Blundell, Richard; Costa Dias, Monica

Working Paper

Alternative approaches to evaluation in empirical microeconomics

cemmap working paper, No. CWP26/08

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Blundell, Richard; Costa Dias, Monica (2008) : Alternative approaches to evaluation in empirical microeconomics, cemmap working paper, No. CWP26/08, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2008.2608

This Version is available at:
https://hdl.handle.net/10419/64802

cemmap
centre for microdata methods and practice

Alternative approaches to evaluation in empirical microeconomics

Richard Blundell
Monica Costa Dias

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP26/08

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Alternative Approaches to Evaluation in Empirical Microeconomics

Richard Blundell*

Monica Costa Dias&

Abstract

This paper reviews a range of the most popular policy evaluation methods in empirical microeconomics: social experiments, natural experiments, matching methods, instrumental variables, discontinuity design and control functions. It discusses the identification of both the traditionally used average parameters and more complex distributional parameters. In each case, the necessary assumptions and the data requirements are considered. The adequacy of each approach is discussed drawing on the empirical evidence from the education and labor market policy evaluation literature. We also develop an education evaluation model which we use to carry through the discussion of each alternative approach. A full set of STATA datasets are provided free online which contain Monte-Carlo replications of the various specifications of the education evaluation model. There are also a full set of STATA .do files for each of the estimation approaches described in the paper. The .do-files can be used together with the datasets to reproduce all the results in the paper.

* University College London and Institute for Fiscal Studies
Address: Institute for Fiscal Studies, 7 Ridgemount Street, London WC1E 7AE. R.blundell@ucl.ac.uk
http://www.ucl.ac.uk/~uctp39a/.
& Universidade do Porto and Institute for Fiscal Studies
Address: Institute for Fiscal Studies, 7 Ridgemount Street, London WC1E 7AE. monica_d@ifs.org.uk,
http://www.ifs.org.uk.
Acknowledgements: We would like to thank the editor and referees as well as graduate students and researchers at UCL, IFS and the COST A23 meeting in Paris 2008 for their helpful comments. This research is part of the program of work at the ESRC Centre for the Microeconomic Analysis of Public Policy at the Institute for Fiscal Studies. We would like to thank the ESRC for financial support. The usual disclaimer applies. The data used in this article can be obtained from Costa Dias, Institute for Fiscal Studies, 7 Ridgemount Street, London WC1E 7AE, monica_d@ifs.org.uk.

I. Introduction

The aim of this paper is to examine alternative evaluation methods in microeconomic policy analysis and to lay out the assumptions on which they rest within a common framework. The focus is on application to the evaluation of policy interventions associated with welfare programs, training programs, wage subsidy programs and tax-credit programs. At the heart of this kind of policy evaluation is a missing data problem. An individual may either be subject to the intervention or may not, but no one individual can be in both states simultaneously. Indeed, there would be no evaluation problem of the type discussed here if we could observe the counterfactual outcome for those in the program had they not participated. Constructing this counterfactual in a convincing way is a key ingredient of any serious evaluation method.

The choice of evaluation method will depend on three broad concerns: the nature of the question to be answered; the type and quality of data available; and the mechanism by which individuals are allocated to the program or receive the policy. The last of these is typically labeled the "assignment rule" and will be a key component in the analysis we present. In a perfectly designed social experiment, assignment is random. In a structural microeconomic model, assignment is assumed to obey some rules from economic theory. Alternative methods exploit different assumptions concerning assignment and differ according to the type of assumption made. Unless there is a convincing case for the reliability of the assignment mechanism being used, the results of the evaluation are unlikely to convince the thoughtful skeptic. Just as an experiment needs to be carefully designed, a structural economic model needs to be carefully argued.

In this review we consider six distinct, but related, approaches: (i) social experiment methods, (ii) natural experiment methods, (iii) discontinuity design methods, (iv) matching

methods, (v) instrumental variable methods and (vi) control function methods. The first of these approaches is closest to the "theory" free method of a clinical trial, relying on the availability of a randomized assignment rule. The control function approach is closest to the structural econometric approach, directly modeling the assignment rule in order to fully control for selection in observational data.[1] The other methods can be thought of lying somewhere in between often attempting to mimic the randomized assignment of the experimental setting but doing so with non-experimental data. Natural experiments exploit randomization to programs created through some naturally occurring event external to the researcher. Discontinuity design methods exploit "natural" discontinuities in the rules used to assign individuals to treatment. Matching attempts to reproduce the treatment group among the non-treated, this way re-establishing the experimental conditions in a non-experimental setting, but relies on observable variables to account for selection. The instrumental variable approach is a step closer to the structural method, relying on exclusion restrictions to achieve identification. Exactly what parameters of interest, if any, can be recovered by each method will typically relate to the specific environment in which the policy or program is being conducted.

In many ways the *social experiment method* is the most convincing method of evaluation since it directly constructs a control (or comparison) group which is a randomized subset of the eligible population. The advantages of experimental data are discussed in papers by Bassi (1983,1984) and Hausman and Wise (1985) and were based on earlier statistical experimental developments (see Cockrane and Rubin 1973; Fisher 1951). Although a properly designed social experiment can overcome the missing data problem, in economic evaluations it is frequently difficult to ensure that the experimental conditions have been met. Since programs are typically voluntary, those individuals "randomized in" may decide not to participate in the treatment. The measured program impact will therefore recover an

"intention to treat" parameter, rather than the actual treatment effect. Further, unlike in many clinical trials, it is not possible to offer the control group a placebo in economic policy evaluations. Consequently individuals who enter a program and then are "randomized out" may suffer a "disappointment" effect and alter their behavior. Nonetheless, well designed experiments have much to offer in enhancing our knowledge of the possible impact of policy reforms. Indeed, a comparison of results from non-experimental data can help assess appropriate methods where experimental data is not available. For example, the important studies by LaLonde (1986), Heckman, Ichimura and Todd (1998), Heckman et al. (1998), Heckman, Smith and Clements (1997) use experimental data to assess the reliability of comparison groups used in the evaluation of training programs. An example of a well conducted social experiment is the Canadian Self Sufficiency Project (SSP) which was designed to measure the earnings and employment responses of single mothers on welfare to a time-limited earned income tax credit program. This study has produced invaluable evidence on the effectiveness of financial incentives in inducing welfare recipients into work (see Card and Robbins 1998).

The *natural experiment approach* attempts to find a naturally occurring comparison group that can mimic the properties of the control group in the properly designed experiment. This method is also often labeled "difference-in-differences" since it is usually implemented by comparing the difference in average behavior before and after the reform for the eligible group with the before and after contrast for a comparison group. This approach can be a powerful tool in measuring the average effect of the treatment on the treated. It does this by removing unobservable individual effects and common macro effects by relying on two critically important identifying assumptions of (i) *common time effects across groups*, and (ii) *no systematic composition changes within each group*. The evaluation of the "New Deal for the Young Unemployed" in the UK is a good example of a policy design suited to this

approach. It was an initiative to provide work incentives to unemployed individuals aged 18 to 24. The program is mandatory and was rolled out in selected pilot areas prior to the national roll out. The Blundell et al. (2004) study investigates the impact of this program by using similar 18-24 years old in non-pilot areas as a comparison group.

The *discontinuity design method* exploits situations where the probability of enrollment into treatment changes discontinuously with some continuous variable. For example, where eligibility to an educational scholarship depends on parental income falling below some cut-off or achieving a specific test score. It turns out to be convenient to discuss this approach in the context of the instrumental variable estimator since the parameter identified by discontinuity design is a local average treatment effect similar to the parameter identified by IV but not necessarily the same. We contrast the IV and discontinuity design approaches.

The *matching method* has a long history in non-experimental evaluation (see Heckman, Ichimura and Todd 1997; Rosenbaum and Rubin 1985; Rubin 1979). The aim of matching is simple: to line-up comparison individuals according to sufficient observable factors to remove systematic differences in the evaluation outcome between treated and non-treated. Multiple regression is a simple linear example of matching. For this "selection on observables" approach, a clear understanding of the determinants of assignment rule on which the matching is based is essential. The measurement of returns to education, where scores from prior ability tests are available in birth cohort studies, is a good example. As we document below, matching methods have been extensively refined and their properties examined in the recent evaluation literature and they are now a valuable part of the evaluation toolbox. Lalonde (1986) and Heckman, Ichimura and Todd (1998) demonstrate that experimental data can help in evaluating the choice of matching variables.

The *instrumental variable method* is the standard econometric approach to endogeneity. It relies on finding a variable excluded from the outcome equation but which is also a determinant of the assignment rule. In the simple linear constant parameter model, the IV estimator identifies the treatment effect removed of all the biases which emanate from a non-randomized control. However, in "heterogeneous" treatment effect models, in which the impact parameter can differ in unobservable ways across individuals, the IV estimator will only identify the average treatment effect under strong assumptions and ones that are unlikely to hold in practice. Work by Imbens and Angrist (1994) and Heckman and Vytlacil (1999) provided an ingenious interpretation of the IV estimator in terms of local treatment effect parameters. We discuss these developments.

Finally, the *control function method* directly analyses the choice problem facing individuals deciding on program participation. It is, therefore, closest to a structural microeconomic analysis. The control function approach specifies the joint distribution of the assignment rule and treatment. It uses the specification of the assignment rule together with an excluded "instrument" to derive a control function which, when included in the outcome equation, fully controls for endogenous selection. This approach relates directly to the selectivity estimator of Heckman (1979).

As already noted, structural micro-econometric simulation models are perfectly suited for ex-ante policy simulation. Blundell and MaCurdy (1999) provide a comprehensive survey and a discussion of the relationship between the structural choice approach and the evaluation approaches presented here. A fully specified structural model can be used to simulate the parameter being estimated by any of the non-experimental estimators above. Naturally, such a structural model would depend on a more comprehensive set of prior assumptions and will be less robust to the structural assumptions. However, results from

evaluation approaches described above can be usefully adapted to assess the validity of a structural evaluation model.

We provide a running example of a structural model of schooling choices to assess each of the proposed econometric non-experimental methods on their ability to recover the returns to education. In the model, individuals differ with respect to educational attainment, which is partly determined by a subsidy policy and partly determined by other factors. This "workhorse" model of education and earnings is used to generate a simulated dataset. Each estimator is then applied to the simulated data with the intention of recovering (some average of) the returns to education. In such a controlled experiment, the true education effects and choices are perfectly understood. Such insight can be used to reveal where each estimator fails and to understand and compare the treatment effect parameters identified in each case. The specification of the education model is described in full detail in the appendix. A full set of STATA datasets are provided online which contain 200 Monte-Carlo replications of all the data sets used in the simulated models. There are also a full set of STATA .do files available online for each of the estimators described in this paper. The .do-files can be used together with the datasets to reproduce all the results in the paper.

The rest of paper is organized as follows. In the next section we ask what are we trying to measure in program evaluation.[2] We also develop the education evaluation model which we carry through the discussion of each alternative approach. Sections III to VIII are the main focus of this paper and present a detailed comparison of the six alternative methods of evaluation we examine here. In each case we use a common framework for analysis and apply each non-experimental method to the education evaluation model. The order in which we discuss the various approaches follows the sequence described above with one exception: we choose to discuss discontinuity design after instrumental variables in order to relate the approaches together. Indeed an organizing principle we use throughout this review is to relate

the assumptions underlying each approach to each other, so that the pros and cons of each can be assessed in common environment. Finally, in section IX we provide a short summary.

II. Which Treatment Parameter?

II.A. Average Treatment Effects

Are individual responses to a policy homogeneous or do responses differ across individuals? If the responses differ, do they differ in a systematic way? The distinction between homogenous and heterogeneous treatment responses is central to understand what parameters alternative evaluation methods measure. In the homogeneous linear model, common in elementary econometrics, there is only one impact of the program and it is one that would be common to all participants and non-participants alike. In the heterogeneous model, the treated and non-treated may benefit differently from program participation. In this case, the average treatment effect among the treated will differ from the average value overall or on the untreated individuals. Indeed, we can define a whole distribution of the treatment effects. A common theme in this review will be to examine the aspects of this distribution that can be recovered by the different approaches.

To ground the discussion, we consider a model of potential outcomes. As is common to most evaluation literature in economics, we consider an inherently static selection model. This is a simple and adequate setup to discuss most methods presented below. Implicitly it allows for selection and outcomes to be realized at different points in time but excludes the possibility of an endogenous choice of time of treatment. Instead, the underlying assumption is that treatment and outcomes are observed at fixed (if different) points in time and thus can be modeled in a static framework. The time dimension will be explicitly considered only when necessary.

A brief comment about notation is due before continuing with the formal specification of the model. In here and throughout the whole paper, we reserve Greek letters to denote the unknown parameters of the model and use upper case to denote vectors of random variables and lower case to denote random variables.

Now suppose one wishes to measure the impact of treatment on an outcome, *y*. Abstract for the moment from other covariates that may impact on *y*; such covariates will be included later on. Denote by *d* the treatment indicator: a dummy variable assuming the value one if the individual has been treated and zero otherwise. The potential outcomes for individual *i* are denoted by y_i^1 and y_i^0 for the treated and non-treated scenarios, respectively. They are specified as

$$(1)\quad \begin{aligned} y_i^1 &= \beta + \alpha_i + u_i \\ y_i^0 &= \beta + u_i \end{aligned}$$

where β is the intercept parameter, α_i is the effect of treatment on individual *i* and *u* is the unobservable component of *y*. The observable outcome is then

$$(2)\quad y_i = d_i y_i^1 + (1 - d_i) y_i^0$$

so that

$$y_i = \beta + \alpha_i d_i + u_i .$$

This is a very general model as no functional form or distributional assumptions on the components of the outcome have been imposed for now. In what follows we show how different estimators use different sets of restrictions.

Selection into treatment determines the treatment status, *d*. We assume this assignment depends on the available information at the time of decision, which may not completely reveal the potential outcomes under the two alternative treatment scenarios. Such information is summarized by the observable variables, *Z*, and unobservable, *v*. Assignment to treatment is then assumed to be based on a selection rule

(3) $d_i = \begin{cases} 1 & \text{if } d_i^* \geq 0 \\ 0 & \text{otherwise} \end{cases}$

where d^* is a function of Z and v

(4) $d_i^* = g(Z_i, v_i)$.

A popular specification for the selection rule is based on the assumption of a linear index:

(5) $d_i^* = 1(Z_i\gamma + v_i \geq 0)$

where γ is the vector of coefficients.

In this general specification, we have allowed for a heterogeneous impact of treatment, with α varying freely across individuals.[3] Estimation methods typically identify some average impact of treatment over some sub-population. The three most commonly used parameters are: the population average treatment effect (ATE), which would be the average outcome if individuals were assigned at random to treatment, the average effect on individuals that were assigned to treatment (ATT) and the average effect on non-participants (ATNT). If it is the impact of treatment on individuals of a certain type as if they were randomly assigned to it that is of interest, then ATE is the parameter to recover. On the other hand, the appropriate parameter to identify the impact of treatment on individuals of a certain type that were assigned to treatment is the ATT. Using the model specification above, we can express these three average parameters as follows

(6) $\alpha^{ATE} = E(\alpha_i)$

(7) $$\begin{aligned} \alpha^{ATT} &= E(\alpha_i | d_i = 1) \\ &= E(\alpha_i | g(Z_i, v_i) \geq 0) \end{aligned}$$

(8) $$\begin{aligned} \alpha^{ATNT} &= E(\alpha_i | d_i = 0) \\ &= E(\alpha_i | g(Z_i, v_i) < 0). \end{aligned}$$

An increasing interest on the distribution of treatment effects has led to the study of additional treatment effects in the recent literature (Bjorklund and Moffitt 1987; Imbens and Angrist 1994; Heckman and Vytlacil 1999). Two particularly important parameters are the local average treatment effect (LATE) and the marginal treatment effect (MTE). To introduce them we need to assume that the participation decision, *d* is a non-trivial function of *Z*, meaning that it is a non-constant function of *Z*. Now suppose there exist two distinct values of *Z*, say Z^* and Z^{**}, for which only a subgroup of participants under Z^{**} will also participate if having experienced Z^*. The average impact of treatment on individuals that move from non-participants to participants when *Z* changes from Z^* to Z^{**} is the LATE parameter

$$\alpha^{LATE}\left(Z^*, Z^{**}\right) = E\left(\alpha_i \left| d_i\left(Z^{**}\right) = 1, d_i\left(Z^*\right) = 0\right.\right)$$

where $d_i\left(Z\right)$ is a dichotomous random variable representing the treatment status for individual *i* drawing observables *Z*.

The MTE measures the change in aggregate outcome due to an infinitesimal change in the participation rate,

$$\alpha^{MTE}\left(p\right) = \frac{\partial E\left(y\left|p\right.\right)}{\partial p}.$$

Bjorklund and Moffitt (1987) were the first to introduce the concept of MTE, which they interpreted as being the impact of treatment on individuals just indifferent about participation if facing observables *Z* where *Z* yields a participation probability $p = P\left(d = 1\left|Z\right.\right)$. Under certain conditions, to be explored later, the MTE is a limit version of LATE.

All these parameters will be identical under homogeneous treatment effects. Under heterogeneous treatment effects, however, a non-random process of selection into treatment may lead to differences between them. However, whether the impact of treatment is homogeneous or heterogeneous, *selection bias* may be present.

II.B. The selection problem and the assignment rule

In non-experimental settings, assignment to treatment is most likely not random. Collecting all the unobserved heterogeneity terms together we can rewrite the outcome equation (2) as

$$(9)\quad \begin{aligned} y_i &= \beta + \alpha^{ATE} d_i + \left(u_i + d_i\left(\alpha_i - \alpha^{ATE}\right)\right) \\ &= \beta + \alpha^{ATE} d_i + e_i. \end{aligned}$$

Non-random selection occurs if the unobservable term *e* in (9) is correlated with *d*. This implies that *e* is either correlated with the regressors determining assignment, *Z*, or correlated with the unobservable component in the selection or assignment equation, *v*. Consequently there are two types of non-random selection: *selection on the observables* and *selection on the unobservables*. When selection arises from a relationship between *u* and *d* we say there is *selection on the untreated outcomes* as individuals with different untreated outcomes are differently likely to become treated. If, on the other hand, selection arises due to a relationship between α and *d* we say there is *selection on the (expected) gains*, whereby expected gains determine participation.

The result of selection is that the relationship between *y* and *d* is not directly observable from the data since participants and non-participants are not comparable. We will see later on that different estimators use different assumptions about the form of assignment and the nature of the impact to identify the treatment parameter of interest. Here we just illustrate the importance of some assumptions in determining the form and importance of selection by contrasting the homogeneous and heterogeneous treatment effect scenarios. Throughout the whole paper, the discussion revolves around the identification of different treatment effects' parameters by alternative methods while sample analog estimators will also be presented.

Under homogeneous treatment effects, selection bias occurs if and only if *d* is correlated with *u* since the outcome equation is reduced to

$$y_i = \beta + \alpha d_i + u_i$$

where α is the impact of treatment on any individual and, therefore, equals α^{ATE} and any of the other parameters defined above since α is constant across the population in this case. The OLS estimator will then identify

$$\mathrm{E}\left[\hat{\alpha}^{OLS}\right] = \alpha + \mathrm{E}\left[u_i \middle| d_i = 1\right] - \mathrm{E}\left[u_i \middle| d_i = 0\right]$$

which is in general different from α if *d* and *u* are related.

The selection process is expected to be more severe in the presence of heterogeneous treatment effects, when correlation between *e* and *d* may arise through *u* (selection on non-treated outcomes) or the idiosyncratic gains from treatment, $\alpha_i - \alpha^{ATE}$ (selection on gains). The parameter identified by OLS in this case is

$$\mathrm{E}\left[\hat{\alpha}^{OLS}\right] = \alpha^{ATE} + \mathrm{E}\left[\alpha_i - \alpha^{ATE} \middle| d_i = 1\right] + \mathrm{E}\left[u_i \middle| d_i = 1\right] - \mathrm{E}\left[u_i \middle| d_i = 0\right].$$

Note that the first term, $\alpha^{ATE} + \mathrm{E}\left[\alpha_i - \alpha^{ATE} \middle| d_i = 1\right]$, is the ATT. Thus, in the presence of selection on (expected) gains OLS will identify the ATT if *d* and *u* are not related.

II.C. A "running" evaluation example: returns to education

Throughout this review we will use a dynamic model of educational choice and returns to education to illustrate the use of each of the non-experimental methods. The model is solved and simulated under alternative conditions, and the simulated data is then used to discuss the ability of each method to identify informative parameters. In all cases the goal is to identify the returns to education.

This simulation exercise can be viewed as a perfectly controlled experiment. With full information about the nature of the decision process and the individual treatment effects,

we can understand the sort of problems faced by each method in each case. It also provides a thread throughout the paper that allows for comparisons between alternative estimators and an assessment of their relative strengths and weaknesses.

At this stage we will focus on the role of selection and heterogeneous effects in the evaluation problem. In the model, individuals differ with respect to a number of factors, both observable and unobservable to the analyst. Such factors affect the costs of and/or expected returns to education, leading to heterogeneous investment behavior. To study such interactions, we use a simpler version of the model by abstracting from considerations about the policy environment. This will be introduced later on, in section IV.E, in the form of a subsidy to advanced education. The model is described in full detail in Appendix 1.

We consider individuals indexed by *i* facing lifetime earnings *y* that depend, among other things, on education achievement. Individuals are heterogeneous at birth with respect to ability, θ, and family environment or family background, *z*. Their lives are modeled in two stages, age one and two. We assume there are only two levels of education, basic and advanced. The educational attainment is represented by the dummy variable *d* where *d=1* for advanced education and *d=0* for low education. In the first stage of life (age one), the individual decides whether to invest in high education based on associated costs and expected gains. It is assumed that the (utility) cost of education, *c*, depends on the observable family background, *z*, and the unobservable (to the researcher) *v*,

(10) $c_i = \delta_0 + \delta_1 z_i + v_i$

where (δ_0, δ_1) are some parameters.

In the second stage of life (age two) the individual is working. Lifetime earnings are realized, depending on ability, θ, educational attainment, *d*, and the unobservable *u*. We assume that *u* is unobservable to the researcher and is (partly) unpredictable by the individual

at the time of deciding about education (age one). The logarithm of lifetime earnings is modeled as follows

(11) $\ln y_i = \beta_0 + \beta_1 x_i + \alpha_0 d_i + \alpha_1 \theta_i d_i + u_i$

where *x* is some exogenous explanatory variable, here interpreted as region, (β_0, β_1) are the parameters determining low-skilled earnings and (α_0, α_1) are the treatment effect parameters for the general and ability-specific components, respectively.

The returns to high education are heterogeneous in this model for as long as $\alpha_1 \neq 0$, in which case such returns depend on ability. The individual-specific return on log earnings is $\alpha_i = \alpha_0 + \alpha_1 \theta_i$.

We assume θ_i is known by individual *i* but not observable by the econometrician. The educational decision of individual *i* will be based on the comparison of expected lifetime earnings in the two alternative scenarios

$$\mathrm{E}\left[\ln y_i \middle| x_i, d_i = 1, \theta_i, v_i\right] = \beta_0 + \beta_1 x_i + \alpha_0 + \alpha_1 \theta_i + \mathrm{E}\left[u_i \middle| v_i\right]$$
$$\mathrm{E}\left[\ln y_i \middle| x_i, d_i = 0, \theta_i, v_i\right] = \beta_0 + \beta_1 x_i + \mathrm{E}\left[u_i \middle| v_i\right]$$

with the cost of education in equation (10). Notice that we are assuming that *z* does not explain the potential outcomes except perhaps indirectly, through the effect it has on educational investment.

The *assignment (or selection) rule* will therefore be

$$d_i = \begin{cases} 1 & \text{if } \mathrm{E}\left[y_i \middle| x_i, d_i = 1, \theta_i, v_i\right] - \mathrm{E}\left[y_i \middle| x_i, d_i = 0, \theta_i, v_i\right] > \delta_0 + \delta_1 z_i + v_i \\ 0 & \text{otherwise} \end{cases}$$

so that investment in education occurs whenever the expected return exceedsthe observed cost.

In this simple model, the education decision can be expressed by a threshold rule. Let $\tilde{v}$ be the point at which an individual is indifferent between investing and not investing in

education. It depends on the set of other information available to the individual at the point of deciding, namely (x, z, θ). Then $\tilde{v}$ solves the implicit equation

$$\tilde{v}(x_i, z_i, \theta_i) = \mathrm{E}\left[y_i \middle| x_i, d_i = 1, \theta_i, \tilde{v}(x_i, z_i, \theta_i) \right] - \mathrm{E}\left[y_i \middle| x_i, d_i = 0, \theta_i, \tilde{v}(x_i, z_i, \theta_i) \right] - \delta_0 - \delta_1 z_i .$$

If tastes for education and work are positively related, *v* measures distaste for education and *u* measures unobserved productivity levels that are positively related with taste for work, then *v* and *u* are expected to be *negatively* correlated. This then means that, holding everything else constant, the higher *v* the higher the cost of education and the smaller the expected return from the investment. As *v* increases it will reach a point where the cost is high enough and the return is low enough for the individual to give up education. Thus, an individual *i* observing the state space (x_i, z_i, θ_i) will follow the decision process,

$$(12)\ d_i = \begin{cases} 1 & \text{if } v_i < \tilde{v}(x_i, z_i, \theta_i) \\ 0 & \text{otherwise} \end{cases}$$

and this implies that educated individuals are disproportionately from the low-cost/high-return group.

The data and estimation routines are freely available online, as desribed in Appendix 2. The datasets include life-cycle information under 3 alternative policy scenarios: unsubsidized advanced education, subsidized advanced education and subsidized advanced education when individuals unaware of the existence of a subsidy one period ahead of deciding about the investment. A set of STATA .do files run the alternative estimation procedures using each of the discussed methods and to produce the Monte-Carlo results. The .do-files can be used together with the datasets to reproduce all the results in the paper.

