

Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment*

Carlos A. Flores[†] Alfonso Flores-Lagunes[‡]

Preliminary Draft: January 18, 2007

Abstract

Estimation of the effect of a treatment or intervention on a given outcome is an important topic in economics and many other sciences. Sometimes researchers want to go a step further and try to understand the mechanisms through which the given treatment affects the outcome. Although some recent applications in economics have looked at this question, their parameters of interest are usually not clearly defined and the assumptions needed for a causal interpretation are rarely stated explicitly. In this paper we fill this void by considering identification and estimation of causal mechanisms of a treatment and causal effects of a treatment net of such mechanisms. We provide precise definitions of our parameters of interest using the potential outcomes framework and motivate their usefulness for policy. Our parameters result in an intuitive decomposition of the total effect of a treatment on an outcome into the causal mechanism and the causal effect net of that mechanism. We then investigate conditions under which these causal effects are identified and can be estimated. We start by assuming a randomly assigned treatment, and then consider the case in which selection into the treatment is assumed to be random conditional on a set of covariates. We also state conditions under which the common approach of directly controlling for the observed value of the post-treatment or mechanism variable can be interpreted as estimating a causal average net treatment effect. We close with two empirical applications that illustrate the concepts and methods discussed in this paper.

Key words and phrases: causal inference, post-treatment variables, principal stratification.

JEL classification: C13

*We want to thank Kei Hirano, Hilary Hoynes and participants at the 2006 annual meeting of the Society of Labor Economists, the University of Miami Labor Lunch, 2006 Midwest Econometrics Group Meeting, and 2006 Latin American Meetings of the Econometric Society for useful comments and discussions. All errors are our own.

[†]Department of Economics, University of Miami. caflores@miami.edu

[‡]Department of Economics, University of Arizona. alfonso@eller.arizona.edu

1 Introduction

A central topic in economics and many other sciences is the estimation of causal effects from treatments or interventions. Estimation of causal effects is important for understanding the effects of policy interventions and to better shape the development of future policies and programs. Perhaps the main framework to analyze causal relationships is the so-called potential outcomes framework (Rubin, 1974; Holland, 1986). The basic idea of such framework is to explicitly consider the potential outcomes that would be observed if it were possible to apply each treatment to each of the units simultaneously. The potential outcomes framework allows the researcher to clearly define the parameters of interest and state the necessary assumptions for identification of causal effects. Many of the estimation methods for causal effects currently used in economics and other sciences use the potential outcomes framework to define estimands of interest and develop estimation strategies.¹

Typically, the main purpose in the estimation of causal effects of a treatment or intervention on a particular outcome is to estimate its total impact, that is, the causal effect of reception of the treatment on the outcome. The average treatment effect is the most common parameter in the literature and has first-order relevance from a policy perspective. In addition, it is often of interest to estimate a causal mechanism through which the treatment or intervention works, and a causal effect of the treatment on the outcome *net* of this mechanism. Knowledge of these two causal concepts allows a better understanding of the treatment and, as a result, can be used for policy purposes in the design, development, and evaluation of interventions. This paper analyzes the identification and estimation of the average causal mechanism through which a treatment or intervention affects an outcome of interest and the average causal effect of the treatment net of this mechanism. Based on the potential outcomes framework, we precisely define our estimands of interest, consider different assumptions that can be employed in their identification and estimation, and analyze other related parameters mentioned elsewhere in the literature.

To motivate the importance of understanding the process through which a treatment works, briefly consider the following two empirical applications that are employed later as illustration of the methods developed in the paper. The first application is a social experiment to evaluate the effect of a job training program on earnings, in which individuals are

¹For an extensive review of estimation methods in the context of the evaluation of active labor market programs, see Heckman, LaLonde and Smith (1999) and Imbens (2004).

randomized into treatment and control groups for the reception of training. In this random assignment example, we consider the so-called lock-in effect of a training program on earnings as a causal mechanism of the treatment, i.e., the labor market experience lost due to the time spent in a training program. In this case we want to know what part of the causal effect of training on earnings is lost because of lost experience. If the lock-in effect of the training program is large, thereby resulting in a large negative causal effect on subsequent earnings, then a policy implication is to consider changing the way in which the program is administered to shorten the time it takes to complete the training or to offer the training while on the job.

The second empirical application uses observational data to consider the effect of smoking during pregnancy on the birth weight. The causal mechanism we consider is gestation time. In this case the goal is to determine what part of the causal effect of smoking during pregnancy on birth weight works through a shorter gestation. If it were determined that the causal effect of smoking during pregnancy on birth weight works mainly through a shorter gestation time (as opposed to working through a low intrauterine growth), then medical procedures that help delay birth may be determined to be helpful. In each of these two examples interest lies in determining the importance of a causal mechanism through which a particular treatment works.

Recently, there has been some research in the economics literature that looks at the effects of a treatment on an outcome net of one or more mechanisms. For example, Dearden et al. (2002) and Black and Smith (2004) look at the effect of school quality on labor market outcomes. Both studies point out that part of the effect of school quality on labor market outcomes may work through increasing years of schooling. Dearden et al. (2002) present results of the effect of school quality on wages with and without controlling for schooling in order to measure the total impact of school quality on wages as well as the effect that works through higher educational attainment. Similarly, Black and Smith (2004) use propensity matching methods with and without including years of education in the propensity score specification. In another paper, Simonsen and Skipper (2006) estimate the effect of motherhood on wages in Denmark. They study the total effect of having children on wages, and they take into account various mechanisms through which motherhood may affect wages. For instance, they consider as a possible channel the sector of employment because, as they point out, Denmark's public sector is known to have higher benefits regarding maternity

leave and more flexible working conditions than the private sector. Other channels they consider are working experience and occupation. They also analyze the effects of motherhood on wages for specific values of the mechanism variable; in particular, they estimate sector-specific (private versus public) treatment effects. They use a propensity matching approach and discuss assumptions needed to estimate the total effect of motherhood on wages, the effect of motherhood on wages net of the mechanisms and the sector-specific treatment effects previously mentioned. As a final example, Ehrenberg et al. (2006) look at the channels or mechanisms through which the Andrew W. Mellon Foundation's Graduate Education Initiative (GEI) affected the attrition and graduation probabilities of PhD students in various academic departments during the 1990s.²

The previous examples show the importance of estimating total effects as well as effects net of one or more mechanisms. Unfortunately, a common problem in this literature is that many times the parameters to be estimated are not clearly defined or are defined within the context of the estimation procedure used (e.g., OLS, matching) and, most importantly, the assumptions needed for a causal interpretation are not always made explicit.³ In contrast, this paper uses a potential outcomes framework (also known as Rubin's causal model) to clearly define our parameters of interest and then state assumptions under which they can be estimated. An advantage of this approach is that it allows us to separate the definition of the estimands from the techniques used to estimate them.

In order to give a causal interpretation to our parameters of interest, we use the concept of principal stratification introduced in Frangakis and Rubin (2002) for estimating treatment effects controlling for a post-treatment variable. In the potential outcomes framework, a causal effect must be a comparison of potential outcomes for the same group of individuals under treatment and control. The basic idea behind Frangakis and Rubin (2002) is to define this same group of individuals based also on the potential values of the post-treatment variable. As stressed in Rubin (2005), when drawing causal inferences is very important to keep the distinction between observed values of a variable and the potential values they represent.

²In 1991 the Andrew W. Mellon Foundation launched the GEI to improve the structure and organization of the PhD programs in the humanities and social sciences. Some of the channels considered by Ehrenberg et al. (2006) are more financial support to graduate students, more course and seminar requirements, higher quality advising, among others.

³An exception is Simonsen and Skipper (2006) who separate the definition of their parameters of interest from their estimation strategy. However, as we will discuss later in the paper, they did not explicitly consider the problem that the observed value of the mechanism variable represents two different potential variables depending on the treatment received; and they did not make explicit the assumption of a constant average net treatment effect.

In the examples from the literature mentioned above the distinction between observed values of the post-treatment variable or mechanism (e.g., observed years of schooling) and which potential values they reflect (e.g., schooling if in a high quality school versus schooling if in a low quality school) is absent. Therefore, comparing treated and control individuals on the basis of observed values of the post-treatment variable may yield misleading conclusions, as illustrated in Mealli and Rubin (2003), Rubin (2004), Rubin (2005), and further discussed later in this paper.

