E(Y|T_1 = 1, T_2 = 0, X) - E(Y|T_1 = T_2 = 0, X)}$$

$$\boldsymbol{\beta_2 = E(Y|T_1 = 0, T_2 = 1, X) - E(Y|T_1 = T_2 = 0, X)}$$

The coefficient $\boldsymbol{\beta_3}$ captures the interaction effect of the two programs. More precisely:

$$\boldsymbol{E(Y|T_1 = T_2 = 1, X) - E(Y|T_1 = T_2 = 0, X) = \beta_1 + \beta_2 + \beta_3}$$

Thus, if $\boldsymbol{\beta_3 = 0}$, the effect of simultaneously participating in both programs is simply the sum of the individual effects $(\boldsymbol{\beta_1 + \beta_2})$, while $\boldsymbol{\beta_3} > 0$ indicates the presence of a positive interaction effect.

Two recent evaluations of STIP in Argentina and Chile evaluate the effectiveness of multiple treatments. In the Argentinean case, Castillo et al. (2011) compares the impact of support for process innovation and support for product innovation on employment and employment composition. The study adopts a standard identification strategy based on a combination of PSM and DD. To compare the two possible treatments, the propensity score is estimated through a multinomial logit where dependent variable takes value zero when the firm receives no treatment; value 1 when the firm receives support for product innovation; value 2 when the firm receives support for process innovation. On the basis of this model, the authors define a common support for each treatment through a pair-wise comparison with non-participant firms. The study finds that the program was able to create more and better jobs. It also finds that while the effect of the two types on support on employment is similar, the impact on real wages generated by support for product innovation is more than double the one generated by process innovation.

Following the same approach, Alvarez (2011) not only compares the effectiveness of two Chilean STIP (FONTEC and FONDEF), but also the effect of participating in both programs. The study finds the programs have a positive effect on both employment and productivity growth. It also finds that these effects are heterogeneous across programs: while FONTEC has a positive impact on productivity and employment, but not on wages, FONDEF has had positive impacts on productivity, employment and also on wages. Finally, the study also finds evidence of reinforcing positive effects from both programs, though restricted to the measures of productivity.

6.4 Heterogeneity of Impact

A frequently analyzed case of impact heterogeneity occurs when the effect of the programs may vary across subgroups. As mentioned in previous examples, one may be interested in evaluating whether research funding has differential effects for men and women, or whether the impact of public credit is different for small and large firms. A simple linear regression framework may help understand how to account for these factors. Let the model be:

$$\boldsymbol{Y_i = \alpha + \beta_i T_i + \gamma X_i + \varepsilon_i}$$

where the impact of the program $\boldsymbol{\beta_i}$ is not constant across units. For instance, $\boldsymbol{Y}$ could be some measure of research productivity and $\boldsymbol{T}$ indicates whether researcher i receives funding. Finally, assume for simplicity that $\boldsymbol{X}$ is a dummy variable indicating gender ($\boldsymbol{X_i = 1}$ if female). The case described in the previous paragraph where the impact is suspected to differ between men and women would correspond to:

$$\boldsymbol{\beta_i = \beta + \beta_X X_i}$$

which means that the impact of the program is the sum of an common effect for both groups, $\boldsymbol{\beta}$, and the differential effect for women, $\boldsymbol{\beta_X}$. Combining these two equations leads to:

$$\boldsymbol{Y_i = \alpha + \beta T_i + \beta_X X_i * T_i + \gamma X_i + \varepsilon_i}$$