II.C.1. Homogeneous treatment effects

Homogeneous treatment effects occur if the returns are constant across the population, that is either $\alpha_1 = 0$ or $\theta_i = \theta$ over the whole population. In this case, the outcome equation (11) reduces to,

$$\ln y_i = \beta_0 + \beta_1 x_i + \alpha_0 d_i + u_i$$

and $\alpha^{ATE} = \alpha^{ATT} = \alpha^{ATNT} = \alpha_0$ while α_0 also equals α^{LATE} and α^{MTE} for any choice of *z*. In this case, the selection mechanism simplifies to $\tilde{v}(z_i)$.

If, in addition, *v* and *u* are mean independent, the selection process will be exclusively based on the cost of education. In this case, OLS will identify the true treatment effect α_0.

II.C.2. Heterogeneous treatment effects

Under heterogeneous treatment effects, education returns vary and selection into education will generally depend on expected gains. This causes differences in average treatment parameters. The ATE and ATT will now be,

$$\alpha^{ATE} = \alpha_0 + \alpha_1 \mathrm{E}(\theta_i)$$
$$\alpha^{ATT} = \alpha_0 + \alpha_1 \mathrm{E}\left[\theta_i \middle| v_i < \tilde{v}(x_i, z_i, \theta_i)\right].$$

If α_1 is positive, then high ability individuals will have higher returns to education and the threshold rule $\tilde{v}$ will be increasing in θ. This is the case where higher ability individuals are also more likely to invest in education. It implies that the average ability among educated individuals is higher than the average ability in the population because $\tilde{v}$ will be increasing in θ and so $\mathrm{E}\left[\theta_i \middle| v_i < \tilde{v}(x_i, z_i, \theta_i)\right] > \mathrm{E}[\theta_i]$. In this case it is also true that $\alpha^{ATT} > \alpha^{ATE}$.

Assuming θ is not observable by the analyst, the outcome equation (11) can be re-written as,

$$\ln y_i = \beta_0 + \beta_1 x_i + \left[\alpha_0 + \alpha_1 \mathrm{E}(\theta_i)\right] d_i + \left[u_i + \alpha_1 d_i \left(\theta_i - \mathrm{E}(\theta_i)\right)\right].$$

and OLS identifies

$$\begin{aligned}\mathrm{E}\left[\widehat{\left(\alpha_0 + \alpha_1 \mathrm{E}(\theta_i)\right)}^{OLS}\right] &= \left[\alpha_0 + \alpha_1 \mathrm{E}(\theta_i)\right] + \alpha_1 \mathrm{E}\left[\theta_i - \mathrm{E}(\theta_i) | d_i = 1\right] + \mathrm{E}\left[u_i | d_i = 1\right] - \mathrm{E}\left[u_i | d_i = 0\right] \\ &= \alpha_0 + \alpha_1 \mathrm{E}\left[\theta_i | d_i = 1\right] + \mathrm{E}\left[u_i | d_i = 1\right] - \mathrm{E}\left[u_i | d_i = 0\right].\end{aligned}$$

This is the ATT if (u, v) and (u, z) are two pairs of mean independent random variables, while the ATE will not be identified by OLS.[4] Indeed, as will become clear from the discussion below, the ATE is much harder to identify.

III. Social Experiments

III.A. Random assignment

Suppose that an evaluation is proposed in which it is possible to run a social experiment that randomly chooses individuals from a group to be administered the treatment. If carefully implemented, random assignment provides the correct counterfactual, ruling out bias from self-selection. In the education example, a social experiment would randomly select potential students to be given some education while excluding the remaining individuals from the educational system. In this case, assignment to treatment would be random, and thus independent from the outcome or the treatment effect.

By implementing randomization, one ensures that the treated and the non-treated groups are equal in all aspects apart from the treatment status. In terms of the heterogeneous treatment effects model we consider in this paper and described in equations (1)-(2), randomization corresponds to two key assumptions:

R1: $\mathrm{E}\left[u_i \middle| d_i = 1\right] = \left[u_i \middle| d_i = 0\right] = \left[u_i\right]$

R2: $\mathrm{E}\left[\alpha_i \middle| d_i = 1\right] = \left[\alpha_i \middle| d_i = 0\right] = \left[\alpha_i\right]$**.**

These randomization "assumptions" are required for recovering the average treatment effect (ATE).

Experiments are frequently impossible to implement. In many cases, such as that of education policies in general, it may not be possible to convince a government to agree to exclude/expose individuals from/to a given treatment at random. But even when possible, experimental designs have two strong limitations. First, by excluding the selection behavior, experiments overlook intention to treat. However, the selection mechanism is expected to be strongly determined by the returns to treatment. In such case, the experimental results cannot be generalized to an economy-wide implementation of the treatment. Second, a number of contaminating factors may interfere with the quality of information, affecting the experimental results. One possible problem concerns drop-out behavior. For simplicity, suppose a proportion *p* of the eligible population used in the experiment prefer not to be treated and when drawn into the treatment group decide not to comply with treatment. Non-compliance might not be observable, and this will determine the identifiable parameter.

To further explore the potential consequences of non-compliance, consider the research design of a medical trial for a drug. The experimental group is split into treatments, who receive the drug, and controls, who receive a placebo. Without knowing whether they are treatments or controls, experimental participants will decide whether to take the medicine. A proportion *p* of each group will not take it. Suppose compliance is unrelated with the treatment effect, α_i. If compliance is not observed, the identifiable treatment effect parameter is,

$\tilde{\alpha} = (1 - p)\mathrm{E}\left[\alpha_i\right]$

which is a fraction of the ATE. If, on the other hand, compliance is observable, the ATE can be identified from the comparison of treatment and control compliers.

Unfortunately, non-compliance will unevenly affect treatments and controls in most economic experiments. Dropouts among the treated may correspond to individuals that would not choose to be treated themselves if given the option; dropouts among the controls may be driven by many reasons, related or not to their own treatment preferences. As a consequence, the composition of the treatment and control groups conditional on (non-)compliance will be different. It is also frequently the case that outcomes are not observable for the drop-outs.

Another possible problem results from the complexity of contemporaneous policies in developed countries and the availability of similar alternative treatments accessible to experimental controls. The experiment itself may affect experimental controls as, for instance, excluded individuals may be "compensated" with detailed information about other available treatments, which in some cases is the same treatment but accessed through different channels. This would amount to another form of non-compliance, whereby controls obtain the treatment administered to experimental treatments.

III.B. Recovering the average return to education

In the education example described in section II.B, suppose we randomly select potential students to be enrolled in an education intervention while excluding the remaining students. If such experiment can be enforced, assignment to treatment would be totally random and thus independent from the outcome or the treatment effect. This sort of randomization ensures that treated and non-treated are equal in all aspects apart from treatment status. The randomization hypothesis (R1) and (R2) would be,

- $\mathrm{E}\left[u_i \middle| d_i = 1\right] = \left[u_i \middle| d_i = 0\right] = \left[u_i\right]$ - no selection on untreated outcomes; and

- $\mathrm{E}\left[\theta_i \middle| d_i = 1\right] = \left[\theta_i \middle| d_i = 0\right] = \left[\theta_i\right]$ - no selection on idiosyncratic gains.

These conditions are enough to identify the average returns to education in the experimental population using OLS,

$$\mathrm{E}\left[\widehat{\left(\alpha_0 + \alpha_1 \mathrm{E}\left(\theta_i\right)\right)}^{OLS}\right] = \alpha_0 + \alpha_1 \mathrm{E}\left(\theta_i\right)$$

which is the ATE.[5]

IV. Natural Experiments

IV.A. The difference-in-differences (DID) estimator

The natural experiment method makes use of naturally occurring phenomena that may induce some form of randomization across individuals in the eligibility or the assignment to treatment. Typically this method is implemented using a before and after comparison across groups. It is formally equivalent to a difference-in-differences (DID) approach which uses some naturally occurring event to create a "policy" shift for one group and not another. The policy shift may refer to a change of law in one jurisdiction but not another, to some natural disaster which changes a policy of interest in one area but not another, or to a change in policy that makes a certain group eligible to some treatment but keeps a similar group ineligible. The difference between the two groups before and after the policy change is contrasted – thereby creating a DID estimator of the policy impact.

In its typical form, DID explores a change in policy occurring at some time period *k* which introduces the possibility of receiving treatment for some sub-population. It then uses longitudinal data, where the same individuals are followed over time, or repeated cross section data, where samples are drawn from the same population before and after the

intervention, to identify some average impact of treatment. We start by considering the evaluation problem when longitudinal data is available.

To explore the time dimension in the data, we now introduce time explicitly in the model specification. Each individual is observed before and after the policy change, at times $t_0 < k$ and $t_1 > k$, respectively. Let d_{it} denote the treatment status of individual *i* at time *t* and d_i (without the time subscript) be the treatment group to which individual *i* belongs to. This is identified by the treatment status at $t = t_1$:

$$d_i = \begin{cases} 1 & \text{if } d_{it} = 1 \\ 0 & \text{otherwise} \end{cases}$$

The DID estimator uses a common trend assumption to rewrite the outcome equation (2) as follows

$$y_{it} = \beta + \alpha_i d_{it} + u_{it}$$
(13) where $\mathrm{E}\left[u_{it} \middle| d_i, t\right] = \mathrm{E}\left[n_i \middle| d_i\right] + m_t.$

In the above equation, *n*is an unobservable individual fixed effect and *m* is an aggregate macro shock. Thus, DID is based on the assumption that the randomization hypothesis ruling out selection on untreated outcomes (R1) holds in first differences

$$\mathrm{E}\left[u_{it_1} - u_{it_{10}} \middle| d_i = 1\right] = \left[u_{it_1} - u_{it_{10}} \middle| d_i = 0\right] = \left[u_{it_1} - u_{it_{10}}\right].$$

This assumption does not rule out selection on the unobservables but restricts its source by excluding the possibility of selection based on transitory individual-specific effects. Also, it does not restrict selection on idiosyncratic gains from treatment that would mimic the randomization hypothesis (R2). As a consequence, and as will be seen, it will only identify ATT in general.

Under the DID assumption we can write,

$$\text{(14)}\ \mathrm{E}\left[y_{it} \middle| d_i, t\right] = \begin{cases} \beta + \mathrm{E}\left[\alpha_i \middle| d_i = 1\right] + \mathrm{E}\left[n_i \middle| d_i = 1\right] + m_t & \text{if } d_i = 1 \text{ and } t = t_1 \\ \beta + \mathrm{E}\left[n_i \middle| d_i\right] + m_t & \text{otherwise.} \end{cases}$$

It is now clear that we can eliminate both β and the error components by sequential differences

$$\begin{aligned}\alpha^{ATT} &= \mathrm{E}\left[\alpha_i \middle| d_i = 1\right] \\ &= \left\{\mathrm{E}\left[y_{it} \middle| d_i = 1, t = t_1\right] - \mathrm{E}\left[y_{it} \middle| d_i = 1, t = t_0\right]\right\} \\ &\quad - \left\{\mathrm{E}\left[y_{it} \middle| d_i = 0, t = t_1\right] - \mathrm{E}\left[y_{it} \middle| d_i = 0, t = t_0\right]\right\}.\end{aligned} \tag{15}$$

This is precisely the DID identification strategy. The sample analog of equation (15) is the DID estimator:

$$\hat{\alpha}^{DID} = \left[\bar{y}_{t_1}^1 - \bar{y}_{t_0}^1\right] - \left[\bar{y}_{t_1}^0 - \bar{y}_{t_0}^0\right] \tag{16}$$

where $\bar{y}_t^d$ is the average outcome over group *d* at time *t*. DID measures the excess outcome change for the treated as compared to the non-treated, this way identifying the ATT, $\mathrm{E}\left[\hat{\alpha}^{DID}\right] = \alpha^{ATT}$.

Notice that, the DID estimator is just the first differences estimator commonly applied to panel data in the presence of fixed effects. This means that an alternative way of obtaining $\hat{\alpha}^{DID}$ is to take the first differences of (13) to obtain

$$y_{it_1} - y_{it_0} = \alpha_i d_{it} + \left(m_{t_1} - m_{t_0}\right) + \left(o_{it_1} - o_{it_0}\right)$$

where *o* represents the transitory idiosyncratic shocks. Under the DID assumptions, the above regression equation can be consistently estimated using OLS. Notice also that the DID assumption implies that the transitory shocks, o_{it}, are uncorrelated with the treatment variable. Therefore, the standard within groups panel data estimator is analytically identical to the DID estimator of the ATT under these assumptions (see Blundell and MaCurdy 1999).

From (14) it follows that repeated cross-sectional data would be enough to identify ATT for as long as treatment and control groups can be separated before the policy change, in period $t = t_0$. Such information is sufficient for the average fixed effect per group to cancel out in the before after differences.

IV.B. A DID Application: The New Deal Gateway in the UK

As an example, the DID approach has been used to study the impact of the "New Deal for the Young Unemployed", a UK initiative to provide work incentives to individuals aged 18 to 24 and claiming Job Seekers Allowance (UI) for six months. The program was first introduced in January 1998, following the election of a new government in Britain in the previous year. It combines initial job search assistance followed by various subsidized options including wage subsidies to employers, temporary government jobs and full time education and training. Prior to the New Deal, young people in the UK could, in principle, claim unemployment benefits indefinitely. Now, after six months of unemployment, young people enter the New Deal "Gateway", which is the first period of job search assistance. The program is mandatory, including the subsidized options part, which at least introduces an interval in the claiming spell.

The Blundell et al. (2004) study investigates the impact of the program on employment in the first 18 months of the scheme. In particular it exploits an important design feature by which the program was rolled out in certain pilot areas prior to the national roll out. A before and after comparison can then be made using a regular DID estimator. This can be improved by a matching DID estimator as detailed in section V.E. The pilot area based design also means that matched individuals of the same age can be used as an alternative control group.

The evaluation approach consists of exploring sources of differential eligibility and different assumptions about the relationship between the outcome and the participation decision to identify the effects of the New Deal. On the "differential eligibility" side, two potential sources of identification are used. First, the program being age-specific implies that using slightly older people of similar unemployment duration is a natural comparison group.

Second, the program was first piloted for three months (January to March 1998) in selected areas before being implemented nation-wide (the "National Roll Out" beginning April 1998). The same age group in non-pilot areas is not only likely to satisfy the quasi-experimental conditions more closely but also allows for an analysis of the degree to which the DID comparisons within the treatment areas suffer from both general equilibrium or market level biases and serious substitution effects. Substitution occurs if participants take (some of) the jobs that non-participants would have got in the absence of treatment. Equilibrium wage effects may occur when the program is wide enough to affect the wage pressure of eligible and ineligible individuals.

The study focuses on the change in transitions from the unemployed claimant count to jobs during the Gateway period. It finds that the outflow rate for men has risen by about 20 percent as a result of the New Deal program. Similar results show up from the use of within area comparisons using ineligible age groups as controls and also from the use of individuals who satisfy the eligibility criteria but reside in non-pilot areas. Such an outcome suggests that either wage and substitution effects are not very strong or they broadly cancel each other out. The results appear to be robust to pre-program selectivity, changes in job quality and different cyclical effects.

IV.C. Weaknesses of DID

IV.C.1. Selection on idiosyncratic temporary shocks: "Ashenfelter's dip"

The DID procedure does not control for unobserved temporary individual-specific shocks that influence the participation decision. If o is not unrelated to d, DID is inconsistent for the estimation of ATT and instead approximates the following parameter

$$\mathrm{E}\left[\hat{\alpha}^{DID}\right] = \alpha^{ATT} + \mathrm{E}\left[o_{it_1} - o_{it_0} \middle| d_i = 1\right] - \mathrm{E}\left[o_{it_1} - o_{it_0} \middle| d_i = 0\right]$$

To illustrate the conditions such inconsistency might arise, suppose a training program is being evaluated in which enrolment is more likely if a temporary dip in earnings occurs just before the program takes place - the so-called "Ashenfelter's dip" (see Ashenfelter1978; Heckman and Smith 1999). A faster earnings growth is expected among the treated, even without program participation. Thus, the DID estimator is likely to over-estimate the impact of treatment.

Another important example of this is in the use of tax reforms as natural experiments to measure the responsiveness of taxable income to changes in marginal tax rates (see Lindsey 1987; Feenberg and Poterba 1993, for repeated cross sections applications; Feldstein 1995, for a panel data application). Much of this literature has focused on individuals with high income, who generally experience more pronounced changes in the tax schedule and are arguably more responsive to changes in marginal tax rates. These papers have relied on DID to identify the elasticity of taxable income to marginal tax rates by comparing the relative change in taxable income of the highest income group with that of other (generally also high income) groups. However, in a recent discussion of these studies Goolsbee (2000) shows that reactions in anticipation of the tax changes that shift income across tax years around the policy change may account for most, if not all, of the responses identified in the earlier literature. This implies that previous estimates may be upward biased by disregarding such behavioral responses. Non-tax related increases in income inequality contemporaneous to the tax reforms has also been suggested to partly explain the identified effects, suggesting the occurrence of differential macro trends to be discussed below.

IV.C.2. Differential macro trends

Identification of ATT using DID relies on the assumption that treated and controls experience common trends or, in other words, the same macro shocks. If this is not the case,

DID will not consistently estimate the ATT. Differential trends might arise in the evaluation of training programs if treated and controls operate in different labor markets. For example, unemployment in different age groups is often found to respond differently to cyclical fluctuations. In particular, unemployment among the youngest is generally more volatile, responding more strongly to changes in macro conditions and thus exhibiting more pronounced rises and drops as the economy evolves.

Figure 1 illustrates what is meant by common trends. It refers to the New Deal study described above and compares treated and controls over time with respect to the outflows from unemployment. The common trends assumption holds when the curves for treated and controls are parallel. In our example, the curves are nearly parallel over most of the period. The only important exception is at the beginning of the observable period. The graph suggests that the common trends assumption on both control groups considered in the study is broadly valid.

The possibility of differential trends motivates the "differential trend adjusted DID estimator". Suppose we suspect that the common trend assumption of DID does not hold but can assume that selection into treatment is independent of the temporary individual-specific effect, o_{it}, under differential trends

$$\mathrm{E}\left[u_{it}\middle|d_i = d, t\right] = \mathrm{E}\left[n_i\middle|d_i = d\right] + q^d m_t$$

where q^d is a scalar allowing for differential macro effects across the two groups (d represents the group and is either one or zero).

The DID estimator now identifies

$$\mathrm{E}\left[\hat{\alpha}^{DID}\right] = \alpha^{ATT} + \left(q^1 - q^0\right)\mathrm{E}\left[m_{t_1} - m_{t_0}\right]$$

which does not recover the true ATT unless $q^1 = q^0$, in which case we are back to the standard DID assumption.

Given the availability of data, one possible solution is to compare the trends of treated and controls historically, prior to the intervention. Historical, pre-reform data can help if there exists another time interval, say (t_*, t_{**}) (with $t_* < t_{**} < k$, over which a similar macro trend has occurred. In that case, by comparing the DID estimate of the impact of treatment contaminated with the bias from differential trend with the estimate of the differential trend over (t_*, t_{**}) one can separate the true impact of treatment from the differential trend.

More precisely, suppose one finds a pre-reform period, (t_*, t_{**}) for which the differential macro trend matches the bias term in the DID estimator, $\left(q^1 - q^0\right)\left(m_{t_1} - m_{t_0}\right)$. That is, $\left(q^1 - q^0\right)\left(m_{t_{**}} - m_{t_*}\right) = \left(q^1 - q^0\right)\left(m_{t_1} - m_{t_0}\right)$.

This means that there is a point in history where the relative conditions of the two groups being compared, treatments and controls, evolves similarly to what they do in the pre-post reform period, (t_0, t_1). Together with the absence of policy reforms that affect the outcome *y* during (t_*, t_{**}), this condition allows one to identify the bias term $\left(q^1 - q^0\right)\left(m_{t_1} - m_{t_0}\right)$ by applying DID to that pre-reform period. The impact of treatment can now be isolated by comparing DID estimates for the two periods, (t_0, t_1) and (t_*, t_{**}). This is the differentially adjusted estimator proposed by Bell, Blundell and Van Reenen (1999), which will consistently estimate ATT,

(17) $\hat{\alpha} = \left\{\left[\bar{y}^1_{t_1} - \bar{y}^1_{t_0}\right] - \left[\bar{y}^0_{t_1} - \bar{y}^0_{t_0}\right]\right\} - \left\{\left[\bar{y}^1_{t_{**}} - \bar{y}^1_{t_*}\right] - \left[\bar{y}^0_{t_{**}} - \bar{y}^0_{t_*}\right]\right\}$.

It is likely that the most recent cycle is the most appropriate, as earlier cycles may have systematically different effects across the target and comparison groups. The similarity of subsequent cycles, and thus the adequacy of differential adjusted DID, can be accessed in the presence of a long history of outcomes for the treatment and control groups.

IV.D. DID with Repeated Cross-sections: compositional changes

Although DID does not require longitudinal data to identify the true ATT parameter, it does require similar treatment and control groups to be followed over time. In particular, in repeated cross-section surveys the composition of the groups with respect to the fixed effects term must remain unchanged to ensure before-after comparability. If before-after comparability does not hold, the DID will identify a parameter other than ATT. We will illustrate this problem within our running education example.

IV.E. Non-linear DID models

A restrictive feature of the DID method is the imposition of additive separability of the error term conditional on the observables, as specified in equation (13). Recent studies have proposed ways of relaxing this assumption. In their analysis of the New Deal for the Young People, Blundell et al. (2004) noted that linearity in the error term can be particularly unrealistic when the outcome of interest is a dummy variable. In such case, the DID method can conceivably predict probabilities outside the [0,1] range. Instead, the authors suggest using the popular index models and assuming linearity in the index. Unfortunately, DID loses much of its simplicity even under a very simple non-linear specification.

To extend DID to a non-linear setting, suppose the outcome equation is now:

(18) $y_{it} = 1(\beta + \alpha_i d_{it} + u_{it} > 0)$

where $1(A)$ is the indicator function, assuming the value one if A is true and zero otherwise. As before,

$u_{it} = n_i + m_t - o_{it}$

and the DID assumption holds,

$$\mathrm{E}\left[u_{it}\middle|d_i,t\right]=\mathrm{E}\left[n_i\middle|d_i\right]+m_t$$

where d_i represents the treatment group. Additional assumptions are required. We assume *o* follows a distribution F where F is invertible.[6] Denote by F^{-1} the inverse probability rule. We simplify the model further by assuming a common group effect instead of allowing for an individual-specific effect: it is assumed that $n_i=n_d$ for *d*=0,1 being the post-program treatment status of individual *i*.[7]

Under these conditions and given a particular parametric assumption about the shape of F, say normal, one could think of mimicking the linear DID procedure by just running a probit regression of *y* on *d* and dummy variables for group and time (and possibly other exogenous regressors *x*) hoping this would identify some average of the treatment parameter α. One could then average the impact on *y* over the treated to recover the average treatment effect on the treated (the individual impact would depend on the point of the distribution where the individual is before treatment).

Unfortunately, this is not a valid approach in general. The problem is that the model contains still another error component which has not been restricted and that, under general conditions, will not fulfill the probit requirements. To see this, notice we can re-write model (18) as follows:

$$y_{it}=1\left(\beta+\alpha^{ATE}d_{it}+d_{it}\left(\alpha_i-\alpha^{ATE}\right)+n_d+m_t-o_{it}>0\right)$$

where $d_{it}\left(\alpha_i-\alpha^{ATE}\right)$ is part of the error term. Standard estimation methods would require a distributional assumption for $\left(\alpha_i-\alpha^{ATE}\right)$ and its independence from the treatment status.

Instead of imposing further restrictions in the model, we can progress by noticing that under our parametric setup,

$$\mathrm{E}\left[y_{it}^0\middle|d_i=d,t\right]=\mathrm{F}\left(\beta+n_d+m_t\right)$$

where, as before, $\left(y^0, y^1\right)$ are the potential outcomes in the absence and in the presence of treatment, respectively. But then the index is recoverable given invertibility of the function F, $\beta + n_d + m_t = \mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^0 | d_i = d, t\right]\right)$.

Using this result it is obvious that the trend can be identified from the comparison of non-treated before and after treatment:

(19) $m_{t_1} - m_{t_0} = \mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^0 | d_i = 0, t_1\right]\right) - \mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^0 | d_i = 0, t_0\right]\right)$.

Moreover, given the common trend assumption it is also true that, would we be able to observe the counterfactual of interest, $\mathrm{E}\left[y_{it}^0 | d_i = 1, t_1\right]$

(20) $m_{t_1} - m_{t_0} = \mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^0 | d_i = 1, t_1\right]\right) - \mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^0 | d_i = 1, t_0\right]\right)$.

But then (19) and (20) can be combined to form the unobserved counterfactual as follows:

$$\mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^0 | d_i = 1, t_1\right]\right) = \mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^0 | d_i = 1, t_0\right]\right) + \left\{\mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^0 | d_i = 0, t_1\right]\right) - \mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^0 | d_i = 0, t_0\right]\right)\right\}.$$

Let the average parameter of interest be α^{ATT}, which measures the average impact among the treated on the inverse transformation of the expected outcomes. Then[8]

$$\begin{aligned}\alpha^{ATT} &= \mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^1 | d_i = 1, t_1\right]\right) - \mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^0 | d_i = 1, t_1\right]\right) \\ &= \left\{\mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^1 | d_i = 1, t_0\right]\right) - \mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^0 | d_i = 1, t_0\right]\right)\right\} - \\ &\quad \left\{\mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^0 | d_i = 0, t_1\right]\right) - \mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^0 | d_i = 0, t_0\right]\right)\right\}.\end{aligned}$$

Rearranging, the missing counterfactual is

$$\mathrm{E}\left[y_{it}^0 | d_i = 1, t_1\right] = \mathrm{F}\left\{\mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^1 | d_i = 1, t_1\right]\right) - \alpha^{ATT}\right\}.$$

Using this expression, the ATT can be estimated by replacing the expected values by their sample analogs,

$$\widehat{ATT} = \bar{y}_{t_1}^1 - \mathrm{F}\left\{\mathrm{F}^{-1}\left(\bar{y}_{t_1}^1\right) - \hat{\alpha}^{ATT}\right\}$$

where

$$\hat{\alpha}^{ATT} = \left[\mathrm{F}^{-1}\left(\bar{y}_{t_1}^{1}\right) - \mathrm{F}^{-1}\left(\bar{y}_{t_0}^{1}\right) \right] - \left[\mathrm{F}^{-1}\left(\bar{y}_{t_1}^{0}\right) - \mathrm{F}^{-1}\left(\bar{y}_{t_0}^{0}\right) \right].$$

Recently, Athey and Imbens (2006) developed a general non-linear DID method specially suited for continuous outcomes: the "changes-in-changes" (CIC) estimator.[9] The discussion of this method is outside the scope of this paper (we refer the interested reader to the original paper by Athey and Imbens, 2006).

