VATT-TUTKIMUKSIA
90
VATT-RESEARCH REPORTS

Takis Venetoklis

PUBLIC POLICY EVALUATION: INTRODUCTION TO QUANTITATIVE METHODOLOGIES

Valtion taloudellinen tutkimuskeskus
Government Institute for Economic Research
Helsinki 2002

ISBN 951-561-410-4
ISSN 0788-5008

Valtion taloudellinen tutkimuskeskus
Government Institute for Economic Research
Hämeentie 3, 00530 Helsinki, Finland
Email: takis.venetoklis@vatt.fi

Oy Nord Print Ab
Helsinki, June 2002

TAKIS VENETOKLIS. Public policy evaluation: Introduction to quantitative methodologies. Helsinki, VATT, Valtion taloudellinen tutkimuskeskus, Government Institute for Economic Research, 2002 (ISSN 0788-5008, No. 90). ISBN 951-561-410-4.

Abstract: This paper is a survey which describes and explains in non-technical terms the logic behind various methodologies used in conducting retrospective quantitative evaluations of public policy programs. The programs usually have as their main targets firms or individuals who benefit from direct subsidies and/or training. It is hypothesised that because of the technical nature of quantitative evaluations, some of the public officials to whom these evaluations are intended, may find them too complex to comprehend fully. Hence, those officials might disregard them up front, or form a biased opinion (positive or negative) or even accept the results on their face value. However, because all evaluations are subjective by definition, the public officials should have some basic knowledge on the logic behind the design and context of evaluations. Only then, can they judge themselves on their worth, and consequently decide to what degree they will take into account their findings and recommendations. The paper initially discusses the issues of accountability and causality and then introduces policy evaluation as a two phase process: First, *estimations* are made on the potential impact of the policy in question and then a *judgement* is passed on the worth of the impacts estimated, through a cost benefit analysis. The estimations in turn, comprise of two related areas: the *design* of the evaluation and the *model specification*. In designs, one has to consider whether counterfactual populations are included or not and whether the impact variables are in cross-sectional or longitudinal format. In model specifications the evaluator must decide which independent control variables he will include in the regression model so as to account for selection bias. In cost benefit analysis decisions have to be made as to whether the analysis will be made at partial equilibrium or general equilibrium level and whether the judgements formulated will be based purely on efficiency grounds or using just distributional criteria as well. The paper recommends among others, that (a) public policy evaluations should establish clear rules of causation between the public intervention and the potential impact measured, (b) limitations in the estimation and cost benefit analysis phase must be explicitly stated and (c) retrospective evaluations should be conducted at closer intervals after the end of the intervention so as to reduce the external heterogeneity generated due to the time lag between the results produced and the on-going programs.

Key words: Public policy evaluation, quantitative methods, cost benefit analysis

Tiivistelmä: Tämä raportti on katsaus, jossa kuvataan julkisen sektorin ohjelmien arviointitutkimuksissa käytettyjen eri kvantitatiivisten tutkimusmenetelmien taustalla olevaa logiikkaa. Ohjelmat on pääasiassa kohdistettu yrityksille ja henkilöille, jotka hyötyvät suorista tuista tai koulutuksesta. Kvantitatiivisten evaluointien menetelmät voivat muodostavat liian suuren ymmärtämiskynnyksen osalle niistä virkamiehistä, joille raportit on tarkoitettu. Tämän takia nämä virkamiehet saattavat jättää ne lukematta tai kieltävät johtopäätökset, mikä muodostaa ennakko-oletusten mukaisen vääristyneen kuvan tai mahdollisesti hyväksyvät tulokset kiinnittämättä huomiota niiden arvoon. Koska arviointitutkimukset ovat määritelmällisesti subjektiivisia, virkamiehillä tulisi olla perustiedot niiden suunnittelussa käytetystä logiikasta. Vain siten he voivat päätellä, miten arviointi on tehty ja missä määrin sen tuloksia ja suosituksia voidaan ottaa huomioon. Raportin alussa pohditaan tili- ja vastuuvelvollisuuteen ja kausaliteettiin liittyviä kysymyksiä. Tämän jälkeen esitellään arviointitutkimus kaksivaiheisena prosessina: ensin tehdään estimointi kyseisen ohjelmien seurauksista ja sitten arvioidaan toimenpiteen onnistumista kustannushyötyanalyysia käyttäen. Estimointi puolestaan sisältää kaksi osaa: arviointimenetelmän suunnittelu ja arviointimallin spesifiointi. Arviointitutkimusta suunniteltaessa tutkija punnitsee sisällytetäänkö evaluoinnin piiriin toimenpiteen vaikutuksen ulkopuolelle jäävä joukko ja käytetäänkö vaikutusta mittaavista muuttujista poikkileikkaus- vai paneeliaineistoa. Mallia tarkennettaessa hän ratkaisee mitä kontrollimuuttujia regressiomalliin sisällytetään korjaamaan otosvirhettä. Kustannushyötyanalyysissä joudutaan valitsemaan tehdäänkö analyysi yleisellä vai osittaisella tasapainomallilla ja tehdäänkö arviot puhtaasti tehokkuusperusteisesti vai käyttäen myös jakopoliittisia kriteerejä. Raportissa suositellaan että (a) ohjelmaevaluaatiota varten tulisi muodostaa ja vakiinnuttaa selvät kausaalisuhteita koskevat säännöt intervention ja niiden vaikutusten mittaamiseksi, (b) estimoinnissa ja kustannushyötyanalyysissa rajoitukset olisi tutkimuksessa selkeästi tuotava esiin, (c) jälkikäteisevaluointi tulisi tehdä nopeammin uudistuksen voimaantulon jälkeen, jotta nykyisten ohjelmien ja evaluoinnin tulosten välisestä aikaviiveestä johtuvaa ulkoista heterogeenisuutta voidaan vähentää.

Asiasanat: Arviointitutkimus, kvantitatiiviset metodit, kustannushyötyanalyysi

Table of contents

1. Introduction[1]

I begin with two simple policy examples.

Example 1

Suppose we have a group of unemployed individuals who are given training and then they are told that they should look for a job (or are placed in firms to obtain on-the-job experience/training). Six months after the completion of the training program we check again their employment situation, and find that 40% of the group are still working.

Are we to conclude that 40% of those that were unemployed before the training, found work *because* of the training they went through?

The treatment (the training) certainly pre-existed of the effect (the increase in employment). But in order to make a comprehensive evaluation one needs to take into account many other things. For example, how can we know that these individuals would not have found a job anyway without the treatment (the training they went through?) That is, how can we isolate the effect (the fact that 40% of these people are working) so it can be attributed solely on the treatment?

What if 30% of the trained were still at work just a week before we measured the post treatment employment levels of the group, but were then laid off due to reasons beyond our control?

What if 20% of those participants in the training program will get a job in the next month (that is after our effect measurement?)

But let us move even further and suppose that we have found that the pure impact of the training program was indeed 40%. Is this a “good” effect? An acceptable figure? How are we to qualify it? How are we to judge its magnitude?

If the training program cost for example € 3000 per participant and the whole administrative costs were another € 2000 per participant is that an acceptable cost to society in general and/or to the public policy makers/implementers in particular?

What about the net costs to the public sector which are reduced because of the taxes these employed individuals pay whilst at work and the savings made by not having to pay unemployment compensation?

How should one value changes in individual utility (the benefit one gains from being employed after a long term of idleness)? What is the value of the psychological improvement that this individual experiences? Can this even be quantified?

Is society altruistic (or indifferent) enough to redistribute some of its resources (tax money) for such programs although at fist glance they do not seem to generate tangible monetary returns? What is the threshold against which this can be measured?

Example 2

In most countries there are many public policies which are geared to promoting investments, R & D and employment of firms. Let us assume that a firm applies to get a grant for such purposes, and receives some money (a business subsidy) through a Ministry/Agency to partly finance an investment.

Assume further, that just before the distribution of the subsidies we acquire the financial statements of that particular firm for the last 3 years prior to the current one and record her profits each year. Two years after the receipt of the subsidy we take a look again at the financial statements of that firms and record once more its yearly profits after the receipt of the subsidy. We notice that the profit growth rate before the receipt of subsidies is at 15% per year and after the receipt it jumps to 20%. Can this 5% increase in profit growth be attributed to the receipt of the subsidy?

[1] The paper has benefited from the comments of Jaakko Kiander, Aki Kangasharju and Heikki Räisänen. All opinions expressed and potential mistakes found in the text, are the sole responsibility of the author.

If the subsidy coverage in the financing of the project, is 50% how can we account for this in our impact estimation? Have the general business conditions been more favourable or less favourable after the receipt of subsidies versus the period before the receipt?

How have the profits of other firms that have not received subsidies and are in the same sector, operating in the same geographical region as our firm in question faired during the same periods (pre and post subsidies)?

After a careful analysis we find that one Euro of extra subsidies increases the firm's profits by 1.3 Euro. On the other hand if we measure at a general level the amount of administrative expenses directly attributed to the whole business program we find that on average we have one half Euro (0.5) of costs. What kind of judgement do we pass on as evaluators for the program as such? Do we conclude that the program has been successful in its impact, but unsuccessful in its strategic implementation? Do we conclude the opposite?

Purpose

Results and recommendations produced through public policy evaluations - especially those unfavourable to the policy under investigation - are frequently not used instrumentally[2] by the policy decision makers, planners and implementers. The reasons are surely numerous, including some which run in the spheres of bureaucratic, political or even interest group theories. Yet another hypothesis could be that because of the technical nature of quantitative evaluations part of the intended audience simply does not have the expertise to actually judge the worth of the results produced. If some of the public administrators and decision makers do not comprehend how the results were produced (the methodologies applied), consequently they can not judge the soundness and quality of the policy recommendations presented. It is also frequent, that due to overloaded work schedule, officials do not read carefully the contents of the studies but just glance the abstract and the conclusions. Indeed, the technical nature of some evaluations enhances this behaviour. Thus, biased opinions on the results are formulated, some are disregarded up front, where as others are accepted on their face value.

However, one needs to have a critical approach to any evaluation study, because by definition - as we shall soon discuss -, evaluations are based on subjective criteria and assumptions. The reader (who may also happen to be a decision policy maker) should have a basic knowledge on certain methodological issues so that he can judge himself the worth of the evaluation presented to him and then decide to what degree he will take into account the findings and recommendations of the evaluation.

The main purpose of this paper is to describe and explain the logic behind several quantitative methodological issues linked to the conduct of evaluations of public program interventions. The programs have as their main target unit, firms or individuals who benefit from direct public subsidies and/or training programs geared towards them. In other words, the material in the paper suits evaluations on policies dealing with labour and business subsidies government interventions.

Another purpose is to put forward some ideas on how just this type of evaluation can be utilised in a more effective way. The paper includes recommendations (a) on the areas that the reader of a technical evaluation study should focus upon, and (b) on the timing that it should be conducted.

The approach is non-technical. Mathematical formulas and statistical models are presented only when their absence would obscure even more the description of the relevant issue. They are nevertheless unavoidable and one should at least attempt to comprehend their basic features, since they will always be present in technical quantitative evaluation reports.

Scope

The conduct of public policy evaluation is a complex issue. To narrow things down, I focus the discussion on evaluations which are (a) *retrospective*, (b) *quantitative*, (c) are based on *non-experimental methods*, (d) are conducted at *micro level* and (e) utilise *secondary data sources*. The reasons why I have chosen to deal with just this type of evaluations, having just these type of characteristics, will be discussed in the last section of the paper.

[2] Instrumental use means that evaluation results and recommendations on a particular policy intervention are actually used for action which somehow effects elements of that policy (Vedung, 2002).

Contents

In Section 2 one finds issues relevant the logic behind the need for public policy evaluation. Why does one conduct evaluation in the first place? In this respect the notions of accountability and the retrospective nature of evaluation are examined.

The questions raised in the two examples earlier, will be discussed in more detail later on. However, it is evident that for someone to conduct an evaluation of a public intervention he needs:

- to identify causal relationships between the treatment (the public intervention) and the effect (the potential impact)
- to estimate the true and net impact of the treatment by isolating it from other confounding factors and
- to pass a judgement on the worth of this isolated "netted" impact through a cost benefit analysis

Sections 3, 4 and 5 follow this approach. Section 3 discusses causality issues. When one conducts public policy evaluation he measures a potential impact. Is the alleged impact *caused* by the intervention? How does one checks that this is so? What are the criteria? Section 4 is the longest of the study. It covers among others, several designs used in estimating policy impacts. Since we are dealing with quantitative evaluations, we are bound to discuss regression models. Here however, I focus on the design logic and the specification of the models rather than on their statistical and mathematical properties. In Section 4.5 I present a fictional example and show how in practice some issues previously mentioned are handled in empirical work. Section 5 deals with cost benefit analysis, that is, methods which can assist the evaluator put a value on the previously estimated impacts and thus pass a judgement of the policy evaluated.

Section 6 concludes with a discussion on the scope of the paper as well as thoughts and recommendations on how technical evaluations can be utilised in a more effective way. In the Appendix one finds the formal notation of several equations relevant to the non-experimental estimation designs of Section 4.3.2.

Audience

The paper is geared towards public sector officials who are involved directly or indirectly with the planning, implementation, and evaluation of public policies. However, the audience is not limited solely to public officials; students and researchers may find the material of the paper useful. Especially, the references cited contain many basic but also some recent methodological advances in the field of public policy evaluation.

The words *Intervention* and *Treatment* are used interchangeably in the text and mean the same thing. The same applies to *Impact* and *Effect*. Although *Policy* is a somewhat wider notion than *Program,* they are also used interchangeably.

2. Why evaluate and what does evaluation mean for this paper[3]?

This section attempts to demonstrate the basic reasons why one conducts public policy evaluations. In the course of this brief analysis, the retrospective meaning to evaluation is brought forward as well.

The formation of policy accountability

A simple assumption generally acceptable is that most citizens, tax payers that is, think and behave rationally. Hence, they would like to see that their tax moneys are spent by their democratically elected government effectively and efficiently in whatever policies the government adopts and finances. Put it differently, citizens, tax payers would like to have an *account* of these policy expenditures. If a policy proves inefficient[4], the citizens would want it altered so it can become efficient; or maybe they would rather see those funds shifted to other policies; or they would even like some of those funds to be returned indirectly to them through lower taxes.

But in a representative democracy, the notion of accountability is needed by others as well. Think, for instance of a system where we have as members the Citizens, the Parliament, the Government, the Ministries, the local units of the ministries and the final target of an adopted policy (Figure 1). When the policy is voted by the Parliament and appropriations are reserved for it, between each of these aforementioned members of the system evolves a so called Principal - Agent relationship. This relationship is more common in economics and contract law where it is formally described and defined. Here, I would like to use it informally, as a conceptual framework to explain the needs that arise within such a system. In this relationship, the Principal delegates some of his own operations as well as the promotion and safeguarding of his interests to the Agent because it is practically impossible to do all this himself due to time, information or resource constraints. In our system, the relationship is of dual direction in some cases and of one direction in some others. The Parliament for example can be thought of as an Agent of the citizens by whom it is elected, but also as a Principal of the Government whom it controls. In turn the Government could be a Principal of its Ministries but an Agent of the Parliament, and so forth. At the two ends of the system we have one-directional relationships. The citizens are only Principals of the Parliament and the final targets of the policy can only be Agents of the Ministry and its units.

Because of this relationship, the need for *accountability* we discussed above comes into the picture. Simply put the Principal at all levels would like to know how his assignments, how his orders *have been carried out* by his Agent. He wants to know whether the Agent has exercised the delegated powers given to him, whether he has exercised his duties properly.

Policy accountability through evaluation

To have this accountability substantiated, the Principal must conduct some *evaluation*, he must research systematically, find out what *has happened* and then pass some judgement on the policy in question.

Assuming as stated earlier that the Principal thinks and behaves rationally, these decisions and judgements are optimised only if the Principal possesses valid, reliable and comprehensive information on how the policy has faired; that is whether it did well or not, what were its weaknesses, what were its strengths, and so forth. How is this reliable information produced? It is produced through sound methods with which the Principal can gather data and methods with which he can analyse the data gathered.

Evaluations however are not conducted for the sake of accountability only. For example, the organisations and their public officials that are involved in the planning and implementation of policies, whether acting as Principals or Agents, would also want some feedback which would assist them in improving their on going policy operations or the planned operations for the future[5].

---

[3] Part of this section is based on Venetoklis (2002).

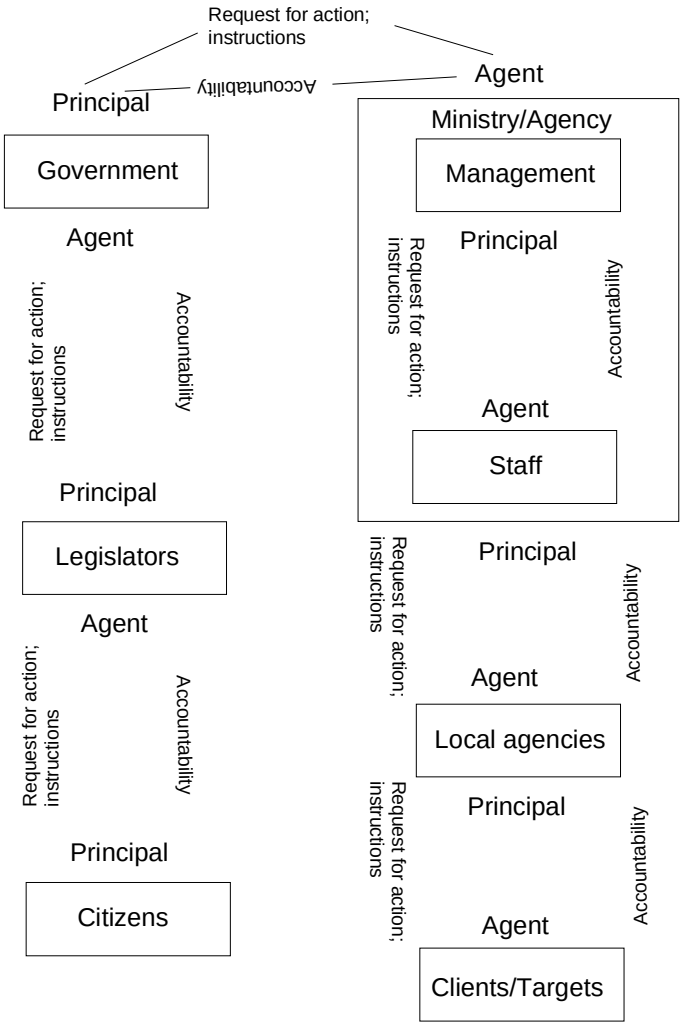
[4] The notion of inefficiency (and efficiency) is not used here based solely on results from a financial cost benefit analysis; that is from an analysis showing a positive or negative net flow of funds linked to that particular policy. The notion of (in)efficiency is used at a more general level denoting the acceptance or rejection of a policy using many other criteria as well. We shall discuss this approach again in Section 5.3.