The definition of our parameters intuitively decompose the average (total) treatment effect into the average causal mechanism and the average causal effect net of that mechanism. A key insight for our decomposition is to think about the following ideal situation. Suppose we are interested on the effect of a randomly assigned treatment T on an outcome Y , and want to learn what part of that effect is through a mechanism S . Ideally, we would perform a new experiment in which the new (counterfactual) treatment is the same as the original one but blocks the effect of T on S , or, in other words, sets the value of the mechanism variable S at the level it would have been if this individual were a control under the original treatment. Under this new experiment, it would be straightforward to look at the effect of T on Y net of the mechanism S by comparing the average outcomes of the individuals that took this new treatment and the ones in the control group. Therefore, intuitively, what one is trying to do when estimating treatment effects net of a mechanism is to learn about a different treatment from the one we have at hand. This motivates the difficulty in estimating treatment effects net of a mechanism since, unfortunately, the commonly available data may provide little information about this alternative experiment. Given the usual trade-offs between data availability and assumptions, it is not surprising that for estimation of causal net treatment effects one needs to make stronger-than-usual assumptions and evaluate their plausibility in any particular application. It is important to point out that this is not a problem of the approach presented in the paper, but it only reflects the difficulty of answering the question we are asking with the available data.

We present two different approaches for estimation of our parameters of interest and explicitly state the assumptions needed under each approach. The first approach is based on estimation of the causal average net effect for a particular subpopulation: those individuals for which the treatment does not affect the mechanism variable. The second approach is based on the imposition of functional form assumptions relating the potential outcomes

of interest. We present each of these approaches for the case in which the treatment is randomly assigned and for the case in which selection into the treatment is based on a set of observable covariates. Finally, we discuss a set of assumptions within our framework under which the usual approach of controlling for the observed value of the mechanism variable can be interpreted as an average causal net effect.

The paper is organized as follows. Section 2 reviews some related literature. Section 3 presents the general framework and defines our parameters of interest. Section 4 analyzes the identification and estimation of our parameters. Section 5 presents the results from two empirical applications that illustrate the methods discussed in this paper. Concluding remarks are provided in the last section of the paper.

2 Review of Relevant Literature

Our goal is to analyze the estimation of two related effects: a *causal mechanism* through which a treatment or intervention affects an outcome of interest, and the *causal* effect of the treatment *net* of this causal mechanism. To achieve this goal we employ and build on an extensive literature on estimation of causal effects using the potential outcomes framework, and more specifically, on literature related to the estimation of causal effects when adjusting for covariates that are affected by the treatment. The latter literature relates to our goal since estimating the causal mechanism through which a treatment affects an outcome implies looking at variables that are observed after the treatment and that are affected by it.⁴

One of the earlier papers that touch on the topic of adjusting for covariates that are affected by a treatment is Rosenbaum (1984). This paper specifies conditions under which controlling for variables affected by the treatment yields the total average treatment effect (*ATE*), reaching the general conclusion that estimators adjusting for such variables are generally biased for the *ATE*. Trivially, the only instance when a causal *ATE* can be identified is when the concomitant variable is not affected by the treatment. Rosenbaum (1984) also defines a new parameter, the "net treatment difference" (*NTD*) that is estimated simply by adjusting for the observed value of the concomitant variable. However, Rosenbaum (1984) points out that the *NTD* has no causal interpretation, although it is argued to "provide insight into the treatment mechanism". More recently, Imbens (2004) also warns about similar

⁴Hereafter we refer to the variables that are affected by the treatment as "post-treatment variables" or "concomitant variables".

pitfalls when controlling for post-treatment variables affected by a treatment when interest lies in the identification of ATE , while Lechner (2005) specifies more explicit conditions to assess the "endogeneity bias" introduced when controlling for variables influenced by the treatment. It is important to stress that in these papers the parameter of interest is the ATE , and not the average treatment effect net of a post-treatment variable.

More recently, Frangakis and Rubin (2002) introduced the concept of principal stratification for the identification of causal effects controlling for post-treatment variables in a number of settings. Principal stratification builds on the concept of potential values and the central idea of comparing "comparable" individuals to obtain causal effects. In short, in order to identify causal effects controlling for a post-treatment variable, it is necessary to consider not only the potential values of the outcome of interest, but also the potential values of the post-treatment variable under the different treatments available. In general, causal effects can be identified by comparing individuals with the same potential values of the post-treatment variable under each of the treatments under consideration (they call such a group a "principal strata"). Such causal effects are called "principal effects" by Frangakis and Rubin (2002). In this paper we make use of this concept and define our estimands to have a causal interpretation based on the idea of principal effects.

Some of the work closer to ours is in Mealli and Rubin (2003) and Rubin (2004). Both of them motivate the use of principal stratification to clarify and analyze the discussion of "direct" versus "indirect" causal effects, which answer questions similar to the ones we deal with here. A direct effect corresponds to a causal effect of a treatment on an outcome net of a post-treatment variable, while an indirect effect corresponds to the causal effect of a treatment on an outcome that is "mediated" by another variable (e.g. a mechanism). Mealli and Rubin (2003) discuss the application of principal stratification to analyze the assumptions needed to estimate direct versus indirect effects in the context of the temporal causal relationships between health and socioeconomic status analyzed by Adams et al. (2003). Their main goal is only to illustrate that the use of principal stratification clarifies the concepts of causality in this framework, and that other methods that ignore potential values of variables influenced by treatment status can potentially lead to misleading causal conclusions. Further discussion and illustration about direct and indirect effects using principal stratification is provided in Rubin (2004).

Even though the concepts of direct and indirect effects as discussed in Mealli and Rubin

(2003) and Rubin (2004) are similar to the causal mechanism and causal net effects we define and analyze here, there are important differences. First, the concepts of direct and indirect effects are loosely defined and motivated in those papers, while the parameters to be presented here are precisely defined. Moreover, we intuitively decompose the (total) ATE into two effects (mechanism and net effect), giving a precise meaning to the causal mechanism effect that we argue is relevant for policy purposes. Second, as we explain below, the concept of direct effect is a special case of our causal net effect for a specific subpopulation. Finally, we formally discuss identification and estimation methods under different assumptions and present empirical applications to illustrate them, which none of the other papers do.

Another strand of literature related to our work is that of Robins and Greenland (1992) and Petersen, Sinisi and van der Laan (2006) in the field of epidemiology (see also references therein). Robins and Greenland (1992) make a similar distinction of direct and indirect effects and present conditions under which they can be estimated; whereas Petersen, Sinisi and van der Laan (2006) discuss the related concepts of "controlled" and "natural" direct effects.⁵ While these concepts are related to ours in some respects, there are important differences as well. Most notably, these papers do not employ the concept of principal stratification we employ here; rather, they base their assumptions for identification on the observed values of the post-treatment variable. Therefore, they do not explicitly address the fact that such observed values reflect two different variables, each corresponding to the potential value under treatment and control. In our view, this obscures the assessment of the plausibility of the assumptions and, as discussed in Rubin (2005) and below, this approach may lead to invalid causal conclusions.⁶

⁵The two concepts differ in what the counterfactual is considered to be when obtaining the direct effect. A "controlled" direct effect is defined when the counterfactual value of the post-treatment variable is set at a specific value; while the "natural" direct effect is when such counterfactual is set to what would be obtained if the particular unit had not been treated.

⁶There is yet another literature related to the current paper that is concerned with the estimation of dynamic causal effects, such as Robins (1986) and more recently Lechner and Miquel (2005). In those papers, the identification of causal effects from sequences of interventions is analyzed. Accounting for the possibility of a dynamic selection process implies making assumptions about the dependence of both the sequence of treatments and the final outcome of interest on intermediate outcomes. In the current paper, we concentrate on a static model of causal effects but consider the causal mechanism through which the treatment impacts the outcome of interest and the causal effect net of this mechanism.