IV.F. Using DID to estimate returns to education

Since formal education occurs earlier in the life-cycle than labor market outcomes, it is generally not possible to evaluate the returns to education using earnings of treated and controls before and after the treatment. However, in the presence of an exogenous change in the environment leading to potential changes in education decisions, one may be able to identify some policy interesting parameter from the comparison of different cohorts. To explore this possibility, we consider the simulation model discussed before but now extended to include an education subsidy.

In this extension to the simple life-cycle model of education investment and working first discussed in section II.B, the individual lives for three periods corresponding to basic education, advanced education and working life. These are denoted by age being zero, one and two, respectively. The new education subsidy is available to individuals going into advanced education. Eligibility to subsidised education depends on academic performance during basic education as described in more detail below.

At birth, age zero, individuals are heterogeneous with respect to ability, θ, and family background, z. At this age all individuals complete basic education and perform a final test. The score on this test depends on innate ability (θ) and the amount of effort (e) the individual decides to put in preparing for it:

(21) $s_i = \gamma_0 + \gamma_1 \theta_i e_i + q_i$

where q is the unpredictable part of the score and (γ_0, γ_1) are some parameters. Effort e carries some utility cost, as described in Appendix 1. The (stochastic) payoff to this effort is the possibility of obtaining a subsidy to cover (part of) the cost of advanced education if scoring above a minimum level $\underline{s}$.

The remaining of the individual's life follows as explained before in section II.B. At age one the individual decides whether to invest in advanced education. In the presence of a subsidy to advanced education the cost is

(22) $c_i = \delta_0 + \delta_1 z_i - 1(s_i > \underline{s})S + v_i$

where $1(A)$ is the characteristic function assuming the value one if proposition A is true and zero otherwise.

Finally, age two represents the working life and earnings are realized. The logarithm of earnings are defined as in equation (11):

$\ln y_i = \beta_0 + \beta_1 x_i + \alpha_0 d_i + \alpha_1 \theta_i d_i + u_i$.

IV.F.1 Specification details and true parameters

For future reference and comparison purposes, we now present a couple of important model parameters along with the true effects. Further specification details can be found in Appendix 1.

The main set of estimates refers to the case where the unobservables in the cost of education (equation (22)) and in the earnings equation (equation (11)) are negatively correlated with the correlation coefficient being -0.5. In all cases we will also consider the uncorrelated model, where all selection occurs in the observables. Eligibility to education subsidy is determined by the test score with individuals scoring above $\underline{s} = 4$ being eligible.

The exogenous explanatory variable, *x*, is assumed to be discrete, in the present case assuming only two values, zero and one. The exogenous determinant of the cost of education, *z*, is assumed to follow a truncated normal distribution in the interval [-2,2]. The unobservable level of ability, θ is also assumed to be a truncated normal but this time in the interval [0,1].

Table 1 presents the true ATE and ATT as well as a range of statistics characterizing the selection process in the presence of an education subsidy. Other parameters and other cases not systematically discussed throughout the paper will be introduced only when relevant. All numbers result from simulations for an economy with an education subsidy under the assumption that individuals are perfectly informed about funding opportunities at birth.

Numbers in table 1 represent two alternative cases depending on whether the correlation between the unobservables *u* and *v* is negative (column 1) or zero (column 2). In both cases, the presence of strong selection mechanisms is quite evident. Rows 1 and 2 display the ATE and ATT respectively, and these are markedly different with the average participant benefiting more from advanced education then the average individual. The presence of selection is also suggested by the numbers in rows 4 to 11 as the average of each variable among the educated is very different from that over the whole population.

Row 5 shows that being eligible to subsidized education is a stronger determinant of participation if the unobservables are uncorrelated. This is expected since individuals have more information about potential gains from education in the presence of correlation between the residuals and use this information in their education decision. Rows 6 to 9 show that both ability and family background are strong determinants of participation and more so in the absence of correlation between the residuals. Region is much less important, as displayed in rows 10 and 11.

IV.F.2. Sources of variability to apply DID

This DID application relies on the occurrence of a policy change and availability of individual information collected before and after the policy change. We denote these cohorts by the time of their education decision, namely t=0 or t=1 for whether the education decision is taken before or after the policy change, respectively. We then explore two sources of variability. The first assumes the subsidy is first piloted in one region, say x=1, while the old policy remains active in the rest of the country (region x=0). The second uses the policy eligibility rule, namely the cutoff point in terms of test score($\underline{s}$), to define treatment and comparison groups. The considered policy change is the introduction of a subsidy for advanced education.

We first discuss the use of pilot studies, with the subsidy being available at time t=1 in region x=1 but not in region x=0 or at time t=0. The question now is: Can we explore this regional change in policy to learn about the returns to education using DID?

We start by noticing that enrollment into education is not solely determined by the subsidy. Some individuals decide to enroll into education even if no subsidy is available or if not eligible, while some eligible individuals will opt out despite the presence of the subsidy. Thus, there will be some educated individuals even when and where the subsidy is not available. To put it shortly, there is non-compliance. As a result, the ATT will not be identified in general. Instead, what may be identified is the average impact for individuals who change their educational decisions in response to the subsidy.

To estimate the returns to education among individuals that change education status in response to the subsidy, we further assume a monotonicity condition - that the chance of assessing subsidized education does not lead anyone to give up education. Instead, it makes education more attractive for all eligibles and does not change the incentives to invest in education among non-eligibles.[10]

Define the treatment and control groups as those living in regions affected (x=1) and not affected (x=0) by the policy change and suppose we dispose of information on education attainment and earnings of both groups before and after the policy change. We can then compare the two regions over time using DID.

Designate by $\overline{\ln y}_{xt}$ the average log earnings in region x at time t. As before, d_{it} is a dummy variable indicating whether individual i in cohort t has acquired high education, and we define the probabilities

$$p_{xt} = P\left(d_{it} = 1 \middle| x, t\right)$$

where again i indexes individuals, x represents the region (x=0,1) and t represents time (t=0,1). Thus, p_{xt} is the odds of participation in region x at time t. The monotonicity assumption, stating that education is at least as attractive in the presence of the subsidy, implies that $d_{i1} \geq d_{i0}$ for all i in region x=1 and, therefore, $p_{11} \geq p_{10}$. In the control region we assume $d_{01} = d_{00}$ for simplicity, ruling out macro trends.[11]

Assuming the decomposition of the error term as in equation (13),

$$u_{it} = n_i + m_t + o_{it}$$

yields, under the DID assumptions,

$$\mathrm{E}\left[\overline{\ln y}_{11} - \overline{\ln y}_{10}\right] = \left(m_1 - m_0\right) + \left(p_{11} - p_{10}\right)\mathrm{E}\left[\alpha_i \middle| d_{i1} = 1, d_{i0} = 0, x_i = 1\right].$$

The above expression suggests that only the impact on the movers may be identified. Similarly,

$$\mathrm{E}\left[\overline{\ln y}_{01} - \overline{\ln y}_{00}\right] = \left(m_1 - m_0\right)$$

since individuals in the control region do not alter their educational decisions. Thus, under the DID assumption we identify,

$$\text{(23)}\quad \mathrm{E}\left[\hat{\alpha}^{DID}\right] = \left(p_{11} - p_{10}\right)\mathrm{E}\left[\alpha_i \middle| d_{i1} = 1, d_{i0} = 0, x_i = 1\right].$$

Equation (23) shows that the mean return to education on individuals moving into education in response to the subsidy can be identified by dividing the DID estimator by the proportion of movers in the treated region, $p_{11} - p_{10}$. This is the LATE parameter.

Not correcting for the proportion of movers implies that a different parameter is estimated: the *average impact of introducing an education subsidy on the earnings of the treated*, in this case the individuals living in region one. This is a mixture of a zero effect for individuals in the treated region that do not move and the return to education for the movers.[12]

Under homogeneous treatment effects, all average parameters are equal and thus ATE and ATT are also identified. However, under heterogeneous treatment effects only the impact on the movers can be identified and even this requires especial conditions. In this example we have ruled out movers in the control regions. If other conditions differentially affect the educational decisions in non-treated regions before and after the policy intervention, movements are expected among the controls as well. Whether the monotonicity assumption mentioned above holds for the control group or not depends on the circumstances that lead these individuals to move. For simplicity, assume monotonicity holds in control areas such that $d_{i1} \geq d_{i0}$ for *i* in the control region. DID will then identify

$$\mathrm{E}\left[\hat{\alpha}^{DID}\right] = \left(p_{11} - p_{10}\right)\mathrm{E}\left[\alpha_i \middle| d_{i1} = 1, d_{i0} = 0, x_i = 1\right] + \left(p_{01} - p_{00}\right)\mathrm{E}\left[\alpha_i \middle| d_{i1} = 1, d_{i0} = 0, x_i = 0\right].$$

Now the ability to single out the impact of treatment on a subset of the movers (movers in region *x*=1 net of movers in region *x*=0) depends on two additional factors: *(i)* that movers in region *x*=1 in the absence of a policy change would have the same returns to education as movers in region *x*=0, which typically requires that they are similar individuals; and *(ii)* that different proportions of individuals move in the two areas.

Now suppose that instead of a pilot study, we are exploring the use of a global, country-wide policy change. Instead of using treated and non-treated regions, one can think of using the eligibility rules as the source of randomization. The treatment and control groups

are now composed of individuals scoring above and below the eligibility threshold $\underline{s}$, respectively. Let $\tilde{s}$ denote eligibility: $\tilde{s}$ is one if $s \geq \underline{s}$ and is zero otherwise. Again, we assume data is available on two cohorts, namely those affected and unaffected by the policy change.

The use of the eligibility rule instead of regional variation suffers, in this case, from one additional problem: the identification of the eligibility group before the introduction of the program. The affected generations will react to the new rules, adjusting their behavior even before their treatment status is revealed (which amounts to becoming eligible to the subsidy). In our model, future eligibility can be influenced in anticipation by adjusting effort at age zero. As a consequence, a change in the selection mechanism in response to the policy reform will affect the size and composition of the eligibility groups over time. This means that eligibles and non-eligibles are not comparable over time and since we are confined to use repeated cross-sections to evaluate the impact of education, it would exclude DID as a valid candidate to the present evaluation exercise when the only source of variation to be explored is eligibility.

This problem has been identified by Abbring and van den Berg (2003) for evaluation studies in the presence of dynamic decision processed. Individuals may react in anticipation of treatment, in an attempt to explore the policy rules. If the rules change, anticipatory behavior may also change, thus rendering individuals with similar characteristics incomparable when such characteristics are affected by the endogenous selection behavior that is not explicitly modeled. Reactions in anticipation to treatment are generally not observable and tend to change over time. Their occurrence may create a problem similar to the Ashenfelter dip described above as their potential impact on the outcome will be absorbed by the transitory unobservable component. Treated and controls with similar pre-treatment

characteristics and outcomes will be inherently different as observables are endogenously affected by the individuals prospects about treatment.

In our example, individuals may react to the new subsidy by increasing effort in the test, raising test performance on average and increasing the odds of becoming eligible to subsidized education. Thus, the ability distribution of eligibles will be affected by the policy change, not only the educational choice.

IV.F.3. Monte-Carlo results: DID

To illustrate the ability of DID to estimate the impact of treatment, we ran a Monte Carlo simulation. This simulation exercise is a completely controlled experiment, for which the exact treatment effects are known and the selection mechanisms fully understood. Such information provides a more insightful view over the evaluation problems when the aim is to learn about returns to education and allows for a direct comparison of different methods in what concerns to the nature of their assumptions, the nature of the identified parameters and their robustness to certain violations of the underlying assumptions.

In this DID application we tried different assumptions, depending on: *(i)* Whether or not the policy is experimented in some parts of the country before being nationally implemented; *(ii)* Whether or not the post-intervention generation is informed about the new rules at the moment of taking the test and defining eligibility and *(iii)* Whether or not the unobservables v and u are correlated. In each case, we estimate the impact of education using DID correcting and not correcting for non-compliance among the treated. Thus, the parameters that we aim to identify are the *aggregate effect of the subsidy on the treatment group*, where treatment is either being in the pilot region or being eligible to subsidized education, and the *effect of education on the movers* where movers are individuals that become educated only if subsidized.

Table 2 displays the true parameters, DID estimates and respective bias for the case where the unobservables *u* and *v* are negatively related.[13] In producing these estimates we explore two sources of differential eligibility: region and test score. In the case of *region*, we assume the policy is first implemented in region one and compare earnings and education take up in region one (treated) with those in region zero (controls). In the case of *test score*, we explore the eligibility rule in terms of test score by comparing individuals scoring above the threshold (treated) with those scoring below the threshold (controls) over both regions.

The first three columns in table 2 show true parameters and DID estimates under the assumption that individuals are totally aware of the policy environment when making their investment decisions. Results in rows 1 and 2 show that a pilot study design identifies the true effect both in aggregate terms, as the impact of the policy change on the experimental population (row 1), and in terms of impact on the movers or LATE, as the returns to education on agents that change educational attainment in response to the introduction of the subsidy. This result is expected since the population in the treated and control regions does not change over time in response to the policy.

Rows 3 and 4 show that using eligibility with an expected policy fails to identify the true parameter. Again this result could be anticipated since optimizing individuals will adjust their behavior in advance to becoming eligible in an attempt to affect their eligibility status. As a consequence, the composition of treated and control groups will change together with the introduction of the policy.

The next three columns, 4 to 6, display similar results when the individuals are unaware of the policy at the moment of making decisions that may affect their eligibility status. Thus, they will not have the required information to act in advance and affect eligibility and the composition of treated and control groups will remain unchanged over the transition period (from before to after the policy change), even when eligibility is used as the

source of variation. In this case, the eligibility criterion will correctly identify the true parameters (rows 3 and 4).

An unexpected policy change will lead very few individuals to change their education investment and, given the present design, all the movers will be concentrated among the eligibles. Thus, the proportion of movers among the treated will be much smaller if region instead of eligibility is used as the source of variation to construct treated and control groups. The consequence is that considerable variation is introduced when correcting the DID estimate since it amounts to dividing by a very small number (see equation (23)). The numbers in row 2, columns 4 to 6, show how such feature is translated in large bias for the DID estimator of the LATE parameter in this case.

For comparison purposes we notice that the corresponding true ATT presented in table 1 is 0.453. This is closer to the LATE parameters when the policy environment changes unexpectedly. If the agents have prior knowledge of the new policy, the most able but with least resources (or higher costs) will try harder to obtain a subsidy as this is will bring them very high returns. If, on the other hand the subsidy is not expected, individuals will not act to attain eligibility and the movers will be less concentrated among very-high-cost/very-high-ability individuals.

V. Matching Methods

V.A. The matching estimator (M)

The underlying motivation for the matching method is to reproduce the treatment group among the non-treated, this way re-establishing the experimental conditions in a non-experimental setting. Under assumptions we will discuss below, the matching method constructs *the* correct sample counterpart for the missing information on the treated outcomes had they not been treated by pairing each participant with members of the non-treated group.

The matching assumptions ensure that the only remaining relevant difference between the two groups is program participation.

Matching can be used with cross-sectional or longitudinal data. In its standard formulation, however, the longitudinal dimension is not explored except perhaps on the construction of the matching variables. We therefore abstract from time in this discussion but will consider the appropriate choice of the matching variables in what follows.

As a starting point we incorporate observable regressors *X* in the outcome equation in a reasonably general way. The covariates *X* explain part of the residual term *u* in (1) and part of the idiosyncratic gains from treatment:

$$(24)\quad \begin{aligned} y_i^1 &= \beta + u(X_i) + \alpha(X_i) + \left[\left(u_i - u(X_i)\right) + \left(\alpha_i - \alpha(X_i)\right)\right] \\ y_i^0 &= \beta + u(X_i) + \left(u_i - u(X_i)\right) \end{aligned}$$

where $u(X)$ is the predictable part of y^0, $\left(u - u(X)\right)$ is what is left over of the error *u* after conditioning on *X*, $\alpha(X)$ is some average treatment effect over individuals with observable characteristics *X* and α_i is the individual *i* specific effect, which differs from $\alpha(X_i)$ by the unobservable heterogeneity term.

To identify ATT, matching assumes that the set of observables, *X*, contains all the information about the potential outcome in the absence of treatment, y^0, that was available to the individual at the point of deciding whether to become treated, *d*. This means that the econometrician has all the relevant information, namely the information that simultaneously characterizes the participation rule and the non-treated outcome. This is called the Conditional Independence Assumption (CIA) and can be formally stated as follows

$$(25)\quad y_i^0 \perp d_i \,|\, X_i \,.$$

Since all the information that simultaneously characterize y^0 and *d* is in *X*, conditioning on *X* makes the non-treated outcomes independent from the participation status. Thus, treated and

non-treated sharing the same observable characteristics, *X*, draw the non-treated outcome, y^0, from the same distribution.

Within model (24), the CIA can be restated in terms of the unobservable in the non-treated outcome equation,

$$\left(u_i - u\left(X_i\right)\right) \perp d_i \left| X_i \right.$$

or, which is the same,

$$u_i \perp d_i \left| X_i \right.$$

meaning that *u* is independent of participation into treatment or, in other words, that there is no selection on the unobservable part of *u* in (24).

The CIA in (25) obviously implies a conditional version of the randomization hypothesis that rules out selection on the untreated outcomes (R1),

(26) $\mathrm{E}\left[u_i \left| d_i, X_i \right.\right] = \mathrm{E}\left[u_i \left| X_i \right.\right]$.

This weaker version of the CIA is sufficient to estimate the ATT on individuals with observable characteristics *X* using matching. Again, nothing like the randomization hypothesis (R2) is required to identify the ATT, which means that selection on the unobservable gains can be accommodated by matching.

The implication (25) or (26) is that treated and non-treated individuals are comparable in respect to the non-treated outcome, y^0 conditional on *X*. Thus, for each treated observation (y^1) we can look for a non-treated (set of) observation(s) (y^0) with the same *X*-realization and be certain that such $ y^0 $ is a good predictor of the unobserved counterfactual.

Thus, matching is explicitly a process of re-building an experimental data set. Its ability to do so, however, depends on the availability of the counterfactual. That is, we need to ensure that each treated observation can be reproduced among the non-treated. This is only possible if the observables *X* do not predict participation exactly, leaving some room for

unobserved factors to influence the treatment status. This is the second matching assumption, required to ensure that the region of *X* represented among participants is also represented among non-participants. Formally, it can be stated as follows

(27) $P\left[d_i = 1 \middle| X_i\right] < 1$.

Given assumptions (26) and (27), we can now define the matching estimator. Let *S* represent the subspace of the distribution of *X* that is both represented among the treated and the control groups. *S* is known as the common support of *X*. Under(27), *S* is the whole domain of *X* represented among the treated. The AT over the common support *S* is

$$\alpha^{ATT}(S) = \mathrm{E}\left[y^1 - y^0 \middle| d = 1, X \in S\right] = \frac{\int_S \mathrm{E}\left[y^1 - y^0 \middle| d = 1, X\right] d\,\mathrm{F}_{X|d}\left(X \middle| d = 1\right)}{\int_S d\,\mathrm{F}_{X|d}\left(X \middle| d = 1\right)}$$

where $\mathrm{F}_{X|d}$ is the cumulative distribution function of *X* conditional on *d* and $\alpha^{ATT}(S)$ is the mean of impact on participants with observable characteristics *X* in *S*.

The matching estimator is the empirical counterpart of $\alpha^{ATT}(S)$. It is obtained by averaging over *S* the difference in outcomes among treated and non-treated with equal *X*-characteristics using the empirical weights of the distribution of *X* among the treated. Formally, the matching estimator of the ATT is

(28) $\hat{\alpha}^M = \sum_{i \in T} \left\{ y_i - \sum_{j \in C} \tilde{w}_{ij} y_j \right\} w_i$

where *T* and *C* represent the treatment and comparison groups respectively, $\tilde{w}_{ij}$ is the weight placed on comparison observation *j* for the treated individual *i* and w_i accounts for the re-weighting that reconstructs the outcome distribution for the treated sample.

Identification of ATE requires a strengthened version of assumption (26) because the correct counterfactual needs to be constructed for both the treated and the non-treated.

This means that both u_i and α_i need to be (mean) independent of *d* conditional on *X*. That is, selection on unobserved expected gains must also be excluded for matching to identify the correct ATE (recovering the second randomization assumption, (R2)). In its weaker version, the CIA is now formally:

$$(29)\quad \begin{aligned} \mathrm{E}\left[u_i \left| d_i, X_i \right.\right] &= \mathrm{E}\left[u_i \left| X_i \right.\right] \\ \mathrm{E}\left[\alpha_i \left| d_i, X_i \right.\right] &= \mathrm{E}\left[\alpha_i \left| X_i \right.\right]. \end{aligned}$$

Estimation of ATE also requires a modification of the overlapping support assumption (27) to ensure that both the treated and the non-treated are represented within the alternative group. Formally,

$$(30)\quad 0 < P\left[d_i = 1 \left| X_i \right.\right] < 1.$$

Under (29) and (30), the ATE over the common support S is

$$\begin{aligned} \alpha^{ATE}(S) &= \mathrm{E}\left[y^1 - y^0 \left| X \in S \right.\right] \\ &= \frac{\int_S \mathrm{E}\left[y^1 - y^0 \left| X \right.\right] d\,\mathrm{F}_X}{\int_S d\,\mathrm{F}_X(X)} \end{aligned}$$

where now the conditional mean effects are weighted using the distribution of the *X*'s over the whole population, $\mathrm{F}_X(X)$.

The choice of the appropriate matching variables, *X*, is a delicate issue. Too much information and the overlapping support assumption will not hold. Too little and the CIA will not hold. The wrong sort of information and neither of the two assumptions will hold. So what is the right balance?

The appropriate matching variables are those describing the information available at the moment of assignment and simultaneously explaining the outcome of interest. Only this set of variables ensures the CIA holds. However, the same is not necessarily true for the overlapping support assumption. It will not hold when participation is determined with certainty within some regions of the support of *X*. In this case matching will identify a

different parameter, namely the average impact over the region of common support. Typically, but not necessarily, individuals gaining the most and the least from treatment will be excluded from the analysis.

However, it is rarely clear what sort of information is in the information set at assignment. What is clear is that matching variables should be determined before the time of assignment and not after as this could compromise the CIA by having matching variables affected by the treatment status itself. A structural model can shed some light on what the correct set of matching variables should be. However, such models are likely to include some unobservable variables and are more naturally used to motivate Instrumental Variables and Control Function methods described further below. Nevertheless, they can suggest possible variables that capture the key determinants of selection. For example, in studies about the impact of training on labor market outcomes, previous labor market history could contain all the relevant information on the unobservable ability and job-readiness as it is partly determined by such factors.

V.B. Propensity score matching

A serious limitation to the implementation of matching is the dimensionality of the space of the matching variables, *X*. Even if all variables are discrete with a finite domain, the dimensionality of the combined space increases exponentially with the number of variables in *X*, making it virtually impossible to find a match for each observation within a finite (even if large) sample when more than a few variables are being controlled for.

A popular alternative is to match on a function of *X*. Usually, this is carried out on the probability of participation given the set of characteristics *X*. Let *P(X)* be such probability, known as the "propensity score". It is defined as

$$P(X) = P(d = 1 | X).$$

The use of *P(X)* has been motivated by Rosenbaum and Rubin's result on the balancing property of the propensity score (1983, 1984). The authors have shown that if the CIA is valid for *X* it is also valid for *P(X)*:

$$y_i^0 \perp d_i \,|\, X_i \qquad \Rightarrow \qquad y_i^0 \perp d_i \,|\, P(X_i).$$

The balancing property of the propensity score implies that, if *P(X)* is known, it can be used to replace *X* in the matching procedure.[14] But then, knowledge of *P(X)* reduces the matching problem to a single dimension, thus simplifying the matching procedure significantly.

However, *P(X)* is not known in concrete applications and needs to be estimated. Whether the overall estimation process is indeed simplified and the computing time reduced depends on what is assumed about *P(X)*. The popular procedure amounts to employing a parametric specification for *P(X)*, usually in the form of a logit, probit or linear probability model. This solves the dimensionality problem but relies on parametric assumptions. Alternatively, a non-parametric propensity score keeps the full flexibility of the matching approach but does not solve the dimensionality problem.

When using propensity score matching, the comparison group for each treated individual is chosen with a pre-defined criteria (established in terms of a pre-defined metric) of proximity between the propensity scores for treated and controls. Having defined a neighborhood for each treated observation, the next step is the choice of appropriate weights to associate the selected set of non-treated observations to each treated observation. Several possibilities are commonly used. We briefly refer the most commonly applied alternatives and refer the interested reader to Leuven and Sianesi (2003) and Becker and Ichino (2002) for more detailed practical guidelines on alternative matching procedures.