[5] In public policy, accountability has yet another dimension which we shall not deal with here. It is the formal accountability and procedural correctness of the administrative side of the program which is examined through auditing procedures.

Evaluation characteristics

Because of the accountability notion, evaluation in this paper naturally has a retrospective characteristic in it. I do not examine ex-ante (before the intervention) nor ex-nunc (during the implementation of the policy) definitions of evaluation and I do not adopt the methodologies applied during such exercises. Thus, for this paper, *evaluation refers (a) to the systematic retrospective examination of certain indicators on the potential impacts of an implemented public policy intervention as well as (b) the passing of judgements on the worth of the measured impacts.*

Figure 1. Accountability and request for action-instructions: The dual role of principals and agents when a policy is launched



Source: Vedung (1997, p. 107), somewhat amended.
Note: The directions of accountability, and request for action and instructions are as read.

3. Causality

The notion of causality is of paramount importance in policy evaluation and especially in impact estimation. This section examines several issues linked to causality and their relation to the quantitative approach of evaluating public policy interventions. Because econometric models (basically regression models) are the main tools utilised with these quantitative methods, I describe below the logic behind these models. Some of the issues discussed in Section 3.1, especially those referring to causation and regression models will be mentioned again in Section 4, when dealing with quantitative methods used in estimating policy impacts.

Basic definitions

Causality is the relationship between a cause and its effect. Causation is the act which produces an effect. Cause is something that brings about an effect. What does it mean when something is causal? It expresses or indicates a cause[6]. Causality is a slightly more general term than causation. The difference is subtle, however. In what follows, I will be using both definitions interchangeably.

3.1 Causation through regression models

In policy evaluation we are interested in finding the (potential) impact that has actually been produced by a certain policy intervention. In "causality" terms that means finding the relationship between the policy intervention (cause - X) and the impact (effect - Y).

How exactly do econometric models measure this effect in practice and account for causation?[7] The whole idea is to examine whether a change or variation in variable X (the potential cause) has indeed *caused a change* or variation in Y (in the impact variable of interest). The regression model[8] normally used is

$Y_i = \beta_0 + \beta_1 * TREAT_i + (\beta_2..\beta_n) * C_i + \varepsilon_i$, where

Y_i is the effect indicator, TREAT(ment) equals 1 if unit i is exposed to the treatment (the policy) and 0 if it is not, C_i is the set of control variables, $\beta_2..\beta_n$ is a vector (a set) of regression coefficients, and ε_i is the error term with a normal distribution, zero mean and constant variance. The estimate of β_1 of the intervention/treatment variable (TREAT) is the estimate of the average causal effect, adjusted for the effects of the vector of control variables C_i. β_0 is the intercept.

Above, TREAT was a binary variable taking values 0 or 1 only. However there are cases where TREAT can represent a "dose"; that is, it can be a continuous variable. For example, in the case of a training program for unemployed individuals, TREAT can be 1 if the unit (the unemployed person) participated in the training program and 0 if he did not. Also, if continuous, TREAT can represent the number of days/weeks the individual was in training, with those that were not trained represented with 0 training days/weeks. The amount of days/weeks could fluctuate depending on the training and the willingness of the individual to participate in the training.

A similar case could be with a business subsidy program to firms. TREAT could be binary, that is 1 if a firm received subsidies and 0 if it did not. Or it could represent a measure of the amount of subsidies received per firm. The amount would normally vary among firms due to the nature of the project partly financed by the subsidy and the percentage of the total investment cost covered with subsidies. Again as with the non-trained individuals, the firms not receiving subsidies could be represented with 0 (zero) subsidies.

The notion of ceteris paribus - that is holding all other (relevant) factors fixed - is the *crux* of establishing a causal relationship. If we focus on the average, or expected impact, a ceteris paribus analysis means estimating the expected value of Y conditional on (given) TREAT and C, or E(Y|TREAT,C). The vector C denotes a set of control variables that we would like to *explicitly hold fixed* when studying the effect of TREAT on the expected value of Y.

---

[6] Merriam-Webster dictionary (2002).
[7] In Wooldridge (2002) and Reiter (2000) one finds comprehensive, explicit, and easy to follow descriptions.
[8] This format is used when we have cross-sectional data, that is, observations from one or two points in time.

If TREAT is binary we are interested in the average difference in outcome Y, between the treatment and no treatment condition (between conditions when TREAT = 1 and when TREAT = 0), and given that the control variables C are fixed. If TREAT is continuous we are interested in the *partial effect* of TREAT on the average value of Y given the control variables C[9].

The dimension of C (how many control variables we use) is not important. We can use one or more. However, deciding on the list of *proper* controls is not always straightforward. *Using different controls can lead to different conclusions about a causal relationship between Y and TREAT.* This is where establishing causality becomes ambiguous: it is up to *the evaluator* to decide which factors need to be held fixed. If we settle on a list of *relevant* controls to the policy problem at hand, and if all these relevant controls are elements of C and *can be observed*, then estimating the partial effect of Y on TREAT given C, or E(Y|TREAT,C), is relatively straightforward. Unfortunately in policy analysis, many potential control variables are not always observed; some hypothetically relevant control variables even, can never be observed. Nevertheless as will be repeated below, theory plays a central role in the selection of the control variables utilised.

3.2 Effects of causes

It is important to make a distinction here. In public policy evaluation, we are interested in the *effects of causes*, that is, in the impacts of interventions. We are not investigating *the causes of effects*, that is we are not interested in what are the potential factors that *may* influence a certain impact outcome (Holland, 1986)[10],[11]. This may be initially a difficult notion to grasp but one may present it easier using the following phrases. Assume that we are evaluating a business subsidy policy on firms and the impact indicator which interests us is the profit of those firms receiving subsidies.

(a) In examining the *effects of causes* we ask the question: How much has the profit of firms changed due to their receipt of business subsidies?
(b) In examining the *causes of effects* we ask the question: What factors are significant in the change of the profit of firms?

With (a), we examine and measure the impact of a policy intervention (in this example the receipt of subsidies) on an indicator of our choice (the profit growth of firms). The impact might turn out to be big, small, with a positive or negative sign, but nevertheless something which is *a priori* accepted as such. Other potential factors that themselves might influence the profit growth are used as controls. The logic behind the choice of these factors is something we shall discuss later on[12].

With (b), we explore potential factors that might influence the profit growth of firms, among which *one* is the policy intervention I mentioned in (a), the receipt of subsidies. In (b) though, we are not looking a priori for a specific intervention or factor; all potential factors known and available to us are included in the model. However, *after* the estimation we focus our attention *only* on the ones whose coefficients turn out to be statistically significant; we decide that the insignificant factors are not in our interest as such, at least under the circumstances bound by our model. The potential problems with approach (b) is that we may later discover some other factor(s) which, if we include in our model, may turn out themselves significant and reject our previous estimation results.

---

[9] For those with basic knowledge of calculus, this is the derivative of Y on TREAT given C; the formal notation is $\partial E(Y|TREAT,C)/\partial TREAT$.

[10] Holland presented this and other arguments on causation in his seminal 1986 paper; his approach, I shall follow here. However, the reader must be aware that this is just one of many theories from several disciplines dealing with the subject. Another approach for instance claims that causation is established by non-elimination. This approach advocates that if we want to establish that TREAT is a genuine cause of Y we must see that the dependence initially observed between the two (the value so to say of the β coefficient of TREAT can not be eliminated through the introduction of one or more variables (potential other causes, confounders) in our analysis. (see for example Suppes, 1970, cited in Goldthorpe, 2000, p.2); or as the Treasury Board of Canada (2001, p. 14) put it "…making causal inferences about results in evaluation means rejecting or accounting for rival plausible explanations…".

[11] A good starting point for causation theories is of course Holland's paper. In addition one can read the survey type articles of Goldthorpe (2001), Arjas (2001) and Cox and Wermuth (2001). For an advanced presentation of causal parameters as utilised in econometric policy analysis see the survey by Heckman (1999b).

[12] In many public policy evaluation studies that use quantitative methods (econometric models) to measure the policy's impacts, one notices that in the tables where the models are demonstrated, only the parameters which refer to the treatment variable are presented; the rest of the data on the control variables are not. The reason is twofold. One is that in cases where there are many control variables, their listing would clatter the table and make the reading more cumbersome. Most importantly though, exactly because in public policy evaluation we are interested in the effects of the causes (the (a) above), we concentrate and emphasise (show) only data referring to the variable of interest; that is, the parameters and statistics of the treatment/impact/intervention variable.

3.3 Establishing rules for causation

Reading the previous section one comes with the impression that the establishment of causation in a policy impact framework is more of a matter of *relative* rather than of *absolute* confidence. This is indeed the case because we can never observe what would have happened to the treated population had the policy not occurred[13]. This is what Holland (1986) has called as the *fundamental problem of evaluation*.

So, how can we in fact ensure that the relationship we have discovered is indeed causal? With an experimental design[14], this can be assured by allocating the treatment at the beginning of the experiment on a random basis. Providing that we have enough observations in both groups (treated and untreated) their whatever differences are cancelled out and we are left with the net clean impact we are looking for. One need not apply complicated statistical methods once the experiment is conducted under those terms. Given that in the natural sciences, experimental designs are used in studies all the time, Schmidt (2001, p. 22) notes that

> "…generations of natural scientists have been raised with the conviction that, if an experiment needs any statistics, one simply ought to have done a better experiment".

However, in non-experimental[14] designs this random assignment of treatment and control is unobtainable by definition. What is attempted instead, is to model the different factors which themselves might be correlated with the impact measurement and then, based on the ceteris paribus principle (other things being equal) isolate and identify as close as possible the true impact of the intervention in question. How close is that estimate to the true impact depends on how we specify the variables in the model, the functional form that we shall give them, the interactions that we might identify, the type of data we have in our hands, the research design we apply, the statistical tests we use to control for biases due to model specification, etc.

What this all comes to, is that in evaluations of public policies utilising non-experimental quantitative methods of analysis, a lot is based on *subjective judgements and assumptions.* This in turn forces us to be very careful with our conclusions. Hence, a list of additional conditions or "rules" has been devised, which if applicable to the relevant evaluation, enhances the validity and credibility our the results. A well known example of such rules was presented by Hill (1965, cited in Goldthorpe, (2001)). Although the conditions were initially meant for epidemiological experimental studies, they are nonetheless applicable for the types of evaluations we are dealing with in this paper. Hill basically said that if one observes a *dependency* between variable X (independent – treatment) and variable Y (dependent – effect), the dependency is *more likely to be causal if*

- there is already *in theory* a causal explanation of the same relationship between X and Y. This means that the same effect is found somewhere else (in other independently conducted studies/evaluations) but conducted under slightly different conditions. The theoretical framework under which many causal explanations are based, can not be emphasised enough. It is the most frequent starting point of all policy impact evaluations. Especially in the case of policies on firms and individuals, microeconomic theory on their behaviour plays a pivotal role in establishing causal relationships
- the effect observed is a large monotonic one (negative or positive); that is, if the effect is large, it is less likely that there is an alternative explanation via an unmeasured, unaccounted variable
- the intervention is massive and some effect is indeed observed; that is, the X is huge and the change in Y is noticeable
- there are no specific interactions among the variables in the equation; that is the control factors do not turn statistically significant if they are interacted among themselves[15].

A fifth assumption that is implicitly accepted in all causality arguments, is that the independent variable X (the treatment) must have *time precedence* in occurrence from the dependent variable Y (the effect). In regression models even when we compare X and Y values taken the same time (say, year) we assume that

[13] That is, unless we were to discover time travel. Had this been the case, we would first simply let things be (status quo). We would observe the change of the impact indicator of interest on the (would be) treated population from time t to t+ i in the future, and record it. We would then go back at t (travel back in time) and now intervene with the policy treatment. We would then again observe the policy impact indicator till time t+i. The what ever difference of the indicator measured in the two time travels would be the *real absolute causal impact* of the policy. We shall come back to this very important topic in Section 4.

[14] We shall present some research designs (experimental and non-experimental (quasi-experimental)), based on fictional data examples in Section 4. We shall not conduct a thorough econometric analysis per se but just show several central concepts on the logic behind such designs.

[15] Statistically speaking, if this were the case, the effect of the treatment as represented by its β coefficient would change and thus become unstable with each new statistical significant interaction.

the measurement of the X occurs earlier than the measurement of Y. How earlier is not important in this case. What is important is that logically *the (potential) effect can not precede the cause*[16].

In policy analysis, dealing for example with business subsidies and firms, the duration of the time lag is even more relevant because we need to let some time pass before we measure potential effects; and that, in order to let the intervention "evolve". How much time is a good question; we shall return to this in the following sections.

3.4 No causation without manipulation

Another important issue that one needs to remember is that *"there is no causation without manipulation".* (Rubin, 1986). Simply put, the independent variable of interest X (the TREATment/intervention) whose coefficient measures the treatment effect on the treated population, must *vary*. If it does not, we shall not be able to measure any impact because the variable will stay constant all over the observations at hand and thus will not be able to capture potential changes.

At the same time some other characteristics (the control variables "C" above), based on this approach can not be "causing" the effect because they are fixed and not manipulated. For example, this means that gender can not in reality *cause* someone's wage to increase or decrease[17].

3.5 Association/Correlation does not necessarily mean causation

Still another common but usually forgotten fact in quantitative analysis is that association/correlation between the variables X and Y does *not* necessarily mean by itself that X has *caused* Y[18]. Causation needs a *logical interpretation* on what constitutes a cause and what an effect. Two simple examples will make this clearer.

Example 1

Before a vaccine (the Salk vaccine) was developed against polio, a study showed that there were more new cases of polio during weeks when soft drinks sales were high than when they were low. Do soft drinks transmit polio? Fortunately no (for soft drink lovers). A third variable, the season during which the variables *polio frequency* and *soft drink consumption* were measured is related to both of them. Polio epidemics were the worst (high) in the summer months; soft drink sales were/are also high in the hot summer months. (SPSS, 1999a).

Example 2

Some studies have found that children with bigger feet spell better. Should parents, worrying about their children's school achievements, start stretching their kids' feet? Fortunately no (for the poor kids' sake). A third variable, age, plays a role here as well. Children with bigger feet spell better because they are older (one average), and older children (on average again) do spell better because they have gone more years to school (Paulos, 1991).

[16] In Section 4.3.3, I will discuss a case where the measurement of the effect can commence long *before* the cause (!).

[17] It is interesting however, that this very important point of the non-causal nature of fixed characteristics, is sometimes not taken into account even by evaluation research "gurus". Weiss (1998, p. 185, footnote 4) mentions that

> *"An independent variable is one which is assumed to be a cause of other variables (dependent variables). In evaluation, independent variables are characteristics or participants, program process and settings which are assumed to affect the dependent variable, the participant outcomes".*

[18] It is believed that the phrase "correlation does not imply causation" is attributed to the famous British mathematician Karl Pearson (The Pearson correlation coefficient has been named after him).

4. Estimating the impact of policy interventions

I mentioned earlier that in the evaluation of policy interventions one (a) estimates the (potential) impacts of the policy in question and then (b) passes a judgement on the findings. This section deals with impact estimation where as the next section deals with the formulation of the judgement. Impact estimation in turn, using quantitative methods, can be classified into two distinct but interrelated processes: (a) the selection of the *evaluation design* and (b) the *model specification.*

Section 4.1 examines the basic principles of a natural experiment and its links to public policy evaluation. Especially it focuses on those principles that deal with the estimation of the potential impact of the intervention. Section 4.2 discuss how causality, counterfactuality, comparisons and impact estimations - all important notions in public policy evaluation - are related to each other. Sections 4.3 and 4.4 discuss issues on the evaluation design and model specification respectively. Finally, Section 4.5 demonstrates through a fictional example how in practice the decision to chose one design over another and one model specification instead of another change the estimated impact of the policy under scrutiny.

4.1 Example of an experiment and its links to public policy evaluation

Here I discuss a study which, at first glance, may seem unrelated to the theme of public policy evaluation. Nevertheless, the rational and importance behind such reference will soon be evident. The article itself contains much more technical and medical jargon that the one presented below, but my idea is to simply convey the principles based on which the study was conducted.

In a study related to the field of *orthodontics*, Xenakis et al. (1995) examined the relationship in the growth between the lower and the upper jaw in laboratory mice. For this, they basically recorded the growth of the lower jaw when the growth of the upper jaw was deliberately prevented from growing. For the study, 140 mice were used. All mice were chosen to be genetically *identical*. A sub-group of 75 *randomly selected* mice were treated and 65 were left as controls. The treatment basically consisted of surgically inserting special material (glue) in the upper jaw of the treated mice, preventing it from growing at the same pace, had there been no treatment. All mice lived in *stable laboratory conditions* with suitable food and water without restrictions. Their weight was measured at the *beginning* of the experiment and at each *further experimental stage* thereafter. At a certain age after birth, a number of mice were terminated from both experimental and control groups and their jaws' growth was measured. Among others, the results revealed that there was a significant difference in the growth of the lower jaw between the treated and control mice. The authors concluded that the growth of the upper jaw seemed to affect the growth of the lower jaw.

What this experimental study showed are some of the fundamental principles of the impact estimation process. In order to find out whether a treatment has been the cause of an event, one needs to simulate the counterfactual condition of what would happen to the units of the population in question were they not exposed to the policy intervention.

Because we can not observe the same unit concurrently in the two different regimes (It is impossible for one to have received and have not received treatment at the same time), we need to create a population which resembles the treated population in all respects, but the treatment, and use *that* as a substitute (a proxy) for the treated population. This is why Xenakis et al. emphasised that the mice population were *genetically identical* and lived in *stable* laboratory conditions. In addition, we need to create such conditions in order to *control* the impacts of all other potential factors that might have an influence so as to isolate the effect that is caused solely by the treatment. That is why in the above example the mice were measured at different time intervals; the researchers were using their age as a control. Both the counterfactual creation and the selection of controls assist in the precise and unbiased estimation of the net impact of the intervention.

Finally - and this is probably the most important lesson of the experiment - the decision as to which unit (mouse) was to be treated and which not, was done *randomly*. This, in combination with the two previous methods (control variables and creation of population from genetically identical units) more or less guaranteed that what ever differences found between the control and treated groups could be *causally* attributed to the treatment. And that, because the randomisation process - in theory - balanced out whatever unobserved differences the two groups might have had.

In public policy evaluation, however we do not usually have the luxury of a laboratory experiment. The whole process is much more complicated because real persons are involved. That is why we are "doomed" to use other approaches to estimate a net policy impact. It is a fundamental truth nevertheless, that a causal result

can never be estimated with 100% certainty. It is a probabilistic approach which attempts to simulate the conditions of an experiment so as to measure/estimate the impact of interest, with the least possible error.

4.2 Causality, counterfactuality, comparisons and impact estimation

What is the relationship among the notions of causality, counterfactuality, comparisons and impact estimations? Recall that public policy evaluation encompasses two processes: (a) the estimation of the potential impact of the public intervention and then (b) the judgement of the results on their worth. Causality and counterfactuality are linked to the first process.