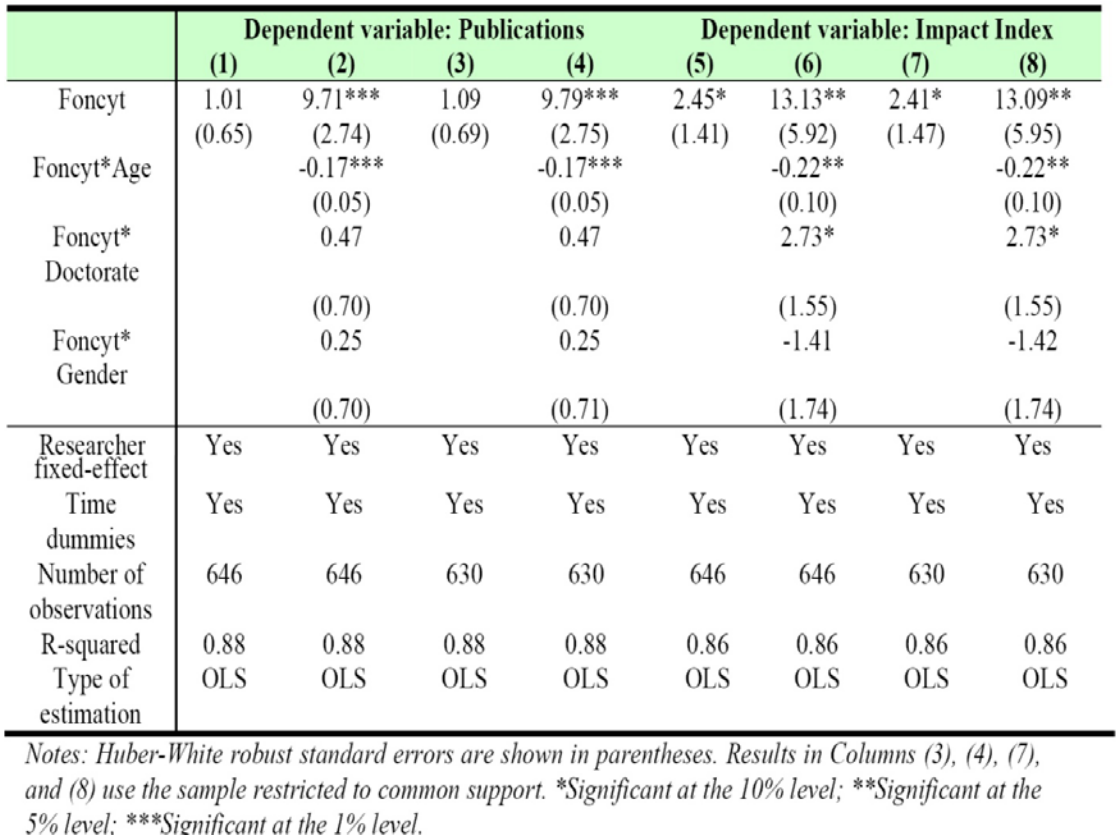
Thus, in this context the heterogeneity of impact may be captured by simply adding an interaction term between the treatment and the subgroups variable. If $\boldsymbol{\beta_X = 0}$ the impact is equal across subgroups, whereas $\boldsymbol{\beta_X} > 0$ would indicate a higher effect for women (and vice versa). The same coefficients can be obtained by running the regression

$$\boldsymbol{Y_i = \alpha + \beta T_i + \varepsilon_i}$$

separately for each subgroup.

For instance, in the paper described in the previous section, Chudnovsky et al. (2006) estimate the differential effects of research funding on researchers by interacting the treatment variable with age, gender and a binary variable for having a Doctorate. Their findings, shown in table 8, reveal that the impact of the program does not differ by gender, but is more important for young researchers. Also, the impact of the subsidy on the quality of research is higher for researchers having a doctorate degree.

Table 8. Heterogeneous Impact of Research Grants



	Dependent variable: Publications				Dependent variable: Impact Index			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Foncyt	1.01	9.71***	1.09	9.79***	2.45*	13.13**	2.41*	13.09**
	(0.65)	(2.74)	(0.69)	(2.75)	(1.41)	(5.92)	(1.47)	(5.95)
Foncyt*Age		-0.17***		-0.17***		-0.22**		-0.22**
		(0.05)		(0.05)		(0.10)		(0.10)
Foncyt* Doctorate		0.47		0.47		2.73*		2.73*
		(0.70)		(0.70)		(1.55)		(1.55)
Foncyt* Gender		0.25		0.25		-1.41		-1.42
		(0.70)		(0.71)		(1.74)		(1.74)
Researcher fixed-effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	646	646	630	630	646	646	630	630
R-squared	0.88	0.88	0.88	0.88	0.86	0.86	0.86	0.86
Type of estimation	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS

*Notes: Huber-White robust standard errors are shown in parentheses. Results in Columns (3), (4), (7), and (8) use the sample restricted to common support. *Significant at the 10% level; **Significant at the 5% level: ***Significant at the 1% level.*

Source: Chudnovsky et al. (2006)

A different way of approaching impact heterogeneity consists of estimating the effects of a program on different parts of the distribution of the outcome using quantile regression methods. Conceptually, this technique is similar to standard linear regression methods but focuses on estimating conditional quantiles instead of conditional means, e.g., the conditional median. This allows evaluating, for instance, whether research funding has different effects on more productive and less productive researchers. Although technical aspects of the estimation involve linear programming and may be cumbersome, quantile regression can be easily performed using several statistical packages (for instance, with the qreg command in Stata). For further details on quantile regression, see Angrist and Pischke, 2009; Koenker and Hallock, 2001, and Koenker and Bassett, 1978.