The *Nearest Neighbor Matching* assigns a weight one to the closest non-treated observation and zero to all others. A widespread alternative is to use a certain number of the

closest non-treated observations to match the treated, frequently the ten closest observations. This reduces the variability of the nearest neighbor estimator and is more reliable specially when the sample of treated individuals is small as each match may significantly affect the results.

Kernel Matching defines a neighborhood for each treated observation and constructs the counterfactual using all control observations within the neighborhood, not only the closest observation. It assigns a positive weight to all observations within the neighborhood and a zero weight to the remaining observations. Different weighting schemes define different estimators. For example, uniform kernel attributes the same weight to each observation in the neighborhood while other forms of kernel make the weights dependent on the distance between the treated and the control being matched, where the weighting function is decreasing in distance. By using more observations per treated, kernel weights reduce the variability of the estimator when compared with nearest neighbor weights and produces less bias then nearest neighbor with many matches per treated. However it still introduces significant bias at the edges of the distribution of $P(X)$. When this is a problem, *Local Linear Matching* will effectively deal with this sort of bias.[15]

Not only do kernel and local linear matching produce more precise estimates than nearest neighbor matching, it is simpler to compute the precision for these estimators. The complexity of propensity score matching requires bootstrapping to be used in computing the standard errors for the effect of treatment. The problem with the nearest neighbor technique is that bootstrapping is not guaranteed to deliver consistent estimates since choosing only one (or a fixed number of) match(es) per treated individual means that the quality of the match does not necessarily improve as the sample (of controls) gets bigger. The same is not true for kernel and local linear matching as with these estimators the sample of matched controls

expands with the sample size (for a thoroughly discussion of bootstrapping see Horowitz 2001).

The general form of the matching estimator is not altered by the sort of weights one decides to apply. As before, it is given by $\hat{\alpha}^M$ in (28).

While propensity score matching is affected by the same problems as fully non-parametric matching in choosing the right set of controlling variables, it also faces the additional problem of finding a sufficiently flexible specification for the propensity score to ensure that the distribution of observables is indeed the same among treated and matched controls. The balancing property of Rosenbaum and Rubin (1983, 1984) ensures that the true propensity score does balance the observables but no similar result exists for the estimated propensity score. In general, one wants to ensure that if the untreated outcome y^0 is mean independent of the participation status conditional on *X* it is also mean independent of the participation status conditional of $\hat{P}(X)$ where $\hat{P}(X)$ is the predicted propensity score based on estimation results. The evaluation literature has proposed a few balancing tests to assess whether the specification for the propensity score is statistically sound. For example, Rosenbaum and Rubin (1985) propose a test based on the comparison of means for each covariate between treated and matched controls. If the difference in means is too large, the test rejects the hypothesis that the samples (of treated and matched controls) are balanced with respect to the covariates when they are balanced with respect to the (predicted) propensity score.

V.B.1. The linear regression model and the matching estimator

In common with the matching estimator, the linear regression model also relies on selection on the observables. It amounts to impose a fully parametric structure to model (24) by assuming that *u* and α are linear functions of *X*:

$$u(X_i) = X_i\pi$$
$$\alpha(X_i) = \xi_0 + X_i\xi_1$$

where (π, ξ_0, ξ_1) are the unknown coefficients. The model can then be written as

(31) $y_i^0 = \beta^0 + X_i\pi^0 + e_i^0$

(32) $y_i^1 = \beta^1 + X_i\pi^1 + e_i^1$

where

$$\beta^d = \beta + d\xi_0$$
$$\pi^d = \pi + d\xi_1$$
$$e_i^d = (u_i - X_i\pi) + d(\alpha_i - \xi_0 - X_i\xi_1)$$

and *d* is the treatment indicator.

Estimation of the ATT requires knowledge of the model for the untreated outcomes only, (31). The strategy to be discussed here consists of comparing observed outcomes among the treated with the predicted counterfactual based on the estimates of equation (31). So in fact this is not a straightforward OLS procedure as it involves a second step to form the predicted non-treated outcomes among the treated ($\hat{\mathrm{E}}\left[y^0 \middle| X, d=1\right]$) and compare them to the observed treated outcomes, y^1.

Instead, many empirical applications use the overall model for observed outcomes *y* to identify some average treatment effect directly from the OLS regression:

$$y_i = \beta + X_i\pi + \alpha(X_i)d_i + e_i \quad \text{where}$$

(33) $$e_i = (u_i - X_i\pi) + d_i(\alpha_i - \alpha(X_i))$$
$$= (u_i - X_i\pi) + d_i(\alpha_i - \xi_0 - X_i\xi_1).$$

However, estimation of (33) is more demanding in ways to be discussed below.

The CIA together with the assumption of exogeneity of the covariates *X*, that is $\mathrm{E}\left[e^0 \middle| X\right] = \mathrm{E}\left[e^0\right]$, ensures that the ATT can be obtained from the OLS predictions of y^0 over the range of *X*'s observed among the treated. The common support assumption (27) is

not required as the parametric specification can be used to extrapolate y^0 outside the observable range of *X* among the controls when predicting the counterfactual for each treated observation.

The imposition of a parametric specification is not as restrictive as it might first seem. In fact, by including many interactions between the variables and higher order polynomials in the (continuous) regressors, one will closely approximate any smooth function y^0 over the domain of observable *X*'s (see the Blundell, Dearden and Sianesi 2005 application, for example). The main requirement is then to use a flexible enough functional form for y^0.

Much more restrictive is the relaxation of the common support assumption. In its absence, the model needs to be extrapolated over unobservable regions of the distribution of *X*, where only the true model in the absence of endogenous regressors can be guaranteed to perform well. Of course, one could always think of imposing the common support assumption within the parametric linear model and estimate the average effect of treatment within regions of *X* simultaneously observed among treated and controls. However, while this is feasible it is rarely done in the context of parametric models given the simplicity of extrapolating outside the observable interval. Most frequently, researchers seem unaware that a common support problem exists.

Another drawback of the parametric linear model is the requirement of exogeneity of *X* in the equation for y^0. In a recent paper, Frolich (2006) noticed that an important and often ignored advantage of non-parametric (as compared to parametric) regression methods is that endogeneity of regressors that are not of main interest (in our case *X*) may not affect the estimated relationship between the regressor of interest (in our case, *d*) and the outcome (*y*). While this is true, a couple of considerations are due in the context of our evaluation problem. First, and as noticed before, non-parametric regression methods, and in particular matching,

can cope with endogeneity of *X* only as long as these covariates are not determined by the regressor of interest, *d*.

The second point is slightly more subtle. Start by noticing that the objective of OLS is not to directly estimate the ATT, which would amount to estimating the regression model (33). In such case the endogenous regressors in *X* not properly instrumented for would contaminate the estimates of all other parameters, including those of $\alpha(X)$. Instead, the OLS approach we are discussing aims at recovering the unobservable counterfactual, $\mathrm{E}\left[y^0 \middle| X, d=1\right]$, using OLS estimates of the parametric model (31).

Consider the simple case where (31) is correctly specified although (some of) the regressors *X* are endogenous and the mean error term is a linear function of *X*: $\mathrm{E}\left[e^0 \middle| X\right] = X\eta^0$ for some set of unknown parameters η^0. Then clearly OLS provides inconsistent estimates of π^0 but (asymptotically) correctly predicts $\mathrm{E}\left[y^0 \middle| X, d=0\right]$ to be $\beta^0 + X\left(\pi^0 + \eta^0\right)$. Under the CIA, the correct counterfactual for a treated individual with observables *X* is recovered by $\hat{\mathrm{E}}\left[y^0 \middle| X, d=1\right] = \widehat{\beta^0}^{OLS} + X\widehat{\left(\pi^0 + \eta^0\right)}^{OLS}$.

The common support assumption is not needed in this case as long as linearity holds over the whole support of *X*.

The more credible case where $\mathrm{E}\left[e^0 \middle| X\right]$ is some unknown, possibly non-linear, function of *X* is more demanding. On the one hand, it strengthens the case for a sufficiently flexible specification y^0 as a function of *X* to compensate for the global nature of the OLS estimator. On the other hand, it casts further doubts on extrapolations outside the observable domain of *X* among the controls, thus calling for a common support restriction to be explicitly imposed.

V.C. Weaknesses of matching

The main weaknesses of matching relate to dataavailability and our ability to select the right information. The common support assumption (27) ensures that the missing counterfactual can be constructed from the population of non-treated. What (27) does not ensure is that the same counterfactual exists in the sample. If some of the treated observations cannot be matched, the definition of the estimated parameter becomes unclear. It is the average impact over some subgroup of the treated, but such subgroup may be difficult to define. The relevance of such parameter depends, of course, on the ability to define the population it corresponds to.

Taken together, assumptions (25) (or (26)) and (27) show how demanding matching is with data: the right regressors X must be observed to ensure that what is left unexplained from y^0 is unrelated with the participation decision; any more than the right regressors will only contribute to make finding the correct counterfactual harder or even impossible. In particular, variables in the decision rule (in Z) but not in X should be excluded from the matching procedure as they only interfere with our ability to (27). To achieve the appropriate balance between the quantity of information at use and the share of the support covered can be very difficult. In a recent paper, Heckman and Lozano (2004) show how important and, at the same time, how difficult it is to choose the appropriate set of variables for matching. Bias results if the conditioning set of variables is not the right and complete one. In particular, if the relevant information is not all controlled for, adding additional relevant information but not all that is required may increase, rather then reduce, bias. Thus, aiming at the best set of variables within the available set may not be a good policy to improve the matching results.

If, however, the right amount of information is used, matching deals well with potential bias. This is made clear by the following decomposition of the treatment effect

$$\mathrm{E}\left[y^1 - y^0 \middle| d=1, X \right] = \left\{ \mathrm{E}\left[y^1 \middle| d=1, X \right] - \mathrm{E}\left[y^0 \middle| d=0, X \right] \right\} - \left\{ \mathrm{E}\left[y^0 \middle| d=1, X \right] - \mathrm{E}\left[y^0 \middle| d=0, X \right] \right\}$$

where the second term on the right hand side is the bias conditional on *X*. Conditional on *X*, the only reason the true parameter, $\alpha^{ATT}(X)$, might not be identified is selection on the unobservable term *u*. However, integration over the common support *S* creates two additional sources of bias: non-overlapping support of *X* and mis-weighting over the common support. Through the process of choosing and re-weighting observations, matching corrects for the latter two sources of bias and selection on the unobservables is assumed to be zero by the CIA.

V.D Using matching to estimate returns to education

In this section we return to education evaluation example. The earnings specification in equation (11) is reproduced here:

$$\ln y_i = \beta_0 + \beta_1 x_i + \alpha_0 d_i + \alpha_1 \theta_i d_i + u_i$$

where *x* is region and can be zero or one. The impact of education on earnings is region-specific given the non-linear form of the earnings equation. In what follows we exclude sorting by region meaning that the distribution of ability does not change by region. So the ATE on log earnings will not depend on region, but the same does not hold with respect to the the ATT due to the selection process.

V.D.1 Monte-Carlo results

We ran some monte-carlo experiments under different assumptions about the relationship between *d* and *u* depending on whether the residuals *u* and *v* are negatively correlated or independent. In the former case there is selection on the unobservables while the

latter constraints selection to occur on the observables. In both alternatives we estimated both the ATT and the ATNT using different sets of conditioning variables. Table 3 details the results for an economy with an education subsidy.

The first two columns in table 3 display matching estimates of the ATT (panel A) and ATNT (panel B) when the unobservables *u* and *v* are negatively correlated. This corresponds to the case where (part of) the selection process occurs on variable unobservable by the researcher and we have seen that matching is incapable of identifying the true parameter under such circumstances as the CIA fails to hold. Columns 1 and 2 just confirm this result in this specific example, independently of the set of matching variables being used.

The numbers in columns 1 and 2 on the table are based on the correlation between *u* and *v* being -0.5. Figure 2 displays the bias in the estimation of the ATT and ATNT for different levels of correlation when the set of matching variables is *x* and (x,θ) for the ATT and ATNT, respectively. The bias is quite considerable even for relatively low levels of correlation, particularly for the ATNT but also for the ATT. When selection on the unobservables is suspected, other methods such as IV and control function are more adequate than matching. These will be discussed in the next sections. Before that, however, we also experimented the use of matching when all selection occurs on the observables. Results are displayed in columns 3 and 4 of table 3.

The correct ATT matching estimator conditions on region (*x*) alone as this is the only variable simultaneously affecting educational investment and earnings in the non-educated state. The numbers in row 2, columns 3 and 4 show that matching identifies the ATT in this case. Including additional, unnecessary matching variables increases bias by reducing the overlapping support in finite samples and increasing variability of the estimator (row 3). In this example, however, the amount of bias introduced by the additional matching variables is small. Excluding the correct matching variable from the set of covariates increases bias

more significantly (row 4). Individuals with the same realization of the covariates other than region may decide differently about education because they expect different gains from the investment due to residing in different regions: individuals in the high-returns region are more likely to participate, but they would also enjoy from higher earnings if remained uneducated than their counterparts in the low-returns region. Thus, when comparing educated and uneducated individuals conditional on covariates excluding *x* we are over-sampling treated from the high-returns region and non-treated from the low-returns/low-earnings region, leading to biased estimates.

Rows 5 to 8 show how much more difficult it can be estimating ATNT than ATT. The CIA justifying matching for the estimation of ATNT requires the *treated outcome* y^1 to be independent of the treatment status *d* conditional on the matching variables. It amounts to ruling out selection on the non-treated outcomes, y^0, *and* gains from treatment, α. y^0 depends on region, *x*, and the gains from treatment depend on ability, θ, and region, *x*. Thus, the correct conditioning set is now (x,θ) and the results on rows 6 to 8 confirm this. Region alone does not solve the selection problem (row 6) but matching is unbiased if ability is added to the conditioning set (row 7). However, ability is rarely available in empirical studies. For this example, row 8 shows that matching on alternative observables reduces the selection problem. This is because the test score is related with ability and is part of the conditioning set. Nevertheless, the bias is always sizeable if ability is not part of the matching variables.

V.E. Combining matching and DID (MDID)

In the presence of longitudinal or repeated cross-section data, matching and DID can be combined to weaken the underlying assumptions of both methods. The CIA is quite strong if individuals are expected to decide according to their forecasted outcome as data is rarely

rich enough to describe the relevant available information. However, the combination of matching with DID as proposed in Heckman, Ichimura and Todd (1997) can accommodate unobserved determinants of the non-treated outcome affecting participation for as long as these are constant over time.

To discuss MDID, we start by decomposing the unobservable term u_i in (24) into a fixed effect (*n*), macro shock (*m*) and an idiosyncratic transitory shock (*o*). MDID can be applied when treated and non-treated are observed over time with at least one observation before and one after the treatment. For simplicity we consider two time periods, (t_0, t_1), where $t_0 < k < t_1$ and *k* is the time of treatment. MDID compares the evolution of treated outcomes with that of non-treated over the observation period (t_0, t_1) and assigns any difference to the impact of treatment. To do so, MDID makes a common trends assumption - had the treated remained non-treated and they would have experienced a change in outcomes equal to that observed among the actual non-treated.

More formally, the model can now be written as

$$(34)\quad \begin{aligned} y_{it}^1 &= \beta + u(X_i) + \alpha(X_i) + \left[\left(n_i + m_t + o_{it} - u(X_i)\right) + \left(\alpha_i - \alpha(X_i)\right)\right] \\ y_{it}^0 &= \beta + u(X_i) + \left(n_i + m_t + o_{it} - u(X_i)\right) \end{aligned}$$

where y_{it}^d is the outcome for individual *i* at time *t* when the his/her treatment status at that time is *d* - it is y^0 when the individual belongs to the non-treated group or when the time is t_0 and is y^1 when the individual is in the treated group and the time is t_1. The MDID assumption states that, conditional on the observables *X*, the evolution of the unobserved part of y^0 is independent of the treatment status. Thus,

$$(35)\quad \left(u_{it_1} - u_{it_0}\right) \perp d_{it_1} \,|\, X_i \,.$$

The main matching hypothesis is now stated in terms of the before-after evolution instead of levels. It means that controls evolve from a pre- to a post-program period in the same way treatments would have evolved had they not been treated. We continue to consider time invariant covariates, *X*, even though MDID explicitly explores the time-series dimension of the data. The discussion on the choice of covariates at the end of V.A, where we argued that the appropriate covariates should reflect the information available to the individual at the time of making a participation decision, explains this choice.

Assumption (35) is not enough to ensure identification of ATT. Just as in the matching case, we also need to impose a common support hypothesis. This will be the same as (27) when longitudinal data is available. If only cross-section data is available we will need to strengthen it to ensure that the treated group can be reproduced in all three control groups characterized by treatment status before and after the program. This version of the common support assumption states that all treated individuals have a counterpart on the non-treated population before and after the treatment

(36) $P\left[d_{it_1} = 1 \middle| X_i, t \right] < 1.$

where $P\left[d_{it_1} = 1 \middle| X_i, t \right]$ is the probability that an individual observed at time *t* with characteristics X_i would belong to the treatment group at time t_1.

The effect of the treatment on the treated can now be estimated over the common support of *X*, call it *S*. The following estimator is adequate to the use of propensity score matching with longitudinal data

$$\hat{\alpha}^{MDID,L} = \sum_{i \in T} \left\{ \left[y_{it_1} - y_{it_0} \right] - \sum_{j \in C} \tilde{w}_{ij} \left[y_{jt_1} - y_{jt_0} \right] \right\} w_i$$

where the notation is similar to what has been used before. With repeated cross-section data, however, matching must be performed over the three control groups: treated and non-treated at t_0 and non-treated at t_1.[16] In this case, the matching-DID estimator would be

$$\hat{\alpha}^{MDID,RCS} = \sum_{i \in T_1} \left\{ \left[y_{it_1} - \sum_{j \in T_0} \tilde{w}^{T}_{ijt_0} y_{it_0} \right] - \left[\sum_{j \in C_1} \tilde{w}^{C}_{ijt_1} y_{it_1} - \sum_{j \in C_0} \tilde{w}^{C}_{ijt_0} y_{it_0} \right] \right\} w_i$$

where (T_0, T_1, C_0, C_1) stand for the treatment and comparison groups before and after the program and $\tilde{w}^{G}_{ijt}$ represents the weight attributed to individual *j* in group *G* and time *t* when comparing with treated individual *i*.

The implementation of the MDID estimator using propensity score matching requires the propensity score to be estimated using the treated and the controls. In the presence of longitudinal data, the dependent variable *d* is set equal to one if the individual is treated and to zero otherwise. The controls are then matched to the treated and the re-weighted sample is use to compute the ATT using DID. In the presence of repeated cross-section data, the dependent variable is set to one if the individual is treated and the period of observation is t_1 and to zero otherwise. Each of the control groups (treated before treatment and non-treated before and after treatment) are then matched to the treated after treatment separately. The overlapping region of support is now composed of the treated to whom a counterfactual is found in each of the three control samples. The three sets of weights can then be used to estimate the ATT using DID (for an empirical application see Blundell et al. 2004).

VI. Instrumental Variables

VI.A. The instrumental variables estimator (IV)

In this section we continue considering the model described by equations (2)-(4) with potential outcomes being partly explained by the observables *X* as in equation (24). The variables *Z* in the decision rule (4) may include all observables *X* in the outcome equation plus additional regressors. For simplicity, we implicitly condition on *X* and omit it from the

discussion below. We also consider only one additional variable in *Z*, which we denote by *z*. Again, the time dimension will not be explicitly considered since longitudinal or repeated cross-section data is not necessarily required to estimate the effect of treatment under the IV assumptions.

In contrast to the matching method, the method of Instrumental Variables (IV) deals directly with selection on the *unobservables*. The IV approach requires the existence of at least one regressor exclusive to the decision rule. In our notation, this is the variable *z*, which is known as the instrument. It affects participation only and so is not in *X*. This is known as the *exclusion restriction*. It implies that the potential outcomes do not vary with *z* and any difference in the mean observed outcomes of two groups differing only with respect to *z* can only be due to consequent differences in the participation rates and composition of the treatment group with respect to potential gains from treatment. When the treatment effect is homogeneous, so that $\alpha^{ATE} = \alpha^{ATT} = \alpha_i = \alpha$, only differences in participation rates subsist and these can be used together with resulting differences in mean outcomes to identify the impact of treatment.

To see this more clearly, we formalize these three assumptions below. The latter assumption states that the treatment effect is homogeneous:

(37) $\alpha_i = \alpha$ for all i.

The first two assumptions define the dependence of the outcome *y* and the participation status *d* on the instrument *z*. They can be stated as:

(38) $\mathrm{P}\left[d = 1 | z\right] \neq \mathrm{P}\left[d = 1\right]$

and

(39) $\mathrm{E}\left[u | z\right] = \mathrm{E}\left[u\right]$.

Under conditions (37) to (39), the instrument *z* is the source of exogenous variation used to approximate randomization. It provides variation correlated with the participation decision only.

Under assumptions (37) and (39), the dependence of *y* on the instrument arises through the index (propensity score) $\mathrm{P}[z] = \mathrm{P}[d = 1|z]$ as follows:

$$\begin{aligned} \mathrm{E}[y_i|z_i] &= \beta + \alpha \mathrm{E}[d_i|z_i] + \mathrm{E}[u_i|z_i] \\ &= \beta + \alpha \mathrm{P}(z_i) + \mathrm{E}[u_i] \\ &= \mathrm{E}[y_i|\mathrm{P}(z_i)]. \end{aligned} \tag{40}$$

Assumption (38) then ensures that two different values of *z* exist that induce variation in *P(z)* and allow for the identification of α. Let (z^*, z^{**}) be such values. Then

$$\mathrm{E}[y_i|z_i = z^*] - \mathrm{E}[y_i|z_i = z^{**}] = \alpha[\mathrm{P}(z^*) - \mathrm{P}(z^{**})]$$

and the treatment effect is identified from the ratio:

$$\alpha = \frac{\mathrm{E}[y_i|z_i = z^*] - \mathrm{E}[y_i|z_i = z^{**}]}{\mathrm{P}(z^*) - \mathrm{P}(z^{**})}. \tag{41}$$

This is the standard IV (or Wald) identification strategy. It is designed to explore discrete instruments or discrete changes in continuous instruments.

In the standard continuous instrument case it is more efficient to explore the whole variation in *z*. The IV conditions discussed above ensure that

$$\begin{aligned} \mathrm{cov}(y,z) &= \alpha \mathrm{cov}(d,z) + \mathrm{cov}(u,z) \\ &= \alpha \mathrm{cov}(d,z) \end{aligned}$$

and the IV estimator is

$$\alpha = \frac{\widehat{\mathrm{cov}}(y,z)}{\widehat{\mathrm{cov}}(d,z)}.$$

VI.B. Weaknesses of IV

A key issue in the implementation of IV is the choice of the instrument. It is frequently very difficult to find an observable variable that satisfies assumption (39), in which case IV is of no practical use. This will happen when participation is mainly driven by the determinants of potential outcomes. In other cases, the instrument has insufficient variation or causes insufficient variation in the propensity score. Instruments with this characteristic are known as weak instruments. Although (38) may still hold with a weak instrument, the consequent (small) size of the denominator in (41) leads to very imprecise estimates of the treatment effect.

Identification using classical IV still relies on the additional homogeneity assumption in equation (37). If (37) does not hold, the exclusion restriction is also unlikely to hold. To see why, notice that the unobservable in the outcome equation is now

$$e_i = u_i + d_i\left(\alpha_i - \alpha^{ATE}\right)$$

and the new exclusion restriction needs to be expressed in terms of *e*:

$$\begin{aligned}\mathrm{E}\left[e|z\right] &= \mathrm{E}\left[u|z\right] + \mathrm{P}(z)\mathrm{E}\left[\alpha_i - \alpha^{ATE}|d=1,z\right] \\ &= \mathrm{E}\left[e\right].\end{aligned}$$

But since *z* explains *d*, the second equality above is generally not satisfied.

The one exception occurs when there is no selection on the idiosyncratic gains. This means that the idiosyncratic gain, $\alpha_i - \alpha^{ATE}$, and the unobservable in the selection rule, *v*, are not related. In such case $\mathrm{E}\left[\alpha_i - \alpha^{ATE}|d=1,z\right] = 0$ and $\mathrm{E}\left[e_i|z\right] = \mathrm{E}\left[e_i\right]$ under (39). Thus, classical IV will still identify ATE (and, which is the same, ATT) if individuals do not have or do not act on unobservable information related to (expected) gains to decide about treatment status.

In the more general case of heterogeneous effects with selection on idiosyncratic gains, IV will not identify ATE or ATT. If individuals are aware of their own idiosyncratic gains from treatment, they will certainly make a more informed participation decision. The resulting selection process breaks the independence between α and *z* conditional on selection since both variables affect the selection process. A change in *z* will drive into treatment individuals with an expected return different from α^{ATE}.

To illustrate the problem, consider the education example we have been using. Assume that the returns to education are partly determined by the child's unobservable ability. Suppose the instrument is some measure of the cost of education (say distance to college) under the assumption that it is uncorrelated with the child's potential earnings and, therefore, ability (this is possibly a strong assumption but serves only our illustration purposes). However, the selection process will create a relationship between distance to college and returns to college education in the data. This is because individuals facing a relatively low cost of education (live closer to college) may be more likely to invest in college education, even if expecting comparatively small returns, then individuals facing higher education costs. Under our simplistic setup, this means that the distribution of ability among college graduates who live far from college is more concentrated on high ability levels than that for college graduates who live close to college. Such compositional differences will then affect the distribution of returns to college in the data for the two groups.

If the homogeneity assumption (37) fails to hold, IV will not generally identify ATE or ATT. This happens because the average outcomes of any two groups differing only on the particular *z*-realizations are different for two reasons: *(i)* different participation rates and *(ii)* compositional differences in the treated/non-treated groups with respect to the unobservables. The later precludes identification of ATE or ATT. However, a different "local" average

parameter can be identified under slightly modified hypothesis - the LATE parameter, to which we now turn.