Causality is the relationship between a cause and its effect; in public policy evaluation it is the relationship between the intervention and the impact. When one establishes the counterfactual and then measures the impact, he implicitly also creates a causal relationship between the two. One can not claim that he has estimated the impact of a public intervention by saying that this impact has not been caused by the intervention. The estimation of the net impact and the creation of causality are two things that go together. The counterfactual on the other hand, assists in the unbiased estimation of the net impact.

Also the making of comparisons is essential in the estimation of a potential impact. In public policy evaluation the comparison activity surfaces by matching the effect with the treatment (intervention) on the effect without the treatment. Finally, we must keep in mind that comparisons are utilised not only in finding the net effect of a program but also in judging the overall goodness of the program; because to judge, one needs to be able to *compare* a certain situation against some other and distinguish the better of the two.

4.3 Designs

Table 1 lists several designs. In literature, they are called *evaluation* designs but a better word would be *estimation* designs because they are used in the estimation of impacts. These designs describe the logic upon which the data (of the population of interest) is selected and then analysed in estimating the policy impacts. The list is not exhaustive; it contains rather only a few designs frequently found in empirical literature dealing with retrospective policy evaluations using quantitative methods[19].

To make things more comprehensible I have used a fictional policy program in all design formats. Let us suppose that the purpose of the evaluation is to pass judgement on the potential impact of a business subsidy program of the government towards firms. One question that we might try to answer is the following: What has been the impact of the receipt of government subsidies (S) on the profitability of firms (P)[20]? The designs found in Table 1 would be used in the estimation of such an impact.

[19] For a more comprehensive coverage on designs see for example Cook and Campbell (1979), Shadish et al. (2002), Rossi et al. (1999), Mohr (1995), Heckman et al. (1999) and Trochim (2002).

[20] Some may argue that the increase in profits of the recipient firms is not a realistic goal of a business subsidies policy. A better goal might have been the increase in the Value Added of firms due to the receipt of subsidies. Nevertheless, for simplicity reasons, I have chosen to use profits as a impact indicator throughout the paper.

Table 1. Selective evaluation designs

	Profit before receipt of subsidies	Receipt of subsidies – Treatment/Intervention	Profit after receipt of subsidies
R1. Classic Randomised Comparison Group Design **Two Groups: Firms, recipients of subsidies, non- recipients** **Random (R) assignment of Treatment** **Cross-sectional observations**			
Treatment group – Firms recipients of subsidies	P^T_{t-1}	S^T_t	P^T_{t+1}
Control Group – Firms non-recipients of subsidies	P^C_{t-1}		P^C_{t+1}
R2. Classic Randomised Comparison Group Design **Two Groups: Firms, recipients of subsidies, non- recipients** **Random (R) assignment of Treatment** **Longitudinal observations**			
Treatment group – Firms recipients of subsidies	…, P^T_{t-3}, P^T_{t-2}, P^T_{t-1}	S^T_t	P^T_{t+1}, P^T_{t+2}, P^T_{t+3}, …
Control Group – Firms non-recipients of subsidies	…, P^C_{t-3}, P^C_{t-2}, P^C_{t-1}		P^C_{t+1}, P^C_{t+2}, P^C_{t+3}, …
NR1. Pre/Post-program Comparison Group Design **One Group: Firms, recipients of subsidies** **Non-Random (NR) assignment of treatment** **Cross-sectional observations**			
Treatment group – Firms recipients of subsidies	P^T_{t-1}	S^T_t	P^T_{t+1}
NR2. Pre/Post-program Comparison Group Design **Two Groups: Firms recipients of subsidies, non- recipients** **Non-Random (NR) assignment of subsidies** **Cross-sectional observations**			
Treatment group – Firms recipients of subsidies	P^T_{t-1}	S^T_t	P^T_{t+1}
Control Group – Firms non-recipients of subsidies	P^C_{t-1}		P^C_{t+1}
NR3. Pre-program/Post-program Comparison Group Design **One Group: Firms recipients of subsidies** **Non-Random (NR) assignment of subsidies** **Longitudinal observations**			
Treatment group – Firms recipients of subsidies	…, P^T_{t-3}, P^T_{t-2}, P^T_{t-1}	S^T_t	P^T_{t+1}, P^T_{t+2}, P^T_{t+3}, …
NR4. Pre-program/post-program Comparison Group Design **Two Groups: Firms recipients of subsidies, non- recipients** **Non-Random (NR) assignment of subsidies** **Longitudinal observations**			
Treatment group – Firms recipients of subsidies	…, P^T_{t-3}, P^T_{t-2}, P^T_{t-1}	S^T_t	P^T_{t+1}, P^T_{t+2}, P^T_{t+3}, …
Control Group – Firms non-recipients of subsidies	…, P^C_{t-3}, P^C_{t-2}, P^C_{t-1}		P^C_{t+1}, P^C_{t+2}, P^C_{t+3}, …

R = Random assignment to (T)reatment and (C)ontrol groups
NR = Non-random assignment to (T)reatment and (C)ontrol groups
T = Treatment group (received intervention/treatment i.e. subsidies)
C = Control group (did not received intervention/treatment i.e. subsidies)
t-1 = It denotes a moment in time *just* before the intervention; does not necessarily mean one year before the intervention. However, if we have as our observation unit the firm, we are most likely to examine the firm's financial statements. Thus we need to check the most recent statement of the firm before the intervention, usually a year before. The same applies for "t-2", "t-3", etc.
t+1 = It denotes a moment in time after the intervention. In cross sectional observations "+1" does not necessarily mean one year after the treatment; it can denote any time after the treatment. Again, with firms as our units, +1 means most likely a year after the receipt of subsidies. The same applies for "t+2", "t+3", etc.
S = Subsidies received (intervention/treatment)
P = Profit

4.3.1 Experimental designs

Designs R1 and R2 are listed for comparison purposes only. Their main characteristic is the *random* assignment of the treatment (intervention - subsidies in this case) among the units of the population of interest (firms). Here, the treatment assignment is *controlled* by the evaluator. We, in fact, have already discussed the R1 design in the "Mice" example in Section 4.1. The R1 utilises cross-sectional observations at two points in time only; once before and once after the intervention. On the other hand, the R2 design differs from R1 in that the pre and post treatment impact measurement is based on a time series (longitudinal) data of both the treated and control groups.

Were we to implement an experimental design, say the R1, as part of an evaluation of a business subsidies policy, we would need to work as follows: We would first have to define our population of interest (all the firms which would be potential recipients of subsidies), measure the effect indicator (the firms' profits - P) *on average* before the intervention, and then randomly distribute subsidies to some of them. We call the firms that would receive subsidies the "Treatment or Treated" group and those that would not, the "Control" group. At some later point in time we would measure the effect indicator once more and calculate the net effect by using the following simple formula:

$$\Delta P_{RandomDiD} = (P^{T}_{t+1} - P^{T}_{t-1}) - (P^{C}_{t+1} - P^{C}_{t-1}) \qquad (1)$$

The superscript refers to the group in question:

- (T)reatment, (C)ontrol

The subscript refers to the time of measurement

- t+1 denotes a moment in time *after* the Treatment (the allocation of subsidies)
- t -1 denotes a moment in time *just before* the allocation of the Treatment

The formula produces the so called *Difference in Difference (DiD)* impact estimator because it measures exactly that; the difference of the differences of the average impact (the profit) between the treated and control groups before and after the treatment.

Such a random experiment would be ideal for policy evaluations but faces some political, ethical and administrative problems. That is because the hypothetical policy, as most of policies, is initially designed to correct some inequalities (market failures), assist firms in their business development, increase employment prospects, etc. If the (free) money is distributed in random, even for experimental purposes, there will most probably be some reservations from the interested parties on its real worth, and its distribution criteria. Why should firm *x* get free funds and not firm *y*? Doesn't this create unfair competition with the blessings of the government?[21] Simply put, it is not easily justified to randomly give public money to some and not to others. Finally, the cost of the experiment itself may prove an even harder obstacle to overcome since money for this purpose could be hard to explain on an ex-ante basis.

4.3.2 Non-experimental designs

The basic difference between the experimental designs R1 and R2 and the non-experimental designs NR1 – NR4 in Table 1 is the method of assigning the treatment. Designs NR1 – NR4 are used in retrospective evaluation studies when the treatment has taken place already in the past. In those cases the treatment has not been distributed randomly, but rather, under certain predefined criteria. This creates the so called "selection bias" which we shall discuss in the section on model specification.

Which is the "best" evaluation/estimation design? In general, from a statistical point of view, the experimental design R1 and even more so the R2, are regarded as the ideal ones[22]. By their random assignment they avoid the selection bias others designs have by definition. If there are enough observations in our data, the "R" designs will most probably produce the best estimates of the policy impact.

However, because of the practical and other problems that are linked to them, we must turn to the other non-experimental designs (some call them quasi-experimental). Is there a design among the NR1 – NR4 that can reduce the bias which exists because of the inequality of the treated and untreated groups? As we shall discuss below and in Section 6.2.1, there can be no definite answer to the question. In most cases, when

[21] The same objections and at an even higher scale have surfaced in experimental studies measuring labour related interventions to individuals (training programs for unemployed, temporary subsidised jobs in firms, etc.).
[22] See also footnote 13 in Section 3.3 earlier.

dealing with policy evaluation, we are bound by the nature of the policy itself and the (un)availability of the data that we are given to analyse.

In the analysis that follows, we need to keep in mind a couple of things: One is, how the *counterfactual* is defined in each of those designs and what links it has to the policy examined. The other is the *timing* of impact measurements and its relevance to the exogenous factors which might influence the impact measurement itself.

In all designs the treatment (policy intervention) is assumed to occur at a single point in time, at t. Nonetheless in reality, a policy is usually implemented over several periods (years) and interventions occur in many time instances. Had I taken this into account, I would have needed to "widen" the graphic representation of the intervention around t. I have not done so, in order not to clatter even more the figures. The logic however, remains the same.

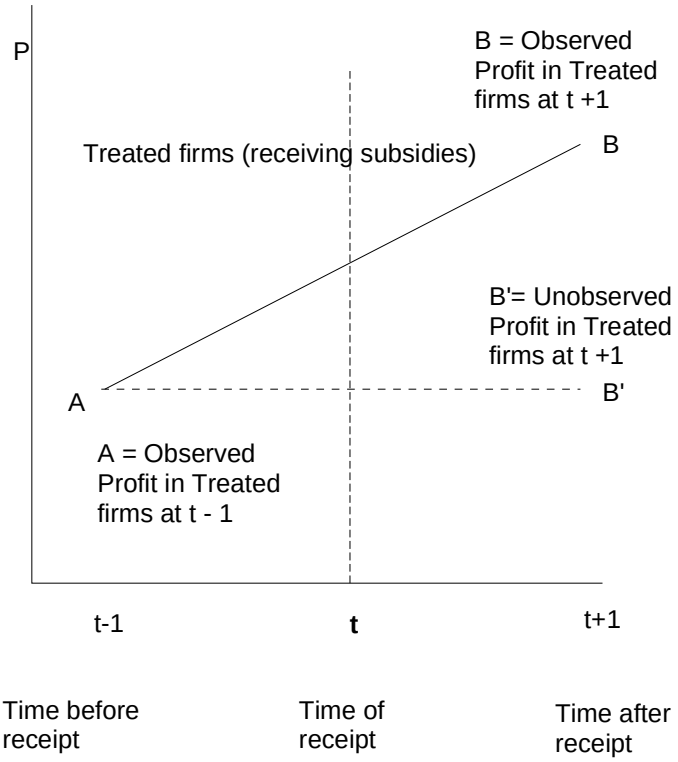
Each design analysis is divided into four sub-sections: The first describes the ideal design through which we would have found the true net impact of the policy. The second describes the design which is normally used for impact estimations. The third presents the assumptions that we make in order have an unbiased impact measurement estimation. The fourth comments on the peculiarities of the respective design.

Design NR1 (One group before after design, cross - sectional)

True impact measurement

In this design (Figure 2)[23], we are using observations only from the treated population; that is profit amounts only from firms that have received subsidies. We measure the profits twice. Once before the receipt of subsidies at time t-1 (point A in Fig. 2) and once after the receipt at time t+1 (point B).

Figure 2: Design NR1



[23] In this, as well as in the rest of the designs in Figures 3, 4 and 5, I show an example where the absolute value and the rate of change of the observed impact are always higher in the Treated population compared to the Control population (the counterfactual). This of course need not always be so; however, in all cases, the estimation rules remain the same.

How do we find the true change of profits (P) that have been generated from the receipt of the treatment (the subsidies, S)? In the ideal case, had we been able to go back in time, we would have just not given the subsidies to the same firms, and record at time t+1 their level of profits, B'. This is the counterfactual condition. The difference between the profits during time t+1 at the observed and counterfactual levels would have given us the true net profits "α" generated by the subsidies received.

$\alpha = B - B'$ (2).

Estimated impact measurement

Unfortunately, we can not observe B'. We can only observe what has happened to the firms' profits before the receipt of subsidies when they were A and after the receipt of subsidies when they are observed to be B. How do we then measure the impact with this design? For the counterfactual amount B', we use as a proxy (substitute), the observed value of the impact *before* the treatment, the A. The net profit impact "β" is now calculated as

$\beta = B - A$ (3)

Assumptions for an unbiased estimated impact measurement

If we want to produce an unbiased estimate of α we should strive to get our estimate equal to the true impact or β = α. From this we get $B - A = B - B'$ because (3) = (2). From both sides B is deleted and we end up with

$A = B'$ (4)

In other words we assume that the observed profits, before the receipts of subsidies, are equal to the unobserved profits after the receipt of subsidies.

Comments

Design NR1 is applicable if one has only a cross section of observations before and after the treatment. Also, it can only be applicable if the policy under evaluation is horizontal, that is it covers all units of the population exposed. If it were vertical, thus covered only part of the population, this would have given us a chance of creating another counterfactual condition. This would happen through a control group which would not have been treated with the intervention. With NR1 we implicitly assume that the impact indicator (the profit) recorded before the intervention (the A) for the firms exposed to the treatment, will continue to be the *same* after the time of intervention for these same firms, *had they not been exposed to the treatment (the B)* .

Another problem that we have with this design surfaces when we try to examine what happens with other factors that possibly influence our impact (profit) measurements. Since we are substituting B' with A, we also assume that the external factors that are existing at time t+1 (when B' is measured) are *the same as* the ones at time t-1, when A is measured. However, this is a pretty heavy assumption because time by definition alters things (if not in a complete controlled, laboratory environment)[24]. For these reasons, Design NR1 is a rather weak one.

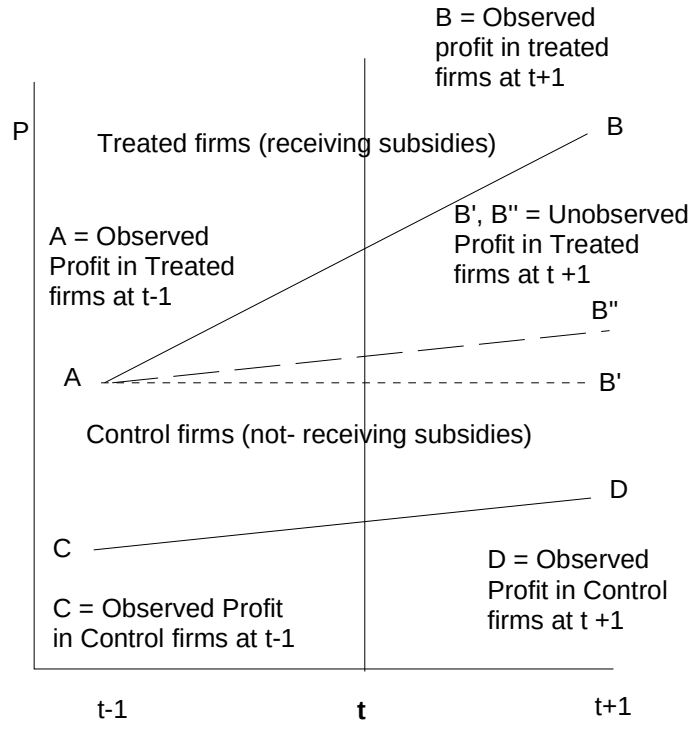
Design NR2 (Two groups before after design [Difference in Difference – DiD], cross – sectional)

True impact measurement

This design, in contrast to Design NR1, uses observations from two populations; from the ones that have been exposed to the treatment and from those that have not (Figure 3). The true impact of this design continues to be α, the same as in (2) above, or $\alpha = B - B'$.

[24] You can see this, with equation 4F in the Appendix. There we say that the expected profits of the firms after the receipt of subsidies at t +1 (although in fact they have not receipt subsidies, thus the profits are unobservable) equal the expected observed profits before the receipt of subsidies at t -1, given the X exogenous factors. But the subscript of the X factors at the left hand side and the right hand side of the equation is not the same, indicating different time periods.

Figure 3: Design NR2



Estimated impact measurement

Using the same example, we measure the profits of firms that have received subsidies and the profits of firms that have not received subsidies. We conduct the measurements in two points in time for both groups; once before the intervention at t-1 and once after, at t+1. Design NR2 is the same as the Design R1 - the classic Difference in Difference design -, but the assignment is not done randomly. The difference between Design NR1 and Design NR2 is on how one defines or selects the counterfactual population. In NR1 it was the impact measurement of the treatment group, at time t-1. Here it is based on a group of firms not exposed to the treatment (firms not given subsidies) and the impact is measured both before and after the intervention. The net impact is "β" and calculated as

$\beta = (B - A) - (D - C)$ (5)

Assumptions for an unbiased estimated impact measurement

Here the assumption is that the units not exposed to the Treatment (the firms not getting subsidies), will produce the same impact as the units who were exposed to the treatment, had those exposed units not been exposed to the treatment, or

$D - C = B - B'$ (6)

As with Design NR1, for an unbiased estimate of α, β should equal to α, or (5) = (2). If that is the case, we get

$(B - A) - (D - C) = B - B'$ (7)

Taking away the B from both sides we obtain

$B' = A + (D - C)$ (7a)

But this in fact is a new unobserved level of B that we call B'' or assume that B' = B''. Thus

B'' = A + (D - C) (8)[25]

Comments

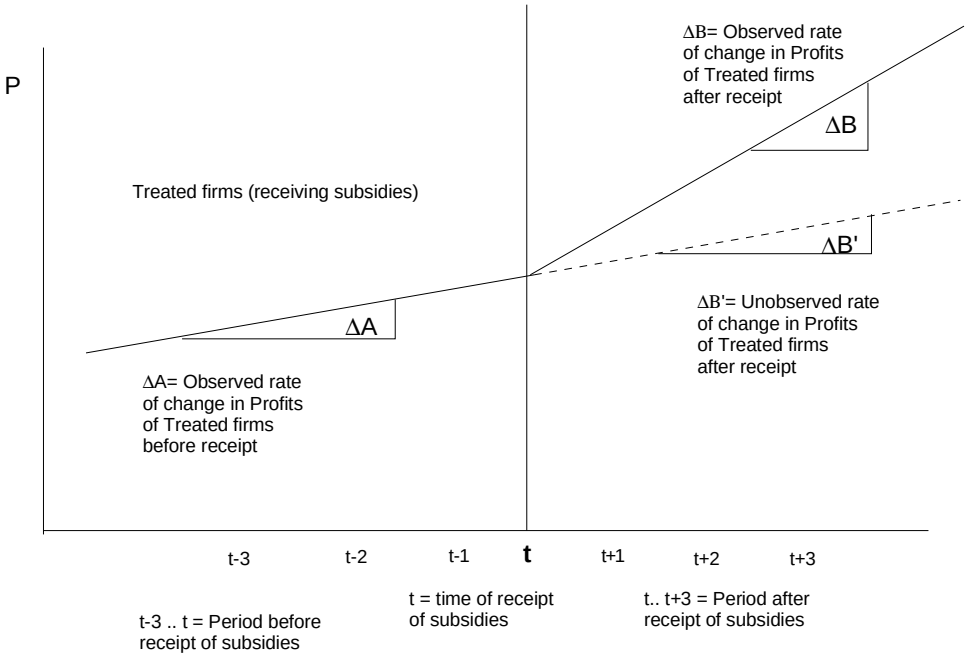
Design NR2 is applicable when we have cross sections of observations before and after the intervention for two groups of units (firms); for those that have received the treatment and for those that have not. To have access to a group of untreated units, the policy evaluated must be vertical, so that it covers only a certain portion of the total population. The counterfactual created from the firms that have not received the subsidies is a better choice than the one which is based solely on the past impact value of the treated population, as in Design NR1. The reason is that since it measures the impact of both the groups at the same time at t-1 and at t+1, the exogenous factors in those times influence equally both measurements at each time period.