3 The Estimands of Interest

3.1 General Framework

In this paper we employ the potential outcomes framework introduced by Neyman (1923) for randomized experiments and later extended by Rubin (1974) to non-randomized settings. As previously mentioned, this approach is now widely used in the program evaluation literature. Assume we have a random sample of size N from a large population. For each unit i in the sample, let $T_i \in \{0, 1\}$ indicate whether the unit received the treatment of interest ($T_i = 1$) or the control treatment ($T_i = 0$). Usually, we are interested on the effect of the treatment T on an outcome Y . Using the potential outcome notation, let $Y_i(1)$ represent the outcome individual i would get if exposed to the treatment, and let $Y_i(0)$ represent the outcome she would get if exposed to the control treatment. In this paper we are particularly interested on analyzing the part of the effect of the treatment T on the outcome Y that works through a mechanism or post-treatment variable S , which we think is affected by the treatment and that also affects the outcome. Since S is affected by the treatment, we must also consider its potential values, denoted by $S_i(1)$ and $S_i(0)$. Hence, $S_i(1)$ represents the value of the post-treatment variable individual i would get if she received treatment, and $S_i(0)$ represents the value under the control treatment. It is important to note that in the analysis that follows we do not need to restrict S to be binary.⁷

Using this notation, the (population) average treatment effect is given by $ATE = E[Y(1) - Y(0)]$.⁸ Thus, we are specifically interested on estimating the part of the ATE that is causally due to the effect of T on Y through the mechanism variable S , and the causal effect of T on Y net of the effect through S .⁹ For each unit, we only get to observe the vector $(T_i, Y_i^{obs}, S_i^{obs})$, where $Y_i^{obs} = T_i Y_i(1) + (1 - T_i) Y_i(0)$ and $S_i^{obs} = T_i S_i(1) + (1 - T_i) S_i(0)$. In words, we only get to observe $Y_i(1)$ and $S_i(1)$ for those units that receive treatment and $Y_i(0)$ and $S_i(0)$ for those units that remain in the control group. This is the so-called

⁷For simplicity, we focus on the case in which we analyze a single mechanism S . However, the discussion and methods can be extended to the case with multiple mechanism variables.

⁸We also adopt the stable unit treatment value assumption (SUTVA) following Rubin (1980). This assumption is common throughout the literature, and it implies that the treatment effects at the individual level are not affected either by the mechanism used to assign the treatment or by the treatment received by other units. In practice, this assumption rules out general equilibrium effects of the treatment that may impact individuals.

⁹Another treatment effect usually analyzed in the literature is the average treatment effect on the treated, which is given by $ATT = E[Y(1) - Y(0)|T = 1]$. For ease of exposition we focus on decomposing the ATE , but the discussion and results can easily be extended to the ATT .

"fundamental problem of causal inference" (Holland, 1986). It is important to stress the fact that the observed value of the post-treatment variable, S^{obs} , represents two different potential variables: $S(1)$ for treated units and $S(0)$ for controls. This point, which is usually overlooked in the economics literature, will be relevant through the rest of the paper.

Before defining our parameters of interest, it is important to be clear about what constitutes a causal effect in this framework. In Rubin's causal model, a causal effect must be a comparison of potential outcomes for the same group of individuals under treatment and control (e.g., Rubin, 2005). As an example, consider the case when we assume that selection into treatment is random conditional on a set of observable covariates X . In this case, by using iterated expectations we can write the *ATE* as $ATE = E\{E[Y(1) - Y(0)|X]\}$. Thus, under this selection on observables assumption, a causal effect is a comparison of the potential outcomes $Y(1)$ and $Y(0)$ for those units with the same vector of covariates X . In this case, units with the same value of X are comparable.

Suppose for the moment that the treatment is randomly assigned. Under our notation, the net treatment difference defined by Rosenbaum (1984) can be written as $NTD = E\{E[Y(1) - Y(0)|S^{obs}]\}$. As pointed out by Rosenbaum, the *NTD* does not have a causal interpretation without further assumptions. The reason for this is that *NTD* compares individuals with the same values of S^{obs} . However, as mentioned before, S^{obs} represents two different potential variables, $S(1)$ and $S(0)$, and therefore individuals with the same value of S^{obs} are not comparable. This point is further illustrated and discussed in Mealli and Rubin (2003), Rubin (2004) and Rubin (2005).¹⁰

Then the question is how to draw causal inferences in the presence of the post-treatment variable S . Frangakis and Rubin (2002) (hereafter FR) introduce a concept for defining causal effects in the presence of post-treatment variables. The basic idea in FR is that one has to define the "same group of individuals" based on the potential values of the post-treatment variable. In FR terminology, the basic principal stratification with respect to post-treatment variable S is a partition of individuals into groups such that, within each group, all individuals have the same vector $\{S(0) = s_0, S(1) = s_1\}$. More generally, any

¹⁰Another way to see the problem of conditioning by S^{obs} is to note that when estimating the *NTD* based on the observed data we are implicitly assuming that the treatment is "randomly assigned" conditional on S^{obs} so that we can write, for instance, $E[Y(1)|S^{obs}] = E[Y(1)|T = 1, S^{obs}] = E[Y^{obs}|T = 1, S^{obs}]$. However, in general we can infer something about the treatment assignment T based on S^{obs} . Hence this assumption fails. Or, in Rubin's (2005) words, "... despite treatment being ignorable ... forcing the conditioning on $[S^{obs}]$ leads to a nonignorable treatment assignment mechanism."

given partition of the individuals into sets which are unions of sets in the basic principal strata is called a principal stratification. For example, a principal stratification can be a partition of individuals into two groups, one for which all individuals have $\{S(0) = 0, S(1) = 1\}$ and the rest.¹¹ A principal effect with respect to a strata is a comparison of potential outcomes within that strata. Since principal strata are not affected by treatment assignment, individuals in that group are indeed comparable and thus principal effects are causal effects. Therefore, while we cannot compare two units with the same values of S^{obs} , we can compare two units with the same values of $S(0)$ and $S(1)$. Since $S(1)$ and $S(0)$ are not affected by T , we can think of them as any other pre-treatment covariates.

Two points are important to mention. First, note that in this setting causal effects are defined within principal strata, that is, we condition on the potential values $S(0)$ and $S(1)$. Since within strata $S(0)$ and $S(1)$ are not affected by treatment assignment, we can think of those causal effects the same way we think about causal effects for a subpopulation with the same values of some given pre-treatment covariate. Second, $S(0)$ and $S(1)$ are never observed simultaneously for any individual, and therefore we are not able to observe the strata to which each individual belongs. As will be further discussed in Section 4, this imposes a challenge when estimating causal effects controlling for post-treatment variables. In the following subsection, we use the concept of principal stratification to define our parameters of interest in order to give them a causal interpretation.

3.2 Definition of Estimands

Using the same notation as in the previous sub-section, we are interested on analyzing how the effect of the treatment T on an outcome Y works through a mechanism or post-treatment variable S . For each individual i in our population we can define the "composite" potential outcomes $Y_i(t, w)$, where the first argument of this expression refers to one of the treatment arms and the second argument is an indicator equal to 1 if potential post-treatment variable $S_i(1)$ is received and 0 if $S_i(0)$ is received instead. For example, the potential outcome $Y_i(1, 0)$ would refer to the outcome individual i would receive if exposed to treatment ($t = 1$) and if she received post-treatment variable level $S_i(0)$. It is important

¹¹FR's idea of principal stratification is closely related to the local average treatment effect interpretation of instrumental variables (e.g., Imbens and Angrist, 1994). For example, in the terminology of Imbens and Angrist (1994), the group of "compliers" is the set of individuals that always comply with their treatment assignment regardless of whether their assignment is to treatment ($T = 1$) or control group ($T = 0$). Therefore, for them $\{S(0) = 0, S(1) = 1\}$, where S is an indicator of actual treatment reception.

to emphasize again that despite the second argument in $Y_i(t, w)$ being a binary indicator, the post-treatment variable S is not required to be binary. Instead, w indicates which potential value of S is used to arrive at the outcome Y , $S_i(0)$ under $w = 0$ and $S_i(1)$ under $w = 1$. Using this notation, we can consider the following composite potential outcomes for any given individual:

1. $Y_i(1, 1)$: this is the outcome the individual would get if she received treatment and post-treatment variable level $S_i(1)$. This potential outcome includes the total effect of receiving treatment on Y (i.e., through S or not). This is exactly the usual potential outcome $Y_i(1)$ under the treatment.
2. $Y_i(0, 0)$: this is the potential outcome when no treatment is received and the post-treatment variable value is $S_i(0)$. This is the outcome the individual would get if the treatment is not given to her and if the value of her post-treatment variable is not altered either. This is exactly the usual potential outcome $Y_i(0)$ under the control treatment.
3. $Y_i(1, 0)$: this is the outcome the individual would receive if she were exposed to the treatment but kept the level of S she would get had not been treated. In other words, it is the outcome the individual would get if we were to give the treatment to her but held the value of her post-treatment variable at $S_i(0)$. Since in this case we are holding the value of S fixed at $S_i(0)$ (i.e., at the level she would get if not treated), this potential outcome includes the effect of T on Y that is *not* through S . This is the key potential outcome we use to define net and mechanism effects below.¹²

Based on these composite potential outcomes, the following three individual-level comparisons are of interest:¹³

- $Y_i(1, 1) - Y_i(0, 0)$: this represents the the usual individual "total treatment effect". For example, the total effect of smoking during pregnancy on birth weight.