VI.C. The LATE parameter

The solution advanced by Imbens and Angrist (1994) is to identify the impact of treatment from local changes in the instrument *z* when the effect is heterogeneous. The rationale is that, under certain conditions, a change in *z* reproduces random assignment locally by inducing individuals to alter their participation status without affecting the potential outcomes, $\left(y^0, y^1\right)$. As with standard IV, the difference in average outcomes between two groups differing only in the realization of *z* results exclusively from the consequent difference in participation rates. Unlike standard IV, the identifiable effect will not correspond to the ATE or the ATT. Instead, it will depend on the particular values of *z* used to make the comparison and the identifiable effect is the average impact on individuals that change their participation status when faced with the change in *z* used to estimate the effect of treatment.

As with classical IV, the validity of an instrument *z* depends on whether it determines participation and can be excluded from the outcomes' equation except through its effect on participation. In an heterogeneous effect framework the exclusion condition requires that: *(i) z* has no joint variation with *v*, or otherwise changes in *z* would not separate changes in the participation rates unrelated to outcomes as simultaneous changes in *v* could be related with changes in the unobservable components of the potential outcomes, particularly gains from treatment; and *(ii) z* is unrelated to the unobserved determinants of potential outcomes.

The LATE assumptions can now be formally established. The first two assumptions are identical to the classical IV assumptions (38) and (39):

(42) $\mathrm{P}\left[d = 1 \middle| z\right] \neq \mathrm{P}\left[d = 1\right]$

(43) $\mathrm{E}\left[u|z\right]=\mathrm{E}\left[u\right]$.

However LATE requires stronger identification assumptions than standard IV to compensate for the relaxation of the homogeneity hypothesis. The additional assumption pertains the relationship between *z* and the remaining unobservables:[17]

(44) $\left(\alpha,v\right)\perp z$.

To simplify the notation, define the random variable $d_i\left(z_i\right)$ which represents the treatment status of individual *i* when drawing $z=z_i$. Thus $d\left(z\right)$ assumes the values one or zero depending on whether the unobservable *v* is in a range that leads to participation or no-participation given *z*, respectively (that is, $d\left(z\right)=1\left(g\left(z,v\right)>0\right)$. Under the notation of the selection rule in (3)-(4), assumption (44) ensures that

$$\begin{aligned}\mathrm{P}\left[d=1|z\right]&=\mathrm{P}\left[\mathrm{g}\left(z,v\right)>0|z\right]\\&=\mathrm{P}\left[\mathrm{g}\left(z,v\right)>0\right]\\&=\mathrm{P}\left[d\left(z\right)=1\right]\\&=\mathrm{P}\left[z\right]\end{aligned}$$

where the second equality means that *z* is exogenous in the selection rule. Furthermore, joint independence of idiosyncratic gains and *v* from *z* also guarantees that

$$\mathrm{E}\left[\alpha|z,d=1\right]=\mathrm{E}\left[\alpha|d\left(z\right)=1\right].$$

Taken together, the three LATE assumptions are sufficient to ensure that $d\left(z\right)$ contains all the information in *z* that explains *y*:

$$\begin{aligned}\mathrm{E}\left[y_i|z\right]&=\beta+\mathrm{P}\left[d_i=1|z\right]\mathrm{E}\left[\alpha_i|z,d_i=1\right]\\&=\beta+\mathrm{P}\left[d_i\left(z\right)=1\right]\mathrm{E}\left[\alpha_i|d_i\left(z\right)=1\right].\end{aligned}\tag{45}$$

This means that all relevant information in *z* results from what can be inferred from *z* and *d* about the location of *v*, the unobservable in the selection rule that is correlated with the unobserved components of the outcomes and that is at the root of the selection problem.

We now use this result to compare the observed outcomes at two distinct values of the instrument *z*, say $\left(z^{*}, z^{**}\right)$:

$$\begin{aligned}\mathrm{E}\left[y_i \middle| z=z^{**}\right]-\mathrm{E}\left[y_i \middle| z=z^{*}\right] &= \mathrm{P}\left[d_i\left(z^{**}\right)=1\right]\mathrm{E}\left[\alpha_i \middle| d_i\left(z^{**}\right)=1\right]-\mathrm{P}\left[d_i\left(z^{*}\right)=1\right]\mathrm{E}\left[\alpha_i \middle| d_i\left(z^{*}\right)=1\right] \\ &= \mathrm{P}\left[d_i\left(z^{**}\right)=1, d_i\left(z^{*}\right)=0\right]\mathrm{E}\left[\alpha_i \middle| d_i\left(z^{**}\right)=1, d_i\left(z^{*}\right)=0\right]- \\ &\quad \mathrm{P}\left[d_i\left(z^{**}\right)=0, d_i\left(z^{*}\right)=1\right]\mathrm{E}\left[\alpha_i \middle| d_i\left(z^{**}\right)=0, d_i\left(z^{*}\right)=1\right].\end{aligned}$$

The intuition behind the above expression is that any change in the average outcome *y* when *z* changes is solely due to changes in the treatment status of a subset of the population. The last equality shows two treatment parameters that one may be willing to identify: the impact of treatment on the treated under z^{**} but not treated under z^{*} and the impact of treatment on the treated under z^{*} but not treated under z^{**}. In practice there are frequently strong arguments to eliminate one of the alternatives. For example, it may be the case that every participant at z^{*} also participates at z^{**} but not the reverse. This is the substance of the monotonicity assumption, the last of the LATE assumptions. Formally:

(46) $d\left(z\right)$ is a *monotonic* function of *z*.

Suppose (46) holds. In particular, suppose *d* is increasing in *z* and $z^{**} > z^{*}$ so that $\mathrm{P}\left[d\left(z^{**}\right)=0, d\left(z^{*}\right)=1\right]=0$.

In such case $\mathrm{P}\left[d\left(z^{**}\right)=1, d\left(z^{*}\right)=0\right]=\mathrm{P}\left(z^{**}\right)-\mathrm{P}\left(z^{*}\right)$ and

$$\mathrm{E}\left[y_i \middle| z^{**}\right]-\mathrm{E}\left[y_i \middle| z^{*}\right]=\left[\mathrm{P}\left(z^{**}\right)-\mathrm{P}\left(z^{*}\right)\right]\mathrm{E}\left[\alpha_i \middle| d_i\left(z^{**}\right)=1, d_i\left(z^{*}\right)=0\right]$$

and this equation can be rearranged to obtain the LATE parameter:

$$\begin{aligned}\alpha^{LATE}\left(z^{*}, z^{**}\right) &= \mathrm{E}\left[\alpha_i \middle| d_i\left(z^{**}\right)=1, d_i\left(z^{*}\right)=0\right] \\ &= \frac{\mathrm{E}\left[y_i \middle| z^{**}\right]-\mathrm{E}\left[y_i \middle| z^{*}\right]}{\mathrm{P}\left(z^{**}\right)-\mathrm{P}\left(z^{*}\right)}\end{aligned} \quad (47)$$

The first equality clarifies the meaning of the LATE parameter: it measures the impact of treatment on individuals that move from non-treated to treated when *z* changes from z^* to z^{**}.

The LATE approach can also be illustrated within our running example on education investment. As before, suppose *z* is a measure of cost, say distance to college, with participation assumed to become less likely as *z* increases. To estimate the effect of college education, consider a group of individuals that differ only in *z*. Among those that invest in further education when distance *z* equals z^* some would not do so if $z = z^{**}$ where $z^* < z^{**}$. In this case, LATE measures the impact of college education on the "movers" by assigning any difference on the average outcomes of the two groups to the different enrollment rates caused by the difference in the cost of investing.[18]

VI.C.1. The LATE assumptions

The independence assumptions (43) and (44) are required to establish the result (45) on which derivation of LATE hinges. It states that *z* is independent of the unobservable components in the outcomes and participation rules, namely u_i (mean independent), $\alpha_i - \alpha^{ATE}$ and v_i. This means that *z* should not affect the observed outcomes through any effect on the potential outcomes or any relation with the unobserved components of the model. While the former is easy to understand the later requires some explanation. Suppose *z* is related with *v* in the participation rule and *v* is related with *u* in the outcome equation. Then the potential outcome will generally be related with *z*.

The education example can be used again to illustrate the conditions under which the independence assumption may not apply. As before, suppose *z* is a measure of the cost of education, say distance to college. A family that values education may take some steps to facilitate the investment, for example by taking it into account when deciding about residence.

Such family may also be particularly interested in encouraging learning and developing a taste for it in the children. So children raised in such environment may benefit both from lower education costs *and* higher taste for education. The later is unobservable, included in *v*, and is likely to be related with the future taste for working, also unobservable and included in *u*. In this case, although *z* has no direct impact on potential outcomes, the selection of home location will create a dependence between *z* and the potential outcomes arising through a dependence between *v* and *u*. That is, if *z* is not exogenous in the participation equation then two groups with different realizations of *z* will represent two different populations in terms of the distribution of *v* and, thus, two different populations in terms of the distribution of *u* (a similar argument could be constructed in relation to α). In such case, even if one could observe both potential outcomes they would not be independent of *z* in the data.

The monotonicity assumption is required for interpretation purposes. It is usually justified on theoretical grounds as it is generally unverifiable.[19] Under monotonicity of *d* with respect to *z*, the LATE parameter measures the impact of treatment on individuals that move from non-treated to treated as *z* changes. If monotonicity does not hold, LATE measures the change in average outcome caused by a change in the instrument, which is due to individuals moving *in and out* of participation, but cannot separate the effect of treatment on individuals that move in from that on individuals that move out as a consequence of a change in *z* (see Heckman 1997).

Notice that the LATE assumptions are local: they only need to hold locally, for the specific values of *z* used in the estimation process. As a consequence, LATE is a local parameter, specific to the population defined by the instrument. This is further discussed in the next section.

VI.C.2. What does LATE measure?

Although analytically very similar to the IV estimator in (41), LATE is intrinsically different since it does not represent ATT or ATE. LATE depends on the particular values of z used to evaluate the treatment and on the particular instrument chosen. The group of "movers" is not in general representative of the whole treated or, even less, the whole population. Whether the parameter is of policy interest or not depends on the instrument and the specific values of the instrument used in the estimation (see, for example, the discussion in Heckman, Lalonde and Smith 1999). When a discrete variable, namely a change in policy, is used to instrument participation, LATE will measure the effect of treatment on individuals changing their treatment status in response to the policy change. In this case, LATE focuses on an important subpopulation and may provide an important measure of the impact of the policy. If, on the other hand, a continuous variable measuring some individual characteristic is used to instrument participation, LATE will generally be much less informative.

In our education example, notice that we discussed two alternative instruments to measure a local effect. In the first case, in the context of DID, we used a change in policy to measure the impact of education on individuals moving into education. DID differs from the standard LATE estimator based on a change in policy only by allowing the aggregate conditions to vary over time (although it requires treated and controls to be similarly affected by the market conditions). In the second case, we discussed the use of family background or cost of education to instrument participation. Clearly, the former is much more informative for the policy maker then the later. The estimated parameter based on our continuous variable will depend on the specific values being compared, may not represent a specific population that can be easily targeted and is more likely to raise arguments about the validity of the instrument (just as illustrated by the discussion in the previous section).

VI.D. The Marginal Treatment Effect

Heckman and Vytlacil (1999, 2001, 2006) and Carneiro, Heckman and Vytlacil (2005) reinterpret the local IV methods and the local treatment effect parameters within a selection model. These authors consider the estimation of the impact of treatment over the whole distribution of a continuous instrument. To do so, they use infinitesimal changes in the participation probabilities to measure the limit of LATE as the change in these probabilities becomes arbitrarily small. As the whole distribution of local treatment effects is possibly identified, all more aggregate parameters can also be recovered by integration over the distribution of the probability of participation.

These authors consider a version of the selection model (3)-(4) which assumes additive separability of the unobservable, *v*. As before, selection follows a latent variable specification where

$$d_i^* = \mathrm{g}(z_i) - v_i$$

and

$$d_i = \begin{cases} 1 & \text{if } v_i \le \mathrm{g}(z_i) \\ 0 & \text{otherwise.} \end{cases}$$

The propensity score as a function of *z* is,

$$\begin{aligned} \mathrm{P}(z_i) &= \mathrm{P}\left[d_i = 1 | z_i \right] \\ &= \mathrm{P}\left[v_i \le \mathrm{g}(z_i) \right] \\ &= \mathrm{F}_v\left(\mathrm{g}(z_i) \right) \end{aligned}$$

where F_v is the distribution function of the unobservable *v*. Heckman and Vytlacil (1999) use a general alternative representation of the above preferences when *v* is an absolutely continuous random variable (meaning that *v* has no mass points). This is obtained by transforming the selection rule

$$v_i \le \mathrm{g}(z_i)$$

by a monotonically increasing function such as F_v to yield:

$F_v(v_i) \leq F_v(g(z_i))$ which can be written as

$\tilde{v}_i \leq P(z_i)$.

Given continuity of *v*, the transformed unobservable $\tilde{v}$ will follow a uniform distribution in [0,1] and the data-equivalent selection model is

(48) $\tilde{d}_i^* = P(z_i) - \tilde{v}_i$

with

(49) $d_i = \begin{cases} 1 & \text{if } \tilde{v}_i \leq P(z_i) \\ 0 & \text{otherwise.} \end{cases}$

The advantage of the latter representation is the connection between the propensity score and the newly defined unobservable $\tilde{v}$: an individual with characteristic *z* is indifferent about participation when drawing $\tilde{v} = P(z)$ and will participate with probability *P(z)* or when drawing $\tilde{v} \leq P(z)$.

Using this representation of the decision process, the marginal treatment effect (MTE) parameter at a point $\tilde{v}^*$ of the distribution of the unobservable can now be written as

$\alpha^{MTE}(\tilde{v}^*) = E[\alpha | \tilde{v} = \tilde{v}^*]$.

This parameter measures the impact of treatment on individuals with unobservable $\tilde{v}^*$ affecting the decision process. Under the LATE assumptions (43)-(44), which are now imported to the analysis of identification and estimation of the MTE parameter, the instrument *z* does not bring further information about the expected gains from treatment after conditioning for $\tilde{v}$. This is because $\tilde{v}$ contains all the information in *d* that may be related with the potential outcomes. So, individuals with the same $\tilde{v}$ but experiencing different values of *z* expect to gain the same from treatment. Thus,

$\alpha^{MTE}(\tilde{v}^*) = E[\alpha | \tilde{v} = \tilde{v}^*, z]$

for any possible value of *z*. In particular, one may try to evaluate the MTE at the point(s) in the distribution of *z* where individuals are indifferent about participation under the assumption that *P(z)* is a non-trivial function of *z*:

$$\alpha^{MTE}\left(\tilde{v}^*\right)=\mathrm{E}\left[\alpha\middle|\tilde{v}=\tilde{v}^*,\tilde{v}^*=\mathrm{P}(z)\right].$$

This specification explains the alternative definition of MTE that is more commonly encountered, namely the average effect of treatment on individuals just indifferent about participation at *P(z)* (this is the definition of Bjorklund and Moffitt 1987 when first introducing the MTE; see also the recent nonparametric application by Moffitt 2007). It is this definition that is explored to identify MTE.

Assumptions (43)-(44) together with the additive separability of *v* can be used to show:

$$\begin{aligned}\mathrm{E}\left[y\middle|z\right]&=\beta+\mathrm{P}(z)\mathrm{E}\left[\alpha\middle|z,d=1\right]\\&=\beta+\mathrm{P}(z)\mathrm{E}\left[\alpha\middle|\tilde{v}\le\mathrm{P}(z)\right]\\&=\mathrm{E}\left[y\middle|\mathrm{P}(z)\right].\end{aligned}$$

Using this formulation, the LATE parameter can expressed as:

$$\begin{aligned}\alpha^{LATE}\left(z^*,z^{**}\right)&=\frac{\mathrm{E}\left[y\middle|z^{**}\right]-\mathrm{E}\left[y\middle|z^*\right]}{\mathrm{P}\left(z^{**}\right)-\mathrm{P}\left(z^*\right)}\\&=\frac{\mathrm{E}\left[y\middle|\mathrm{P}\left(z^{**}\right)\right]-\mathrm{E}\left[y\middle|\mathrm{P}\left(z^*\right)\right]}{\mathrm{P}\left(z^{**}\right)-\mathrm{P}\left(z^*\right)}\\&=\alpha^{LATE}\left(\mathrm{P}\left(z^*\right),\mathrm{P}\left(z^{**}\right)\right)\end{aligned}$$

Within this framework, LATE measures the impact of treatment on individuals with unobservable characteristics $\tilde{v}$ in the interval $\left[\mathrm{P}\left(z^*\right),\mathrm{P}\left(z^{**}\right)\right]$. Again, this parameter does not change with the particular values of *z* selected for as long as $\tilde{v}$ remains on the interval $\left[\mathrm{P}\left(z^*\right),\mathrm{P}\left(z^{**}\right)\right]$.

The MTE can be defined from LATE by considering an arbitrarily small interval in $\tilde{v}$. The limit can also be taken on the estimator of LATE to define an estimator for MTE. Notice that the definition of LATE in (47) determines the LATE estimator by identifying the movers using the instrument *z*. This will not be independent of *z* because it relies on the specific variation used to assess the impact, which will determine the specific population of movers under scrutiny (or the margin to which the parameter corresponds to).

The Local Instrumental Variables (LIV) is precisely an estimator of the MTE obtained by taking the limit of the LATE estimator (47) as $\mathrm{P}\left(z^{*}\right)$ becomes arbitrarily close to $\mathrm{P}\left(z^{**}\right)$:

$$\alpha^{LIV}\left(\mathrm{P}\left(z\right)\right)=\frac{\partial \mathrm{E}\left[y\middle|\tilde{v}=\mathrm{P}\left(z\right)\right]}{\partial \mathrm{P}\left(z\right)}$$

This estimator also depends on *z* in the sense that, for a particular selected value of the instrument, the specific margin at which MTE is being estimated is the specific value of the unobservable at the indifference point, namely $\tilde{v}=\mathrm{P}\left(z\right)$. Comparatively to LATE, however, the use of MTE is usually associated with the intention to recover the full distribution of treatment effects. MTE uses a continuous instrument to recover the whole (or an interesting part) of the distribution of participation probabilities from zero to one for as long as all individuals have strictly positive probabilities of being treated and non-treated (that is, for as long as *z* is able to move all individuals in and out of treatment).

If data are rich enough to explore changes in treatment status over the whole distribution of $\tilde{v}$ then all the average parameters, namely ATE, ATT, ATNT and LATE, can be expressed as averages of MTE using different weights (see Appendix 3for details). For example, the estimation of ATT using MTE with a continuous instrument *z* requires the space of $\tilde{v}$, [0,1], to be finely discretized using the distribution of *P(z)*. Estimation may use some non-parametric regression procedure to identify the slope of *y* with respect to *P(z)* at each of

the points on the grid - say a Local Quadratic Regression. This is the MTE at each point $\tilde{v}$. The ATT among individuals with a probability of participation equal to p, $\alpha^{ATT}(p)$, may then be obtained by integrating the MTE's over the space of $\tilde{v}$ up to p - these are the participants among those with a probability of participation equal to p. The overall ATT may now be obtained by integrating $\alpha^{ATT}(p)$ over the whole distribution of p (see Carneiro and Lee 2007; Carneiro Heckman and Vytlacil 2005, for details on the implementation procedure).

However, data may not be rich enough to allow for the estimation of MTE over the whole distribution of $\tilde{v}$, in which case LATE may be the best available option. This is clearly the case when the instrument is binary as, for example, a specific change in policy, in which case LATE is particularly suited to identify a parameter of interest and the MTE approach is imply not possible as it relies on a continuous instrument.

VI.E. Using IV to estimate returns to education

Under the IV conditions, the variables in the selection process which do not enter the outcome equation may be used to instrument the endogenous participation variable. Within the returns to education example, this amounts to use the variable(s) determining education investment to instrument education attainment when ability is not observed. In our simple model this means that family background (z) is a valid instrument while the test score (s) is not since it is correlated with ability, which directly affects earnings.

Table 4 displays estimates of the ATT using standard and local IV. All numbers are for an economy with subsidized advanced education available to agents performing well at the basic education level. We consider both negatively correlated (columns 1 and 2) and independent (columns 3 and 4) unobservables, u and v.

Rows 2 and 3 in the table display classical IV estimates. We expect these to be biased as the homogeneity assumption (37) is not met. Given this, the estimator based on the instrument *z* does surprisingly well (row 2). As expected, the invalid instrument *s* produces more bias (row 3).

Similar *local* IV estimates are presented in rows 4 and 5 and show an interesting pattern. Under uncorrelated residuals, the instrument *z* used with local IV correctly identifies the ATT, just as expected (columns 3 and 4). However, the bias is considerably larger in the case of correlated unobservable terms even when education is instrumented with *z* (columns 1 and 2). To understand the source of bias in the case of correlated residuals, notice that the local IV technique estimates the ATT by integrating the MTE over the population of participants (see Appendix 3). The MTE at $\tilde{v}$ measures the impact of treatment on individuals that draw this specific value for the unobservable component of the selection rule. Because this unobservable contains all information about potential outcomes in the selection rule, changing *z* conditional on $\tilde{v}$ will not change (expected) gains. So the MTE can, in particular, be interpreted as the impact on individuals with observable characteristics *z* that make them indifferent about participation at $\tilde{v}$. These are the individuals that draw $\tilde{v} = \mathrm{P}(z)$.

Estimation of the MTE relies on these movers to feed a (local) IV estimator. However, the ATT cannot be recovered if *P(z)* is not observed to vary over the whole unit interval. More precisely, the identification of the ATT will be affected in the absence of observations for *P(z)* in $\left[0, \underline{p}\right]$ for some $\underline{p}$ significantly larger than zero. In this case, we know that individuals experiencing $\tilde{v} < \underline{p}$ will always prefer to participate within the observable range of *P(z)*. But then we never observe these individuals at their indifference point between participation and non-participation. Unfortunately, these individuals are unlikely to be a random sample of the population: they prefer to participate even at low levels of *P(z)*, which may indicate they expect to earn more from participation than most of the

population. Thus, the estimated effect will not be the ATT but the average treatment for individuals indifferent between participation and non-participation at the values of $\tilde{v}$ in the observable interval, $\left[\underline{p} > 0, \overline{p}\right]$.

The lack of support affects the results in columns 1 and 2, when the disturbances are correlated, where only values of *P(z)* above 0.06 are observable. The expected outcome is that the obtained estimates are downward biased. In the uncorrelated disturbance case, however, the range of observable *P(z)* starts very close to zero. In this case, displayed in columns 3 and 4, we are able to identify the impact of education even among individuals that show a strong preference towards education. As predicted, estimates based on LIV for correlated residuals tend to be lower than the true value of this parameter (row 4, column 1 and 2) while uncorrelated residuals produce unbiased estimates (row 4, columns 3 and 4).[20]

The above discussion is closely related to the literature on the ability of IV to produce interpretable parameters (see Heckman 1997; Heckman and Vytlacil 1998). The local parameters estimated by IV depend on the ability of the used instrument to induce a change of treatment status in each individual. Even the estimation based on the MTE, which explicitly attempts to run over the whole distribution of $\tilde{v}$ and uses the correct weights to aggregate the local parameters, may produce estimates that are not global. Instead, such estimates may dependent on the particular instrument being used and apply only to the subpopulation of individuals that would switch treatment status at observable values of the instrument. Whether or not the identified parameter is of interest depends on the specific policy/evaluation question.

A final remark concerns the use of an invalid instrument such as the test score, *s*. Both estimators in table 4 are considerably less biased in the presence of a positive subsidy. The reason for this lies on the individual's response to the introduction of an education subsidy. Contrary to the no subsidy scenario, many individuals will make a positive effort to

score better on the test if a subsidy is available. Such effort is related to *z*. Thus, the relationship between *s* and θ will be reduced while *z* will become related with the test score due to the endogenous effort decision. Although still an invalid instrument, *s* will now incorporate more exogenous variation that is related with participation, which helps in the identification of the true effect.

VII. Discontinuity Design

VII.A. The discontinuity design estimator (DD)

Certain non-experimental policy designs provide sources of randomization that can be explored to estimate treatment effects under relatively weak assumptions. This is really the motivation for the natural experiment approach discussed earlier. However, a special case that has attracted recent attention occurs when the probability of enrollment into treatment changes discontinuously with some continuous variable *z*. The variable *z* is an observable instrument, typically used to determine eligibility. It is, therefore, in matrix *Z* in the selection model (3)-(4). The discontinuity design estimator (DD) uses the discontinuous dependence of *d* on *z* to identify a local average treatment effect even when the instrument does not satisfy the IV assumptions discussed before. Instead of some exclusion or independence assumption like (39) and (44), DD relies on a continuous relationship between the instrument *z* and all the determinants of the outcome except participation in treatment. Any discontinuity in *y* as a function of *z* is, therefore, attributed to a discontinuous change in the participation rate as a function of *z*. As will be discussed, the parameter identified by DD is a local average treatment effect like the LATE parameter discussed under IV but is not necessarily the same parameter.[21]

As before, we assume participation *d* is determined by *z* and the unobservables *v* in a completely flexible way: $d = 1\left(g(z,v) \geq 0\right)$. The dependence of *d* on *z* means that the participation probability changes with *z*. The main source of identification used by DD is a discontinuity in such probability at a given point in the distribution of *z*. The discontinuity may be *sharp* or *fuzzy*, depending on whether participation is a deterministic function of *z* or not. We now discuss these two cases.

VII.A.1. The sharp design

The most popular case, although empirically less frequent, is what is known by *sharp design*. This occurs when *z* fully determines participation on the basis of a threshold, z^*. The treated (non-treated) are individuals with values of *z*, say, above (below) the threshold. In this case, participation status changes at z^* for all individuals, from being deterministically equal to zero to being deterministically equal to one.