Although a potential weakness of quantile regression is that it relies on the conditional independence assumption, which in many contexts can be difficult to accept, relatively recent literature discusses possible extensions of this method. For instance, Abadie, Angrist and Imbens (2002), and Chernozhukov and Hansen (2005) present a combination of quantile regression methods and instrumental variables, and Harding and Lamarche (2009) adapt this methodology to panel data.

Firpo (2007) develops a semiparametric method to compute two additional parameters of interest, quantile treatment effects (QTE) and quantile treatment effects on the treated (QTT), which are defined as follows:

$$\boldsymbol{QTE = \Delta_{\tau} = q_{1,\tau} - q_{0,\tau}}$$

$$\boldsymbol{QTT = \Delta_{\tau|T=1} = q_{1,\tau|T=1} - q_{0,\tau|T=1}}$$

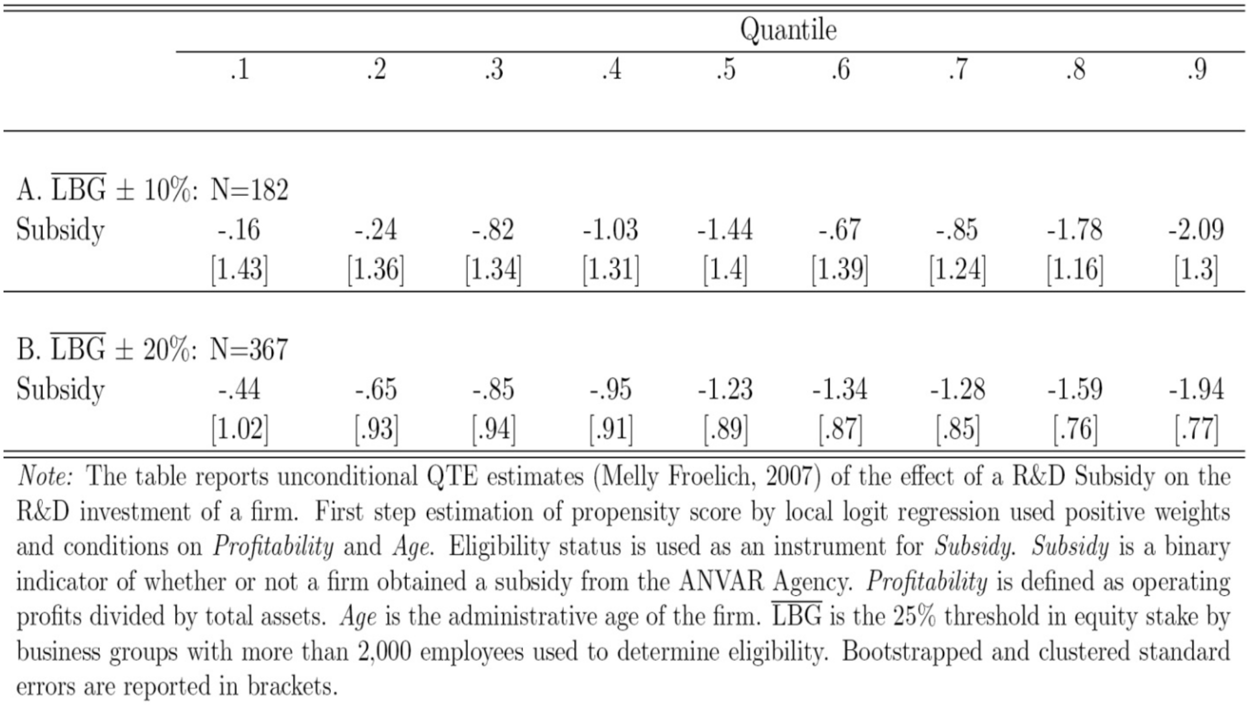
where $\boldsymbol{q_{1,\tau}}$ and $\boldsymbol{q_{0,\tau}}$ represent the $\boldsymbol{\tau}$-th quantile under each potential treatment status, and $\boldsymbol{\tau \in (0, 1)}$. These parameters are analogous to ATE and ATT, but focusing on quantiles instead of averages.