¹²For completeness, note that $Y_i(0, 1)$ is the potential outcome the individual would get when the treatment is not given to her but she receives a value of the post-treatment variable equal to $S_i(1)$. In this case the potential outcome includes the effect on the outcome when no treatment is given but we give to the individual the level of S she would get had she been treated.

¹³There are other three possible comparisons based on these composite potential outcomes: $Y_i(0, 1) - Y_i(0, 0)$; $Y_i(1, 1) - Y_i(0, 1)$ and $Y_i(1, 0) - Y_i(0, 1)$. However, they are not relevant for the present purposes, and thus we do not consider them in the sequel.

- $Y_i(1, 0) - Y_i(0, 0)$: this difference gives the effect of T on Y when the value of the post-treatment variable is held constant at $S_i(0)$. Hence, it is the part of the effect of T on Y that is *not* due to a change in S caused by the treatment. For example, the effect of smoking during pregnancy on birth weight that is *not* due to a change in gestation time caused by smoking. We call this the *individual causal net effect*.
- $Y_i(1, 1) - Y_i(1, 0)$: this difference gives the effect of a change in S , which is *due* to T , on the outcome Y . Here we hold constant all other ways in which T may affect Y , since $Y_i(1, 0)$ already considers the effect of T on Y through other channels. For example, this difference shows the effect of a change in gestation time due to smoking on birth weight, holding all other effects of smoking during pregnancy fixed. We call this the *individual causal mechanism effect*.

Given these comparisons, we can decompose the individual *total* treatment effect or *ITTE* as:

$$ITTE = Y_i(1, 1) - Y_i(0, 0) = \underbrace{[Y_i(1, 1) - Y_i(1, 0)]}_A + \underbrace{[Y_i(1, 0) - Y_i(0, 0)]}_B. \quad (1)$$

Hence, at the individual level, the total effect is decomposed into the part of the effect due to a change in S because of a change in T (term A or mechanism effect); and the part of the effect holding S fixed at $S(0)$ (term B or net effect). In terms of the empirical applications to be used later, *ITTE* is the individual total effect of smoking during pregnancy (training) on birth weight (earnings). Term A is the part of the effect *due* to a change in gestation (experience) *due to* smoking during pregnancy (receiving training). Term B is the effect of smoking during pregnancy (training) if we hold gestation (experience) at its potential value in the absence of the treatment, i.e. the effect *net* of the level of gestation (experience).

In a similar way, we can decompose the population average total treatment effect (*ATE*) as:

$$ATE = E[Y(1, 1) - Y(1, 0)] + E[Y(1, 0) - Y(0, 0)]. \quad (2)$$

Similar to the decomposition in (1), the first term reflects the part of the average treatment effect that is due only to a change in S because of a change in T , and the second part shows the part of the average effect holding S fixed at $S(0)$. It is clear from this decomposition that we need to make treatment comparisons adjusting for the post-treatment variable

S . This is not a straightforward task because the post-treatment variable is affected by the treatment, and one has to be careful in defining the estimands to avoid losing their causal interpretation. Thus, in order to causally interpret our parameters of interest, we employ the concept of principal stratification described in the previous subsection and condition on the principal strata $\{S(0) = s_0, S(1) = s_1\}$.

Using iterated expectations, we can write the ATE after controlling for $S(0)$ and $S(1)$ as

$$ATE = E \{E[Y(1, 1) - Y(0, 0)|S(0) = s_0, S(1) = s_1]\} = E[\tau(s_0, s_1)] \quad (3)$$

where the outer expectation is taken over $S(0)$ and $S(1)$ and we let $\tau(s_0, s_1) = E[Y(1, 1) - Y(0, 0)|S(0) = s_0, S(1) = s_1]$. Then, using the same decomposition as in (2) we have:

$$\begin{aligned} ATE &= E \{E[Y(1, 1) - Y(1, 0)|S(0) = s_0, S(1) = s_1]\} \\ &\quad + E \{E[Y(1, 0) - Y(0, 0)|S(0) = s_0, S(1) = s_1]\}. \end{aligned} \quad (4)$$

We thus define the causal average net treatment effect or *CANTE* as:

$$CANTE = E \{E[Y(1, 0) - Y(0, 0)|S(0) = s_0, S(1) = s_1]\} \quad (5)$$

and the causal average mechanism treatment effect or *CAMTE* as:

$$CAMTE = E \{E[Y(1, 1) - Y(1, 0)|S(0) = s_0, S(1) = s_1]\} .^{14} \quad (6)$$

In the remaining of this section we provide some discussion about these estimands, and in the next section we concentrate on different assumptions and conditions upon which *CANTE* and *CAMTE* can be identified and estimated.

3.3 Discussion of Estimands

As mentioned in the introduction, an intuitive way to think about our estimands is the following. We can think of $Y(1, 0)$ as the potential outcome of an alternative counterfactual experiment in which the treatment is the same as the original one but holds S fixed at $S_i(0)$ for each individual i . That is, the alternative treatment blocks the effect of T on S . In

¹⁴Although we could have used the terms "direct effect" and "indirect effect" to define our parameters, we prefer our names for two reasons. First, they differ in important ways from direct effects as defined in Mealli and Rubin (2003) and Rubin (2004), as discussed later in section 3.4. Second, our names make clear that these effects are considered with respect to a particular mechanism S . On the other hand, strictly speaking, a "pure" direct effect would have to net out all possible mechanisms through which the treatment may affect the outcome.

this case, for any given individual, her causal net treatment effect is the difference between the outcome of this alternative treatment, $Y_i(1, 0)$, and $Y_i(0, 0)$ from the original control treatment. Similarly, her causal mechanism effect is given by the difference in the outcomes of the original treatment and the alternative one. Since this counterfactual experiment is seldom available, our estimands will pose special challenges for their identification and estimation in the next section.

An important property of our decomposition is as follows. Note that the definition of *CANTE* involves the comparison of the outcome $Y(1, 0)$ from the alternative experiment to the outcome $Y(0, 0)$ under control, and thus it includes not only the part of the *ATE* that is totally unrelated to the mechanism variable S , but also the part of the *ATE* that results from a change *in the way* S affects Y . That is, even though the level of S is held fixed at $S(0)$ under the alternative experiment, the treatment may still affect the way S affects the outcome, and this is part of *CANTE* in our definition. In other words, *CANTE* directly answers the following question: what would be the new average outcome if we change the treatment to be the same as the original one but we hold S fixed at $S(0)$? This is an important distinction since, for answering this question, it is not necessary to know how S affects Y . If one is interested on the alternative question of what would be the outcome from a new treatment that holds S fixed at $S(0)$ *and* also holds constant the way S affects Y , then we would need a different approach from the one we present here (and additional assumptions). On the other hand, we argue below that our definitions of *CANTE* and *CAMTE* are the relevant parameters from a policy point of view in most instances.

To illustrate this point consider the following example. If participants in a training program lose substantial labor market experience due to the program and this negatively affects their earnings, a policy maker may want to change the training program to be the same as the original one but without affecting labor market experience (i.e., holding experience fixed at their $S(0)$ level). In this case, the effect of this new training program would include not only the part of the effect of the program on earnings that is totally unrelated to experience, but it would also include the effect of the training program on how experience affects wages, i.e., the effect of the training program on the "returns to experience". *CANTE* takes this into account, correctly answering what the effect of this alternative treatment (policy) on the outcome would be.