Thus the probability of participation changes discontinuously at z^* from zero to one. The identification condition with sharp design can be stated as follows,

$$\begin{array}{l} \lim_{z \to z^{*-}} \mathrm{P}\left(d = 1 | z\right) = \mathrm{P}\left(z^{*-}\right) = 0 \\ \lim_{z \to z^{*+}} \mathrm{P}\left(d = 1 | z\right) = \mathrm{P}\left(z^{*+}\right) = 1 \end{array} \quad (50)$$

where, to simplify the notation, $\mathrm{P}\left(z^{*-}\right)$ ($\mathrm{P}\left(z^{*+}\right)$)represents the limit of the propensity score ($\mathrm{P}\left(d = 1 | z\right) = \mathrm{P}\left(z\right)$) as *z* approaches z^* from below (above). Both limits are assumed to exist.

The fact that participation is locally a deterministic function of *z* means that individuals do not contribute to the decision process.[22] The sharp design implies that the decision process is exogenously determined by *z* and all the selection is on the observables. Thus, the impact of treatment is probably independent from the selection process, at least

locally. Although selection occurs only on the observables, matching is not feasible given the absence of overlap between treated and controls once *z* is included in the set of covariates. Instead of the common support assumption used in matching, DD is based on the additional hypothesis of continuity of the remaining determinants of outcomes as functions of *z* at z^*. Under sharp design, all that is required is continuity of y^0 at z^* to ensure that the non-treated on one side of the threshold are the correct counterfactual for the treated on the opposite side. Within our model of outcomes, (1)-(2), this is equivalent to the condition

(51) $\mathrm{E}\left[u_i \middle| z^{*+}\right] = \mathrm{E}\left[u_i \middle| z^{*-}\right]$

where $\mathrm{E}\left[u_i \middle| z^{*+}\right]$ and $\mathrm{E}\left[u_i \middle| z^{*-}\right]$ are the limits of $\mathrm{E}\left[u_i \middle| z\right]$ when *z* approaches z^* from above and below, respectively.

Under assumptions (50) and (51), any observed discontinuity in *y* at z^* results exclusively from the discontinuity in the participation rate. The DD parameter is in this case:

$$\alpha^{DD}\left(z^*\right) = \mathrm{E}\left[y_i \middle| z^{*+}\right] - \mathrm{E}\left[y_i \middle| z^{*-}\right]$$

where $\mathrm{E}\left[y_i \middle| z^{*+}\right]$ and $\mathrm{E}\left[y_i \middle| z^{*-}\right]$ are the limits of $\mathrm{E}\left[y_i \middle| z\right]$ when *z* approaches z^* from above and below, respectively. $\alpha^{DD}\left(z^*\right)$ measures the impact of treatment on a randomly selected individual with observable characteristics *z* just above z^*:

$$\alpha^{DD}\left(z^*\right) = \mathrm{E}\left[\alpha \middle| z^{*+}\right].$$

This is not necessarily the same as $\mathrm{E}\left[\alpha \middle| z^*\right]$ because nothing was said about the continuity of this object at z^*. If we now impose this additional assumption

(52) $\mathrm{E}\left[\alpha \middle| z^{*+}\right] = \mathrm{E}\left[\alpha \middle| z^{*-}\right]$

the DD parameter can be more naturally interpreted as being the impact of treatment on a randomly selected individual at the cutoff point (z^*):

$\alpha^{DD}\left(z^*\right) = \mathrm{E}\left[\alpha \middle| z^*\right]$.

There are a few examples of economic studies that fall in the category of sharp design. They typically involve some exogenously imposed eligibility rule with a cut-off point. One example is the New Deal evaluation discussed above. Among other things, eligibility is based on age. Eligibles are those individuals that have not completed 25 years of age when reaching six months in unemployment. De Giorgi (2005) used this rule to estimate the impact of the New Deal on the oldest participants using a sharp regression discontinuity approach. Results in De Giorgi (2005) confirm the findings in Blundell et al. (2004) that employment probabilities increase by 4 to 5 percentage points among the treated as a consequence of treatment. Another recent empirical application by Card and Shore-Sheppard (2004) studies the impacts of expansions of the Medicaid system to cover children in low income families where the eligibility rules were based on age or date of birth.

VII.A.2. The fuzzy design

Possibly more common in economics is the *fuzzy design*. It refers to the situation in which the conditional probability of participation, $\mathrm{P}\left(d=1\middle|z\right)=\mathrm{P}\left(z\right)$, is discontinuous at z^* but does not completely determine participation so the jump at z^* is of smaller magnitude than one. A fuzzy design occurs when dimensions other than *z*, particularly unobserved dimensions, also affect participation. In the general fuzzy design case, participation and non-participation occur on both sides of the threshold. Thus, assumption (50) needs to be adjusted accordingly

(53) $\mathrm{P}\left(z^{*-}\right) \neq \mathrm{P}\left(z^{*+}\right)$.

To illustrate a possible fuzzy design, consider our education example and suppose a subsidy is available for individuals scoring above a certain threshold in a test. The university intake will include both subsidized and unsubsidized individuals. However, the threshold-rule is expected to create a discontinuity in the probability of enrollment given the discontinuous change in the cost of education at the threshold.

Just as in the sharp design case, identification of the treatment effect parameter requires continuity of the remaining determinants of the outcomes. But since treated and non-treated exist on both sides of the threshold in a general fuzzy design, continuity of the outcomes is now required for both y^0 and y^1. Thus, both assumptions (51) and (52) are required. Under (51) to (53), any discontinuity of *y* at z^* can only be linked to the discontinuous change in the participation rate at that point.

Unfortunately, regression discontinuity under fuzzy design loses much of its simplicity and appeal. The additional problem here is that only a subpopulation moves treatment status at the discontinuity point and the selection of movers is likely to be related with potential outcomes. This is a similar problem to that discussed under LATE, where the distribution of the unobservables in the selection rule could conceivably be related with *z*, thus rendering the two comparison groups different with respect to unobservables possibly related with potential outcomes. Fuzzy DD relies on the following additional local (mean) independence assumption to identify a local treatment effect parameter:

(54) $\mathrm{E}(\alpha_i | d, z) = \mathrm{E}(\alpha_i | z)$ for z in a small neighborhood of z^*.

This assumption rules out selection on idiosyncratic gains at the local level. It states that the mean gain from treatment for a population with a fixed value of *z* does not depend on the treatment status. Condition (54) is required for fuzzy DD to closely reproduce sharp DD, where selection is locally excluded given the deterministic participation rule. But it is a strong

assumption, even at the local level, as it locally excludes the possibility of unobserved factors related with gains locally determining participation along with *z*.

In terms of our running education example, where *z* is the test score determining eligibility to subsidized education, assumption (54) excludes the possibility of ability factors not totally captured in the test score to affect participation and outcomes simultaneously conditional on the test score.

Under assumption (54), the conditional mean outcome *y* at a point *z* close to z^* (specifically, $z \in \left[z^{*-}, z^{*+} \right]$) can be written as:

$$\mathrm{E}\left[y_i | z \right] = \beta + \mathrm{E}\left[\alpha_i | z \right] \mathrm{P}(z) + \mathrm{E}\left[u_i | z \right].$$

The additional (dis)continuity assumptions (51)-(53) will suffice to identify the DD parameter:

$$(55)\ \alpha^{DD}\left(z^*\right) = \frac{\mathrm{E}\left[y_i \middle| z^{*+} \right] - \mathrm{E}\left[y_i \middle| z^{*-} \right]}{\mathrm{P}\left(z^{*+}\right) - \mathrm{P}\left(z^{*-}\right)}.$$

As before, $\alpha^{DD}\left(z^*\right)$ is the local average treatment effect, $\mathrm{E}\left(\alpha_i | z = z^*\right)$. Under (54) it measures the mean impact of treatment on a randomly selected individual with characteristic z^*. This is an average treatment effect at the local level since selection on idiosyncratic gains is locally excluded.

The local continuity and independence assumptions recover randomization under discontinuity in the odds of participation at the discontinuity point. The independence assumption is precisely a local version of randomization on gains assumption (R2), meaning that ATE is identifiable locally by DD. Note also that, under the independence assumption (54), ATE and ATT are locally equal. Randomization on untreated outcomes (corresponding to assumption (R1)) is not guaranteed to hold but instead the error term for the non-treated, *u*,

is required to be a continuous function of *z* at z^*. Continuity ensures that it vanishes by differencing on the limit, thus ceasing to be a problem.

The DD estimator is the sample analog of (55):

$$(56)\ \hat{\alpha}^{DD}\left(z^*\right)=\frac{\overline{y}^{+}-\overline{y}^{-}}{\hat{\mathrm{P}}\left(z^{*+}\right)-\hat{\mathrm{P}}\left(z^{*-}\right)}.$$

where $\overline{y}^{+}$ and $\overline{y}^{-}$ are sample averages of the observed outcomes at each side of the threshold and $\hat{\mathrm{P}}\left(z^{*+}\right)$ and $\hat{\mathrm{P}}\left(z^{*-}\right)$ are estimators of the participation probability at each side of the threshold.

A non-parametric version of DD is simple to implement. It only requires running non-parametric regressions of *y* and *d* on *z* locally, separately on each side of the discontinuity point. The predicted limits can then be used to estimate the impact of treatment using expression (56) (for more details and alternative procedures see van der Klaauw 2008 and citations therein).

A simple special case

One case that is empirically relevant is that of a treatment only available but not mandatory on one side of the threshold (say z^{*+}). This is the case, for example, of the Swedish Youth Practice, a subsidized employment program available for unemployed individuals under the age of 25.[23] Participation is not compulsory among eligibles but is not possible for anyone aged 25 or above. This is a special case of fuzzy design which turns out to be identical to the sharp design in terms of necessary identification assumptions (for more details see Battistin and Rettore 2007). The simplicity of this case stems from the fact that, at the non-eligible side of the threshold (say z^{*-}) the expected outcome conditional on *z* is $\beta+\mathrm{E}\left[u\middle|z^{*-}\right]$, which does not depend on gains. Thus, under the continuity assumption

$$\alpha^{DD}\left(z^{*}\right)=\mathrm{E}\left[\alpha_i\left|d_i=1,z^{*+}\right.\right]$$
$$=\frac{\mathrm{E}\left[y_i\left|z^{*+}\right.\right]-\mathrm{E}\left[y_i\left|z^{*-}\right.\right]}{\mathrm{P}\left(z^{*+}\right)}.$$

Under the additional continuity assumption (52), the DD parameter is more naturally interpreted as the impact of treatment on participants at the margin,

$$\alpha^{DD}\left(z^{*}\right)=\mathrm{E}\left[\alpha_i\left|d_i=1,z^{*}\right.\right].$$

VII.B. The link between DD and IV

Interestingly, we have discussed the average treatment effect at a local level before, under IV. This was the LATE parameter or, when taking the limits using a continuous instrument, the MTE. To understand the similarities and differences between DD and local IV we consider the fuzzy design case and notice that both methods will identify the same parameter in a sharp design framework, namely the mean effect of treatment on a randomly selected individual among the treated close to the eligibility cutoff point.

The fuzzy design case is slightly more complex. DD relies on continuity and the local independence assumption in equation (54). The latter determines the parameter identified by DD as being the average impact of treatment on a randomly selected individual with a value of *z* at the threshold.

In turn, LATE relies on the independence assumptions (43)-(44) and on the monotonicity assumption (46). Under these conditions, LATE identifies the average impact of treatment on a randomly selected individual from the group of agents that change participation status as the value of the instrument changes from z^{*-} to z^{*+}.

The empirical estimates of LATE and DD when applied to the same neighborhood of z^{*} coincide exactly, what differs is the interpretation. The preferred interpretation should

be justified on the grounds of the specific application and policy design. If individuals are believed to have no decision power at the local level, then estimates may represent local effects on randomly selected individuals (DD interpretation). Alternatively, if the policy provides clear participation incentives on one side of the threshold and individuals are expected to make informed participation decisions, then a local impact on the movers becomes a more credible interpretation (LATE).[24]

VII.C. Weaknesses of DD

An obvious drawback of DD is its dependence on discontinuous changes in the odds of participation. In general this implies that only a local average parameter is identifiable. As in the binary instrument case of local IV, the DD analysis is restricted to the discontinuity point dictated by the design of the policy. As discussed before under LATE with continuous instruments, the interpretation of the identified parameter can be a problem whenever the treatment effect, α, changes with z.

To illustrate these issues, consider the context of our educational example. Suppose a subsidy is available for individuals willing to enroll in high education for as long as they score above a certain threshold $\underline{s}$ in a given test. The subsidy creates a discontinuity in the cost of education at the threshold and, therefore, a discontinuity in the participation rates. On the other hand, the test score, s, and the returns to education, α, are expected to be (positively) correlated if both depend on, say, ability. But then, the local analysis will only consider a specific subpopulation with a particular distribution of ability which is not that of the whole population or of the treated population. That is, at best the returns to education are estimated at a certain margin and other more general parameters cannot be inferred.

However, we could also suspect that neither the DD nor the LATE assumptions hold in this example. The former requires local independence of the participation decision from the

potential gains conditional on the test score. But at any given level of the test score there is a non-degenerate distribution of ability levels. If higher ability individuals expect to gain more from treatment and are, for this reason, more likely to participate, then the local independence assumption of DD (54) cannot be supported. The latter requires exogeneity of the instrument (test score) in the decision rule. But again, if both the test score and the gains from treatment depend on ability and individuals use information on expected gains to decide about participation, then the instrument will not be exogenous in the selection rule. However, while this may be a serious problem to the use of LATE more generally, the infinitesimal changes considered here will reduce its severity when applied in the context of DD.

Related with the previous comment, there is also the possibility that individuals manipulate z in order to avoid/boost the chances of participation. Such behavior would in general invalidate the DD and LATE identification strategies by rendering the two comparison groups incomparable with respect to the distribution of unobservables. In a recent paper, Lee (2005) notices that this is only a problem if the individuals can perfectly control z, thus positioning themselves at one or the other side of the threshold at will. But even with imperfect control over z, the interpretation of the identified parameter may change, particularly when a change in policy is being explored. This is because an endogenous reaction to the eligibility rule by manipulating z will affect the composition of the group on the neighborhood of the threshold. Thus, the estimated effects may correspond to a margin very different from what would be expected in the absence of such behavior.

In the context of the education illustration being recurrently used in this paper, notice that agents will certainly react to the existence of an education subsidy and to the eligibility rules. If eligibility is based on a test-score they may put extra effort on preparing for it if willing to continue investing in education. Thus, for sure the group of individuals scoring above the eligibility threshold will differ from that of individuals scoring below not

only due to ability but also because preparation effort is endogenously selected. However, the ex-ante random nature of the test-score implies that any two groups scoring just nfinitesimally differently will be compositionally identical as such small differences in test scores are random. Thus, the local DD/LATE will still identify the treatment effect at the threshold margin. However, the introduction of a subsidy policy will affect effort in preparation for the test and thus change the distribution of test-scores and the composition of students in the neighborhood of the threshold. Thus, the post-policy marginal student in terms of eligibility will not be the same as the pre-policy one, particularly perhaps with respect to ability, and this will determine the estimated impact.

A final downside of DD, which is also common with local IV relates to the implementation stage. By restricting analysis to the local level, the sample size may be insufficient to produce precise estimates of the treatment effect parameters.

VII.D. Using DD to estimate the returns to education

Estimation using the discontinuity design method is only possible when a discontinuous change in participation can be used. Within the returns to education example, such discontinuity is introduced by the criterion defining eligibility to subsidized advanced education. This is based on the test score and classifies as eligibles those individuals scoring above $\underline{s} = 4$.

Table 5 displays the monte carlo results using discontinuity design to estimate the impact of education at the eligibility threshold. We present estimates under the assumption of no correlation and negative correlation between the error components in the selection and outcome equations.

We expected this method to work well, its robustness arising from the weak set assumptions on which it is based. However, being a local method based on a small number of

observations, it can become very sensitive to small changes in a few observations. Moreover, DD with imperfect compliance requires dividing by the difference in participation rates as shown in equation (55). If the difference is relatively small, it introduces important variation in the estimation process. For these reasons, some care should be exerted during estimation regarding the choice of weights and bandwidth.

Table 5 displays DD estimates using two of the most commonly applied kernel weights, Epanechnikov (row 1) and Gaussian (row 2), together with a range of possible bandwidths. Bias is measured against the local ATE, the parameter that DD identifies in ideal conditions. For comparison purposes, notice that the ATT as reported in table 1 is 0.453, very close to the local ATE of 0.469. This is a feature of this example and does not necessarily extend to other cases. Quite on the contrary, the local parameter estimated by DD can be very different from the ATT. Like LATE, DD is especially well suited to address the questions related with potential extensions of the policy to a wider group by partially relaxing the eligibility rules.

Rows 1 and 2 in table 5 show that there is considerable variation in the amount of bias by shape of the kernel function and bandwidth. In general, small bandwidths perform better in terms of bias but can introduce large variability on the estimates. In our example, smaller bandwidths are better for the Gaussian kernel weights (row 2), which use all the observable domain to estimate the effect of treatment. Epanechnikov weights, however, work better with slightly bigger bandwidths in this example as it concentrates more weight in fewer observations, becoming more sensitive to small variations particularly on the participation probabilities.

VIII. Control Function Methods

VIII.A. The Control Function Estimator (CF)

When selection is on the unobservables, one attractive approach to the evaluation problem is to take the nature of the selection rule (3)-(4) explicitly into consideration in the estimation process (see Heckman 1976). The control function estimator (CF) does exactly this, treating the endogeneity of *d* as an omitted variable problem.

Consider the outcome equation (9) together with the selection rule (3)-(4). In this section we again abstract from time as it is not a determinant factor for CF. For simplicity of notation and exposition, we also drop the regressors *X* in the outcome equation (considered under matching), implicitly assuming that all the analysis is conditional on the observables *X*.

CF is based on the assumption that all relevant selection information is contained in *v*. It can be formally stated as

(57) $(u,\alpha) \perp (d,Z) | v$.

This assumption states that, were we be able to control for *v*, *d* would be exogenous in the outcome equation (see, for example, Blundell and Powell 2003, 2004). Assumption (57) also states that *Z* is independent of (u,α) conditional on *v*, an exclusion restriction in the spirit of IV but explicitly stating that any impact of *Z* on potential outcomes arises through what it says about the unobservable *v* together with the observable participation status. Thus, changes in *Z* lead to changes in the distribution of *observable* outcomes that can be rooted to changes in the composition of the treated population since *Z* has no effect on the distribution of *potential* outcomes.

Often it is only a weaker conditional mean restriction that is required. After conditioning on other possible regressors in the outcome equation, *X*, (or, alternatively, if *d* is

additively separable from *X*) all that is required is mean independence of *u* from *d* and *Z* conditional on *v*:

$$(58)\quad \begin{aligned} \mathrm{E}\left[u|v,d,Z\right] &= \mathrm{E}\left[u|v\right] = \mathrm{h}_u(v) \\ \mathrm{E}\left[\alpha|v,d,Z\right] &= \mathrm{E}\left[\alpha|v\right] = \mathrm{h}_\alpha(v) \end{aligned}$$

where $(\mathrm{h}_u, \mathrm{h}_\alpha)$ is a pair of functions of *v*, the control functions.

CF is close to a fully structural approach in the sense that it *explicitly* incorporates the decision process in the estimation of the impact of the treatment. Unlike IV, which conditions on a fitted value of the endogenous regressor, CF conditions on both the endogenous regressor and the error term from the decision equation. The problem is how to specify and identify the unobservable term, *v*.

If *d* is a *continuous* variable and the decision rule g is known and invertible, then *d* and *Z* are sufficient to identify *v*. In such case, *v* is a deterministic function of (*d*,*Z*), making conditioning on *v* equivalent to conditioning on (*d*,*Z*), which is observable.[25] The regression equation would then be

$$\begin{aligned} \mathrm{E}\left[y_i|v_i\right] &= \beta + d_i\,\mathrm{E}\left[\alpha_i|v_i\right] + \mathrm{E}\left[u_i|v_i\right] \\ &= \beta + d_i\,\mathrm{E}\left[\alpha_i|Z_i,d_i\right] + \mathrm{E}\left[u_i|Z_i,d_i\right] \end{aligned}$$

where the remaining error

$$\begin{aligned} r_i &= \left(u_i - \mathrm{E}\left[u_i|v_i\right]\right) + d_i\left(\alpha_i - \mathrm{E}\left[\alpha_i|v_i\right]\right) \\ &= \left(u_i - \mathrm{E}\left[u_i|Z_i,d_i\right]\right) + d_i\left(\alpha_i - \mathrm{E}\left[\alpha_i|Z_i,d_i\right]\right) \end{aligned}$$

is mean independent of (*d*,*Z*). The above equations suggest a solution to the endogeneity of *d*, namely to explicitly include $\mathrm{E}\left[u_i|v_i\right] = \mathrm{E}\left[u_i|d_i,Z_i\right]$ in the the regression equation while simultaneously specifying α as a flexible function of *v* (or (*Z*,*d*)) to avoid further parametric assumptions.[26]

However, if *d* is discrete, and in particular if it is a dummy variable, all that can be identified under typical assumptions is a threshold for *v* as a function of *d* and *Z*. This is made clear from the parametric specification for the selection rule in (5), where all that can be inferred when the parameters γ are known is whether *v* is above or below $-Z\gamma$ depending on whether *d*=1 or *d*=0. The following additional index restriction and exogeneity of *Z* on the selection rule univocally identifies the participation region

(59) $d_i^* = 1\left(\mathrm{g}\left(Z_i\right) + v_i \geq 0\right)$ with $Z \perp v$.

Together with condition (58), assumption (59) ensures that the conditional expectations of $\left(u, \alpha\right)$ on (*Z*,*d*) are functions of the index only. Under assumptions (58) and (59), the regression equation is

$$\begin{aligned} \mathrm{E}\left[y_i \middle| d_i, Z_i\right] &= \beta + d_i \,\mathrm{E}\left[\alpha_i \middle| d_i = 1, Z_i\right] + \mathrm{E}\left[u_i \middle| d_i, Z_i\right] \\ &= \beta + d_i \alpha^{ATT}\left(Z_i\right) + d_i \,\mathrm{E}\left[u_i \middle| v_i > -\mathrm{g}\left(Z_i\right)\right] + \left(1 - d_i\right)\mathrm{E}\left[u_i \middle| v_i < -\mathrm{g}\left(Z_i\right)\right]. \end{aligned} \tag{60}$$

OLS can be used to estimate (60) if the terms $\mathrm{E}\left[u_i \middle| v_i > -\mathrm{g}\left(Z_i\right)\right]$ and $\mathrm{E}\left[u_i \middle| v_i < -\mathrm{g}\left(Z_i\right)\right]$ are known and $\alpha^{ATT}\left(Z\right)$ follows a flexible specification. These are functions of $\mathrm{g}\left(Z\right)$ alone and thus can be estimated if $\mathrm{g}\left(Z\right)$ is known.[27] Notice that to estimate the ATT we can weaken assumption (61) and write $u \perp \left(d, Z\right)\middle| v$.

Early applications of the control function approach use a parametric assumption on the joint distribution of the error terms, *u* and *v*, and a functional form assumption for the decision rule. The most commonly encountered set of assumptions impose joint normality and linearity. The selection model of outcomes becomes:

$$\begin{aligned} y_i &= \beta + \alpha_i d_i + u_i \\ d_i &= 1\left(Z_i\gamma + v_i \geq 0\right) \\ \left(u, v\right) &\sim N\left(0, \Sigma\right). \end{aligned}$$

The mean independence assumption (58) is sufficient to identify ATT in this context. Under this parametric specification, (58) can be re-written as

$$(62)\quad \begin{aligned} \mathrm{E}\left[u \middle| d=1, Z\right] &= \rho\lambda_1\left(Z\gamma\right) \\ \mathrm{E}\left[u \middle| d=0, Z\right] &= \rho\lambda_0\left(Z\gamma\right) \end{aligned}$$

where $\rho = \sigma_u \operatorname{corr}(u,v)$, σ_u is the standard error of *u*, and the control functions are (adopting the standardization $\sigma_v = 1$ where σ_v is the standard error of *v*):

$$\lambda_1\left(Z\gamma\right) = \frac{n\left(Z\gamma\right)}{\Phi\left(Z\gamma\right)} \quad \text{and} \quad \lambda_0\left(Z\gamma\right) = \frac{-n\left(Z\gamma\right)}{1-\Phi\left(Z\gamma\right)}$$

for *n* and Φ to stand for the standard normal pdf and cdf, respectively. Thus, joint normality implies that the conditional expectation of *u* on *d* and *Z* is a known function of the threshold, $Z\gamma$, that determines the assignment propensity: $\mathrm{P}\left(d_i = 1 \middle| Z_i\right) = \mathrm{P}\left(v_i \geq -Z_i\gamma \middle| Z_i\right)$.

This model is typically estimated using the Heckit procedure, Heckman (1976, 1979). This is a two-step estimator. The first step generates predictions of the control functions specified above from a regression of *d* on *Z*. The second step estimates the enlarged outcome equation by OLS:

$$(63)\quad y_i = \beta + d_i\alpha^{ATT}\left(Z_i\right) + d_i\rho\lambda_1\left(Z_i\hat{\gamma}\right) + \left(1-d_i\right)\rho\lambda_0\left(Z_i\hat{\gamma}\right) + r_i$$

where *r* is what remains of the error term in the outcome equation and is mean independent of *d*:

$$r_i = \left[u_i - \hat{\mathrm{E}}\left(u_i \middle| d_i, Z_i\right)\right] + d_i\left[\alpha_i - \hat{\mathrm{E}}\left(\alpha_i \middle| d_i = 1, Z_i\right)\right].$$

It is clear from the regression equation (63) that only the ATT can be identified when the impact of treatment is heterogeneous:

$\alpha^{ATT}\left(Z_i\right) = \alpha^{ATE}\left(Z_i\right) + \mathrm{E}\left[\alpha_i - \alpha^{ATE}\left(Z_i\right) \middle| d_i = 1, Z_i\right]$. Most empirical applications drop the dependence of the average effect α^{ATT} on *Z*. This may affect the consistency of the estimated

treatment effect parameter since $\mathrm{E}\left[\alpha_i - \mathrm{E}\left(\alpha_i \middle| d_i = 1\right) \middle| d_i = 1, Z_i\right]$ is generally different from zero.[28]

VIII.B Weaknesses of CF

The relative robustness of the classical parametric CF method comes from the structure it imposes on the selection process. This makes this approach particularly informative for the policy maker by allowing for selection on the unobservables and supporting the extrapolation of results to alternative policy scenarios. However, this same feature has been strongly criticized for being overly restrictive. A number of semi-parametric CF estimators have been proposed that deal (at least partially) with this problem (e.g. see the review by Powell 1994 and also Ahn and Powell 1993; Andrews and Schafgans 1998; Das, Newey and Vella 2003).