Design NR3 (One group before after design, with trends)

True impact measurement

Design NR3 is another version of Design NR1 where we analyse observations from the treated population only (from the firms that have received subsidies only). In Design NR3 there are several measurements before and after the treatment instead of just one. This enables us to account for trends in the impact measurement (the profits) in the periods before the treatment, the ΔA, and after the treatment, the ΔB (Figure 4).

Figure 4: Design NR3



The true impact measurement, say the rate of change of profits due to a receipt of subsidies at time t, can be achieved, as you recall, only by moving backwards in time which is impossible. In Figure 4, notice that we depict the counterfactual (what would have happened to the rate of change of profits of the firms had they not received subsidies) by ΔB'. The net rate of change of profits of the firms due to the receipt of subsidies is then the difference between what it is with subsidies and what it is without, or

α = ΔB – ΔB' (9)

[25] This in turn would give us the unbiased estimate of the profits as α = B - B''

Estimated impact measurement

As with the previous designs we can not observe ΔB'; we can only observe what was the rate of growth of the firms' profits before the receipt of subsidies. Thus, we take this rate and use it as a proxy to estimate the impact of subsidies, β on the rate of change of the firms' profits.

$\beta = \Delta B - \Delta A$ (10)

Assumptions for an unbiased estimated impact measurement

To have an unbiased estimator of the true impact, the estimate needs to equal the true impact or $\beta = \alpha$. From this we get $\Delta B - \Delta A = \Delta B - \Delta B'$, because (10) = (9). By deleting from both sides the ΔB we end up with

$\Delta B' = \Delta A$ (11)

This implies that the unobserved rate of change in profits after the receipt of subsidies for the treated firms, is equal to the observed rate of change before the receipt of subsidies.

Comments

Design NR3 suffers from the same problems as Design NR1. It does not account for differences in the observable factors between the period before and after the receipt of subsidies. It just assumes that the factors are the same pre and post intervention[26]. Also it assumes that the rate of change of profits during the period before the receipt of subsidies will continue with the same rate after the receipt of subsidies and based on this, it measures the counterfactual. Nevertheless, it does account for trends, and it is preferable to Design NR1 if we have enough periodic observations pre and post intervention, and if the policy is applied horizontally.

Design NR4 (Two groups before after design [Difference in Difference – DiD], with trends)

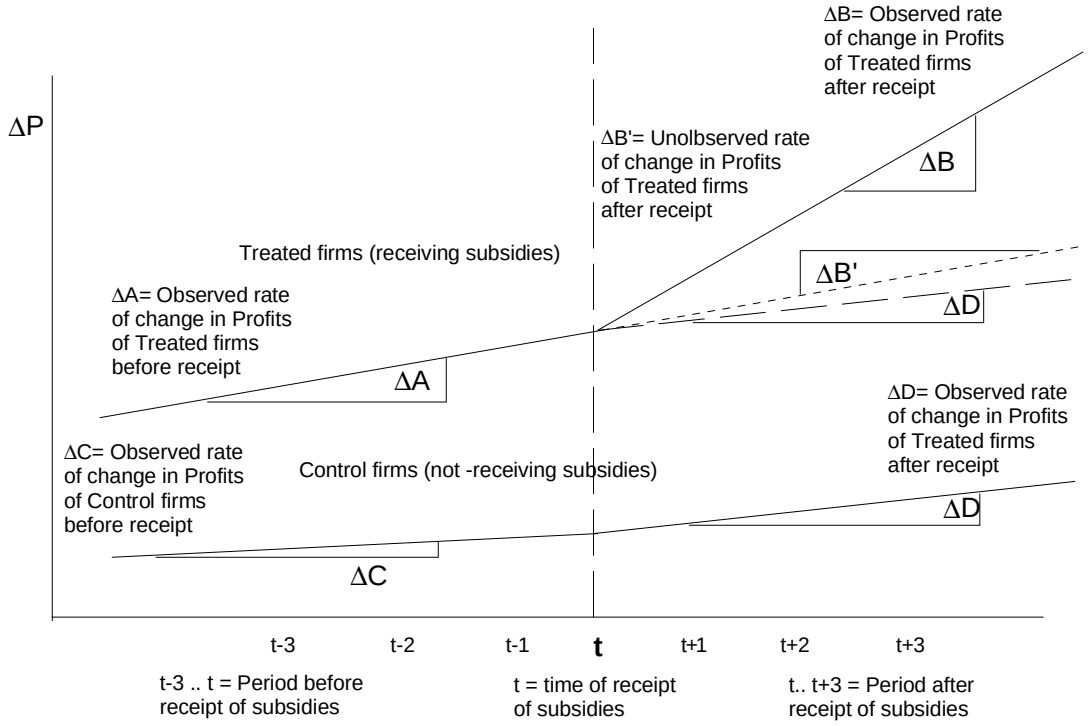
True impact measurement

Design NR4 is the equivalent of Design R2 without the random assignment of treatment. Design NR4 is a combination of NR2 and NR3 Designs. It accounts for the net rate of change of the impact (the profits) using measurements taken from two populations, the Treated and the Control, the latter being used as a counterfactual. As with Design NR2, it can be used only if the program in question is not applied horizontally, but vertically, so that it does not cover all units of the potential population. The other obvious prerequisite is that observations from both populations before and after the interventions are available for several periods.

As with Design NR3, the true impact measurement is found by first calculating the rate of profit change for the Treated population during the period after the receipt of subsidies at time t. This is shown in Figure 5 as ΔB. Using our imaginary method, we go back in time and see what the rate of profit change would have been, had those firms not received the subsidies. This is depicted by ΔB'. The impact of subsidies for the recipient firms is the difference of the two rates or $\alpha = \Delta B - \Delta B'$.

[26] See the analysis of the Difference in Difference estimation in Section 4.4.5. and equation 11F in the Appendix.

Figure 5: Design NR4



Estimated impact measurement

ΔB' can not be observed, just assumed. To actually measure the rate of change for profits, we apply a Difference in Difference estimation. That is, we deduct the rates of profits change of the control firms before and after the intervention (the counterfactual), from the respective rates of profits change of the treated firms, or

$\beta = (\Delta B - \Delta D) - (\Delta A - \Delta C)$ (12)

Assumption for an unbiased estimated impact measurement

Because we deal with two sets of firms and we wish to have an unbiased estimator of the true impact, we make the following assumption: The difference between the control and the treated groups in the rate of change of the indicator of interest (the profits) during the period before the intervention equals the difference of the rate of change of the indicator of interest during a period after the intervention, again between the two groups, *in the absence of the intervention*. That is,

$\Delta D - \Delta C = \Delta B - \Delta B'$ (13)

We shall now follow the analysis as in Design NR2. With all designs, we should aim that our estimate β is a close as possible to the true impact net impact α, or β = α. From this and since we aim that (11) = (12) we have

$\Delta B - \Delta B' = (\Delta B - \Delta D) - (\Delta A - \Delta C)$ (14)

By deleting ΔB from both sides we get

$\Delta B' = \Delta A + (\Delta D - \Delta C)$ (15)

This is a new unobserved level of ΔB' which we call ΔB''

$\Delta B' = \Delta A + (\Delta D - \Delta C)$ (16)[27]

[27] This would give as the unbiased estimate of the change of the rate in profits as $\alpha = \Delta B - \Delta B''$

Comments

Design NR4 accounts for the counterfactual and at the same time for *patterns* of the indicator of interest before and after the intervention. If we can control for observable (and unobservable) factors and we are lucky enough to have enough longitudinal observations from both treated and untreated groups before the treatment and after the treatment (i.e. in panel format) we should apt for Design NR4 because it will probably produce the best estimates of the true impact. Finally, the estimation should be enhanced by utilising statistical methods that we shall discuss in the following section (i.e. by building regression models with all observable factors that we think have an influence on the effect in addition to the variable representing the policy impact, alone or with interactions, utilise time series techniques to account for gradual changes in the effect such as fixed effects regressions with panel data, apply matching or instrumental variable techniques to account for selection bias, etc.).

4.3.3 The role of timing in estimating impacts

The timing of estimating impacts in a retrospective evaluation is very crucial. When do we start measuring? When do we stop? There are several factors that need to be taken into account. One is the indicator of interest measured and the target population of the intervention. In certain cases, it is clear when one starts measuring, but in some others it is not. Take for instance an evaluation of a training program for a group of people. If the evaluator is interested in whether the program has helped its participants to find jobs, he will certainly check this *after the* training program is completed. Whether one will take into account the participants' previous employment situation as well as that of another group of people who did not participate, is of course a choice of methodology. The point put forward is that, to evaluate this case, one has to start checking for an impact *after* the intervention.

However, in other cases one may begin measuring the potential impact of an intervention much *before* the actual intervention takes place. This is in contrast to the rules mentioned in Section 3.3 according to which there must be a clear *precedence* of the intervention from the effect in order to establish causation. Consider for example, the public program that distributes business subsidies to firms. We have decided to check for a potential impact by measuring the profit growth within those firms. To measure for growth one needs at least two observations. The profit at time t -1, representing a time *just before or just after*[28] the intervention and then the profit at a time t+1 in the future. The growth is then found as profit at t+1 less profit at t-1.

Do we begin measuring the profit at t-1 only just before or just after the receipt of subsidies? That is, very close to the intervention event? Not necessarily. The firm might have started changing its behaviour, thus influencing its profit figures, even *before* the actual receipt of subsidies. For instance, *anticipating* the receipt of subsidies, the firm might have begun investing earlier. This in turn, could have affected profits even before all the subsidies were received. When exactly one starts measuring, is a problem which is not easily solved. The most common answer would be "it depends"[29].

What about the time t+1? When do we conduct the final measurement? How long after the end of the intervention? Of course we need to have a certain amount of time lag so that the intervention has time to "affect", but again there is no clear answer. It depends on the policy under scrutiny.

In the case of the training program when one looks at a point in time, it might well be that, had the measurement been conducted a few weeks earlier or later, the result (the measured impact indicator) might have been different. In order to get a full picture of the impact, one needs time series and panel data to examine these periods. Generally, what finally determines these timing issues is the availability, content and format of the data at hand, rather than the optimal design envisioned by the idealist evaluator[30].

4.4 Model specification methods

If we are to implement those non-experimental designs (NR1 - NR4), we have to assume that the two populations, the treated and the control are equal before the intervention. However, the designs as such are not sufficient, to give us the best unbiased estimate of the measured impact. Indeed as we have noted

[28] Theoretically, whether the measurement *begins* before or after the intervention is of no significance per se as long as we do not leave any substantial time lag between the measurement and the intervention event.
[29] For an extensive discussion on the topic of timing, see Venetoklis (2001a).
[30] In the final section I will discuss yet another dimension of timing linked to pubic policy evaluation.

above, one implicitly sees that if suitable data is available (we have multiple observations of the unit of interest on different times) and if the policy under scrutiny permits (it is vertical and not horizontal, thus covers only part of the population of interest) we can come closer to reducing the impact bias through the use of a counterfactual population. But because we are not assigning the treatment randomly, differences between the control and treated populations are by definition still there and some of them can never even be observed.

What we have to do then is to account for such differences. There are several ways of doing so, and I have classified them as methods for model specification. This means finding/creating certain (control) variable(s) in our models and selecting the population to be analysed in such a way as to balance these differences.

4.4.1 Observable control variables

The selection of observable control variables in our regression equation is the first step in reducing these differences. The choice of control variables depends on many factors, but basically one builds the model having as a guide (a) the theory explaining the behaviour of the target group of the policy intervention, (b) the availability of data and (c) the variables utilised in other studies measuring the same or a similar policy. It must be kept in mind however, that model building is not an exact science, and it depends on many other issues as well which we shall not cover here (i.e. the functional format of the variables at hand, the availability of proxy variables if an important variable is missing from the data, the distribution of the continuous variables and their outliers, the potential interactions between variables, the sub groupings of categorical variables, etc.).

4.4.2 Simple matching technique

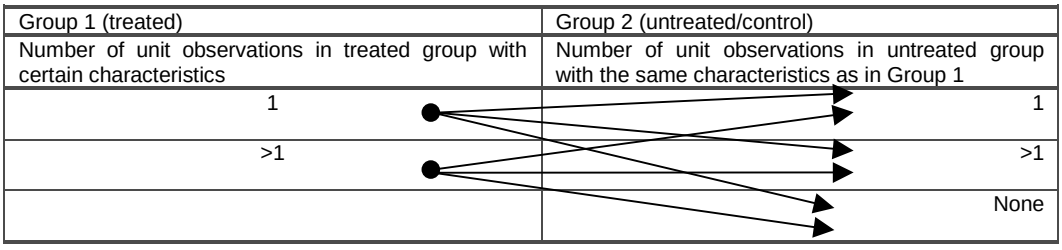
One can also match treated and non treated units on some of their characteristics. Taking the observed characteristics of the treated units as a base, one can attempt to find the units in the untreated group which have the same characteristics as the ones defined in the treated group.

The method is initially appealing but there are some practical problems that can easily prove the attempt fruitless. The reason is simple. The more characteristics one uses as a base, the more untreated observations one needs to be able to find and match them against the treated group.

Thus, as the observable characteristics (factors) of the treated group grow in number (horizontally - across) and in sub-categories/strata (vertical - within), the probability of finding an equivalent observation in the non-treated group diminish quickly even if one has “rich” data in abundance. To give an idea, just in the general case of binary factors, the number of control units is 2^p, where p is the number of binary characteristics of the treatment units. Take for example the situation where we have only 3 binary control variables. We should find at least 2^3 observations in the population of the control units so that all the treatment units have at least one match with the same observable characteristics/factors. If the p increases to 15 or 20 (a usual situation with many recorded factors) the number of observations in the control group could grow to over a million!

There are also problems in the number of observations within the clusters of equal characteristics. That is, one might be confronted with the following situation, depicted in Figure 6.

Figure 6. Simple matching technique: matching combinations between observations of treatment and control groups



Attempting to match the observations of group 1 with the ones of group 2 entails some decisions on the selection criteria for matching. We have the following 6 possibilities:

(1,1), (1, >1), (1,None), (>1,1), (>1,>1), (>1, None).

We must disregard the cases where there are no equivalent units found in group 2; that is (1,None) and (>1, None). In the three cases where there are more than one observations in either or both groups (1,>1),

(>1,1), (>1,>1) we would take the average of those units for what ever treatment and impact indicator we measure. Finally, the most straight forward case is the one to one matching (1,1).

Until now we have assumed that the matched variables are categorical. However, a common problem is that the matching conditions just mentioned are even more difficult to achieve when the matched variables are continuous. It is (relatively) easier to find, for example, firms who are in the same sector(s) than find firms who have the same sales figure at a given financial year. We can of course define a margin, say, 10% over or under the respective sales figures of the treated group but by doing so, we are moving away from the whole matching idea. In any case when it comes to firms and individuals the heterogeneity is so vast, that we are most likely to fail unless we limit ourselves to a few matching conditions.

To conclude, in order to do simple matching, one must usually have data in abundance and must apply the matching selection based on some criteria which are usually implemented through coding. Nevertheless, we see how easily observations can be dropped out, especially if we have many characteristics as our base. And in any case, other problems still exist. We have yet to account for the possibility that groups 1 and 2 differ in characteristics which are *unobserved* (or not recorded in our database). And that, in turn, means that the estimated impact of our treatment indicator may still be biased.

4.4.3 Matching with propensity scoring[31]

A technique to overcome the aforementioned problems is called propensity scoring. The propensity scoring is an indicator (a number) which depicts *the conditional probability of being assigned (or not) to a particular treatment*. By conditional we refer to a set of characteristics/factors that can predict such an assignment.

In practice what is done is that we estimate a probability model (i.e. through a logistic regression) to predict the likelihood that units in our data are assigned to a treatment group. In the case of business subsidy programs we would estimate the probability of firms being given subsidies versus those that were not[32]. In the probability model we can use as many variables (characteristics, factors) that we believe matter in the assignment process. For example, in a business subsidy policy framework, these factors could be the financial state of the subsidised firm, its location, the type of investment in question, the estimated jobs created/sustained, the business analyst handling the application, the frequency with which the subsidised firm has received subsidies in the past, etc. After we produce this score per unit (firm), we then identify subgroups of firms from the unsubsidised group which happen to have *a similar propensity score* as the firms that were subsidised. We could achieve this, using the following pseudo-algorithm:

1. We build a parsimonious logit function (as simple as possible, yet representative model) to estimate the propensity score, say through a logistic regression model. From this we get a score for each unit.
2. We sort the data at hand according to the estimated score of each unit, ranking the observations from lowest to highest.
3. We stratify all observations into several groups so that the respective scores within each stratum (group) between treated and control units are close enough (not different, at a given statistical significance level).
4. We conduct the statistical test as follows: for all covariates (observed factors/characteristics of our units) the hypothesis (H0) is that the differences in means across treated and control units are not significantly different from zero.
5. If all the factors between treated and control units for all strata (groups) are balanced (not significantly different from zero), we stop and proceed with the evaluation designs mentioned earlier.
6. If any factor "X" is not balanced for some groups, we divide the unbalanced groups in smaller sub-groups and conduct the comparison again (recursively) until this factor is balanced as well.
7. If factor X is not balanced in all groups, we modify the logit function (the model - see step 1) by adding interaction terms (factor X * factor Z) or higher order terms of factor X (i.e. X^2) and conduct the comparisons again from step 2 (recursively).