Including the part of how S affects Y (e.g. "returns" to S) in *CANTE* is more relevant

from a policy perspective, compared to a parameter that only accounts for the part of the effect that is totally unrelated to S . The reason is that a policy maker typically has some degree of control over S , while very rarely over how S affects Y .¹⁵ In the previous example, the administrators of a training program have some degree of control over the level of labor market experience that might be lost due to the time spent in training (e.g. offering training while on the job or shortening the time to completion of the program), while it seems unlikely that they could influence the (potentially) different returns to experience that the market awards to trained versus non-trained individuals. This same line of reasoning applies to our other empirical application: doctors may be able to influence the gestation time (e.g. using specific drugs) but they are unlikely to influence how gestation time impacts birth weight. This idea is tied to the notion of a "treatment" that underlies Rubin's causal model, in which treatments are seen as interventions which can be potentially applied to each individual (e.g., Holland, 1986). Hence, our definition of $CANTE$ looks at the effect of changing or altering the level of the post-treatment variable given to an individual, while taking into account that a policy maker is seldom able to affect the way in which S affects Y .

As a final remark, to gain more intuition into $CANTE$ and $CAMTE$, we present a simple but illustrative exercise analyzing them in the following two extreme cases. First, consider the situation in which all the effect of a treatment works exclusively through S for the entire population. In this case, intuitively, $CANTE$ should be zero and $CAMTE$ should equal ATE . Using our notation, in this case $Y(1,0) = Y(0,0)$ and thus it follows from (5) that $CANTE = 0$ (i.e. there is no effect of T on Y net of S). Similarly,

$$CAMTE = E \{E[Y(1,1) - Y(1,0)|S(0) = s_0, S(1) = s_1]\} = E[\tau(s_0, s_1)] = ATE,$$

that is, all the effect of T on Y is through the mechanism given by S .

Second, consider the situation in which none of the effect of T on Y is through S , in which case $CANTE$ should equal ATE and $CAMTE$ should be zero. This situation can arise due to two reasons: either S does not affect Y (even though S may be affected by T) and thus $\{S(1), S(0)\}$ are independent of $\{Y(1), Y(0)\}$; or T simply does not affect S and thus $S(1) = S(0)$. Regardless of which reason is at play, the consequence is the same: $Y(1,1) = Y(1,0)$ and thus

$$CANTE = E \{E[Y(1,0) - Y(0,0)|S(0) = s_0, S(1) = s_1]\} = E[\tau(s_0, s_1)] = ATE$$

¹⁵One potentially relevant case where the policymaker might have some degree of influence is when general equilibrium effects due to the treatment are present.

and $CAMTE = 0$ from (6).

3.4 Relation of the Estimands to Other Parameters in the Literature

Our definition of $CANTE$ is different in nature from the net treatment difference (NTD) defined in Rosenbaum (1984). NTD defines an estimand that simply adjusts for an observed post-treatment variable and has no causal interpretation if the post-treatment variable is affected by the treatment. $CANTE$ explicitly accounts for the possibility that the post-treatment variable is affected by the treatment using its potential values; and effectively decomposes the ATE into the causal effect of the treatment on the outcome net of the post-treatment variable and the causal effect that is attributed to the post-treatment variable (mechanism). At a conceptual level, $CANTE$ and $CAMTE$ are easier and clearer to interpret than the NTD .

We now compare our estimands to the concepts of direct and indirect effects in Mealli and Rubin (2003) and Rubin (2004). Both their concepts and our estimands rely on the idea of principal stratification by Frangakis and Rubin (2002) and thus can be interpreted as causal effects. Mealli and Rubin (2003) define a direct effect as a comparison of $Y(1)$ and $Y(0)$ within the stratum for which $S(0) = S(1) = s$. In other words, using our notation in (3), Mealli and Rubin's direct effect can be written as $DE(s) = \tau(s, s)$. Given that in this strata we have $S(0) = S(1) = s$, then $Y(1, 1) = Y(1, 0)$ and therefore the $DE(s)$ is the same as $CANTE$ in (5) defined for this particular subpopulation. More generally, we can define the average direct effect as $ADE = E[\tau(s, s)]$, which is the average of the direct effects over the possible values of S . Unless $CANTE$ is constant in the population, the average direct effect will differ from $CANTE$. Moreover, direct effects do not represent a natural decomposition of the ATE into a direct and an indirect effect in the way $CANTE$ and $CAMTE$ do because they ignore all the individuals for which $S_i(1) \neq S_i(0)$. Finally, note that the definition of the direct effect effectively rules out a mechanism effect, since it is only defined for the subpopulations for which there is no such effect. In this regard, our parameters seem to be more general and relevant for policy purposes.

We offer the following summarizing remarks about our parameters. First, $CANTE$ and $CAMTE$ are defined within principal strata and thus are always causal effects, as opposed to the NTD which generally does not have a causal interpretation. Second, the definition of

CANTE applies to the entire population and not just to a particular subpopulation, which is the case for the direct effect (*DE*). As a result, *CANTE* can be used in decomposing the average total effect of the treatment on the outcome in an intuitive way. Based on these differences, we believe that *CANTE* and *CAMTE* are conceptually more intuitive and also more relevant for policy purposes.

3.5 Alternative Definitions to *CANTE* and *CAMTE*

We have defined *CANTE* in (5) as the average difference between $Y(1, 0)$ and $Y(0, 0)$, where $Y(1, 0)$ can be regarded as the potential outcome from an alternative treatment which is the same as the original one but holds the post-treatment variable S fixed at $S(0)$. In some applications, however, the alternative treatment of interest (the counterfactual) may be different. For instance, one may want to look at the effect of a new treatment which is the same as the original one but holds the value of S fixed at a particular level \bar{s} . That is, in this case we set $S(1) = \bar{s}$ for all individuals. In the birth weight example, one may want to look at the effect of smoking during pregnancy on birth weight holding the level of gestation fixed at a "normal" level, e.g. $\bar{s} = 9$ months. Let $Y(1, \bar{s})$ denote the potential outcome of this new treatment with the value of the post-treatment variable fixed at \bar{s} . Then, we can define the corresponding parameter as

$$\alpha = E \{E[Y(1, \bar{s}) - Y(0, 0) | S(0) = s_0, S(1) = s_1]\} \quad (7)$$

As in (5), this parameter is defined conditional on the principal strata in order to maintain a causal interpretation. One could then decompose the total *ATE* as

$$\begin{aligned} ATE &= E \{E[Y(1, 1) - Y(1, \bar{s}) | S(0) = s_0, S(1) = s_1]\} \\ &\quad + E \{E[Y(1, \bar{s}) - Y(0, 0) | S(0) = s_0, S(1) = s_1]\}. \end{aligned} \quad (8)$$

The second term in (8) equals α and gives the effect of the treatment T on Y if the value of the post-treatment variable were fixed at \bar{s} ; whereas the first term represents the effect on Y from a change in S due to a change caused by the treatment, relative to a value of S equal to \bar{s} . The parameter α in (8), without conditioning on principal strata, is what Petersen, Sinisi and van der Laan (2006) call a "controlled" direct effect. The main difference between our work and the controlled direct effects in Petersen, Sinisi and van der Laan (2006) is that we condition on principal strata, and this has important implications for the estimation of the parameters of interest, as will be discussed further in section 4.

The current discussion illustrates that a key point when defining net effects is to have clear what the counterfactual experiment (treatment) of interest is. This counterfactual treatment determines the potential outcome we want to learn about. Similarly, it is important that this counterfactual treatment is defined based on potential values of the post-treatment variable in order to preserve a causal interpretation. To identify and estimate these alternative net effect definitions, such as the parameter α in (8), the general framework described in the next section can be employed.