In its non-parametric setup, the CF approach has been shown to be equivalent to the LATE approach. We now turn to this comparison and notice that while such advances deal with the main criticism to CF, they also reduce the usefulness of the CF approach to inform about possible policy changes.

VIII.C. The link between CF and IV

There are two key assumptions underlying the selection model specified in the previous section: *(i)* the parametric assumption on the joint distribution of unobservables and *(ii)* the linear index assumption on the selection rule. As noted above, important recent developments have proposed new semi-parametric estimators that relax these assumptions. More recently, Vytlacil (2002) has shown that the LATE approach can be seen as an

application of a selection model. To see this, we first compare the two methods and then briefly discuss the equivalence result of Vytlacil.

Non-parametric CF relies on assumptions (57) and (59) when *Z* has a non-zero impact on participation to establish and estimate the regression equation as specified in (60), which we repeat here to highlight the differences to IV:

$$\mathrm{E}\left[y_i \middle| d_i, Z_i\right] = \beta + d_i \mathrm{E}\left[\alpha_i \middle| d_i = 1, Z_i\right] + \mathrm{E}\left[u_i \middle| d_i, Z_i\right].$$

The three CF assumptions can be equivalently written as

- *d* is a non-trivial function of *Z*;
- *Z* is independent of (u, α, v) and
- Index restriction: $d_i = 1\left(\mathrm{g}\left(Z_i\right) + v_i \geq 0\right)$.

In turn, the LATE approach is based on the following regression model

$$\mathrm{E}\left[y_i \middle| Z_i\right] = \beta + \mathrm{P}\left(d_i = 1 \middle| Z_i\right)\mathrm{E}\left[\alpha_i \middle| d_i = 1, Z_i\right]$$

based on the LATE assumptions discussed in section VI.B. We repeat them here:

- *d* is a non-trivial function of *Z*;
- *Z* is independent of (u, α, v) and
- Monotonicity assumption: as with LATE, let $d_i\left(Z\right)$ be the random variable representing individual *i* treatment status if drawing *Z*; the monotonicity assumption states that $d\left(Z^*\right) \geq d\left(Z^{**}\right)$ (or $d\left(Z^*\right) \leq d\left(Z^{**}\right)$ for all individuals.

Both sets of assumptions have been discussed before and so we skip any more comments on them here. Instead we notice that the first and second assumptions in each set are equivalent, but not the third one. The difference is that LATE does not impose additive separability in the selection model to identify the treatment effect. It does not require any functional form or distributional assumptions, instead relying on the general form for the

decision process as specified in (3)-(4) together with the monotonicity assumption. The additive separability of the unobservable term in the selection rule implies the monotonicity assumption of LATE since the decision process is based on a threshold rule: $g\left(Z^{*}\right)$ is either greater or smaller than $g\left(Z^{**}\right)$ and so everyone that participates under the lowest one will also participate under the highest one.

The reverse implication, however, is not necessarily true. However, the LATE assumptions are equivalent to the CF assumptions if taken all together. Vytlacil (2002) shows that under the LATE assumptions it is always possible to construct a selection model $\tilde{d}\left(Z\right)$ of the type described in the third CF assumption and such that $\tilde{d}\left(Z\right)=d\left(Z\right)$ almost everywhere. This means that under the LATE assumptions stated above, we can always find a selection model satisfying an index restriction that rationalizes the data at hand. This equivalence result shows that the LATE approach can be seen as an application of a non-parametric version of the CF method.

Also notice that the local IV method of Heckman and Vytlacil (1999 and 2001) discussed earlier withdraws the monotonicity assumption of LATE and is instead based on the additive separability of the selection rule, as in (59). Thus, it is explicitly a CF method.

VIII.D. Using CF to estimate the returns to education

Table 6 displays estimates of the ATT using the fully parametric CF approach applied to our running returns to education example. For the non-parametric CF estimates, we re-direct the reader to section VI.E, where local IV is discussed. The specification used in the estimation assumes that the outcome depends linearly on education and region. Education take-up is modeled as a probit linear on the covariates listed in column 1.

In this estimation exercise we consider two alternative scenarios depending on the availability of an advanced education subsidy. In both cases, the selection decision is based on the contemporaneous information that affect either the cost of education or its expected returns. This information includes family background (*z*), region of residence (*x*), ability (θ) and the unobservable in the decision process *v*. The selection process is more complex when a subsidy is available and starts earlier in the life-cycle, at the time the agent decides how much effort to put on the preparation for the test. In such case, the test score is also a relevant piece of information in deciding about education as it determines eligibility to the subsidy.

Rows 2 to 4 in the table display estimates based on the typically observable information using alternative, increasingly flexible specifications. While not directly relevant, adding the test score significantly reduces bias in the estimation of the ATT for an economy with no education subsidy (see rows 2 and 3, columns 3 and 4). The reason for this is that the test score is a good proxy to ability as it is determined by it and ability is the typically omitted regressor responsible for part of the selection process. The same is not true for an economy with a subsidy, where the test score is a relevant part of the selection information as it determines access to subsidized education. Moreover, adding ability itself and controlling for the form of the eligibility rule (rows 5 and 6) does not help in reducing bias.

Overall, the bias produced by the parametric CF estimator is not large but persists over the different specifications we tried, the reason being the adopted specifications. Having the correct specifications is important to ensure that the exogeneity requirements do indeed hold. But this is generally a difficult task, even in a simple model as the one discussed here. In this example, the individual decides to enroll in advanced education if (see equation (12)):

$$\left[\exp\left(\beta_0+\beta_1 x+\alpha_0+\alpha_1\theta\right)-\exp\left(\beta_0+\beta_1 x\right)\right]\mathrm{E}\left[\exp\left(u\right)|v\right]>\mathrm{c}\left(z,s\right)+v$$

As a consequence, the selection mechanism does not follow the index structure imposed in our parametric specification and includes an error term which is not normally distributed even

though the unobservables (*u*,*v*) follow a joint normal distribution. Mis-specification also plagues the log-earnings equation, with returns to education being modelled as constant instead of allowing for dependence on ability as in the true model, ($\alpha_0 + \alpha_1\theta$).

IX. Summary

This paper has presented an overview of alternative methods for the evaluation of policy interventions at the microeconomic level. The choice of appropriate evaluation method has been shown to depend on three central considerations: the policy parameter to be measured, the data available and the assignment mechanism by which individuals are allocated to the program or receive the policy. Through studying a combination of the econometric underpinnings and the actual implementation of each method we hope to have convinced the reader that no method dominates. Indeed the requirements placed on the design of any evaluation to fully justify the use of *any* of the standard evaluation methods are typically difficult to satisfy.

One key to the appropriate choice of method has been shown to be a clear understanding of the "assignment rule" or the mechanism by which assignment of individuals are allocated to the policy or program. In a sense this is a precursor to the choice of appropriate evaluation method. At one end of the spectrum, in a perfectly designed social experiment, assignment is random and at the other end of the spectrum, in a structural microeconomic model, assignment is assumed to obey some (hopefully plausible) model of economic choice. Perfect experimental designs and fully plausible structural allocation theories that command wide acceptability are rare. We have shown how alternative methods exploit different assumptions concerning assignment and differ according to the type of assumption made.

Unless there is a convincing case for the reliability of the assignment mechanism being used, the results of the evaluation are unlikely to convince the thoughtful skeptic. Just as an experiment needs to be carefully designed, a structural economic model needs to be convincingly argued.

We have also seen that knowledge of the assignment mechanism alone is not enough. Each method will have a set of possible evaluation parameters it can recover. That is, even if the arguments behind the assumed assignment rule is convincing, any particular method will typically only permit a limited set of policy questions to be answered. For example, we have seen that ex-ante evaluations that seek to measure the impact of policy proposals place inherently more stringent demands on the research design than ex-post measurements of existing policies. Similarly, measuring distributional impacts rather than simple average impacts typically also rests on stronger assumptions. Even where the randomization assumption of an experimental evaluation is satisfied and is fully adopted in implementation, the experiment can only recover a limited set of parameters. In the end any reasonable evaluation study is likely to adopt a number of approaches, some being more robust but recovering less while others answering more complex questions at the cost of more fragile assumptions.

Appendix 1

A simple dynamic model of investment in education

Consider an economy of heterogeneous individuals indexed by *i* facing lifetime earnings *y* that depend on the highest level of education achieved. We distinguish between two levels of education, low and high. The prototypical individual in this model lives for three periods, which we denote by age being zero, one or two. At age zero all individuals are in school. At age one some individuals will enrol in college and at age two all individuals are

working. The problem of the individual is to decide optimally about educational investment when there is uncertainty about the future returns to the investment. We will now explain the model in more detail.

At birth (age zero) each individual is characterized by three variables, which we denote by (θ, x, z). For interpretation purposes, we assume θ measures ability and is observable to the individual but unobservable to the econometrician. *z* is observable to the individual and econometrician and measures the conditions faced by the individual while young that affect the cost of education. It will be interpreted as some measure of family background or some measure os cost like distance to college. Finally, *x* is another observable variable to both the individual and econometrician. It measures market conditions and we interpret it as region. All three variables are assumed to remain unaltered throughout the individuals life.

Based on this information, the individual decides at age zero about the level of effort in school. Combined with ability θ, the endogenous effort will determine performance in school. This is measured as a score in a test and is denoted by *s*:

(64) $s_i = \gamma_0 + \gamma_1 \theta_i e_i + q_i$

where e is effort, q is the unpredictable component of the score and (γ_0, γ_1) are the parameters.

The test score is revealed in the next period, at age one, after the effort choice being made. Depending on its value, it may give access to subsidized education if such subsidy exists. Eligibility is defined on a threshold rule: students scoring above $\underline{s}$ will be eligible while students scoring below this level will not.

Investment in high education has a (utility) cost, denoted by *c*. *c* depends on the individual's characteristics as well as on the test score if an education subsidy is available. In the presence of a subsidy, *c* is defined as

(65) $c_i = \delta_0 + \delta_1 z_i - 1(s_i > \underline{s})S + v_i$

where *s* is the education subsidy available to eligible individuals, the function $1(A)$ is the characteristic function, assuming the value one if *A* is true and zero otherwise, *v* is the unpredictable part of the cost of education and (δ_0, δ_1) are the parameters.

The decision of whether or not to invest in education occurs at age one. The test score, *s*, and the unpredictable part of the cost of education, *v*, are revealed at the start of this period and used to inform the decision process. Thus, the precise cost of education is known at the time of deciding about the investment. What is not known with certainty at this stage is the return to education as it depends on an unpredictable component as viewed from age one. Only at age two is this uncertainty resolved, when the individual observes lifetime earnings. These are specified as:

(66) $\ln y_i = \beta_0 + \beta_1 x_i + (\alpha_0 + \alpha_1 \theta_i) d_i + u_i$

where *y* is earnings, *d* is a dummy variable representing the education decision, $(\alpha_0, \alpha_1, \beta_0, \beta_1)$ are the parameters of the earnings function and *u* is the unpredictable component of earnings. Notice that the returns to education are not known in advance, at age one, because *u* is unknown and *y* is nonlinear in its arguments.

We now formalize the problem of the individual in a dynamic framework. The individual chooses effort at age zero to maximize lifetime utility. His/her choice is conditional on how effort affects the test score (equation (64)) and the impact of the test score on the cost of education (equation (65)). It can be formalized as

(67) $\mathrm{V}_{0i}(\theta_i, z_i, x_i) = \max_{e_i} \left\{ -\lambda e_i + \rho \mathrm{E}_{s,v} \left[\mathrm{V}_{1i}(\theta_i, z_i, x_i, s_i, v_i) \right] \right\}$

where V_{ai} represents the discounted value of present and future utility for individual *i* when aged *a*, ρ is the discount factor and the index in the expectations operator lists the random variables at the moment of selecting effort, with respect to which the expected value is to be

computed. From the above equation the optimal level of effort is a function of θ, *z* and *x*, $e^{*}(\theta, z, x)$.

The problem of the individual at age one is that of choosing the educational level without knowing the returns to the investment with certainty. Conditional on the (known) form of the earnings equation(66), the problem can be formalized as

$$(68)\ \mathrm{V}_{1i}(\theta_i, z_i, x_i, s_i, v_i) = \max_{d_i}\left\{-c_i d_i + \rho \mathrm{E}_u\left[y_i(\theta_i, d_i, x_i, u_i)|v_i\right]\right\}$$

where we allow for *v* and *u* to be related and thus condition the expected value on *v*.

Under the model specification in equation (68), the education decision follows a reservation rule defined in the cost of education. The optimal decision is a function of the information set at age one, $d = d^{*}(\theta, z, x, s, v)$. Formally:

$$(69)\ d_i = \begin{cases} 1 & \text{if } \mathrm{E}\left[y_i|d_i=1, x_i, \theta_i, v_i\right] - \mathrm{E}\left[y_i|d_i=0, x_i, \theta_i, v_i\right] > c_i \\ 0 & \text{otherwise.} \end{cases}$$

Finally, at age two the individual works and collects lifetime earnings as defined in equation (66). There is no decision to be taken at this stage.

Average parameters

The impact of high education on the logarithm of earnings for individual *i* is $\alpha_i = \alpha_0 + \alpha_1 \theta_i$. We can use this expression to specify the ATE on log earnings as

$$\begin{aligned} \alpha^{ATE} &= \alpha_0 + \alpha_1 \mathrm{E}[\theta_i] \\ &= \alpha_0 + \alpha_1 \int_{\Theta} \theta \, \mathrm{f}_{\theta}(\theta) d\theta \end{aligned}$$

where $\mathrm{f}_{\theta}(\theta)$ is the probability density function of θ and Θ is the space of possible realizations or domain of θ.

In a similar way, the ATT on log earnings is just

$$\alpha^{ATE} = \alpha_0 + \alpha_1 \mathrm{E}\left[\theta_i | d_i = 1\right].$$

However, it is now more difficult to derive the exact expression $\mathrm{E}\left[\theta_i \middle| d_i = 1\right]$ as it depends on the endogenous individuals' choices. To do this, we will assume that *v* and *u* are not positively correlated, thus $\mathrm{corr}(u,v) \leq 0$. In particular, we take *u* to be a linear random function of *v*,

$$u_i = \mu v_i + r_i$$

where $\mu \leq 0$ is the slope parameter and *r* is a iid shock. In this case, the reservation policy described in equation (69) in terms of the cost of education *c* can now be expressed in terms of the unobservable component, *v*. We denote it by $\tilde{v}$ and note that it is a function of the variables known at age one that impact either on the cost of education or on the expected future earnings. Thus, $\tilde{v}(\theta, z, x, s)$ but since $s = \gamma_0 + \gamma_1 \theta e(\theta, z, x) + q$ it is equivalent to write it as $\tilde{v}(\theta, z, x, q)$. The reservation policy $\tilde{v}$ fully characterizes the educational decision: whenever the individual draws a shock $v > \tilde{v}$ the decision will be not to participate while the opposite happens when $v < \tilde{v}$. Thus, the decision rule (69) can be re-written as,

$$d_i = \begin{cases} 1 & \text{if } \mathrm{v}_i < \tilde{v}(\theta_i, z_i, x_i, q_i) \\ 0 & \text{otherwise.} \end{cases}$$

Conditional on the set of variables (θ, z, x, q), the size of the population investing in education will be given by,

$$\begin{aligned} \mathrm{P}\left[d = 1 \middle| \theta, z, x, q\right] &= \mathrm{F}_v\left(\tilde{v}(\theta, z, x, q)\right) \\ &= \int_{-\infty}^{\tilde{v}(\theta,z,x,q)} \mathrm{f}_v(v) dv \end{aligned}$$

which is just the cumulative density function of *v* at the reservation point, $\tilde{v}(\theta, z, x, q)$. Notice that to derive the above expression it is being assumed that the *v* is independent of (θ, z, x, q).

We can now integrate ability over the whole educated population to obtain $\mathrm{E}\left[\theta \middle| d = 1\right]$:

$$\mathrm{E}\left[\theta|d=1\right]=\int_{\Theta}\int_{D(z)}\int_{D(x)}\int_{-\infty}^{+\infty}\theta\,\mathrm{F}_{v}\left(\tilde{v}\left(\theta,z,x,q\right)\right)\mathrm{f}_{(\theta,z,x,q)}\left(\theta,z,x,q\right)dqdxdzd\theta$$

where $\left(\Theta,D(z),D(x)\right)$ stand for the domains of $\left(\theta,z,x\right)$ and $\mathrm{f}_{(\theta,z,x,q)}\left(\theta,z,x,q\right)$ is the joint density function of $\left(\theta,z,x,q\right)$.

Parameters used in the simulations

- Discount parameter: $\rho=1$;
- Utility cost of effort to prepare test: $\lambda=0.9$;
- Test score (equation (64)): $\gamma_0=1.0$, $\gamma_1=2.5$, $q\sim\mathrm{N}(0,1)$;
- Cost of education (equation (65)): $\delta_0=3.0$, $\delta_1=-1.2$, $\underline{s}=4.0$, $S=2.5$, $v\sim\mathrm{N}(0,1)$;
- Earnings (equation (66)): $\beta_0=0.7$, $\beta_1=0.3$, $\alpha_0=0.01$, $\alpha_1=0.7$, $u=\mu v+r$ where $\mu=-0.5$ in the correlated case and $\mu=0.0$ in the non-correlated case, $r\sim\mathrm{N}\left(0,\sigma^2=0.75\right)$ in the correlated case and $r\sim\mathrm{N}(0,1)$ in the non-correlated case;
- State variables: $\theta\sim\mathrm{N}\left(0.5,\sigma=0.25\right)$ truncated at zero and one, $z\sim\mathrm{N}(0,1)$ truncated at -2 and 2, x is a Bernoulli with p=0.4.

Appendix 2: STATA Datasets and .do files

The individual life-cycle model of education investment and earnings, described in Appendix 1, is used to construct simulated data under alternative assumptions about the policy environment and the nature of the decision process. The two main **STATA datasets**, MCdta-corr.dta and MCdta-nocorr.dta, contain 200 Monte-Carlo replications of datasets of 2000 simulated observations each. The first dataset represents the case of unobservable (to the econometrician) information related to expected gains used to inform individual decisions on education investment. The second dataset represents the case where all relevant information

used to decide about education is observable. Both datasets include life-cycle information under 3 alternative policy scenarios: unsubsidized advanced education, subsidized advanced education and subsidized advanced education when individuals unaware of the existence of a subsidy one period ahead of deciding about the investment.

Two additional auxiliary datasets are also available, required only for the DID estimator, MCdta-corr-noS.dta and MCdta-nocorr-noS.dta. These also contain 200 Monte-Carlo replications of datasets of 2000 observations each. They are used together with MCdta-corr.dta and MCdta-nocorr.dta, respectively, to construct repeated cross-sections of education and earnings information corresponding to periods before (MCdta-corr-noS.dta and MCdta-nocorr-noS.dta) and after (MCdta-corr.dta and MCdta-nocorr.dta) a policy change amounting to the introduction of a subsidy to advanced education. Both datasets contain information for an economy with no education subsidy only.

A set of **STATA .do files** was created to run alternative estimation procedures using each of the discussed methods and to produce the Monte-Carlo results. There are two .do-files for each method, labelled "name-of-the-method".do and "name-of-the-method"-programs.do. The former contains the the main routine, which defines the dataset and variables being used, calls the estimation routines and displays the results. The latter contains two main estimation routines (accompanied by other auxiliary routines in some cases). The first routine implements the respective estimator in a given dataset for a certain set of variables provided by the user. The second routine repeatedly applies the estimator procedure in the first one to a series of datasets to produce the Monte-Carlo results.

The .do-files can be used together with the datasets to reproduce all the results in the paper.

Appendix 3

Average treatment parameters

All the average parameters can be expressed as averages of the MTE using different weights. Consider the ATT. Participants at any point *p* of the distribution of $\tilde{v}$ are those that draw $\tilde{v} < p$. Thus,

$$\alpha^{ATT}(p) = \int_0^p \alpha^{MTE}(\tilde{v}) \mathrm{f}_{\tilde{v}}(\tilde{v} | \tilde{v} < p) d\tilde{v}$$
$$= \frac{1}{p} \int_0^p \alpha^{MTE}(\tilde{v}) d\tilde{v}$$

where the second equality results from the fact that $\tilde{v}$ is uniformly distributed. Integrating over all the support of *p* yields the overall ATT,

$$\alpha^{ATT} = \int_0^1 \alpha^{ATT}(p) \mathrm{f}_p(p | d = 1) dp$$
$$= \int_0^1 \int_0^p \alpha^{MTE}(\tilde{v}) \frac{\mathrm{f}_{p|d}(p | d = 1)}{p} d\tilde{v} dp.$$

Similarly, the ATE, ATNT\ and LATE are,

$$\alpha^{ATE} = \int_0^1 \int_0^1 \alpha^{MTE}(\tilde{v}) \mathrm{f}_p(p) d\tilde{v} dp$$
$$\alpha^{ATNT} = \int_0^1 \int_p^1 \alpha^{MTE}(\tilde{v}) \frac{\mathrm{f}_{p|d}(p | d = 0)}{1 - p} d\tilde{v} dp$$
$$\alpha^{LATE}(p^*, p^{**}) = \frac{1}{p^{**} - p^*} \int_{p^*}^{p^{**}} \alpha^{MTE}(\tilde{v}) d\tilde{v}.$$

References

Abadie, Alberto, Joshua Angrist, and Guido Imbens. 2002. "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings." *Econometrica* 70(1): 91-117.

Abbring, Jaap, and Gerard van den Berg. 2003. "The Nonparametric Identification of Treatment Effects in Duration Models." *Econometrica* 71(5): 1491-1517.

Ahn, Hyungtaik, and James Powell. 1993. "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism." *Journal of Econometrics* 58(1): 3-29.

Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics* 60(1): 47-57.

Ashenfelter, Orley, and David Card. 1985. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *Review of Economics and Statistics* 67(4): 648-660.

Athey, Susan, and Guido Imbens. 2006. "Identification and Inference in Nonlinear Difference-In-Differences Models." *Econometrica* 74(2): 431-497.

Bassi, Laurie. 1983. "The Effect of CETA on the Post-Program Earnings of Participants." *Journal of Human Resources* 18(4): 539-556.

Bassi, Laurie. 1984. "Estimating the Effects of Training Programs with Nonrandom Selection." *Review of Economics and Statistics* 66(1): 36-43.

Battistin, Eric, and Enrico Rettore. 2008. "Ineligibles and Eligible Non-Participants as a Double Comparison Group in Regression Discontinuity Designs." *Journal of Econometrics* forthcoming.

Becker, Sascha, and Andrea Ichino. 2002. "Estimation of average treatment effects based on propensity scores." *The Stata Journal* 2(4): 358-377.

Bell, Brian, Richard Blundell, and John Van Reenen. 1999. "Getting the Unemployed Back to Work: An Evaluation of the New Deal Proposals." *International Tax and Public Finance* 6(3): 339-360.

Bjorklund, Anders, and Robert Moffitt. 1987. "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models." *Review of Economics and Statistics* 69(1): 42-49.

Blundell, Richard, Monica Costa Dias, Costas Meghir, and John Van Reenen. 2004. "Evaluating the Employment Impact of a Mandatory Job Search Assistance Program." *Journal of the European Economics Association*, 2(4): 596-606.

Blundell, Richard, Lorraine Dearden, and Barbara Sianesi. 2005. "Evaluating the Impact of Education on Earnings: Models, Methods and Results from the NCDS." *Journal of the Royal Statistical Society* Series A, 168(3): 473-512.

Blundell, Richard, Alan Duncan, and Costas Meghir, C. 1998. "Estimating Labour Supply Responses using Tax Policy Reforms." *Econometrica* 66(4): 827-861.

Blundell, Richard, and Thomas MaCurdy. 1999. "Labor Supply: A Review of Alternative Approaches." In *Handbook of Labour Economics*, eds. Orley Ashenfelter, and David Card, Volume 3: 1559-1695. Amsterdam: Elsevier Science.

Blundell, Richard, and James Powell. 2003. "Endogeneity in Nonparametric and Semiparametric Regression Models." In *Advances in Economics and Econometrics*, eds. Mathias Dewatripont, Lars Hansen, and Stephen Turnsovsky, 294-311. Cambridge: Cambridge University Press.

_____. 2004. "Endogeneity in Semiparametric Binary Response Models." *The Review of Economic Studies* 71(3): 581-913.

Card, David, and Philip Robins, P. 1998. "Do Financial Incentives Encourage Welfare Recipients To Work?." *Research in Labor Economics* 17(1): 1-56.

Card, David, and Lara Shore-Sheppard. 2004. "Using Discontinuous Eligibility Rules to Identify the Effects of the Federal Medicaid Expansions on Low-Income Children." *The Review of Economics and Statistics* 86(3): 752-766.