Propensity score may be viewed as a very elaborate form of matching, one which simply uses many more matching criteria than the simple matching approach described in Section 4.4.2.[33].

[31] This introductory presentation is based on Rizzo (2001) and Dehejia and Wahba (1998, p. 20).

[32] More preferably, versus those firms that initially applied for subsidies but were then rejected. In doing so, we reduce the selection bias of firms which were not interested in applying for subsidies in the first place. On the other hand, we are faced with a reduced amount of potential control units since we limit ourselves to those that were rejected. For example the rejection rate of firms applying for business subsidies through the Ministry of Trade and Industry (the KTM) is about 1 in 3 only.

[33] There is theoretical and empirical literature in abundance on the matching method in general, and the propensity score method in particular. Empirical studies are evaluating mostly labour related subsidy programs. The mathematically literate can check for example Lechner (2000, 2001), Augurzky (2000), Kluve et al. (2001), Dehejia and Wahba (1998), Heckman et al. (1997, 1998). The definite reference on matching techniques (including the propensity score) is Rosenbaum (2002).

4.4.4 Instrumental variables[34]

I mentioned above that in regressions with non-random assignment, we are faced with the situation where we may have data that does not include all the factors that may influence the dependent variable (the impact of the policy intervention).

The use of instrumental variables is one way of dealing with the issue. The approach has two steps. First the treatment (the policy intervention) variable is used as a dependent variable in a regression equation which has at its right hand side two types of variables: (a) some variables taken from our observed data and (b) an additional variable called an "instrument". This variable has the characteristic that it is strongly correlated with the dependent variable of the first step (the treatment variable) but it is not directly correlated with the impact variable (the impact variable which we think is being caused by the policy intervention) or the error term. In the second step, we take the predicted value of the policy intervention (the treatment variable) generated from running the regression during the first step and use it as our main independent variable of interest in another regression model. Only then does the latter model use the policy impact variable as the dependent variable of interest.

In essence what we are trying to do is to *simulate* a randomisation process. Thus we are using variables that can predict well the intervention (treatment) variable but not the (potential) policy effect. The usual problem is that we can not find easily these instruments for all kinds of policy analyses[35]. And even worse, if we do find such an instrument, its quality might be so poor that the impact estimates might come out even more far apart from reality, than in the case of having not been used.

The instrumental variable method is similar in logic to the propensity score method. Both have two steps: initially they calculate a variable which is then placed in the final model estimating the policy impact. The difference is that the variable placed in the final impact regression model is calculated in a dissimilar way[36].

4.4.5 Difference in Difference estimation (DiD)

The Difference in Difference estimation method was mentioned briefly in the section on evaluation designs. Since it is a very popular and widely used quantitative method for policy impact estimation, I elaborate on it further. I discuss in detail only the *static cross-section* version of the method where one has two sets of observations for each group, before and after the policy intervention. However, the "better" version, where one has a panel data set[37] is easily applied since it is based on the same logic as the static two period cross sectional method.

The need for this method stems from our inability to account for all the potential factors that might influence the variability of the dependent variable of interest (the policy impact). We are limited to only the observable variables gathered in our data set. Had we known with certainty all the variables that influence the dependent variable of interest and had we been able to record their values, we would simply use them in the right side of the regression equation and feel confident that our estimate coefficient would be unbiased. But it is impossible to know *all* the potential exogenous influential factors. Thus, our estimates will be biased unless we somehow account for there unobservable factors.

We can classify all these unobservable factors into two types: those that stay constant (fixed) through out the policy (program) period examined, and those that *vary* during the same time. There is little we can do with the *varying* unobservable factors. However, with some data manipulation on the existing observable variables, we can *eliminate* the unobservable *constant* variables. Hence we can get rid of the potential bias caused by them.

How is this done? Recall from Section 3.1 that our basic regression model is[38]

$Y_{it}=\beta_{0t} + \beta_{1t}*TREAT_{it} + \varepsilon_{it,}$ with t = 1, 2 and where

[34] This introductory analysis is based on SPSS (1999b).
[35] For example can you come out with a variable that is *strongly* correlated with a subsidy given to a firm, but *not correlated* with the firm's profit growth?
[36] Yet another estimation method is the so called Heckman selection estimator (Heckman, 1979). We shall not elaborate on this method, but there is ample theoretical and empirical literature on the topic.
[37] Panel dataset means having data which contains measurements of a certain policy impact indicator for some periods before during and after the government policy intervention of the *same* units of interest.
[38] Here for the sake of simplicity I do not include other control observable variables in the equation.

Y_i is the effect indicator, TREAT(ment) equals 1 if unit i is exposed to the treatment (the policy) and 0 if it is not[39], t denotes the period examined (1 is the period before and 2 the period after the intervention) and ε_i is the error term with a normal distribution, zero mean and constant variance. The estimate of β of the intervention/treatment variable (TREAT) is the estimate of the average causal effect. The error term can then be assumed to consist of two amounts; the unobservable fixed variables (φ) and the unobservable varying ones (u), or $\varepsilon_{it} = \varphi_{it} + u_{it}$. φ then, captures all the factors that influence the Y impact indicator but stay *fixed* through the period examined. u on the other hand represents the *varying* factors that influence the impact variable of interest.

Because we assume that φ is correlated with the TREAT variable, if we do not delete it, our regression model will produce a biased coefficient β of TREAT. To account for this, because φ is constant over time, we subtract the data over the two periods we have at hand, t = 1 and t = 2. We can write the equations for the two periods as

$Y_{i2} = \beta_{02} + \beta_1 * TREAT_{i2} + \varphi_{i2} + u_{i2}$, for the second period (t = 2) and
$Y_{i1} = \beta_{01} + \beta_1 * TREAT_{i1} + \varphi_{i1} + u_{i1}$, for the first period (t =1).

But because φ is constant ($\varphi_{i2} = \varphi_{i1}$), when we subtract the second equation from the first we get

$(Y_{i2} - Y_{i1}) = (\beta_{02} - \beta_{01}) + \beta_1 * (TREAT_{i2} - TREAT_{i1}) + (u_{i2} - u_{i1})$ or
$\Delta Y_i = \delta_0 - \beta_1 * \Delta TREAT_i + \Delta u_i$, where

Δ denotes the change from one period to the other, from t = 1 to t = 2; the difference of the intercept terms of the two periods ($\beta_{02} - \beta_{01}$) is denoted by δ_0. The unobserved fixed effects φ do not appear anymore in the equation, because they have been differenced away, cancelled out. Of course we assume that the rest of the unobservables varying factors represented by Δu_i are not correlated with the TREAT variable (the intervention).

When we have more than two periods of observations, say three, we simply subtract all the values of the second period from the third and the first from the second. The same applies when we have more than three periods. The other way of doing this, is to average all the values of the periods at hand and subtract each period's value from the calculated mean[40].

4.5 Numerical example estimating the impact of a fictional business subsidy policy

Introduction and logic

This section shows practically some of the designs and model specifications listed in the previous sub-sections. Due to space constraints I do not cover all the aforementioned designs and the model specifications; rather, I concentrate on the most basic issues such as the counterfactuality, the inclusion of control variables, the handling of outliers and finally the utilisation of the treatment variable both in "dose" and in binary format. The basic idea is to give an overview of how the impact results can indeed change given the certain assumptions, designs and model specifications we choose to utilise.

Table 2 presents the data with which I demonstrate all this. The table consists of 44 fictional records on firms[41]. The purpose of this evaluation example is to measure whether the receipt of business subsidies has had any impact on the profit of the recipient firms and if so by how much.

[39] There are cases where TREAT is continuous, thus can represent a "dose".
[40] A comprehensive description of these methods is found in Wooldridge (2000, chapters 13 and 14). The methods may seem tedious and complicated, but nowadays all this is done automatically by statistical software programs especially designed for such analyses. For example the statistical software STATA has an option in its regressions functions called "fe" - (f)ixed (e)ffects.
[41] The idea for this example is taken from Schmidt (2001). In his paper, he also uses a fictional example which measures the impact of training participation in the future employment of individuals. His impact indicator is the difference between the employment status before the treatment (before participation in the training) and after participation. Thus, it can either be 0 (no effect) or 1 (positive effect). This approach assumes that the impact (if there is any impact) is homogenous across the participants. This facilitates the measurements because the evaluator does not - among other things - have to worry about the effect of potential outliers in the effects and in the causes. In the current example, I show how the impact is influenced by outliers and how one can deal with such a situation.

The calculations are based on a simple Ordinary Least Squares (OLS) regression. I have not attempted to discuss, say, problems of multicolliniarity or heteroscedasticity or the transformation of the variables in other functional formats (log-linear, log-log, linear-log, etc) in order for the coefficients to become more robust and the estimator of interest β to become BLUE (Best Linear Unbiased Estimator). For the correct robustness of the results, one can consult an introductory econometrics text book and apply different techniques with different statistical software available, checking for the assumptions which are common in this approach. As I mentioned earlier the values of this example are fictional, so whatever the tests, they may still not make any sense in a real world environment.

The emphasis of the analysis is in the *logic* behind the evaluation design and the model specification. My main objective is to demonstrate that the coefficient of the treatment indicator β produced through these econometric/statistical methods, indeed *varies* based (a) on the design one chooses to follow and (b) on the model specification (the types of variables) utilised in the regressions.

List of variables

TREATED (Column (1)), takes only two values and shows whether the firm in question received any subsidies (TREATED = 1) or not (TREATED = 0).

P_B_T (Column (2)) , is the (P)rofit level of the firm measured just (B)efore the (T)reatment (before the receipt of subsidies).

P_A_T (Column (3)), is respectively the (P)rofit of the firm measured (A)fter the (T)reatment (after the receipt of subsidies).

SECTOR (Column (4)), indicates in which industrial SECTOR the firm belongs to. In this example I have used Manufacturing =1 and High tech = 2.

P_DIFF (Column (5)), is simply the (P)rofit (DIFF)erence measured at firm level, before and after the receipt of the treatment (the subsidies); that is (5) = (3) - (2).

SUBSIDY (Column (6)), is the amount of SUBSIDIES the firm has received (that is, when TREATED = 1). For firms who have not participated in the program (TREATED=0) the amount is zero (0).

Table 2. Fictional data used in example measuring the impact of a business subsidy policy

UNIT No	TREATED (given subsidies) Yes = 1, No = 0	P_B_T Profit Before Treatment	P_A_T Profit After Treatment	SECTOR Manuf = 1, High T = 2	P_DIFF Difference in profit (growth)	Amount of SUBSIDY
	(1)	(2)	(3)	(4)	(5) = (3)-(2)	(6)
1	1	40	50	2	10	20
2	1	50	61	2	11	18
3	1	35	43	1	8	15
4	1	44	60	2	16	23
5	1	38	50	2	12	22
6	1	34	38	1	4	8
7	1	41	54	2	13	17
8	1	58	71	2	13	17
9	1	44	53	1	9	11
10	1	31	47	2	16	28
11	1	41	51	2	10	19
12	1	51	62	2	11	18
13	1	36	44	1	8	6
14	1	45	61	2	16	21
15	1	35	39	1	4	7
16	1	39	51	2	12	15
17	1	42	55	2	13	29
18	1	59	72	2	13	24
19	1	45	54	1	9	10
20	1	32	48	2	16	30
21	0	37	42	1	5	0
22	0	47	54	1	7	0
23	0	32	37	1	5	0
24	0	41	48	1	7	0
25	0	35	41	1	6	0
26	0	38	46	1	8	0
27	0	55	63	1	8	0
28	0	41	46	1	5	0
29	0	48	55	1	7	0
30	0	33	37	1	4	0
31	0	42	50	1	8	0
32	0	56	63	1	7	0
33	0	42	48	1	6	0
34	0	29	37	1	8	0
35	0	31	33	1	2	0
36	0	29	37	1	8	0
37	0	42	46	1	4	0
38	0	36	43	1	7	0
39	0	28	33	2	5	0
40	0	50	54	2	4	0
200*	1	50	250	2	200	32
201*	1	250	50	1	-200	10
400*	0	60	260	2	200	0
401*	0	260	60	1	-200	0

* = Outliers

Model

The model which I am using is the classic one[42]:

$Y = \kappa + \beta X$ or $Y = \kappa + \beta X + \gamma Z$, where

Y represents the dependent variable of interest, the measured impact. It is the profit difference (P_DIFF, a continuous variable in Column (5)). X is the independent variable of interest, and is represented in two ways: (a) as a subsidies amount (SUBSIDY, a continuous variable in Column (6)) and (b) as a binary variable, denoting whether the firm has received subsidies or not (TREATED, in Column (1)). Z is a control variable. It is the firm's industrial sector (SECTOR, a categorical variable -here binary- in Column (4)).

κ is the constant term indicating the value of Y (P_DIFF), if X (SUBSIDY or TREATED) and Z (SECTOR) are not affecting at all its value, that is their coefficients are zero. β is the coefficient which shows the actual (potential) effect of the subsidies on the profit difference of firms in ceteris paribus conditions (other things being equal). Finally, γ is the coefficient of the Z (SECTOR).

[42] See also Section 3.1 on causation and regression models, and Section 4.4.5 on the Difference in Difference (DiD) estimation.

Analysis

We begin by examining the data through some descriptive statistics. Table 3 lists the mean values of the impact indicator of interest, the change in profits of firms. I have listed the means initially based on whether the firm received subsidies or not (was treated or not) and then adding its industrial sector as a control. Remember we are interested in finding out whether the receipt of subsidies (the treatment) has had any effect on the profit (growth[43]) of the firm.

To demonstrate the effect of outlier values I have also included the same means but now adding selectively in the calculations, the outlier values of observations no. 200 (profit difference = 200) , 201 (-200), 400 (200) and 401 (-200).

The profit means are classified into 8 groups (A - H) based on what outliers are included in the calculations. The outliers seem to play an important role in this example, where the number of observations is relatively small and the deviation of the outliers from the rest of the observations is quite big. However, as the number of observations increase, the outliers' "influence" decreases. In any case, if the evaluator confirms that the extra high or low results are indeed valid (and not due to erroneous input), he can examine them separately for any particular characteristics. This examination, which could be vertical and detailed (i.e. qualitative in nature) might produce valuable insights as to some peculiarities of specific units which react exceptionally strong to the treatment administered through the whatever policy in question. Again, my main intention is not to focus on the effect of outliers per se, but to concentrate more on *how the means change* after we control for the treatment and moreover, after we include the sector as well.

The data in the table was deliberately constructed in such a way so that the firms which received subsidies have on average higher profit growth compared to the ones that did not. The same applies to the firms which were in the high tech sector; they also have, on average, a higher profit growth compared to their counterparts in the manufacturing industry. As can be seen, the impact means indeed change dramatically in all groups depending on whether one calculates the mean based on (a) all firms, (b) based on those that have received subsidies or not, (c) based on which industrial sector the firm belongs to, and finally (d) based on whether one takes into account several outlier values.

[43] Growth in this case can denote both positive and negative growth.

Table 3. Mean profit difference (growth) based on values before and after the treatment (before and after the receipt of subsidies)

Group	Treatment	Control (sector)	Outliers included (Obs no)	Obs	Mean Difference in profit
A	All firms	No	No	40	8,63
	0	No	No	20	6,05
	1	No	No	20	11,20
	0	1	No	18	6,22
	0	2	No	2	4,50
	1	1	No	6	7,00
	1	2	No	14	13,00
B	All firms	No	200, 201, 400, 401	44	7,84
	0	No	200, 201, 400, 401	22	5,50
	1	No	200, 201, 400, 401	22	10,18
	0	1	200, 201, 400, 401	19	-4,63
	0	2	200, 201, 400, 401	3	69,67
	1	1	200, 201, 400, 401	7	-22,57
	1	2	200, 201, 400, 401	15	25,47
C	All firms	No	200	41	13,29
	0	No	200	20	6,05
	1	No	200	21	20,19
	0	1	200	18	6,22
	0	2	200	2	4,50
	1	1	200	6	7,00
	1	2	200	15	25,47
D	All firms	No	201	41	3,54
	0	No	201	20	6,05
	1	No	201	21	1,14
	0	1	201	18	6,22
	0	2	201	2	4,50
	1	1	201	7	-22,57
	1	2	201	14	13,00
E	All firms	No	400	41	13,29
	0	No	400	21	15,29
	1	No	400	20	11,20
	0	1	400	18	6,22
	0	2	400	3	69,67
	1	1	400	6	7,00
	1	2	400	14	13,00
F	All firms	No	401	41	3,54
	0	No	401	21	-3,76
	1	No	401	20	11,20
	0	1	401	19	-4,63
	0	2	401	2	4,50
	1	1	401	6	7,00
	1	2	401	14	13,00
G	All firms	No	200, 401	42	8,21
	0	No	200, 401	21	-3,76
	1	No	200, 401	21	20,19
	0	1	200, 401	19	-4,63
	0	2	200, 401	2	4,50
	1	1	200, 401	6	7,00
	1	2	200, 401	15	25,47
H	All firms	No	201, 400	42	8,21
	0	No	201, 400	21	15,29
	1	No	201, 400	21	1,14
	0	1	201, 400	18	6,22
	0	2	201, 400	3	69,67
	1	1	201, 400	7	-22,57
	1	2	201, 400	14	13,00

In Table 4, I regress the difference in profit (P_DIFF) on the treatment variable of interest which is either in a binary (TREATED) or continuous (SUBSIDY) format. Note how the impact measured by the β coefficient in Columns (5) and (6), changes depending on whether the counterfactual is taken into account, whether the industrial sector is used as control and finally how the results are distorted by accounting or not for outlier impact values.

The estimation designs applied are either NR1 or NR2 (see Table 1) depending on whether we incorporate a control group (the counterfactual) or not in the models[44]. In Column (5), the Treatment variable in its binary

[44] I would have also liked to demonstrate a similar example, now using longitudinal data, but due to space constraints I did not pursue the idea further.

format, where as in Column (6) it's in its continuous format. Column (5) indicates how much on average the profit of firms differ by comparing those that have received and those that have not received subsidies. Column (6) on the other hand, shows how much the profit of firms change when we increase the receipt of subsidies to a firm by one. This is the "dose" estimation and in general the logic behind it is that "more is better".

Compare for example the β coefficient values (Column (6)) of models 1 and 3 in each of the Groups A - H. Model 1 takes, theoretically, into account the counterfactual situation by measuring the impact on profit on firms which have *and* have not received subsidies; model 3 measures the impact on profit solely on the subsidised firms. Because the subsidised firms have on average more profit growth than the unsubsidised ones (see the respective means in Table 3), the coefficients in model 3 are positively biased. But this is just one estimation and indeed does not take into account other factors which might also effect the impact. The situation is not so straight forward when for instance one includes in the model the industrial sector and selectively some outlier values as in models 2 and 4.