4 Identification and Estimation of the Parameters of Interest

In this section we discuss identification and estimation of the parameters *CANTE* and *CAMTE* defined in section 3.2. In order to highlight the main issues for identification and estimation of our parameters of interest, we start by considering the case in which the treatment is randomly assigned. This case is important in its own right given the existence of social experiments in economics, such as the one used in our first empirical application. Next, we discuss the case in which the treatment is not randomly assigned but is assumed to be random given a set of observed covariates. This later assumption is typically known in the literature as the unconfoundedness or selection on observables assumption. We present two identification strategies under each of these treatment-assignment mechanisms. In the last subsection we discuss a set of assumptions under which the usual approach of controlling for S^{obs} (e.g., a regression of Y^{obs} on T and S^{obs}) can be interpreted as a causal average net effect.¹⁶

Regardless of the mechanism used to assign the treatment, identification and estimation of our parameters faces two challenges. First, we have to take into account that for each unit under study only one of the potential values of the post-treatment variable is observed. Hence, the observed value of the post-treatment value, S^{obs} , represents two different potential values, $S(1)$ for treated units and $S(0)$ for controls units. This in turn implies that the principal strata $\{S(0) = s_0, S(1) = s_1\}$ needed for a causal interpretation of *CANTE* is never observed. The second challenge one has to tackle is the fact that the key potential outcome needed for estimation of *CANTE*, $Y_i(1,0)$, is practically never observed— this is

¹⁶In the rest of the section we focus on identification and estimation of *CANTE* since, by definition, we can obtain $CAMTE = ATE - CANTE$.

in contrast to the case of estimation of the *ATE* where the problem is that $Y(1)$ is only observed for treated units and $Y(0)$ is only observed for controls.

Because of these challenges, it is not surprising that the assumptions needed for estimation of *CANTE* (and *CAMTE*) are stronger than the ones needed for estimation of *ATE*. It is important to stress that this is not a problem of our approach, but it only reflects the difficulty of answering the question at hand given the available data. In an ideal situation in which we could perform the counterfactual experiment and observe the outcome from the alternative experiment which holds S fixed at $S(0)$, none of these challenges would arise and estimation of *CANTE* would be straightforward.¹⁷ Hence, in estimating our parameters, what we are implicitly trying to do is to learn about an alternative treatment based on the one we have at hand. As usual in causal inference problems, difficulties arise from a missing data problem and the strength of the assumptions needed to estimate the parameters are a reflection of the usual trade-off between quality of the data and assumptions. Importantly, one must have a clear understanding of the assumptions needed to identify and estimate our parameters in practice, in order to be able to evaluate their plausibility in any particular setting.

Let us start by considering the problem that the principal strata $\{S(0) = s_0, S(1) = s_1\}$ is unobserved. As it is usually the case in econometrics when we need to condition on an unobserved variable, we base our inference on predictions of the potential values of the post-treatment variable. Given a set of covariates X , our goal is to predict $S(0)$ for treated units and $S(1)$ for control units. One can do this in different ways.¹⁸ Suppose we were to use a matching approach based on X , and let $Z_M(i)$ be the set of indices for the M closest matches for unit i in terms of a given distance measure $\|X_i - X_j\|$ and for which $T_i \neq T_j$. As before, let S^{obs} denote the observed value of the post-treatment variable S . Then, we can

¹⁷Under this counterfactual treatment we have that $S(1) = S(0)$ for all units (by construction of the counterfactual treatment), and the potential outcome $Y(1,0)$ would be observed for those who received the treatment.

¹⁸See, for instance, the approaches discussed in Imbens (2004) in the context of predicting potential outcomes in the estimation of average treatment effects.

define $\widehat{S}(1)$ and $\widehat{S}(0)$ for each unit i as:¹⁹

$$\widehat{S}_i(0) = \begin{cases} S_i^{obs} & \text{if } T_i = 0 \\ \frac{1}{M} \sum_{k \in Z_M(i)} S_k^{obs} & \text{if } T_i = 1 \end{cases} \quad (9)$$

and

$$\widehat{S}_i(1) = \begin{cases} \frac{1}{M} \sum_{k \in Z_M(i)} S_k^{obs} & \text{if } T_i = 0 \\ S_i^{obs} & \text{if } T_i = 1 \end{cases} \quad (10)$$

Therefore, for estimation of *CANTE* we have to rely on the predicted values $\widehat{S}(1)$ and $\widehat{S}(0)$ to condition on principal strata. The reliance on predicted values of unobserved variables to identify and estimate parameters of interest is not uncommon in econometrics or statistics. For example, propensity score approaches require the estimation of the propensity score (Rosenbaum and Rubin, 1983), and the widely used sample selection model of Heckman (1979) requires estimation of a selection term or inverse Mills ratio under normality.²⁰ In our case, the fact that we actually observe the values of $S(0)$ for controls and $S(1)$ for treated individuals helps us not only to get a better predicted value of the potential values of the post-treatment variable, but it also allows us to empirically evaluate how well we are predicting them.

Given that we have to base our inferences about *CANTE* on the predicted values of $S(1)$ and $S(0)$, one key assumption for identification and estimation of *CANTE* is that we are able to compare individuals within the same *observed* (predicted) strata. That is, we will need to assume that individuals with the same predicted value of the post-treatment variable under treatment and control are comparable. We will further discuss this assumption in the sequel. For now, we note that, as one would expect, the better the covariates help predict the post-treatment potential value, the better we will be able to predict the unobserved principal strata. Also, note that regardless of the treatment-assignment mechanism, for identification and estimation of our parameters, the covariates X must play the role of helping us predict the missing potential values of the post-treatment variable.

¹⁹Alternatively, one could use a regression function approach to predict $S(1)$ and $S(0)$. Let $\mu_t(x) = E[S(t) | X = x]$ for $t = \{0, 1\}$ be the regression functions of the post-treatment potential values on X . Then, given the estimator $\widehat{\mu}_t(x)$ of these regression functions, we would define $\widehat{S}(1)$ and $\widehat{S}(0)$ for each unit i as $\widehat{\mu}_1(x)$ and $\widehat{\mu}_0(x)$, respectively. In principle, for each unit one could estimate only the missing potential value of the post-treatment variable and set the other equal to the observed one. However, in this case one may worry about systematic differences between predicted and observed values.

²⁰It is important to point out, however, that for estimation using the propensity score it is advantageous to use the estimated propensity score as opposed to the true propensity score in order to achieve the semi-parametric efficiency bound (Hirano, Imbens and Ridder, 2003). Nevertheless, the correct specification of the propensity score is usually unknown to the researcher.

The second challenge we face for estimation of *CANTE* is that the key potential outcome in the definition of *CANTE*, $Y_i(1, 0)$, is in principle unobserved for all units. This term reveals the difficulty in estimating both the causal mechanism and the causal effect net of such mechanism: while the pairs $Y(1, 1), S(1)$ and $Y(0, 0), S(0)$ are observed for some individuals (those treated and those not treated, respectively), the counterfactual $Y(1, 0)$ that is crucial in both definitions is in principle never observed. This is in contrast to the case of estimation of total treatment effects (e.g. *ATE*). Despite this serious missing data problem, we can still impose assumptions under which *CANTE* and *CAMTE* can be identified from the available data. These assumptions will be necessarily stronger than those needed for estimation of the *ATE*. Consequently, it is important to make those assumptions explicit in order to gauge their plausibility in any given application.

In what follows, we present two different sets of assumptions under which *CANTE* and *CAMTE* can be identified. The first identification strategy discussed below is based on the fact that we do have information on $Y(1, 0)$ for a specific subpopulation, which is the group of individuals for which the treatment does not affect the post-treatment variable S . For this subpopulation we have $Y(1, 1) = Y(1, 0)$, and this helps us identify our parameters under some assumptions. The second identification strategy is based on the idea that the observed values of $Y(1, 1)$ and $Y(0, 0)$ may provide valuable information about $Y(1, 0)$. Under some functional form assumptions, we can predict the missing outcome $Y(1, 0)$ and thus identify and estimate our parameters. In the following two sections we describe these strategies under the assumption of a randomly assigned treatment. Subsequently, we consider the case in which we assume the treatment is randomly assigned conditional on a set of observable covariates.