Without going into technical details, the main idea of the methodology is the following. One way to estimate an ATE under the Conditional Independence Assumption is to compute the conditional average effect $\boldsymbol{E(Y_1 - Y_0|X)}$ and then averaging over the distribution of $\boldsymbol{X}$, which yields an estimate of $\boldsymbol{E(Y_1 - Y_0)}$. This is what matching estimators do, for example. However, it is not possible to calculate unconditional quantile treatment effects by averaging the results for the conditional quantiles, since the mean of the quantiles is not the quantile of the mean. Firpo (2007) shows how to define appropriate weights based on the propensity score to compute QTE and QTT without calculating conditional quantiles.

Frölich and Melly (2009) extend the results by Abadie, Angrist and Imbens (2002) and Firpo (2007) to allow for endogenous treatment choice using instrumental variables in the estimation of unconditional QTE.

Serrano-Velarde (2008) applies the method by Frölich and Melly (2009) to study the impact of R&D subsidies on investment decisions of firms. Among other things, the author estimates unconditional quantile treatment effects under heterogeneity and finds that soft loans crowd-out firm R&D investment, especially at the upper tail of the unconditional distribution. These results are shown in Table 9

Table 9. Quantile Treatment Effects
Dependent Varibale: Private R&D Investment (log)



	Quantile								
	.1	.2	.3	.4	.5	.6	.7	.8	.9
A. $\overline{\text{LBG}} \pm 10\%$:	N=182								
Subsidy	-.16	-.24	-.82	-1.03	-1.44	-.67	-.85	-1.78	-2.09
	[1.43]	[1.36]	[1.34]	[1.31]	[1.4]	[1.39]	[1.24]	[1.16]	[1.3]
B. $\overline{\text{LBG}} \pm 20\%$:	N=367								
Subsidy	-.44	-.65	-.85	-.95	-1.23	-1.34	-1.28	-1.59	-1.94
	[1.02]	[.93]	[.94]	[.91]	[.89]	[.87]	[.85]	[.76]	[.77]

Note: The table reports unconditional QTE estimates (Melly Froelich, 2007) of the effect of a R&D Subsidy on the R&D investment of a firm. First step estimation of propensity score by local logit regression used positive weights and conditions on *Profitability* and *Age*. Eligibility status is used as an instrument for *Subsidy*. *Subsidy* is a binary indicator of whether or not a firm obtained a subsidy from the ANVAR Agency. *Profitability* is defined as operating profits divided by total assets. *Age* is the administrative age of the firm. $\overline{\text{LBG}}$ is the 25% threshold in equity stake by business groups with more than 2,000 employees used to determine eligibility. Bootstrapped and clustered standard errors are reported in brackets.

Serrano-Velarde (2008)

6.5 Externalities and General Equilibrium Effects

A given STIP can have side effects. For instance, a subsidy that favors certain types of research and firms may put excluded projects and firms at a competitive disadvantage. During the time they receive public support, the beneficiaries can outpace the followers in a "winner takes all" game. Conversely, there can also be positive spillovers from supported projects to other projects via the transmission of knowledge between firms or because of rent spillovers. In the former case, firms that do the same kind of research (which can be proxied by patenting in similar patent classes or having the same kind of research expertise) are likely to benefit from each other's research. In the latter case, a firm may indirectly benefit from an STIP if it produces inputs for the production of new products or processes of downstream supported innovative firms, or if it has complementary qualifications that are needed by the firms that are directly supported by an STIP. The likely presence of externalities should also be taken into account when devising randomized experiments or when creating appropriate control groups (see Angelucci and Di Maro 2010). A measure like R&D tax credits may only contribute positively to welfare once externalities are taken into account (e.g., see Parsons and Phillips 2007).

Besides externality effects on the performance of other firms or agents in the economy, an STIP can also have general equilibrium effects that should enter the welfare calculation. For example, an R&D grant can raise the wages of scientists and engineers if they are in inelastic supply and thus indirectly increase the cost of doing research in an economy and possibly slow down innovation activity. An intellectual property right, like a patent, leads to a temporary monopoly position with negative effects on competition and consumer surplus. All these factors must be considered when designing the evaluation to properly estimate the different types of impact of a pubic program.

A straightforward way of estimating spillover effects consists on using different control groups. The basic idea is to use a comparison group of non-beneficiary units that may be affected by the spillover effects, and another group of untreated units that is not likely to be affected by these secondary impacts. For instance, one could use a group of untreated firms from the same geographical area (city, region, cluster, etc) than treated firms, and a group of firms from a different area, and compare the results using one and the other comparison group. Then, the difference between these two sets of results will reflect the spillover effects of the program. For further details on estimating spillover effects, see Angelucci and Di Maro (2010).