To conclude, as mentioned in Section 4.3.2, the basic message is clear. In retrospective (non-experimental) quantitative evaluation studies, where one does not apply random treatment assignment, the impact measurement is directly linked to what ever model specification and estimation design the evaluator chooses to implement on his data.

Which is the best design? Which is the best model specification? I shall answer these questions in Section 6.2.1.

Table 4. Regression coefficient of subsidies (binary and continuous variable) indicating the effect of subsidies on profit

Group	Model No.	Treatment All firms (TREATED = 1 & 0) Subsidised only (TREATED = 1)	Control (SECTOR)	Outliers included (Obs no)	Amount of obs (N)	β coefficient of binary variable of interest (TREATED)	β coefficient of continuous variable of interest: SUBSIDY>0 TREATED=1) SUBSIDY=0 TREATED=0)
		(1)	(2)	(3)	(4)	(5)	(6)
A	1	All firms	No	No	40	5,150	0,318
	2		Yes	No	40	2,940	0,288
	3	Subsidised only	No	No	19		0,420
	4		Yes	No	19		0,281
B	1	All firms	No	200, 201, 400, 401	44	4,682	1,170
	2		Yes	200, 201, 400, 401	44	-26,651	-0,692
	3	Subsidised only	No	200, 201	21		3,905
	4		Yes	200, 201	21		4,074
C	1	All firms	No	200	41	14,140	1,203
	2		Yes	200	41	6,465	1,612
	3	Subsidised only	No	200	20		2,568
	4		Yes	200			4,272
D	1	All firms	No	201	41	-4,907	0,267
	2		Yes	201	41	-19,182	-0,654
	3	Subsidised only	No	201			1,963
	4		Yes	201			2,659E-02
E	1	All firms	No	400	41	-4,086	-8,234E-02
	2		Yes	400	41	-19,582	-1,517
	3	Subsidised only	No	400			* (n/a)
	4		Yes	400			* (n/a)
F	1	All firms	No	401	41	14,962	0,742
	2		Yes	401	41	10,763	0,587
	3	Subsidised only	No	401			* (n/a)
	4		Yes	401			* (n/a)
G	1	All firms	No	200, 401	42	23,952	1,596
	2		Yes	200, 401	42	14,236	1,879
	3	Subsidised only	No	200, 401			* (n/a)
	4		Yes	200, 401			* (n/a)
H	1	All firms	No	201, 400	42	-14,143	-0,145
	2		Yes	201, 400	42	-37,962	-2,248
	3	Subsidised only	No	201, 400			* (n/a)
	4		Yes	201, 400			* (n/a)

* Here we do not conduct any estimations (as in the respective models of A-D) because some of the outliers belong to non-subsidised firms.

5. Cost benefit analysis in public policy evaluation

Money is king. This colloquialism is very relevant to public policy evaluation. To rephrase from Section 2, because policies are implemented through programs and because programs need money to run, accountability demands surface theoretically all the way from the tax payers towards the legislators, the government, the ministries, the agencies, their regional/local offices, and finally stop at the final recipients (targets) of the intervention.

The accountability demands may not be of the same strength all along. They are weaker, say, between the final recipients and the local agencies but stronger between the legislators and the government. Regardless of the uneven strength of accountability, the fact that it exists obliges the evaluator to not only estimate correctly the true net impact of the policy intervention, but also to pass a judgement on the estimated impact's worth. This is where the cost benefit analysis comes into the picture. And because the accountability is linked to the budgeted appropriations, one must translate the costs and benefits associated with the policy, *explicitly* into monetary terms.

One problem that the evaluator faces in such analysis is that there is not a clear definition of what can be classified as costs and what as benefits under the context of the public policy in question. Another problem is that even if all costs and benefits are finally identified, some are very difficult or even impossible to measure or to assign them a money value.

In the private sector things are much clearer because cost benefit analysis is normally done on a project or investment level basis. There, the relevant costs and benefits associated with the particular project are easily identified and quantified, since one can use private market price mechanisms (opportunity cost) for valuing them. Profitability is the *yardstick* with which private sector business activities are measured.

The aforementioned problems in the public sector can be approached first by defining at what level we shall conduct the cost benefit analysis. And this because different considerations are used at bottom (micro) level and different ones at a more general (higher) level. In economic jargon for example, cost-benefit analysis at bottom level is called *partial equilibrium* where as at higher level, *general equilibrium* analysis.

We could think that the difference between the two is on the *who* is affected by the policy. In partial equilibrium we measure the net effects on the *treated units only*. In general equilibrium, in addition to those that were treated directly, we take into consideration the effects that occur to *other units* as well. In both cases we still need to define the costs and benefits, in some basic common quantitative denominator (money), make the necessary comparisons and finally pass a judgement on the worth of the policy examined[45].

5.1 Partial equilibrium analysis

A logical point to begin is the general acceptable principle that the benefits of the policy should somehow be higher than the costs involved. If they end up being the opposite way, the concern is that the particular policy wastes society's resources, thus should be terminated, shift the funds to other more beneficial activities or should be altered so it can become beneficial.

Let us first identify on a per unit basis the costs and benefits associated with the public policy evaluated. The most common method applied in private sector projects and some big public sector investments is to define for some specific time period (or even in perpetuity) the net financial *outflows* of the project, the respective net *inflows* and *discount* them with a certain *rate* of return to the so called *present value* (PV) prices. If the discounted inflows (benefits) less the discounted outflows (costs) are positive then the project/policy is in principle acceptable[46].

I shall not discuss here the alternative use of the resources that one might bring forward; that the net present value of an alternative project/policy could be higher than the one under scrutiny. This is a question that would naturally come forward had the PV of the current policy turned out negative or positive but less than a predefined threshold.

[45] The following analysis is at a very elementary level. For a theoretically rich presentation of the general equilibrium approach when evaluating social programmes, see for example Heckman (1999a) and Heckman and Smith (1998).

[46] Here, I have used inflows as benefits and outflows as costs to denote that we are looking the whole process from the point of view of the government. Inflows mean that funds are received by the government and outflows that funds are spent by the government.

Consider again the business subsidy policy as a concrete example and let us attempt to identify the relevant costs and benefits[47].

Costs

Costs could include

- the amount of subsidies the firm has received for a specific project partially financed by the subsidy (S)
- the administrative expenses utilised for the implementation of the policy (A) and
- the opportunity cost of capital (O) that the government would have gained if it had invested the same amount of money, distributed as subsidies, in a risk free investment.

Benefits

As far as benefits are concerned we could identify:

- the net *input fund flows* of the intervention estimated through the first evaluation phase earlier, say the profits (P) and
- the increased taxation that the firm would be paying back to the government due to the increased profits (T)
- some *unidentified benefit (utility)* that usually the subsidised firm or an outside valuer qualifies as such, but which is difficult to put an explicit monetary value on (U).

If we assume that the effect of the policy is felt during a specific time frame, then the Cost Benefit formula becomes:

Net Present Value (NPV) of a policy =
Sum of discounted Benefits *less* sum of discounted Costs, for each period in question, at some discount rate

The formal notation is

$$NPV_{POLICY} = \left(\frac{B_1}{1+r} + \frac{B_2}{(1+r)^2} + \ldots + \frac{B_n}{(1+r)^n}\right) - \left(\frac{C_1}{1+r} + \frac{C_2}{(1+r)^2} + \ldots + \frac{C_n}{(1+r)^n}\right) \quad (10)$$

where,
(B)enefits/inflows = P+T+U
(C)osts/outflows = S+A+O
r = discount rate of return[48], [49]
n = periods within policy time frame

As can be seen from the above elementary analysis, the estimation is based on many assumptions and it can only be a very crude approximation of the true costs and benefits. What happens in practice, is that the evaluator through his model specification attempts initially to examine if the main benefits and costs differ, when comparing the growth of one relative to the other. How is this done?

Recall that in the regression models discussed earlier, the TREAT variable (the intervention variable) can take two formats; a binary and a continuous one. If we use the continuous format and account in our model specification for the counterfactual and for other control factors, the β coefficient of the TREAT indicates how much the effect, say in this case the profits (inflows/benefits) increase with one unit increase of the TREAT (the outflows/costs).

[47] The costs and benefits listed here are just an indicative sample of many possible types.

[48] The standard assumption is that r is 10 percent per year (US Office of Management and the Budget, 1972). Although 10 percent might sound too high, especially in times like today when interest rates are very low, a conservative approach is not rejected right up front. In practice, the evaluator can play with different yields, say r =0.1, 0.07, 0,05, 0,04, which can then apply either uniformly for each period in the time frame, or in different combinations.

[49] Equation (10) assumes that the there is no inflation. To account for this, each term of (10) must be divided by $(1 + \pi)$, $(1 + \pi)^2$, … $(1 + \pi)^n$ respectively, where π denotes the rate of inflation. However in practice one may argue that this is superfluous because when one anticipates inflation in the future, adjusts accordingly (increases) the discount rate of return r. Whatever the approach, the most important thing for these calculations is to be consistent and explicit when applying them. For example, when one compares the NPV of two policies he must use the same logic in calculating their NPV, either based on nominal values (not adjusting for inflation) or on real ones.

In this case, if the β coefficient is higher than 1, it means that the monetary benefits generated by the firms are *higher* that the costs used to generate them by the government. If the coefficient is less that 1 then the opposite occurs. We have then to make a judgement call and say whether we think the magnitude of the β coefficient is big enough to justify the program effect that it measures. If the coefficient is less than 1, then it is most likely that our judgements will be critical (negative) from a simple observable financial cost-benefit point of view[50].

In a more general case, the sign of the β coefficient is also of importance. Until now we have assumed that the sign is positive; that is, as the intervention (TREAT) variable increases, so does the effect variable (Y). However, there may be instances when the sign is negative. If all other statistical assumptions hold and the t-scores end up being statistically significant, this can be a quite worrying signal for the policy decision makers. Then we would have a case where the intervention not only fails to have a positive effect (even inefficient from a pure financial cost benefit angle, the β coefficient being less than 1 but still >0), but we now see that the policy does the contrary to what it was originally designed to achieve[51].

The problem of course is with the "U", the unobservable utility (benefit) assigned by anyone directly exposed to the policy, or even by an outsider. What value can one impose on this "benefit"? For example, many firms admit that the receipt of subsidies may have not been explicitly sound, judging from the firms' financial statements (they show low profits attributed to the subsidies receipts). Despite this however, they claim that the receipt of subsidies has in fact benefited them in many other ways, not explicitly shown in their financial statements (our main and most "objective" data source). They may claim that the fact that the firms received the subsidies (were treated with the intervention) this by itself caused them to invest further, assisted them in not having their profits reduced or even created such business opportunities that the increased profits will eventually materialise at some point in the future[52].

Maybe one argument of the public sector officials for the proliferation of programs, shown to be repeatedly non-efficient and non-effective, is exactly this: they also claim that the recipients themselves think that the program is useful irrespective of evaluations studies which show the opposite. However, one can clearly see an obvious bias in both arguments; they could be interpreted as attempts by both the bureaucrats and the recipients to simply maximise their own utility function.

Take of instance an SME in the manufacturing sector, that has a yearly turnover of EUR 500 000. These firms are very often recipients of subsidies. Assume that the firm receives a subsidy of EUR 30 000 for an investment. If the firm has a profit margin of 20%, to generate the same amount of subsidies from its own operations the firm would have needed to make an additional EUR 150 000 of sales or 30% of extra turnover from its current level. Hence, this can be interpreted as strong evidence of utility maximisation by the recipient firm. Public sector officials on the other hand, can not maximise their utility function through additional sales (profits), as firms or individuals in the private sector. They can attempt however to satisfy their utility function indirectly by maximising their bureau budget and through this, enjoy higher status, prestige, power, etc. (Niskanen, 1971). But budget maximisation in ministries and public agencies could mean in some cases the proliferation of programs and policies which are non-efficient (see more on this in Venetoklis (2001c)).

5.2 General equilibrium analysis

Consider again a business subsidy program to firms. Suppose that some recipient firms have increased their profits which was indeed one of the goals of the program. Nevertheless, in the process some other non-subsidised firms could have become less competitive due to this, have lost some market share and their profits have been reduced. This is in a general equilibrium context a *displacement* effect. Or one realises that some subsidised firms would have gone through with the investment even without the government intervention, which creates a *dead-weight effect*. Or one could even think the opposite; that the process has

[50] See in Section 4, Table 4 where in column (6), I list values of these coefficients in the fictional example. Notice those that are >1 and those that are <1.

[51] For examples of how the β coefficient is interpreted in evaluation of public policies, see Venetoklis (2001a) and in Kangasharju and Venetoklis (2002). Venetoklis (2001a) estimated the impact of business subsidies on the Value Added growth of the recipient firms. Kangasharju and Venetoklis (2002) estimated, among others, the effect of labour related subsidies on the growth of the salary expenditures of the subsidised firms. In both cases the assumption was that the recipient firms on average should have been able to produce a β coefficient of over 1. In other words, the recipient firms were expected to produce an amount of money at least equal to the subsidies received. And that, because the money given to them was free, that is with no obligation to return it. In both studies the β coefficient of the subsidies turned out consistently less than 1, although higher than 0. The authors concluded that, at least from a financial cost benefit point of view, the policies evaluated were ineffective and inefficient.

[52] These arguments are more evident, not in investment, but in R & D subsidies. R & D subsidy recipients and distributors justify them by claiming that they enable them to engage in otherwise risky long term investments with potentially high returns in the future.

had a *positive spillover* effect to other non-subsidised firms who became sub-contractors to the subsidised firms indeed because of their increased activity generated through their subsidised investment.

As another example, consider a labour market program providing training and then job search assistance to long term unemployed individuals. In a general equilibrium context, cost benefit analysis would include in addition to the costs and benefits at individual participant level, the effects to other individuals, non-participants in the program, all of course expressed in monetary terms. If the program does help the participants to get back to the active labour market a side –unintended- effect might be that non-participants short-term unemployed would not find jobs so easily any more. This could be classified as a *displacement* effect. Related to this, is a situation where the ability to hire people whose salaries are subsidised, makes firms replace (some of) those workers they would have hired, with those participating in the program; and this, exactly because of the attractiveness generated by their low labour cost. Still another case could be that firms take advantage of the subsidising jobs program and hire individuals although in the absence of the program they would have still increased their work force. In these two latter cases the firms created, via the intervention, a *dead-weight* effect (Smith (2000), pp. 25-26).

Just as in the partial equilibrium analysis, the increase or decrease of taxation revenue of the government because of changes in the working status of participating and non-participating units (firms/individuals) due to the program must be taken into account. All these costs and benefits should then be aggregated and, as in the partial equilibrium case, discounted to Present Values at a given discount rate.

5.3 Economic efficiency versus distributional justice

I will include here yet another consideration that the evaluator needs to be aware of when conducting cost benefit analysis at general equilibrium level. It is also relevant to the partial equilibrium analysis since one can justify through this, some arguments of the policy intervention recipients and the bureaucrats defending an otherwise inefficient policy.

In welfare economics there is a long tradition of considering social welfare as having (at least) two dimensions; that of *economic efficiency* and that of *distributional justice*. Economic efficiency concerns the size of the total wealth of *all members of society*, while distributional justice concerns the way that this total is *shared* amongst individuals. We could think of the cost benefit analysis we have described till now, as having an economic efficiency character. A policy is economically efficient, if its net present value is positive; otherwise it is inefficient. However, if we were to apply a welfare economics approach, one could say that public policy evaluation should not be judged based on efficiency criteria *only*; that is, it should not be judged *solely* on the differences between net outflows and net inflows generated from the implementation of the policy. In essence this means that if the efficiency cost benefit analysis shows positive net present values for the policy in question, then the judgement is positive. If on the other hand, the net present value is negative, this does not necessarily mean that the policy should be rejected outright. There might be some instances where, regardless of the apparent monetary inefficiency, the policy should nevertheless continue, for what society thinks are *just distributional reasons*.

Again we come to giving a monetary value to “U “, the utility that (some part(s) of) society, the recipients and even bureaucrats assign to the policy. Even if the efficiency analysis shows a negative difference, one might argue that society is nevertheless ready and willing to cover this deficit because of altruistic reasons. In other words, society would apply a more *just distribution* to its resources because she puts a greater implicit value on “U”.

For example, even if a business subsidies policy does not seem to have a positive effect on certain firms in remote regions around the country, the continuation of such a policy might be based on these arguments. With this logic however, if a policy is shown to cause a decrease in economic efficiency, then those who claim that the policy ought to still continue, must defend their arguments. They must show proof that the decrease in efficiency is outweighed by an improvement along some other dimension of social welfare, say by the need to support these firms in remote regions. And this is exactly the reasoning that comes with the distributional justice dimension (Sugden and Williams, 1978, pp. 93-94).

Finally, think of the training program for the unemployed. The government organises such programs to assist those people re-enter the job market. But, especially for the long-term unemployed of a certain age, some of these programs do not seem to produce the desired results (might not help the participants find a permanent job). Nonetheless, these programs are run on a continuous basis. How is this justified? Is it purely due to altruistic needs that society allocates certain of its resources, hypothesising that even the brief introduction of

those participants to training and to temporary subsidised jobs helps them? This reasoning would implicitly increase the “U” to levels higher than the tangible costs (fund outflows) incurred.

This indeed might be one explanation, but not the only one. It may be argued that society, thinking in a purely economically efficient way, invests in these programs so as to *prevent* the participants from generating extra social costs in the future, because of the potential physical and psychological illnesses related to long-term unemployment. Some have even claimed that programs for the unemployed simply *purchase social stability*. In that case the assumption is that if unemployed people stay unemployed too long, they may engage in anti-social activities such as riots and crimes, the cost of which would be by far higher than the programs in question (Rosen, 1995, p. 162).

What approach should the evaluator follow in formulating his final judgement on the policy? Is economic efficiency the best way or should the judgement incorporate distributional justice criteria as well? These and other questions will be answered in Section 6.2.2.

6. Discussion

Every public policy evaluation report includes (or should include) a discussion (a judgement) based on its findings. The discussion may emphasise the strengths and weaknesses of the evaluated policy and furthermore recommend certain actions to improve the areas in which flaws have been identified.

This paper is not an evaluation of a specific public policy per se; it is a descriptive account of some of the issues that are or ought to be found in a evaluation study of this type. In that respect, the discussion that follows includes some recommendations which are applicable, not on a specific public policy, but rather on the conduct and utilisation of the evaluation itself.

6.1 Narrowing the scope of public policy evaluation

I mentioned in the introduction that public policy evaluation is a complex activity. To narrow things down, the issues analysed in this paper are relevant to evaluations with some special characteristics. The discussion has focused on evaluations which are (a) *retrospective,* (b) *quantitative*, (c) are based on *non-experimental methods*, (d) are conducted at *micro level* and (e) utilise *secondary data sources*. I now give reasons why I have chosen to analyse just these types of public policy evaluations.