Carneiro, Pedro, Karsten Hansen, and James Heckman. 2001. "Removing the Veil of Ignorance in Assessing the Distributional Impacts of Social Policy." *Swedish Economic Policy Review* 8(2): 273-301.

_____. 2003. "Estimating Distributions of Counterfactuals with an Application to the Returns to Schooling and Measurement of the Effect of Uncertainty on Schooling Choice." *International Economic Review* 44(2): 361—422.

Carneiro, Pedro, James Heckman, and Edward Vytlacil. 2005. "Understanding What Instrumental Variables Estimate: Estimating the Average and Marginal Return to Schooling." Mimeo, University of Chicago.

Carneiro, Pedro, and Simon Lee. 2007. "Changes in College Enrollment and Wage Inequality: Distinguishing Price and Composition." Mimeo, University College London.

Cochrane, William, and Donald Rubin. 1973. "Controlling Bias in Observational Studies." *Sankyha* Ser. A 35: 417-446.

Das, Mitali, Whitney Newey, and Francis Vella. 2003. "Nonparametric Estimation of Sample Selection Models." *Review of Economic Studies* 70(1): 33-58.

De Giorgi, Giacomo. 2005. "The New Deal for Young People Five Years On." *Fiscal Studies* 26(3): 371-383.

Feenberg, Daniel and James Poterba. 1993. "Income inequality and the incomes of very high income taxpayers: Evidence from tax returns." In *Tax Policy and the Economy*, eds. James Poterba, Volume 7: 145-77. Cambridge: MIT Press.

Feldstein, Martin. 1995. "The effect of marginal tax rates on taxable income: A panel study of the 1986 Tax Reform Act'." *Journal of Political Economy*. 103(3): 551-72.

Fisher, Ronald. 1951. *The Design of Experiments*, 6th edition. London: Oliver and Boyd.

Frölich Markus. 2006. "A Note on Parametric and Nonparametric Regression in the Presence of Endogenous Control Variables." IZA Discussion Paper 2126. Institute for the Study of Labor.

Goolsbee, Austan. 2000. "What happens when you tax the rich? Evidence from executive compensation." *Journal of Political Economy*. 108(2): 352-78

Hahn, Jinyong. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica* 66(2): 315-331.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69(1): 201-209.

Hausman, Jerry, and David Wise, eds. 1985. *Social Experimentation*. Chicago: University of Chicago Press.

Heckman, James. 1976. "The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such methods." *Annals of Economic and Social Measurement* 5, 475-492.

_____. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1): 153-611.

_____. 1997. "Instrumental Variables: A Study of the Implicit Assumptions underlying one Widely used Estimator for Program Evaluations." *Journal of Human Resources* 32(3): 441-462.

Heckman, James, Hideniko Ichimura, and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *The Review of Economic Studies* 64(4): 605-654.

_____. 1998. "Matching as an Econometric Evaluation Estimator." *The Review of Economic Studies* 65(2): 261-294.

Heckman, James, Hideniko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterising Selection Bias Using Experimental Data." *Econometrica* 66(5): 1017-1098.

Heckman, James, Robert LaLonde, and Jeffrey Smith. 1999. "The Economics and Econometrics of Active Labour Market Programs." In *Handbook of Labour*

Economics, eds. Orley Ashenfelter, and David Card, Volume 3A: 1865-2097. Amsterdam: Elsevier Science.

Heckman, James and Salvador Lozano. 2004. "Using Matching, Instrumental Variables and Control Functions to Estimate Economic Choice Models." *The Review of Economics and Statistics* 86(1): 30-57.

Heckman, James, and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." In *Longitudinal Analysis of Labour Market Data*, eds. James Heckman and Burton Singer, 156- 246. New York: Wiley.

_____. 1986. "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes." In *Drawing Inferences from Self-Selected Samples*, ed. Howard Wainer, 63-107. Berlin: Springer Verlag.

Heckman, James, and Jeffrey Smith. 1999. "The Pre-Program Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies." *Economic Journal* 109(457), 313-348.

Heckman, James, Jeffrey Smith, and Nancy Clements. 1997. "Making the Most out of Program Evaluations and Social Experiments: Accounting for Heterogeneity in program Impacts." *Review of Economic Studies* 64(4): 487-536.

Heckman, James, and Edward Vytlacil. 1998. "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return is Correlated with Schooling." *Journal of Human Resources* 33(4): 974-987.

_____. 1999. "Local Instrumental Variables and Latent Variable Models for Identifying the Bounding Treatment Effects." *Proceedings of the National Academy of Sciences* 96(8), 4730-4734.

_____. 2001. "Local Instrumental Variables." In *Nonlinear Statistical Modeling: Essays in Honor of Takeshi Amemiya*, eds. Cheng Hsiao, Kimio Morimune, and James Powell, 1-46. New York: Cambridge University Press.

_____. 2006. "Econometric Evaluation of Social Programs, Part II." In *Handbook of Econometrics*, eds. James Heckman and Edward Leamer, Volume 6: 4875-5134, Amsterdam: Elsevier.

Horowitz, Joel. 2001. "The Bootstrap." In *Handbook of Econometrics*, eds. James Heckman and Edward Leamer, Volume 5: 3159-3228. Amsterdam: Elsevier.

Imbens, Guido, and Joshua Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2), 467-75.

Imbens, Guido and Thomas Lemieux. 2007. "Regression discontinuity designs: A guide to practice." Journal of Econometrics. Forthcoming.

Klaauw, Wilbert. 2008. "Regression-Discontinuity Analysis: A Survey of Recent Developments in Economics." *Review of Labour Economics and Industrial Relations* 22 (2): 219-45.

LaLonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76(4): 604-620

Larsson, Laura. 2003. "Evaluation of Swedish Youth Labor Market Programs." *Journal of Human Resources* 38(4): 891—927.

Lee, David. 2008. "Randomized Experiments from Non-random Selection in U.S. House Elections." *Journal of Econometrics* 142(2): 675-697.

Leuven, Edwin, and Barbara Sianesi. 2003. *PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing.* Statistical Software Components S432001, Boston College Department of Economics.

Lindsey, Lawrence. 1987. "Individual taxpayer response to tax cuts: 1982-1984 with implications for the revenue maximizing tax rate." *Journal of Public Economics* 33(1): 173-206.

Moffitt, Robert. 2007. "Estimating Marginal Treatment Effects in Heterogeneous Populations" , NBER Working Paper 13534.

Newey, Whitney, James Powell and Francis Vella. 1999. "Nonparametric Estimation of Triangular Simultaneous Equations Models." *Econometrica* 67(3): 565-603.

Nickell, Stephen. 1981. "Biases in Dynamic Models with Fixed Effects." *Econometrica* 49(6): 1417-1426.

Powell, James. 1994. "Estimation of Semiparametric Models." In *Handbook of Econometrics,* eds. Robert Engle, and Daniel McFadden, Volume 4: 2443-2521. Amsterdam: North Holland.

Rosenbaum, Paul, and Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41-55.

_____. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79(387): 516-524.

_____. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score." *American Statistician* 39(1): 33-38.

Rubin, Donald. 1979. "Using Multivariate Matched Sampling and regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association* 74(366): 318-329.

Ten Have, Thomas, Michael Elliott, Marshall Joffe, Elaine Zanutto, and Catherine Datto. 2004. "Causal Models for Randomized Physician Encouragement Trials in Treating

Primary Care Depression". *Journal of the American Statistical Association* 99(1): 16-25.

Vytlacil, Edward. 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica* 70(1): 331-341.

Table 1

Monte-Carlo experiment – true effects and selection mechanism

		Correlation between u and v	
		Negative	Zero
		(1)	(2)
(1)	ATE	0.354	0.354
(2)	ATT	0.453	0.471
(3)	Proportion investing in education	0.286	0.344
	Proportion eligible to subsidy:		
(4)	Whole population	0.165	0.276
(5)	Among educated	0.346	0.649
	Mean ability (θ):		
(6)	Whole population	0.492	0.492
(7)	Among educated	0.632	0.658
	Mean family background (z):		
(8)	Whole population	0.039	0.039
(9)	Among educated	0.559	0.582
	Mean region (x):		
(10)	Whole population	0.396	0.396
(11)	Among educated	0.452	0.447

Notes: Results from simulated data using the true individual effects and all the 200 Monte-Carlo replications to represent the population. In whole, results are based on 400,000 simulated individuals. Simulations for economy with subsidized advanced education and individuals totally aware of its availability and eligibility rules at birth.

Table 2

Monte-Carlo experiment – DID estimates and bias

		Expected policy			Unexpected policy		
		True parameter (1)	estimate (2)	bias (3)	True parameter (4)	estimate (5)	bias (6)
	Regional variation: region one is pilot						
(1)	Uncorrected estimates	0.035	0.037	0.044	0.011	0.013	0.133
(2)	Corrected estimates	0.520	0.519	0.002	0.486	0.187	0.615
	Eligibility variation: eligibles score above $\underline{s}$ in test						
(3)	Uncorrected estimates	0.203	0.328	0.612	0.202	0.202	0.000
(4)	Corrected estimates	0.516	0.617	0.196	0.487	0.470	0.035

Notes: Simulated data based on 200 Monte-Carlo replications of 2000 observations each. Simulations for economy with subsidized advanced education. Estimates in columns 1 to 3 (4 to 6) based on the assumption that the post policy generation is fully (not) aware of the availability of the subsidy and eligibility conditions at birth. No time trends were included in the simulations. Results respect to model with negatively related unobservables, u and v. Estimates in rows 1 and 2 use region to construct treatment and control groups for when the policy is piloted ahead of the national roll-out; estimates in rows 3 and 4 use eligibility status to construct treatment and control groups for when all regions introduce the policy simultaneously. Uncorrected DID estimates in rows 1 and 3 are standard DID estimates. Corrected DID estimates in rows 2 and 4 are re-scaled estimates to account for non-compliance (see equation (23)). Bias measured in relative terms as compared to the true parameter (columns 3 and 6).

Table 3

Monte Carlo experiment – Matching estimates of the ATT, ATNT and bias

		corr(u,v)<0		corr(u,v)=0	
		Coefficient (1)	Bias (2)	Coefficient (3)	Bias (4)
	Panel A: ATT				
(1)	True ATT	0.453		0.471	
	Matching estimates based on the covariates:				
(2)	x	0.971	1.145	0.469	0.005
(3)	x,z,s,θ	1.284	1.836	0.506	0.074
(4)	x,s,θ	1.319	1.914	0.567	0.203
	Panel B: ATNT				
(5)	True ATNT	0.293		0.315	
	Matching estimates based on the covariates:				
(6)	x	0.971	2.081	0.469	0.597
(7)	x,θ	1.011	2.209	0.287	0.021
(8)	x,z,s	1.265	3.014	0.358	0.221

Notes: Simulated data based on 200 Monte-Carlo replications of 2000 observations each. Simulations for economy with subsidized advanced education and individuals totally aware of its availability and eligibility rules at birth. Columns 1 and 2 (3 and 4) display results under the assumption of negatively related (independent) residuals u and v. Numbers in rows 1 and 5 are the true parameters while numbers in rows 2 to 4 (6 to 8) are matching estimates of the ATT (ATNT) using alternative sets of conditioning variables. Matching based on the propensity score kernel method with Epanechnikov weights and a bandwidth of 0.05. Bias measured in relative terms as compared to the true ATT or ATNT. x, z, s and θ represent region, family background, test score and innate ability, respectively.

Table 4

Monte Carlo experiment – IV and LIV estimates of the ATT and bias

		corr(u,v)<0		corr(u,v)=0	
		Coefficient (1)	Bias (2)	Coefficient (3)	Bias (4)
(1)	True parameters - ATT	0.453		0.471	
	Classical IV using as instruments:				
(2)	z (family background)	0.418	0.039	0.404	0.120
(3)	s (test score)	0.583	0.343	0.537	0.170
	Local IV using as instruments:				
(4)	z (family background)	0.384	0.152	0.484	0.028
(5)	s (test score)	0.382	0.157	0.401	0.147

Notes: Simulated data based on 200 Monte-Carlo replications of 2000 observations each. Simulations for economy with subsidized advanced education and individuals totally aware of its availability and eligibility rules at birth. Columns 1 and 2 (3 and 4) display results under the assumption of negatively related (independent) residuals u and v. Estimates in rows 2 and 3 (4 and 5) obtained using classical (local) IV with the detailed instruments. Local IV estimates the marginal treatment effect (MTE) over the support of the propensity score based on a local quadratic regression using Epanechnikov kernel weights and a bandwidth of 0.4. Bias measured in relative terms as compared to the true ATT (columns 2 and 4).

Table 5

Monte Carlo esperiment – DD estimates of the local ATE and bias

Kernel weights	True Local ATE (1)	Bandwidth=0.5 coefficient (2)	Bandwidth=0.5 bias (3)	Bandwidth=1.0 coefficient (4)	Bandwidth=1.0 bias (5)	Bandwidth=1.5 coefficient (6)	Bandwidth=1.5 bias (7)
(1) Epanechnikov	0.469	0.306	0.347	0.457	0.025	0.500	0.065
(2) Gaussian		0.477	0.018	0.524	0.117	0.542	0.156

Notes: Simulated data based on 200 Monte-Carlo replications of 2000 observations each. Simulations for economy with subsidized advanced education and individuals totally aware of its availability and eligibility rules at birth. Results refer to negatively related unobservables, *u* and *v*. Estimation of the DD parameter at the eligibility cutoff point (score *s*=4) based on local linear regression using Epanechnikov (row 1) or Gaussian (row 2) kernel weights. Estimates based on alternative values for the bandwidth ranging from 0.5 (columns 2 and 3) to 1.5 (columns 6 and 7). The true local ATE represents the impact of education for agents scoring around the threshold, between 3.99 and 4.01. Bias measured in relative terms as compared to the true local ATE (columns 3, 5 and 7).

Table 6

Monte Carlo experiment – parametric CF estimates of the ATT and bias

		Positive subsidy		No subsidy	
		Coefficient (1)	Bias (2)	Coefficient (3)	Bias (4)
(1)	True parameter - ATT	0.453		0.434	
	CF estimates using the selection variables:				
(2)	(x, z)	0.384	0.152	0.377	0.132
(3)	(x, z, s)	0.519	0.146	0.441	0.017
(4)	(x, z, s)+interactions	0.528	0.166	0.461	0.062
(5)	(x, z, s, θ)+interactions	0.554	0.223	0.569	0.310
(6)	$(x, z, d_{s \geq \underline{s}}, \theta)$+interactions	0.555	0.227	-	-

Notes: Simulated data based on 200 Monte-Carlo replications of 2000 observations each. Simulations for economy with subsidized advanced education and individuals totally aware of its availability and eligibility rules at birth. Results refer to negatively related unobservables, *u* and *v*. Estimates based on the parametric CF approach under the assumption of joint normality of the residuals (Heckit estimator). Variables in the outcomes' equation are education and region. Selection into education is modeled as a linear index model with explanatory variables as described in the first row of the table. *z*, *x*, *s*, $d_{s \geq \underline{s}}$ and θ stand for family background, region, test score, test score above $\underline{s}(= 4)$ (eligible to subsidised advanced education) and ability, respectively. The interactions include second order terms of each continuous variable and the product of each combination of two different variables. Bias in relative terms as compared with ATT (columns 2 and 4).

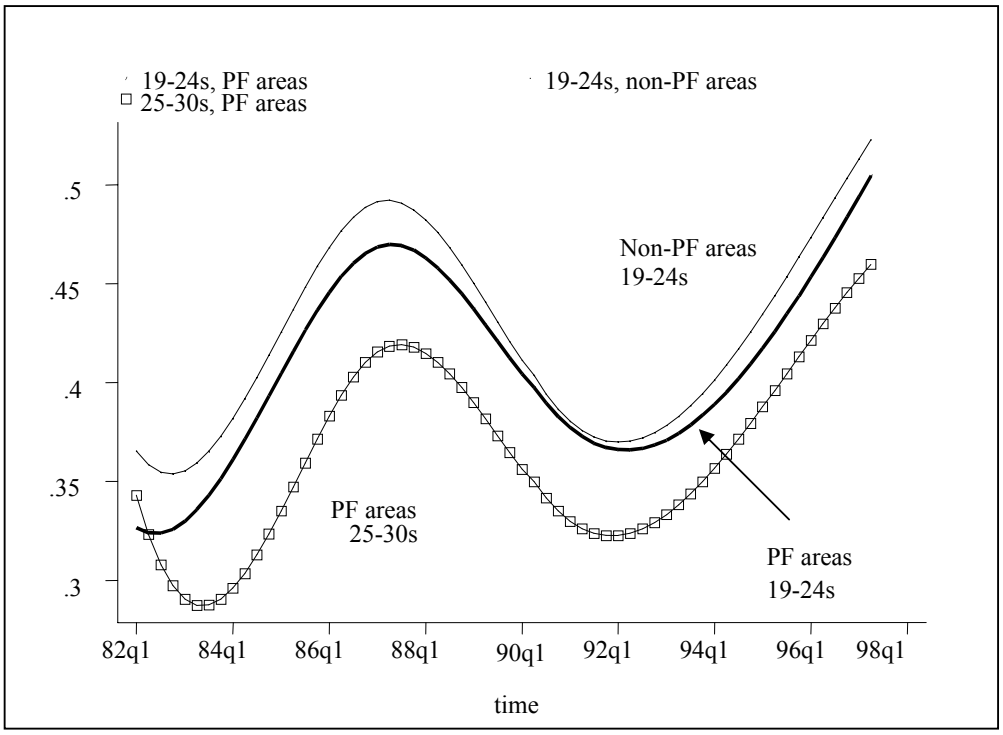

Figure 1

Effect of the NDYP by the end of the 10th month on JSA

Notes: PF stands for "Pathfinder" or "Pilot" areas. Figure plots the probability of leaving unemployment claimant count by age and region of residence. From Blundell et al. (2004).

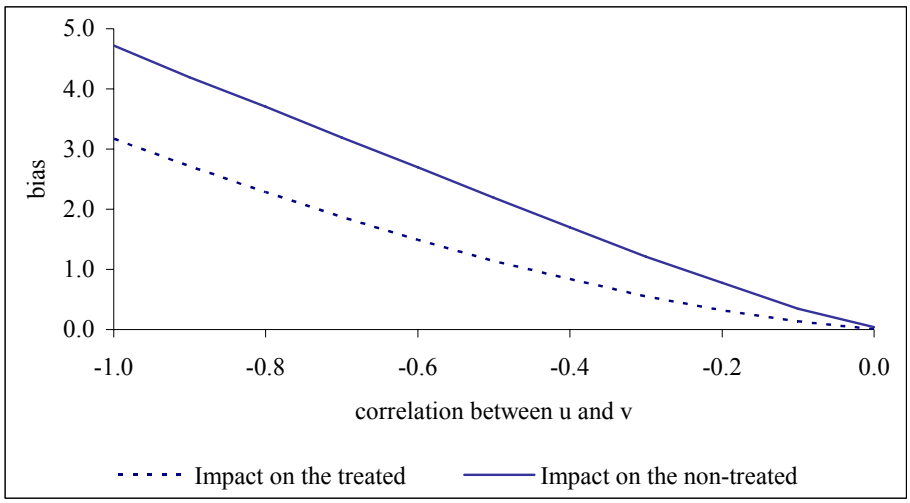

Figure 2

Relative bias in the matching estimator by level of correlation between the unobservables u and v

Note: Bias is measured in relative terms, as a percentage of the true treatment effect.

1 The examination of fully specified structural evaluation models is beyond the scope of this review but for many important ex-ante policy evaluations they are the dominant approach; see Blundell and MaCurdy (1999) for some examples in the evaluation of tax and welfare policy proposals.

2 In the labor market area, from which we draw heavily in this review, the ground breaking papers were those by Ashenfelter (1978), Ashenfelter and Card (1985) and Heckman and Robb (1985, 1986).

3 See, for example, Carneiro, Hansen and Heckman (2001, 2003) for a discussion of the distribution of treatment effects.

4 One could always think of controlling for z in the OLS regression if u and z are not mean independent. This is the motivation of the matching method, which will be discussed later in this paper.

5 For a dichotomous treatment in an experimental setting with identical composition of the treated and controls, the OLS estimator is the difference in the outcomes of treated and controls after the treatment period.

6 More precisely, we are assuming the transitory shocks, *o*, are *iid* continuous random variables with a strictly increasing cumulative density function, F, which is assumed known.

7 This is generally required for non-linear discrete choice models (see Nickell 1981).

8 Notice that α^{ATT} is not the ATT since $\mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^{1}\middle|d_i=1,t_1\right]\right)$ is generally different from the average index for this group and time period (which is $\beta+\alpha^{ATT}+n_1+m_{t_1}$) given the non-linearity of F^{-1} and the heterogenous nature of the treatment effect. To see why notice that,

$$\mathrm{E}\left[y_{it}^{1}\middle|d_i=1,t_1\right]=\int_{D(\alpha)}\mathrm{F}\left(\beta+\alpha+n_1+m_{t_1}\right)d\,\mathrm{G}_{\alpha|d}\left(\alpha\middle|d_i=1\right)$$

where $D(\alpha)$ is the space of possible treatment effects, α, and $\mathrm{G}_{\alpha|d}$ is the cumulative distribution function of α among individuals in treatment group *d*. Applying the inverse transformation yields,

$$\begin{aligned}\mathrm{F}^{-1}\left(\mathrm{E}\left[y_{it}^{1}\middle|d_i=1,t_1\right]\right)&=\mathrm{F}^{-1}\left[\int_{D(\alpha)}\mathrm{F}\left(\beta+\alpha+n_1+m_{t_1}\right)d\,\mathrm{G}_{\alpha|d}\left(\alpha\middle|d_i=1\right)\right]\\&\neq\int_{D(\alpha)}\mathrm{F}^{-1}\left[\mathrm{F}\left(\beta+\alpha+n_1+m_{t_1}\right)\right]d\,\mathrm{G}_{\alpha|d}\left(\alpha\middle|d_i=1\right).\end{aligned}$$

However, it can be used to recover the ATT as exposed in the main text.

9 An extension to the discrete case is also considered by the authors.

10 We discuss this type of monotonicity assumption in more detail later on, along with the LATE parameter.

11 Notice that no other factors differentially affecting education investments of cohorts 0 and 1 are considered.

12 In the presence of non-compliance, this result also applies when other criterion is used define the treatment groups for as long as the treatment status is well defined before and after the policy change.

13 Similar results were obtained for the case where the unobservables are independent and are available from the authors under request.

14 More recently, a study by Hahn (1998) shows that P(X) is ancillary for the estimation of ATE. However, it is also shown that knowledge of P(X) may improve the efficiency of the estimates of ATT, its value lying on the "dimension reduction" feature.

15 For a discussion of non-parametric matching estimators including Kernel and local linear regression methods see, Heckman, Ichimura and Todd (1997).

16 As with the DID estimator, our ability to correctly separate treated from non-treated at t_0 is determinant for the quality of the estimates.

17 A slightly stronger version of assumptions (43)-(44) is frequently imposed: $(u,\alpha,v) \perp z$ or, which is the same $(y^0, y^1, v) \perp z$.

18 Abadie, Angrist and Imbens (2002) extend this approach to the evaluation of *quantile treatment effects*. The goal is to assess how different parts of the outcome's distribution are affected by the policy. As with LATE, a local IV procedure is used, making the estimated impacts representative only for the sub-population of individuals changing their treatment status in response to the particular change in the instrument being considered.

19 However, in a recent study Ten Have et al. (2004) used a medical experiment to study the presence of defiers (individuals that behave against the monotonicity rule) and the consequences of assumptions about this group on estimated effects. In the context of their study they show evidence that defiers may exist and that estimated effects are sensitive to the assumptions about this group of individuals even with only a small number of defiers.

20 Not observing the top of the distribution of *P(z)* does not affect the identification of ATT since agents with $\tilde{v} > \overline{p}$ will never participate for the range of *P(z)* observable. They are always non-participants and for as long as this is also true in the population (as it happens to be the case in our example) they just do not belong to the population of interest for the evaluation of the ATT.

21 For an insightful discussion of DD see Hahn, Todd and Van der Klaauw (2001); more recently, Imbens and Lemieux (2007), provide a detailed discussion of DD together with implementation issues.

22 The possibility that individuals adjust *z* in response to the eligibility criteria in the intent of changing participation status is ruled-out from the DD analysis.

23 See Larsson (2003) for further details on this program.

24 If one was to relax the DD independence assumption (54) for fear of local selection on idiosyncratic gains and replace it for a monotonicity assumption like the LATE condition (46), then it can be showed that additional independence assumptions would be required to establish (55). These additional independence conditions are of a similar sort of those required by LATE (see condition (44)). At the local level, they exclude the possibility of endogeneity of *z* in the selection rule and limit the information content of *z* in explaining gains from treatment to what it reveals about the location of the unobservables in the decision rule together with the observed participation status. For further details see Hahn, Todd and van der Klaauw (2001).

25 For a continuous *d* and invertible decision rule, *v* can be obtained as $v = \mathrm{g}^{-1}\left(d, Z\right)$.

26 Nonparametric estimators of models with continuous *d* (triangular systems of equations) have been proposed in the literature and are simple to implement. See, for example, Newey, Powell and Vella (1999).

27 Notice that identification of these functions usually requires at least one variable in *Z* to be excluded for the outcomes equation (this is part of assumption (57) as discussed earlier). This means that the excluded *z* affects untreated outcomes through the information it contains about *v* when combined with the observed participation status. One exception to this rule is that of a fully parametric model where the decision rule is sufficiently non-linear to separate variation if $\mathrm{E}\left[u \middle| d, Z\right]$ from that on the regressors in the outcomes equation. Although the functional form may identify $\mathrm{E}\left[u \middle| d, Z\right]$, it is strongly advised to avoid relying on this source of identification alone.

28 An additional joint normality assumption between the idiosyncratic gains α_i and the unobservable *v* would further allow for the identification of the ATE but this assumption is not required to estimate the ATT and is not usually imposed.