7. Conclusion

A methodological review must be a continuous work in progress. This guideline is no exception. Although we have made a comprehensive effort to include many of the issues and methodologies related to the impact evaluation of STIP, as pointed out in the introduction, the focus of these methodologies is still on solving the problem of attribution and whether it is possible to say something regarding the effectiveness of these programs.

We recognize that this is just a first step. And integral assessment of STIP also requires the consideration of the efficiency point of view. This can be achieved only through cost-benefit analysis. The methodologies reviewed in this guideline should be seen as "inputs" into the calibration of social cost-benefit analysis later on. In addition, a complete cost-benefit analysis should include not only parameters capturing spillovers effects of these programs (and we say in this guideline how some steps towards this could be taken) but also information on the impacts of these programs on consumer surplus. We have not seen any application of these techniques that solve the attribution problem on the consumer side yet.

On the top of this a further extension of this guideline should include meta-analysis methodologies. This approach might be very useful for the redesign of STIP. To the extent that the evaluation wave of STIP spreads throughout the region, it might become feasible to build datasets of impact evaluations where the unit of analysis is the individual program (rather than the beneficiary) and the control variables are several design attributes of these programs. The present guideline is just the first step in this direction.

References

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

Abadie, A., and Imbens. G. 2008. “On the Failure of the Bootstrap for Matching Estimators”. *Econometrica*. 76(6): 1537-1557.

Almus, M., and Czarnitzki, D. 2003. “The Effects of Public R&D Subsidies on Firms' Innovation Activities: The Case of Eastern Germany”. *Journal of Business & Economic Statistics.* 21(2): 226-236.

Alvarez, R. 2011. Public Programs, Firm Performance and Employment: Evidence from Chile. Mimeographed document.

Angelucci, M. and Di Maro, V. 2010. “Program Evaluation and Spillover Effects”. SPD Working Papers 1003, Inter-American Development Bank, Office of Strategic Planning and Development Effectiveness.

Angrist, J. and Krueger, A. 1991. “Does Compulsory School Attendance Affect Schooling and Earnings?” *The Quarterly Journal of Economics*. 106(4): 979-1014.

Angrist, J. and Pischke, J. S. 2009. *Mostly Harmless Econometrics. An Empiricist’s Companion*. Princeton: Princeton University Press.

Arrow, K. 1962. “Economics Welfare and the Allocation of Resources for Invention,” in R. Nelson, ed. *The Rate and Direction of Inventive Activity*. Princeton: Princeton University Press.

Ashenfelter, O. 1978. “Estimating the Effect of Training Programs on Earnings. *Review of Economics and Statistics*, 60: 47-57.

Banerjee, A. and Duflo, E. 2008. “The Experimental Approach to Development Economics”. NBER Working Paper 14467, National Bureau of Economic Research.

Bhattacharya, S. and Ritter, J. R. 1983. “Innovation and Communication: Signaling with Partial Disclosure”. *Review of Economic Studies*. 50(2): 331-46.

Benavente, J., Crespi, G. and Maffioli, A. 2007. “Public Support to Firm-Level Innovation: An Evaluation of the FONTEC Program”. OVE Working Papers 0507. Inter-American Development Bank, Office of Evaluation and Oversight.

Bérubé, C. and Mohnen, P. 2007. “Are Firms That Received R&D Subsidies More Innovative?”. UNU-MERIT Working Paper Series 015.

Binelli, C. and Maffioli, A. 2006. “Evaluating the Effectiveness of Public Support to Private R&D: Evidence from Argentina”. OVE Working Papers 1106. Inter-American Development Bank, Office of Evaluation and Oversight.

Blanes, J. and Busom, I. 2004. “Who participates in R&D subsidy programs?: The Case of Spanish Manufacturing Firms”. *Research Policy*. 33(10): 1459-1476.

Bond, S. 2002. “Dynamic Panel Data Models: a Guide to Microdata Methods and Practice”. CEMMAP Working Papers CWP09/02, Centre for Microdata Methods and Practice. Department of Economics University College London-Institute for Fiscal Studies.

Bresnahan T. F. and Trajtenberg, M. 1995. “General Purpose Technologies: Engines of Growth?”. NBER Working Papers 4148. National Bureau of Economic Research.

Bronzini, R. and Iachini, E. 2011. “Are Incentives for R&D Effective? Evidence from a Regression Discontinuity Approach”. Temi di discussione (Economic Working Papers) 791. Bank of Italy. Economic Research Department.

Caliendo, M., and Kopeinig, S. 2005. “Some Practical Guidance for the Implementation of Propensity-score Matching”. IZA Discussion Paper 1588. Institute for the Study of Labor.