4.1 Identification and Estimation Based on a Specific Subpopulation

We first consider the case in which individuals are randomly assigned to the treatment. Hence, unless otherwise noted, we keep the following assumption.²¹

Assumption 1 $Y(1, 1), Y(0, 0), Y(1, 0), S(1), S(0) \perp T$

Under this assumption, the treatment received by each individual is independent of her potential outcomes and potential values of the post-treatment variable. Note that Assump-

²¹As in Dawid (1979), we write $X \perp Y$ to denote independence of X and Y .

tion 1 also implies $Y(t, w) \perp T | \{S(1), S(0)\}$ for $t = \{0, 1\}$.²²

A key to identifying *CANTE* given in (5) is to note that we can identify it for a particular subpopulation or principal strata. This subpopulation is the group for which $S_i(1) = S_i(0)$; that is, the subpopulation for which T does not affect S . For this strata we have that $S_i(1) = S_i(0)$ and therefore $Y_i(1, 0) = Y_i(1, 1)$. Hence, $Y_i(1, 0)$ is actually observed for those individuals that belong to this strata and have received treatment.²³ Note that for individuals in this strata, if $Y_i(1, 1) \neq Y_i(0, 0)$ then we can be sure that the causal effect $Y_i(1, 1) - Y_i(0, 0)$ is due to factors different from the change in S caused by T .

For this particular subpopulation or principal strata with $S_i(1) = S_i(0)$, we define the *Local CANTE* (hereafter *LCANTE*) as:

$$LCANTE = E \{E[Y(1, 0) - Y(0, 0) | S(0) = s, S(1) = s]\} = E \{\Delta(s)\} \quad (11)$$

where $\Delta(s) = E[Y(1, 0) - Y(0, 0) | S(0) = s, S(1) = s]$ is the local *CANTE* in the stratum $S(1) = S(0) = s$.²⁴ Indeed, *LCANTE* is similar to the "direct effect" discussed in Mealli and Rubin (2003) and Rubin (2004). More precisely, *LCANTE* equals the average direct effect discussed in the previous section, which is simply the average of the direct effects for all the stratum for which $S(1) = S(0) = s$. In the context of a binary post-treatment variable, *LCANTE* would be the average of the direct effects for the stratum with $S(0) = S(1) = 0$ and $S(0) = S(1) = 1$. It is important to remark that the direct effect used by Mealli and Rubin (2003) and Rubin (2004) is actually a *local CANTE* since it is defined only for a specific subpopulation. As discussed before, in this sense *CANTE* is a more general concept than the direct effect, and it actually decomposes the *ATE* in two intuitive parts.^{25,26}

The key to regard *LCANTE* as a causal effect is to note that it is defined within a principal strata, and thus has a causal interpretation (Frangakis and Rubin, 2002). Moreover, note that for the subpopulation with $S(1) = S(0) = s$ we have that $Y(1, 0) = Y(1, 1)$, and therefore the local causal average *net* treatment effect (*LCANTE*) equals the local average treatment effect (*LATE*) for this subpopulation. There is precedent in the literature

²²See, for instance, Lemma 4 in Dawid (1979).

²³Note that while $Y_i(1, 1) = Y_i(1, 0)$ for the set of people for which $S_i(1) = S_i(0)$, for the rest of the individuals we have that in general $Y_i(1, 0) \neq Y_i(1, 1)$.

²⁴Note that the outer expectation in *LCANTE* is taken over all strata with $S(1) = S(0) = s$.

²⁵We are unaware of work that estimates the direct effect using the potential outcome framework as discussed in Frangakis and Rubin (2002). We do estimate *LCANTE* for our empirical applications below.

²⁶Even if it is the case that there are no individuals for which $S(1) = S(0)$ in the population, under very mild assumptions we can have that for those individuals with $S(1)$ relatively close to $S(0)$, $Y(1, 0)$ is also relatively close to $Y(1, 1)$, and we can use the approach described here to obtain an estimate of *LCANTE*.

on estimation of local average treatment effects. In particular, Imbens and Angrist (1994) interpret instrumental variables (*IV*) estimators as estimators of a local (causal) average treatment effect (*LATE*). A common example is the estimation of a treatment effect when the treatment is randomized but there exists non-compliance, i.e. individuals that do not comply with their treatment assignment. In this case, an *IV* estimator that uses randomization as an instrument for T identifies the (local) average treatment effect on the compliers, that is, the subgroup of individuals that comply with their assignment. In fact, as noted by FR, this idea can be seen as an example of principal stratification in which the local average treatment effect obtained is the one for the stratum formed by those individuals that comply with the treatment assignment. In our case, $\Delta(s)$ is the local average treatment effect for the stratum for which $S(1) = S(0) = s$.

Suppose for the moment that we know which individuals belong to each of the strata $S(1) = S(0) = s$ for all values of s . Recall that Assumption 1 implies that $Y(t, w) \perp T | \{S(1), S(0)\}$ for $t = \{0, 1\}$, and therefore, conditioning on the principal strata $\{S(1), S(0)\}$ controls for all observed and unobserved individual characteristics reflected in the post-treatment variable S . Then, under Assumption 1, *LCANTE* is identified for this group since it can be written as

$$\begin{aligned}
LCANTE &= E\{\Delta(s)\} = E\{E[Y(1, 0) - Y(0, 0) | S(0) = s, S(1) = s]\} \\
&= E\{E[Y(1, 1) | S(0) = s, S(1) = s] - E[Y(0, 0) | S(0) = s, S(1) = s]\} \\
&= E\{E[Y^{obs} | T = 1, S(0) = s, S(1) = s] - E[Y^{obs} | T = 0, S(0) = s, S(1) = s]\}
\end{aligned} \tag{12}$$

where in the last line we have let Y^{obs} denote the observed value of Y for each individual. Therefore, if we see the individuals that belong to the strata $S(1) = S(0) = s$, *LCANTE* is identified since it can be written as a function of observed variables.

Unfortunately, $S(0)$ and $S(1)$ are never observed for the same individual. Hence, we need to base our inference on predicted values of those potential values of the post-treatment variable. Given a set of covariates X , let $\widehat{S}(1)$ and $\widehat{S}(0)$ be the predictions of the potential values of the post-treatment variable based on X . Then, for identification and estimation of *LCANTE* in this setting we need to assume that the potential outcomes are independent of the treatment assignment given the observed strata $\{\widehat{S}(1), \widehat{S}(0)\}$. More formally, we assume:

Assumption 2 $Y(1, 1), Y(0, 0), Y(1, 0) \perp T | \{\widehat{S}(1), \widehat{S}(0)\}$

Assumption 2 implies that the treatment assignment is random conditional on the observed strata. In other words, it implies that once we are within a principal strata with $\{\widehat{S}(1) = s_1, \widehat{S}(0) = s_0\}$ the treatment assignment is ignorable, so that there are no other (observed or unobserved) variables that can help us predict the assigned treatment and that are also correlated with the potential outcomes. This assumption guarantees that we can compare treated and controls units with the same observed values of $\widehat{S}(1)$ and $\widehat{S}(0)$. Therefore, Assumption 2 implies that once we condition on the observed strata, there are no systematic differences between treated and control individuals.

To give a specific example, while random assignment of a training program guarantees that those individuals with the same levels of potential experience $\{S(1), S(0)\}$ are comparable, in practice we have to assume that those individuals with the same *observed* potential experience $\{\widehat{S}(1), \widehat{S}(0)\}$ are comparable. In this case, $\widehat{S}(1)$ and $\widehat{S}(0)$ would be based on observable characteristics such as, for example, education, gender and race. From this discussion, it is clear that the better we are able to predict $S(1)$ and $S(0)$, the more plausible Assumption 2 will be. In the limit, if we are able to perfectly predict $S(1)$ and $S(0)$, then the random assignment nature of the treatment (i.e., Assumption 1) will guarantee that conditioning on the observed principal strata is enough to make causal comparisons.

Following the same steps as in (12) but now controlling for the observed strata, we have that under Assumptions 1 and 2 *LCANTE* is identified from observed data:

$$LCANTE = E \left\{ E[Y^{obs} | T = 1, \widehat{S}(0) = s, \widehat{S}(1) = s] - E[Y^{obs} | T = 0, \widehat{S}(0) = s, \widehat{S}(1) = s] \right\} \quad (13)$$

So far we have discussed estimation of *CANTE* for a special subpopulation, which we called *LCANTE*. We noted that for this observed strata with $\{\widehat{S}(1) = s, \widehat{S}(0) = s\}$ there is no causal mechanism effect by definition, and thus *LCANTE* equals the local *ATE*. If interest lies on the estimation of *CANTE* and *CAMTE*, then we need further assumptions. The following assumption allows the identification of *CANTE* and *CAMTE* once *LCANTE* has been identified.