(a) Retrospective evaluation looks back to what has happened to the recipients of the interventions through a program. From a practical point of view this is the most common and logical way to go and of course is also in harmony with the developed accountability demands we discussed in Section 2. One can conduct ex ante (before the intervention) and ex nunc (during the implementation) evaluations but at the end of the day one wants to know what *has* happened, what *has* been the impact of the program through the intervention. In fact, one may classify an ex nunc evaluation in some sense as a kind of retrospective evaluation as well, because it does look back to see what has happened to the program, "*up to that point in time"*.

(b) I support the conduct of quantitative evaluations (versus qualitative ones) for several reasons. One is that qualitative evaluation may give us a direction of change, where as quantitative evaluation shows the magnitude of change (Chiang, 1974, p. 136). There is no objection that all evaluations are subjective (at the end, evaluators are called to pass judgements) but the subjectivity generated is much less using quantitative than qualitative criteria and measurements.

Martin (1998, p.150) argues, that it is essential for evaluation to develop objectives in specific and quantitative terms. This reduces ambiguity and provides clear targets to aim at. The same logic is adopted by McCloskey (1998, p.4, cited in Schreiner, 2001)). He says that

> *"…quantification is useful, not for its own sake but rather because it helps to make assumptions and judgements explicit. For example, financial cost-benefit analysis must be explicit about the financial costs and benefits included and must either assume away non-financial effects or make an explicit qualitative judgement about them… Some subjectivity is inevitable but excesses occur when judgements rest on unexamined experience, fuzzy logic, or implicit assumptions. Subjectivity is non-transparency; opaque or implicit factors lack inter-personal reliability, and this might let mistakes sneak out… Objectiveness or subjectiveness inheres not in an effect but rather in its measurement. Qualitative benefits and costs are unmeasured, unmeasurable, or measured in units with low interpersonal reliability; quantitative measures have high inter-personal reliability".*

Finally, quantitative evaluation can be applied horizontally; that is, it is easier to conduct when there is an abundance of data since it can utilise all of it. Qualitative evaluation is a vertical exercise where deep and selective analysis is conducted on a small sub-sample of the participating (and non-participating) population. This, in term, limits our ability to make inferences. To put it differently, with quantitative evaluation we can see better the general picture.

(c) Non-experimental methods are linked to the retrospective characteristic of evaluation. I am observing what has happened *after* the application of specific policy measures on a specific target population. Usually the criteria of eligibility for a unit to be treated are predefined in the policy specification and the policy targets are imposed in advance. In other words, we do not have in our hands an experimental design, where the intervention is randomly distributed among several members of the population. It is generally agreed that only through experimental approaches one can find with confidence the true (potential) impact of a public intervention. In practice however, this is difficult to conduct when the target units are as in our case, people

or firms. Laboratory experiments are much easier to apply, but experimental designs in a social framework are very expensive and many times are not practically possible due to ethical, legal or administrative reasons.

(d) The evaluation of a policy at *micro level* is probably the most important type of evaluation. The investigation of what has happened at the bottom end of the policy implementation, to the final recipient (and non-recipient) through the whatever policy measure, is paramount. The so called strategic evaluations at higher levels, are useful only if the evaluation at bottom level is first identified. What good would it do to conduct a strategic evaluation at higher level, find that some general process objectives have been achieved, but then realise that at the bottom (unit) level there has been minuscule or no impact, taking under consideration the amount of funds utilised?[53]

(e) To apply all the above one needs secondary[54] *micro level* data in electronic format. I refer mainly to financial information (balance sheet and profit and loss statements) when it comes to firms or other data on individuals. The secondary nature of the data is emphasised in order to reduce the chance of biased responses (in cases where the recipients of a specific intervention policy are asked to judge *themselves* on its worth). This well known response bias occurs due to a potential benefit that the respondents believe they gain, if they answer in a certain way (i.e. they believe that their chances of getting more of the intervention at a later time increases). Another distortion of the potential impact can happen even in spite of the honest intentions of the respondents. Because of the complexity of the environment in which they operate and the confounding factors that are also involved in determining the potential policy impacts, the respondents are not able to estimate precisely the investigated net impact (see more on this in Venetoklis, 2001b).

The emphasis on micro level data as a basis for policy analysis has been advocated at European Union level as well. The internal market European Council, in its conclusions on the Cardiff economics reform process (2000, cited in CEC (2001)) stressed among others, the importance of monitoring and evaluating the economic effects of the European Union's policies[55], notably through statistical (micro level) data.

Micro level data in electronic format is fortunately in abundance in Finland. Ministries and agencies gather information on firms and individuals participating in different programs, on a daily basis. Financial information of firms in electronic format (the main source giving the most objective picture of a company's status at a specific point in time and through the financial year) is available, with a lag time of only a year or two. Despite that the utilisation of all such information is somewhat restricted due to laws on personal and business confidentiality, with a minimum of data manipulation these problems can be solved[56].

6.2 General issues on evaluation utilisation

6.2.1 The ideal non-experimental design and model specification: Can there ever be one?

Which of the non-experimental designs mentioned in Section 4.3.2 is the best? Intuitively one sees that perhaps having longitudinal instead of cross-sectional measurements gives a better perspective of the patterns in certain variables of interest thus accounts better for the potential impact of the intervention. Also including in the model the counterfactual population, should help estimate better the true impact.

However, when it comes to choosing the actual model specification, that is the selection of variables that are to be used in the right hand side of the regression equation as controls[57], is a matter that is debatable. Why should we create a propensity score indicator and not find and utilise an instrumental variable? Why should we use a matching approach and not a simple list of observable control variables which we think are the best controls for this type of impact evaluation? After all, who can claim that the estimate of one specification is better or worse than the other?

Evaluation is about comparisons. Hence, in evaluating several model specifications we have to find a base, a benchmark to compare them against. How is this done? Researchers have devised a clever way of doing so. We saw earlier that the experimental design[58] is considered ideal for policy analysis, if it can be applied.

---

[53] Proponents of an overall approach to evaluation can read for example Virtanen and Uusikylä (2002); for a classification of evaluation models at different operational levels, see Vedung (1997, p. 36).
[54] By secondary data I refer to data not gathered directly from the source.
[55] This specific recommendation referred to State Aid policies.
[56] Once the identification (name) of a specific unit is deleted and as long as another variable is still included as a key (i.e. modified firm taxation code - Ly tunnus, or personal id code – henkilötunnus) one can link separate databases and analyse them easily.
[57] See also Section 4.4.
[58] In science nothing is take for granted and social experiments are not an exception either. Some argue that despite their obvious advantages due to randomisation, experiments still possess several limitations. I will not dwell on this school of thought. For a

Based on this, researchers have conducted wide based experiments for several public policies. In a frequently quoted labour policy experiment in the US, the policy of training unemployed individuals was evaluated. One goal was to examine how much the intervention (training) effected to incomes of the participants. The chosen individuals were assigned randomly to the Treatment (training) and to the Control (no training) group and their incomes were followed for some time after the training. The impact results produced were used as benchmark values. Next, the researchers re-analysed the data, but now they applied non-experimental methods of analysis. The results of the different models and designs were then compared against the ones of the experimental approach.

Although there has not been absolute agreement, the latest consensus indicates that for non-experimental designs, panel data estimations combined with a propensity score indicator produce results closest to the respective benchmark experimental design; in econometric jargon these methodologies produce “robust” estimations[59]. This is an exciting area in program evaluation, and I am sure we shall read more advances in the field in the years to come[60].

Again and to conclude, this is the ideal case. We might have all the good intentions to apply such a design, but we might be limited by the lack of data, or by the nature of the policy (horizontal – vertical). What ever the case the evaluator must report these limitations and proceed with the best possible design given the data at hand.

6.2.2 Cost benefit analysis: Where should the emphasis be?

Where should we focus our efforts in a cost benefit analysis? Should we conduct it at partial or general equilibrium level? Should we pass our judgements based on the efficiency dimension only or should we take the distributional justice rhetoric also into account?

Devarajan et al. (1997, p. 40) argue that in order for a policy to be worth undertaking, the evaluator must judge that at a minimum, the relevant benefits exceed the relevant costs linked to the particular policy intervention. And that should be valid in both partial and general equilibrium contexts. This means that at a minimum, one should conduct a cost-benefit analysis based on economic efficiency grounds.

The role of the evaluator is to give evidence based on facts, and not formulate policy as such. The final decision on a policy should be left to policy planners and decision makers. They probably have a better account of the parameters needed to judge on the value of ”U” (the unidentified/unquantified utility) and whether the policy can indeed be used as a distributional justice vehicle, even if from a purely efficiency point of view it is not justifiable.

An evaluation should begin at the bottom (at partial equilibrium level). This covers both the estimation of the impact phase and the cost benefit analysis phase using the efficiency dimension. If one finds in the partial equilibrium analysis that costs exceed the benefits by a large margin, it would be useless to try and justify the policy from an efficiency point of view, even at the general equilibrium level. If, on the other hand, the partial equilibrium results are positive, one can attempt a general equilibrium analysis as well, but still based only on efficiency grounds. The distributional justice aspect of the policy should not be touched.

6.2.3 The timing of retrospective evaluation: A utilisation paradox

In retrospective evaluations, the underlying logic is that by conducting them one learns from the past in order to apply this accumulated knowledge in the future. To do so we make some assumptions. One is that the conditions that prevailed during the time that the evaluated policy was implemented will *continue* in the future. Another is that the participants and non-participants will *continue to behave in the future as in the past,* if they are exposed in the future to the same exogenous conditions and policy interventions as during the period evaluated; that is, during the period based on which judgements and proposals were formulated for the new, upcoming program.

---

presentation of arguments for and against social experiments, see for example Heckman and Smith (1995), Smith (2000), and Riccio and Bloom (2001).

[59] For an extensive review of the more advanced evaluation methods and an estimation on their robustness, see Blundell and Dias (2000).

[60] For a presentation of such comparisons see for example LaLonde (1986), Dehejia and Wahba (1998), Smith and Todd (2000), Heckman, Ichimura and Todd (1997), Blundell and Dias (2000).

These two assumptions are *implicit* of course because we know that the state of the world and the behaviour of the actors directly or indirectly related to the program are dynamic. We can not identify behavioural patents consistently nor can we predict external environmental shocks with certainty.

But then the natural counter-argument that is raised is this: Why do we use such evaluations when we know that the conditions based on which their results and their recommendations were made can never be replicated? This is a utilisation paradox. The answer is that we have no better way of gathering knowledge even if it can never be replicated. We must satisfy ourselves with this. And naturally, this knowledge is better than no knowledge at all. But can we improve the quality of this knowledge? I would argue yes, through a *better timing on the conduct of retrospective evaluations*. Let me elaborate.

Evaluations take *time* to conduct. First, one needs to design the evaluation itself, gather the relevant material, analyse it, write the report and disseminate the results to the decision makers. Then the decision makers, in theory at least, have to take under consideration the evaluation's results when themselves design/redesign a program [61].

It is however normal that programs do not run for specific period of time, stop for reflection and continue. The process is dynamic and one program period follows the other without any interruptions. For overall evaluations, there is a *dead time zone* which can be measured in years (up to two or three sometimes) after the end of a specific program till the results of the evaluation are in the hands of the decision makers and planners. But by then the current program is already on its way with probably structured rules and measures which do *not necessarily reflect the lessons from the previous period,* simply because the evaluation results were not available at the time of design. The programs are then designed and implemented with a combination of ad-hoc procedures, feedback from interest groups and information from much older evaluation reports.

There is no question that the time lag paradox between past knowledge and its utilisation on current policies/activities is a phenomenon in all sciences. The difference, in the case of social programs towards individuals and firms, is that the environment as well as the units of interest are so heterogeneous that we should attempt to close the gap with every possible means. Otherwise we may very well end up using obsolete knowledge in designing/altering policies for these target groups.

Can we create an information system which can *reduce* the time lag between the possible utilisation of evaluation results and the period from which they were drawn? If this is achieved, it will decrease the probable exogenous heterogeneity that reigns, and the changes that may be incorporated in a program will reflect better the needs that are designed to fulfil.

In that system, there could be a *continuous flow of information* on estimated impact of the interventions, during the implementation process, directed towards the public officials responsible for designing and implementing the specific program. The program itself must also become more flexible. This means having clauses in the legislation that would allow the planners and decision makers to alter the initial goals as they see fit, and maybe shift funding from one program area to another easily. This could result in certain types of measures to continue, others altered and some even scrapped from the program altogether.

What is the main tool for such a system? It is the *data* referring to certain aspects of the policy and its potential impacts. As mentioned earlier Finland is a pioneer in this respect because, in almost every policy activity one finds in *abundance* material gathered in a systematic fashion, even on a daily basis[62]. Hence, the infrastructure is in place. If we can create the modules with which all this data is analysed and evaluated and then construct the reporting channels through which the results are disseminated to the relevant actors, we have for sure gone a long way. Because, by reducing this *time lag gap*, we automatically reduce the probability that the exogenous factors which prevailed during the last program period have been altered extensively during the coming period in which recommendations are to be taken into account.

[61] I mentioned earlier that for many reasons this usually does not happen.

[62] For example when it comes to business subsidy policies the Ministry of Trade and Industry (the major distributor of subsidies to firms in the country) maintains a database where all information relating to subsidy applications and to the applicant firms are fed on a daily basis. Also the Ministry of Labour (the organisation that is responsible for the planning and implementation of policies combating unemployment, has installed a similar database monitoring system. Through the system, groups of individuals who for example participate in some training program, are followed using a three month window (Räisänen, 2001). When the data is examined and analysed, this type of monitoring is converted into a retrospective evaluation, and conclusions are made on the policy's impacts. This is exactly the kind of continuous information system which is essential for streamlining and improving public policies.

6.2.4 Retrospective quantitative evaluations of public policies: What to look for

Public policy evaluations are by definition subjective. Thus, the main audience of those evaluation reports - the public officials related to the policy under scrutiny - must feel confident that the results and recommendations shown have a strong internal and external validity.

Strong *internal validity* means that the evaluation puts forward solid constructs creating a causal relationship between the policy intervention and the intended impact (effect), if that is indeed observed. As we are never 100% sure that the tested intervention is in fact the one cause for the variation of the effect (the indicator of impact), in the evaluation there must be ample proof that this is most probably the case. This can be advocated through (a) theory, (b) results in similar evaluation studies which resemble the ones in the current one, (c) the inclusion of a counterfactual population if possible, (d) solid design and advanced quantitative methods which delineate the selection bias problem, etc.

As far as *external validity* is concerned, one should be able to infer the reported estimations to the general population, if that is ever applicable. Because of data peculiarities frequently encountered in observational studies of public policies, it may often be the case that external inferences are seldom done, or even needed. In most cases for example, the examined population is *all* under scrutiny[63]. If however there is only a sample of the total exposed population evaluated, then assurances must be made that the sample is indeed a representative one.

There must also be a clear and explicit presentation on the considerations used when conducting a cost benefit analysis. Is the analysis going to be conducted at partial equilibrium level or will one attempt an overall general equilibrium cost benefit estimation? Shall we judge the results from a pure economic efficiency perspective or shall we use just distributional arguments?

What ever the approach followed, the main guidelines are that (a) the evidence collected, the estimations made and the judgements passed should be *as objective as possible*, where as (b) any assumptions made should be *clearly indicated*. The evaluator himself knows best the strengths and weaknesses of his work. Thus, there should also be a section in the evaluation report where all the limitations and constraints under which the analysis was made, are *stated explicitly*. This is what the reader should always look for.

Finally, if possible, the data upon which the estimations and cost benefit calculations were made should be *available* to other researchers for replication. This will add to the credibility of the produced results and recommendations[64].

Treasury of Canada (2001, pp. 27-29) summarises some of the above points as follows:

> *"Objectivity is of paramount importance in evaluative work. Evaluations are often challenged. Objectivity means that the evidence and conclusions can be verified and confirmed by people other than the original authors. Simply stated, the conclusions must follow from evidence. Evaluation information and data should be collected, analysed and presented so that if others conducted the same evaluation and used the same basic assumptions, they would reach similar conclusions. This is more difficult to do with some evaluation strategies than others, especially when the strategy relies heavily on the professional judgement of the evaluator… In particular, it should always be clear to the reader what the conclusions are based on, in terms of the evidence gathered and the assumptions used… When conclusions are ambiguous, it is particularly important that the underlying assumptions be spelled out. Poorly formulated conclusions often result when assumptions used in a study are not stated…"*

[63] I mentioned above that for the type of evaluations I discuss, I assume that there is available data for all units of the populations exposed to the interventions and similarly also data for some for those units that have not been exposed.

[64] As discussed in Section 6.1 this may not be possible due to confidentiality legislation preventing the dissemination of this data to other persons except those directly responsible for its analysis. Nevertheless, this is something that one should aim at. In the US for example, the data of LaLonde's (1986) paper was the base of numerous other studies.

References

Arjas, E. (2001). "Causal analysis and statistics: A social science perspective". In *European Sociological Review*, Vol. 17, No. 1, pp. 59-64.

Augurzky, B. (2000). *Matching the extremes: A sensitivity analysis based on read data.* University of Heidelberg. Department of Economics, Heidelberg (version as in October 2000).

Blundell, R. and Dias, M.C. (2000). "Evaluation methods for non-experimental data". In *Fiscal Studies*, Vol. 21, No. 4, pp. 427-468.

Chiang, A. (1974). *Fundamental methods of mathematical economics*, 2nd ed. Tokyo: McGraw-Hill, Kogakusha.

Commission of the European Communities - CEC (2001). *State Aid Scoreboard*, COM (2001) 412 Final: Brussels (18.7.2001).

Cook , T.D. and Campbell, D.C. (1979). *Quasi-Experimentation.* Chicago: Rand McNally.

Cox, D.R. and Wirmuth, N. (2001). "Some statistical aspects of causality". In *European Sociological Review*, Vol. 17, No. 1, pp. 65-74.

Dehejia, R. and Wahba, S. (1998). *Propensity Score matching methods for non-experimental causal studies*, NBER working paper, No. 6829: Cambridge, MA.

Devarajan, S., Squire, L. and Suthiwart-Narueput, S. (1997). "Beyond Rate of Return: Reorienting Project Appraisal". In *World Bank Research Observer*, Vol. 12, No. 1, pp. 35-46.

European Council (2000). *Conclusions.* 248th Council, Internal Market 16 March 2000, point 20.

Goldthorpe, J. H. (2001) . "Causation, Statistics and Sociology". In *European Sociological Review*, Vol. 17, No. 1, pp.1-20.

Heckman J.J., Ichimura, H. and Todd, P. (1998). "Matching as an econometric evaluation estimator". In *Review of Economic Studies*, Vol. 65, pp. 261-294.

Heckman, J.J and Smith, J. A. (1998). *Evaluating the welfare state*. NBER working paper, No. 6542: Cambridge, MA.

Heckman, J.J. (1979). "Sample selection bias as specification error". In *Econometrica*, Vol. 47, No. 1, pp.153-162.

Heckman, J.J. (1999a). "Accounting for heterogeneity, diversity and general equilibrium in evaluating social programmes". In *The Economic Journal*, Vol. 111, F654-699.