Carboni, O. 2008. “The Effect of R&D Subsidies on Private R&D: Evidence from Italian Manufacturing Data”. Working Paper CRENoS 200815, Centre for North South Economic Research, University of Cagliari and Sassari, Sardinia.

Card, D. and Krueger, A. 1994. “Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania”. *American Economic Review,* 84(4): 772-93.

Chernozhukov, V. and Hansen, C. 2005. “An IV Model of Quantile Treatment Effects”. *Econometrica*. 73(1): 245-261.

Chudnovsky, D., López, A., Rossi, M. and Ubfal, D. 2006. “Evaluating a Program of Public Funding of Scientific Activity. A Case Study of FONCYT in Argentina”. OVE Working Papers 1206. Inter-American Development Bank, Office of Evaluation and Oversight.

Clausen, T. 2008. “Do Subsidies Have Positive Impacts on R&D and Innovation Activities at the Firm Level?” Working Papers on Innovation Studies 20070615. Centre for Technology, Innovation and Culture. University of Oslo.

Cohen, W. and Levinthal, D. A. 1989. “Innovation and Learning: The Two Faces of R&D. Implications for the Analysis of R&D investment”. *Economic Journal*. 99: 569-596.

Crépon, B., Duguet, E. and Mairesse, J. 1998. “Research, Innovation, and Productivity: An Econometric Analysis at the Firm Level”. NBER Working Papers 6696. National Bureau of Economic Research.

Crespi, G and F. Peirano (2007): Measuring Innovation in Latin America: What We Did, Where We Are and What We Want to Do. Mimeographed Document. IDRC-RICYT.

Crespi, G., Criscuolo C., Haskel J. E., and Slaughter, M. 2008. “Productivity Growth, Knowledge Flows, and Spillovers”. NBER Working Papers, 13959, National Bureau of Economic Research.

Crespi, G., Maffioli, A. and Meléndez, M. 2011. “Public Support to Innovation: The Colombian COLCIENCIAS' Experience”. IDB Publications 38498. Inter-American Development Bank.

Criscuolo, C. 2009. “Direct and Indirect Effects of Innovation Policy”. Presentation. Mimeographed Document.

Crump, R., Hotz, J., Imbens, G. and Mitnik, O. 2009. “Dealing with Limited Overlap in Estimation of Average Treatment Effects”. *Biometrika*, 96(1): 187-199.

Czarnitzki, D. 2002. “Research and Development: Financial Constraints and the Role of Public Funding for Small and Medium-sized Enterprises”. ZEW Discussion Papers 02-74, ZEW - Center for European Economic Research.

David, P., Hall, B. and Toole, A. 1999. “Is Public R&D a Complement or Substitute for Private R&D? A Review of the Econometric Evidence”. Working Paper Series 2050924. Department of Economics. Institute for Business and Economic Research. UC Berkeley.

DiNardo, J. and Lee, D. 2010. “Program Evaluation and Research Designs”. NBER Working Papers 16016, National Bureau of Economic Research.

Duguet, E. 2004. “Are R&D Subsidies a Substitute or a Complement to Privately Funded R&D? Evidence from France Using Propensity Score Methods for Non-experimental Data”. Public Economics 0411007. EconWPA.

European Union (2009). *Design and Evaluation of Tax Incentives for Business Research and Development. Good Practices and Future Developments. Final Report.* Expert Group on Impacts of R&D Tax Incentives, Brussels.

Fabegerber J and Verspagen, B. 2002. “Technology Gaps, Innovation Diffusion and Transformation: An Evolutionary Approach”. *Research Policy*. 31: 1291-1304.

Firpo, S. 2007. “Efficient Semiparametric Estimation of Quantile Treatment Effects”. *Econometrica*. 75(1): 259-276.

Freeman, C. (1987). *Technology Policy and Economic Performance: Lessons from Japan*. London: Pinter.

Frölich, M. 2004. “Finite-sample Properties of Propensity-score Matching and Weighting Estimators”. *The Review of Economics and Statistics*. 86(1): 77-90.

Frölich, M. and Melly, B. 2009. “Unconditional Quantile Treatment Effects under *Endogeneity*”. CEMMAP Working Papers CWP32/07, Institute for Fiscal Studies.

Gertler, Sebastian, G., P and E. Schargrodsky. 2005. “Water for Life: The Impact of the Privatization of Water Services on Child Mortality”, *Journal of Political Economy*. 113(1): 83-120.

Giuri, P. and Myriam, M. 2007. “Inventors and Invention Processes in Europe: Results from the PatVal-EU Survey”. *Research Policy*. 36(8): 1105-1106

González, X. and Pazó, C. 2008. “Do Public Subsidies Stimulate Private R&D Spending?”. *Research Policy*. 37(3): 371-389.

Gorg, H. and Strobl. E. 2007. “The Effect of R&D Subsidies on Private R&D”. *Economica*, 74(294): 215-234.

Griliches, Z. 1979. “Issues in Assessing the Contribution of Research and Development to Productivity Growth”. *Bell Journal of Economics*. 10(1):92-116.

Hall, B. and Van Reenen, J. 2001. “How Effective are Fiscal Incentives for R&D? A Review of the Evidence”. *Research Policy*, 29: 449-469.

Hall, B. and Lerner, J. 2009. "The Financing of R&D and Innovation". NBER Working Papers 15325. National Bureau of Economic Research.

Hall, B. and Maffioli, A. 2008. "Evaluating the Impact of Technology Development Funds in Emerging Economies: Evidence from Latin America". NBER Working Papers 13835. National Bureau of Economic Research.

Harding, M. and Lamarche. C. 2009. "A Quantile Regression Approach for Estimating Panel Data Models Using Instrumental Variables". *Economics Letters*. 104(3): 133-135.

Harris, R., Cher Li, Q. and Trainnor, M. 2009. "Is a Higher Rate of R&D Tax Credit a Panacea for Low Levels of R&D in Disadvantage Regions?" *Research Policy*. 38: 192-205.

Heckman, J., Ichimura, H. and Todd, P. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme". *Review of Economic Studies*. 64(4): 605-54.

Heckman, J., Smith, J. and Clements, N. 1997. "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts". *The Review of Economic Studies*. 64(4): 487-535.

Heinrich, C., Maffioli, A. and Vázquez, G. 2010. "A Primer for Applying Propensity-score Matching". SPD Working Papers 1005, Inter-American Development Bank, Office of Strategic Planning and Development Effectiveness.

Hirano, K., and Imbens, G. 2004. "The Propensity Score with Continuous Treatments". Mimeographed Document.

Holland, P. 1986. "Statistics and Causal Inference". *Journal of the American Statistical Association*. 81(396): 945-960.

Imbens, G. 2009. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)". NBER Working Papers 14896, National Bureau of Economic Research.

Imbens, G. and Kalyanaraman, K. 2009. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator". NBER Working Papers 14726. National Bureau of Economic Research.

Imbens, G. and Lemieux, T. 2007. "Regression Discontinuity Designs: A Guide to Practice". NBER Working Papers 13039, National Bureau of Economic Research.

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

Jacob, B. and Lefgren, L. 2007. “The Impact of Research Grant Funding on Scientific Productivity”. NBER Working Papers 13519, National Bureau of Economic Research.

Jaffe, B. 2002. “Building Programme Evaluation into the Design of Public Research-Support Programmes”. *Oxford Review of Economic Policy*. 18(1): 22-34

Janz, N., Ebling, G., Gottschalk, S. and H. Niggemann. 2001. “The Mannheim Innovation Panels (MIP and MIP-S) of the Centre for European Economic Research (ZEW), Schmollers Jahrbuch, Zeitschrift für Wirtschafts- und Sozialwissenschaften 121, 123-129.

Koenker, R. and Bassett, G.1978. “Regression Quantiles”. *Econometrica,* 46(1): 33-50.

Koenker R. and Hallock, K. 2001. “Quantile Regression”. *Journal of Economic Perspectives*. 15(4): 143-156.

Lach, S. 2000. “Do R&D Subsidies Stimulate or Displace Private R&D? Evidence from Israel”. NBER Working Papers 7943. National Bureau of Economic Research.

Lee, D. and Lemieux, T. 2010. “Regression Discontinuity Designs in Economics”. *Journal of Economic Literature*. 48(2): 281-355.

Lerner, J. 1999. “The Government as Venture Capitalist: The Long-Run Impact of the SBIR Program”. *Journal of Business*: 72(3): 285-318.

Lerner, J. 2002. “When Bureaucrats Meet Entrepreneurs: The Design of Effective Public Venture Capital Programmes”. *The Economic Journal*. 112: 73-84.

Lokshin, B. and Mohnen, P. 2010. “How Effective are Level-based R&D Tax Credits? Evidence from the Netherlands”. UNU-MERIT Working Paper Series 040.