Heckman, J.J. (1999b). *Causal parameters and policy analysis in economics: A twentieth century retrospective.* NBER working paper, No. 7333: Cambridge, MA.

Heckman, J.J., Ichimura, H. and Todd, P. (1997). "Matching as an econometric evaluation estimator: Evidence from a Job Training program". In *Review of Economic Studies*, Vol. 64, pp. 605-654.

Heckman, J.J., LaLonde, R. and Smith, J. (1999). "The Economics and Econometrics of Active Labour Market Programs". In Ashenfelter, A. and D. Card, (eds.). *Handbook of Labour Economics*, Vol. 3, Amsterdam: Elsevier Science.

Hill, A.B. (1965). "The environment and disease: association or causation": In *Proceedings of the Royal Society of Medicine,* Vol. 58, pp. 295-300.

Holland, P. (1986). "Statistics and Causal Inference". In *Journal of the American Statistical Association*, Vol. 81, No. 396, pp. 945-960.

Kangasharju, A. and Venetoklis, T. (2002). *Business subsidies and employment of firms: Overall evaluation and regional extensions*. Government Institute for Economic Research – VATT discussion paper, No. 268: Helsinki.

Kluve, J., Lehmann, H. Schmidt, C. M. (2001). *Disentangling Treatment Effects of Polish Active Labour Market Policies: Evidence from Matched Samples.* University of Heidelberg, Department of Economics, Heidelberg (version as at 02.07.01).

LaLonde, R.L. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data". In *The American Economic Review*, Vol. 76, No. 4, pp. 604-620.

Lechner, M. (2000). *Identification and estimation of causal effects of multiple treatments under the conditional independence assumption.* University of St. Gallen. Swiss Institute for International Economics and Applied Economic Research (SIAW), St Gallen (version as in August 2000).

Lechner, M. (2001). *Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labour Market Policies.* University of St. Gallen. Swiss Institute for International Economics and Applied Economic Research (SIAW), St Gallen (version as in January 2001).

Martin, R.J. (1998). *"Western New York MTC uses evaluative information for continuous improvement".* In Shapira, P. and Youtie, J. (eds.). *Manufacturing modernisation: Implications of evaluation results for program improvement and policy development.* School of Public Policy at Georgia Institute of Technology and Georgia Tech Economic Development institute (pp. 149-152): Atlanta, GA.

McCloskey, D.N. (1998). *The Rhetoric of Economics*, 2nd ed. Madison: University of Wisconsin Press.

Merriam-Webster (2002) (At http://www.m-w.com/home.htm as at 03.05.02)

Moffitt, R. (1991). "Program evaluation with non-experimental data". In *Evaluation Review*, Vol. 15, No. 3, pp. 291-314.

Mohr, L.B. (1995). *Impact analysis for program evaluation,* 2nd ed. Thousand Oaks, CA: Sage.

Niskanen, W. (1971). "Bureaucracy and representative government". pp. 3-320, in Niskanen, W. (1994) *Bureaucracy and Public Economics*. Aldershot: Edward Elgar.

Paulos, J.A. (1991). *Beyond numeracy. An uncommon dictionary of mathematics*. London: Penguin books.

Reiter, J. (2000). "Using statistics to determine causal relationships". In *The Mathematical Association of America*, No. 107, pp. 24-32.

Riccio, J.A. and Bloom, H.S. (2001). *Extending the reach of randomised social experiments: New directions in evaluations of America welfare-to-work and employment initiatives.* Manpower Demonstration Research Corporation (MDRC). Working paper on research methodology. (version at in October 2001).

Rizzo, J.A. (2001). *Propensity Scoring Methods: A New Tool for Health Outcomes Analysis Using Retrospective Databases.* November 28, 2001.
(At http://hopes.med.ohio-state-edu/Presentation/Propensity%20Scoring%20Methods.doc as at 3.4.2002)

Rosen, H.S. (1995). *Public Finance*, 4th ed. Chicago: Irwin.

Rosenbaum, P.R. (2002). *Observational studies*, 2nd ed. New York: Springler.

Rossi, P., Freeman, H. and Lipsey, M.W. (1999). *Evaluation.* Beverly Hills, CA: Sage

Rubin, D.B. (1986). "Statistics and Causal Inference: Comment: Which Ifs have Causal Answers". In *Journal of the American Statistical Association*, Vol. 81, No. 396, pp. 961-962.

Räisänen, H. (2001). "Implementation issues in Finland: Experiences, developments and context of labour market policy measures". In *Labour Market Policies and Public Employment Service*, OECD: Paris

Schmidt, C.M. (2001). *Knowing what works. The case for rigorous program evaluation*. University of Heidelberg, Department of Economics discussion paper, No 347: Heidelberg.

Schreiner, M. (2001). *Evaluation and Microenterprise programs.* Washington University. Centre for Social Development : St Louis.

Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002). Experimental and Quasi-Experimental Designs for Generalised Causal Inference. Boston: Houghton-Mifflin.

Smith, J. (2000). *A critical survey of empirical methods for evaluating active labour market programs*. Manuscript, University of Ontario. Department of Economics, Ontario (version as at September 1, 2000).

Smith, J. and Todd, P. (2000*). Does matching overcome LaLonde's critique in observational studies for causal effects?* Manuscript, University of Pennsylvania.

SPSS (1999a). *Base 10.0 Applications guide*: Chicago

SPSS (1999b). *Regression Models 9.0*: Chicago

Sugden, R. and Williams, A. (1978). *The principles of practical cost-benefit analysis.* Oxford: Oxford University Press.

Suppes, P. (1970). *A probabilistic Theory of Causation.* Amsterdam: North Holland.

Treasury Board of Canada (2001). *Program Evaluation Methods: Measurement and Attribution of Program Results. Review, Practices and Studies.* Government Review and Quality Services (At http:// www.tbs-sct.gc.ca/eval/pubs/method/dwn.htm as at 10.4.2002)

Trochim, W.M. (2002) *Research methods knowledge base*. (At http://trochim.human.cornell.edu/kb/design.htm as at 16.04.02).

US Office of Management and Budget (1972*). Discount rates to be used in evaluating time-distributed costs and benefits.* Circular No. A-94 (rev.): Washington, D.C.

Vedung, E. (1997). *Public Policy and Program Evaluation*. New Brunswick: Transaction Publishers.

Vedung, E. (2002). *Utilisation of Evaluation*. Lecture given at VATT, Helsinki on 24.4.2002.

Weiss, C.H. (1998). *Evaluation*. 2nd ed. New Jersey: Prentice Hall.

Venetoklis, T. (2001a). "Impact of business subsidies on growth of firms – Preliminary evidence form Finnish panel data". In Venetoklis, T. (2001). *Business subsidies and bureaucratic behaviour. A revised approach* . Government Institute for Economic Research – VATT, research report No. 83: Helsinki.

Venetoklis, T. (2001b). "Methods applied in evaluating business subsidy programs: A Survey". In Venetoklis, T. (2001). *Business subsidies and bureaucratic behaviour: A revised approach*. VATT-Government Institute for Economic Research, research report No. 83: Helsinki.

Venetoklis, T. (2001c). "Business subsidies and bureaucratic behaviour". In Venetoklis, T. (2001). *Business subsidies and bureaucratic behaviour: A revised approach*. VATT-Government Institute for Economic Research, research report, No. 83: Helsinki.

Venetoklis, T. (2002). "Lectio Praecursoria: Business subsidies and bureaucratic behaviour": In *Hallinnon tutkimus (Administrative Studies)*, Vol. 21, No. 1, pp. 97-100.

Virtanen, P. and Uusikylä, P. (2002). *Julkisten Yritystukien Vaikuttavuusarvioinnin Käsikirja.* Kauppa- ja teollisuusministeriön hallinnonalalla. KTM: Helsinki (In Finnish).

Wooldridge, J. (2000). *Introductory econometrics: A modern approach*. South-Western College Publishing.

Wooldridge, J. (2002). *Econometric analysis of cross-section and panel data*. Cambridge, Massachusetts: The MIT Press.

Xenakis, D., Ronning, O., Kantomaa, T., and Helenius, H. (1995). "Reactions of the mandible to experimentally induced asymmetrical growth of the maxilla in the rat". In *The European Journal of Orthodontics*, Vol. 17, No. 1, pp. 15-24.

Appendix

Formal notations[65] for non-experimental designs

The following, exhibit the formal notations of the non-experimental designs I described earlier. Next to each equation, there is a number with the "F" at its end corresponding to the relevant equation above.

Design NR1 (One group before after design, cross - sectional)

True impact measurement (α)

$$\alpha = E(P^{T}_{t+1} \mid X_{t+1},\ D = 1) - E(P^{T}_{t+1} \mid X_{t+1},\ D = 0) \qquad (2F)$$

Estimated impact measurement

$$\beta = E(P^{T}_{t+1} \mid X_{t+1},\ D = 1) - E(P^{T}_{t-1} \mid X_{t-1},\ D = 1) \qquad (3F)$$

Assumptions for an unbiased estimated impact measurement

We strive for β = α, thus

$$E(P^{T}_{t+1} \mid X_{t+1},\ D = 0) = E(P^{T}_{t-1} \mid X_{t-1},\ D = 1) \qquad (4F)$$

In (2F), (3F) and (4F),

- P denotes the average impact (profit) before (t-1) and after (t+1) the intervention (the subsidies).
- X is an array of observable factors before (t-1) and after (t+1) the intervention.
- D is a binary indicator variable showing whether the population whose impact we measure has been exposed (D=1) or not (D=0) to the treatment; that is if D=1 this denotes that the firms have received subsidies and if D=0 they have not.
- The superscript T denotes the (T)reatment group.

Design NR2 (Two groups before after design [Difference in Difference – DiD], cross – sectional)

True impact measurement (α)

$$\alpha = E(P^{T}_{t+1} \mid X_{t+1},\ D = 1) - E(P^{T}_{t+1} \mid X_{t+1},\ D = 0) \quad (2F \text{ - repeated})$$

Estimated impact measurement (β)

$$\beta = (E(P^{T}_{t+1} \mid X_{t+1},\ D = 1) - E(P^{T}_{t-1} \mid X_{t-1},\ D = 1)) - (E(P^{C}_{t+1} \mid X_{t+1},\ D = 0) - E(P^{C}_{t-1} \mid X_{t-1},\ D = 0)) \qquad (5F)$$

Assumptions for an unbiased estimated impact measurement

$$E(P^{C}_{t+1} - P^{C}_{t-1} \mid X,\ D = 0) = E(P^{T}_{t+1} - P^{T}_{t-1} \mid X,\ D = 0) \qquad (6F)$$

We should aim for β = α, thus

$$E(P^{T}_{t+1} \mid X,\ D = 0) = E(P^{T}_{t-1} \mid X,\ D = 1) + E(P^{C}_{t+1} - P^{C}_{t-1} \mid X,\ D = 0) \qquad (8F)$$

In (2F), (5F), (6F) and (8F),

- P denotes the average impact (profit) before (t-1) and after (t+1) the intervention (the subsidies).
- X is an array of observable factors.
- D is a binary indicator variable showing whether the population whose impact we measure has been exposed (D=1) or not (D=0) to the treatment; that is if D=1 this denotes that the firms have received subsidies and if D=0 they have not.
- The superscript T denotes the (T)reatment group and C denotes the (C)ontrol group.

[65] Part of the formal notations presented, are based on Moffitt (1991).

Design NR3 (One group before after design, with trends)

True impact measurement (α)

$$\alpha = E(\Delta P^{T}_{t+k} \mid X,\ D = 1) - E(\Delta P^{T}_{t+k} \mid X,\ D = 0) \qquad (9F)$$

Estimated impact measurement (β)

$$\beta = E(\Delta P^{T}_{t+k} \mid X,\ D = 1) - E(\Delta P^{T}_{t-k} \mid X,\ D = 1) \qquad (10F)$$

Assumptions for an unbiased estimated impact measurement

Since we wish for β = α, then

$$E(\Delta P^{T}_{t-k} \mid X,\ D = 1) = E(\Delta P^{T}_{t+k} \mid X,\ D = 0) \qquad (11F)$$

In (9F), (10F) and (11F),

- t and k denote different time periods, 1..n, with t>k.
- ΔP denotes the average rate of change of the impact (profits) before (t-k) and after (t+k) periods from the intervention,
- X is an array of observable factors,
- D is a binary indicator variable showing whether the population whose impact we measure has been exposed (D=1) or not (D=0) to the treatment.
- The superscript T denotes the (T)reatment group

Design NR4 (Two groups before after design [Difference in Difference – DiD], with trends)

True impact measurement (α)

$$\alpha = E(\Delta P^{T}_{t+k} - \Delta P^{T}_{t-k} \mid X,\ D = 1) - E(\Delta P^{T}_{t+k} - \Delta P^{T}_{t-k} \mid X,\ D = 0) \qquad (11F)$$

Estimated impact measurement (β)

$$\beta = E(\Delta P^{T}_{t+k} - \Delta P^{T}_{t-k} \mid X,\ D = 1) - E(\Delta P^{C}_{t+k} - \Delta P^{C}_{t-k} \mid X,\ D = 0) \qquad (12F)$$

Assumptions for an unbiased estimated impact measurement

$$E(\Delta P^{C}_{t+k} - \Delta P^{C}_{t-k} \mid X,\ D = 0) = E(\Delta P^{T}_{t+k} - \Delta P^{T}_{t-k} \mid X,\ D = 0) \qquad (13F)$$

We should aim for α = β , thus

$$E(\Delta P^{T}_{t+k} \mid X,\ D = 0) = E(\Delta P^{T}_{t-k} \mid X,\ D = 0) + E(\Delta P^{C}_{t+k} - \Delta P^{C}_{t-k} \mid X,\ D = 0) \qquad (16F)$$

In (11F), (12F), (13F) and 16F),

- t and k denote different time periods, 1..n, with t>k.
- ΔP denotes the average rate of change of the impact before (t-k) and after (t+k) a certain period of the intervention,
- X is an array of observable factors,
- D is a binary indicator variable showing whether the population whose impact we measure has been exposed (D=1) or not (D=0) to the treatment.
- The superscript T denotes the (T)reatment group and C the (C)ontrol group

VATT-TUTKIMUKSIA -SARJASSA ILMESTYNEITÄ

PUBLISHED VATT-RESEARCH REPORTS

61. Korkeamäki Ossi: Valtion palkat yleisiin työmarkkinoihin verrattuna: vuodet 1989 - 1997. Helsinki 2000.
62. Uusitalo Roope: Paikallinen sopiminen ja yritysten työvoiman kysyntä. Helsinki 2000.
63. Milne David – Niskanen Esko – Verhoef Erik: Operationalisation of Marginal Cost Pricing within Urban Transport. Helsinki 2000.
64. Vaittinen Risto: Eastern Enlargement of the European Union. Transition in applicant countries and evaluation of the economic prospects with a dynamic CGE-model. Helsinki 2000.
65. Häkkinen Iida: Muuttopäätös ja aluevalinta Suomen sisäisessä muuttooliikkeessä. Helsinki 2000.
66. Pyy-Martikainen Marjo: Työhön vai eläkkeelle? Ikääntyvien työttömien valinnat työmarkkinoilla. Helsinki 2000.
67. Kyllönen Lauri - Räty Tarmo: Asuntojen hinta-laatusuhde Joensuussa, semiparametrinen estimointi. Helsinki 2000.
68. Kyyrä Tomi: Welfare Differentials and Inequality in the Finnish Labour Market Over the 1990s Recession. Helsinki 2000.
69. Perrels Adriaan: Selecting Instruments for a Greenhouse Gas Reduction Policy in Finland. Helsinki 2000.
70. Kröger Outi: Osakeyhtiöiden verotuksen investointikannustimet. Helsinki 2000.
71. Fridstrøm Lasse – Minken Harald – Moilanen Paavo – Shepherd Simon – Vold Arild: Economic and Equity Effects of Marginal Cost Pricing in Transport. Helsinki 2000.
72. Schade Jens – Schlag Bernhard: Acceptability of Urban Transport Pricing. Helsinki 2000.
73. Kemppi Heikki – Perrels Adriaan – Pohjola Johanna: Kasvihuonekaasupäästöjen alentamisen taloudelliset vaikutukset Suomessa. Vaiheen 1 Loppuraportti. Helsinki 2000.
74. Laine Veli – Uusitalo Roope: Kannustinloukku-uudistuksen vaikutukset työvoiman tarjontaan. Helsinki 2001.
75. Kemppi Heikki – Lehtilä Antti – Perrels Adriaan: Suomen kansallisen ilmasto-ohjelman taloudelliset vaikutukset. Vaiheen 2 loppuraportti. Helsinki 2001.
76. Milne David – Niskanen Esko – Verhoef Erik: Legal and Institutional Framework for Marginal Cost Pricing in Urban Transport in Europe. Helsinki 2001.
77. Ilmakunnas Seija – Romppanen Antti – Tuomala Juha: Työvoimapoliittisten toimenpiteiden vaikuttavuudesta ja ennakoinnista. Helsinki 2001.

78. Milne David – Niskanen Esko – Verhoef Erik: Acceptability of Fiscal and Financial Measures and Organisational Requirements for Demand Management. Helsinki 2001. (Not yet publiched).

79. Venetoklis Takis: Business Subsidies and Bureaucratic Behaviour. Helsinki 2001.

80. Riihelä Marja – Sullström Risto: Tuloerot ja eriarvoisuus suuralueilla pitkällä aikavälillä 1971-1998 ja erityisesti 1990-luvulla. Helsinki 2001.

81. Ruuskanen Petri: Sosiaalinen pääoma – käsitteet, suuntaukset ja mekanismit. Helsinki 2001.

82. Perrels Adriaan – Kemppi Heikki – Lehtilä Antti: Assessment of the Macro-economic Effects of Domestic Climate Policies for Finland. Helsinki 2001. Tulossa.

83. Venetoklis Takis: Business Subsidies and Bureaucratic Behaviour, A Revised Approach. Helsinki 2001.

84. Moisio Antti – Kangasharju Aki – Ahtonen Sanna-Mari: Menestyksen mitta? Vaihtoehtoisia mittareita aluetalouden kehitykselle. Helsinki 2001.

85. Tuomala Juha: Työvoimakoulutuksen vaikutus työttömien työllistymiseen. Helsinki 2002.

86. Ruotoistenmäki Riikka – Babygina Evgenia: The Actors and the Financial Affairs of the Northern Dimension. Helsinki 2002.

87. Kyyrä Tomi: Funktionaalinen tulonjako Suomessa. Helsinki 2002.

88. Räty Tarmo – Luoma Kalevi – Koskinen Ville – Järviö Maija-Liisa: Terveyskeskusten tuottavuus vuosina 1997 ja 1998 sekä tuottavuuseroja selittävät tekijät. Helsinki 2002.

89. Hakola Tuulia: Economic Incentives and Labour Market Transitions of the Aged Finnish Workforce. Helsinki 2002.