Loof, H. and Heshmati, A. 2005. “The Impact of Public Funds on Private R&D Investment: New Evidence from a Firm Level Innovation Study”. Discussion Papers 11862, MTT Agrifood Research Finland.

Inter-American Development Bank. 2008. Industrial Policies in Latin America and the Caribbean. RES Research Projects. http://www.iadb.org/en/research-and-data/project-details,3187.html?id=3776.

Lundvall, B.A. 1992. *National Systems of Innovation: Towards a theory of Innovation and Interactive Learning.* London: Pinter.

Martin, S. and J. Scott (2000). “The Nature of Innovation Market failure and the Design of Public Support for Private Innovation”. *Research Policy*. 29: 437-447.

Navarro J. C. and Llisterri, J. 2010. “The Importance of Ideas: Innovation and Productivity in Latin America”.In C. Pages, editor. *The Age Of Productivity: Transforming Economies from The Bottom Up.* Washington, DC: Inter-American Development Bank. New York: Palgrave Macmillan

Nelson, R. 1959. “The Simple Economics of Basic Scientific Research”. *Journal of Political Economy*. 67: 297-306.

Nelson, Richard R, 1981. “Research on Productivity Growth and Productivity Differences: Dead Ends and New Departures”. J*ournal of Economic Literature, American Economic Association*. 19(3):1029-64.

Nelson, R. 1993. National Innovation Systems: A Comparative Study. Oxford University Press, New York.

Organization for Economic Co-operation and Development, OECD. 2010. R&D Tax Incentives: Rationale, Design and Evaluation. Paris:OECD/Eurostat.

Parson, M. and Phillips, N. 2007. “An Evaluation of the Federal Tax Credit for Scientific Research and Experimental Development”. Working Paper 2007-08, Department of Finance, Canada.

Reiss, P. and Wolak, F. 2007. “Structural Econometric Modeling: Rationales and Examples from Industrial Organization”. In J. Heckman and E. Leamer, editors. *Handbook of Econometrics,* edition 1, volume 6, chapter 64. Elsevier.

Rosenbaum, P. R., and Rubin, D. B. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects”. *Biometrika.* 70: 41–55.

Sagasti, F. 2011. Conocimiento y desarrollo en América Latina: Antecedentes, evolución y perspectivas de las políticas de ciencia, tecnología e innovación. Ciudad de México: Fondo de Cultura Económica.

Serrano-Velarde, N. 2008. “Crowding-Out at the Top: The Heterogeneous Impact of R&D Subsidies on Firm Investment”. Mimeographed Document. European University Institute.

Smith, K. 2006. “Measuring Innovation”. In J. Fagerberg, D. Mowery and R. Nelson, editors. The Oxford Handbook of Innovation, Oxford: Oxford University Press.

Soete, L; B. Verspagen and Ter Weel B. 2010. “Systems of Innovation”. In: B. H. Hall, and N. Rosenberg, editors. *The Economics of Innovation*. Amsterdam: North Holland.

Steinmuller, E. 2010. “Economics of Technology Policy”. In: B. H. Hall, and N. Rosenberg, editors. *The Economics of Innovation*. Amsterdam: North Holland.

Stiglitz, J. and Weiss, A. (1981). “Credit Rationing in Markets with Imperfect Information”. *American Economic Review*. 71(3): 393-410.

Toivanen, O. 2009. “Innovation Policy, Entrepreneurship, and Development: A Finnish View”. UNU-MERIT Working Paper Series 050.

Trajtemberg M. 2006. “Innovation Policy for Development: An Overview”. WP 6-06. Foerder Institute for Economic Research.

Ubfal, D. and Maffioli, A. 2010. “The Impact of Funding on Research Collaboration: Evidence from Argentina”. SPD Working Papers 1006, Inter-American Development Bank, Office of Strategic Planning and Development Effectiveness.

Van Pottelsberghe B. and G. Dominique, 2007. “The Economics of the European Patent System: IP Policy for Innovation and Competition”. ULB Institutional Repository 2013/6183. Universite Libre de Bruxelles.

Van Pottelsberghe B, E. Megally and S. Nysten. 2009. ”Evaluation of Current Fiscal Incentives for Business R&D in Belgium”. Working Paper WP-CEB 03/01. Universite Libre de Bruxelles, Solvay Business School, Centre Emile Bernheim.

Wallsten, S. 2000. “The Effects of Government-Industry R&D Programs on Private R&D: The Case of the Small Business Innovation Research Program”. *RAND Journal of Economics*. 31(1): 82-100.

Wooldridge, J. 2002. Econometric Analysis of Cross Section and Panel Data. MIT Press.