Assumption 3 *CANTE is constant over the population.*

Under this assumption we have that $CANTE = LCANTE$, where $LCANTE$ is for the subpopulation defined in (11), and thus $CANTE$ is identified. In addition, the part of the ATE that is due to the mechanism S is given by $CAMTE = ATE - CANTE$.

A few observations about this identification condition are in order, especially since Assumption 3 seems like a strong assumption. First, we point out that Assumption 3 is weaker than the common and restrictive assumption in the literature of a constant average treatment effect (see, e.g. Heckman, LaLonde and Smith, 1999). Assumption 3 allows for heterogeneous effects of the treatment on the outcome variable, but such heterogeneity is restricted to work only through the mechanism or post-treatment variable S (i.e. through equation (6)). The plausibility of this assumption can be gauged in light of this observation. Second, note that this assumption is similar to the assumption of a constant average treatment effect when estimating ATE using an instrumental variable. In that case, without further assumptions, we can only identify the $LATE$ for the group of individuals who change treatment status in response to a change in the instrumental variable. However, under the assumption of a constant ATE we have that $LATE = ATE$. Importantly, as mentioned before, Assumption 3 allows some heterogeneity, but restricts this heterogeneity to come only through the mechanism. Third, as further discussed in section 4.4, it is important to note that the usual approach that controls directly for S^{obs} implicitly assumes a similar but stronger condition as that in Assumption 3. Finally, we note again that the strong character of this identification condition shows the difficulty of learning about a counterfactual treatment from the information available on the original treatment, following the interpretation given at the beginning of this section.

We now turn our attention to a brief discussion of the actual estimation of the parameters of interest using the identification strategy presented above. For the case when treatment assignment has been random, one way to proceed is as follows: (i) specify and estimate a model (based on X) to predict $S(1)$ and $S(0)$;²⁷ (ii) identify the subpopulation for which $\widehat{S}(1) = \widehat{S}(0)$ based on the predictions of the model; (iii) obtain for this subpopulation $LCANTE$ using (13);²⁸ (iv) under Assumption 3, estimate $CAMTE = ATE - LCANTE$.

²⁷A sensitivity and specification analysis about the model for predicting the missing outcomes can be performed at this stage.

²⁸If the post-treatment variable under consideration is continuous or if the procedure we use to estimate $S(1)$ and $S(0)$ yields a continuous variable (and hence the probability of finding someone with $\widehat{S}(1) = \widehat{S}(0)$ is zero), then one can use nonparametric methods. For example, one can consider a window within which values of $\widehat{S}(1)$ and $\widehat{S}(0)$ are considered to be equal. As usual, such window will tend to zero as the sample size goes to infinity. Alternatively, one could use a kernel function to give higher weight to observations for

Note that estimation of *LCANTE* in step (iii) implies estimating *ATE* for this subpopulation controlling for $\widehat{S}(0) = \widehat{S}(1) = S^{obs}$. To accomplish this, one can use any of the methods described, for example, in Imbens (2004), such as regression or matching. The simplest way to do so is by running a regression for this subpopulation of the observed outcome on treatment status and the observed value of the post-treatment variable. The coefficient on T equals *LCANTE*.

4.2 Identification and Estimation based on $Y(1, 1)$ and $Y(0, 0)$

The main problem in the estimation of *CANTE* in (5) is that we have to make inferences about a potential outcome that is not usually observed, $Y(1, 0)$. In the previous section we used the fact that $Y(1, 0) = Y(1, 1)$ for the subpopulation with $S(1) = S(0)$ to identify *CANTE*. Another approach one can use with available data is to use the information on $Y(1, 1)$ for treated units and $Y(0, 0)$ for controls to learn about $Y(1, 0)$. Assumptions of this type are implicitly made many times in the literature (e.g., Petersen, Sinisi and van der Laan, 2006); however, it is important to make these assumptions explicit in order to gauge their plausibility.

One can use $Y(1, 1)$ and $Y(0, 0)$ in many different ways to try to learn about $Y(1, 0)$. The specific assumption to be made will likely depend on the particular application at hand, and it is desirable to consider alternative assumptions to check the sensitivity of the results to the particular way we model the unknown object $E[Y(1, 0)]$. We present here one such assumption to illustrate the approach.

Suppose we assume that the conditional expectations of the potential outcomes $Y(1, 0)$ and $Y(1, 1)$ have the same functional form in terms of $S(1)$ and $S(0)$, but the former sets $S(1) = S(0)$. For example, let $E[Y(1, 1)]$ be of the form $E[Y(1, 1) | S(1), S(0)] = a_1 + b_1 S(1) + c_1 S(0)$. Then, this assumption implies that $E[Y(1, 0) | S(1), S(0)] = a_1 + b_1 S(0) + c_1 S(0)$. We can state this assumption more generally as follows

Assumption 4 Let $E[Y(1, 1) | S(0) = s_0, S(1) = s_1] = f_1(S(1), S(0))$. Then,²⁹

$$E[Y(1, 0) | S(0) = s_0, S(1) = s_1] = f_1(S(1), S(0)) |_{S(1)=S(0), S(0)}.$$

which $\widehat{S}(1)$ is closer to $\widehat{S}(0)$.

²⁹We use the notation $|_{S(1)=S(0), S(0)}$ to express that a function is to be evaluated at the point $(S(1) = S(0), S(0))$.

We now make a few comments about this assumption.³⁰ First, Assumption 4 directly acknowledges the fact that we are trying to learn about a counterfactual treatment based on the information available on the original treatment, reflecting the difficulty of estimating causal net and mechanism effects with the information at hand. Second, it is important to note that regarding $Y(1, 0)$ as being the outcome of the alternative or counterfactual treatment does not imply Assumption 4. The definition of $Y(1, 0)$ implies that $Y(1, 0) = Y(1, 1)$ for those with $S(1) = S(0)$; however, for those with $S(1) \neq S(0)$ it is not necessarily the case that $Y(1, 0)$ has the same functional form as $Y(1, 1)$ when setting $S(1) = S(0)$. Finally, note that when $f_1(S(1), S(0))$ is a linear function of $S(1)$ and $S(0)$ (e.g., as in a linear regression), Assumption 4 implicitly sets the mean outcome of $Y(1, 0)$ from those individuals in the strata $\{S(1) = s_1, S(0) = s_0\}$ to be equal to the mean outcome in the principal strata where T does not affect S and $S(1) = S(0) = s_0$. That is, it implicitly assigns the mean outcome of $Y(1, 1)$ from those individuals in the strata $S(1) = S(0) = s_0$.³¹

Suppose for the moment that $S(1)$ and $S(0)$ are both observed. Then, under Assumptions 1 and 4 we can write *CANTE* as

$$\begin{aligned}
\text{CANTE} &= E\{E[Y(1, 0) - Y(0, 0)|S(0) = s_0, S(1) = s_1]\} \\
&= E\{E[Y(1, 0)|S(0) = s_0, S(1) = s_1]\} - E\{E[Y(0, 0)|S(0) = s_0, S(1) = s_1]\} \\
&= E\left\{f_1(S(1), S(0))\Big|_{S(1)=S(0), S(0)}\right\} - E\{E[Y^{obs}|T = 0, S(0) = s_0, S(1) = s_1]\}
\end{aligned} \tag{14}$$

where we have that $E[Y(1, 1)|S(0) = s_0, S(1) = s_1] = f_1(S(1), S(0)) = E[Y^{obs}|S(0) = s_0, S(1) = s_1, T = 1]$. That is, the model for $Y(1, 1)$ is estimated based on the treated units, since for them $Y(1, 1)$ is observed.

³⁰As mentioned before, one can use different assumptions depending of the application at hand. For example, in a particular application it may be appropriate to impute the mean value of $E[Y(1, 0)|S(0) = s_0, S(1) = s_1]$ as the average of the mean outcome $Y(1, 1)$ for those individuals in the strata $\{S(0) = S(1) = s_1\}$, and the mean outcome $Y(1, 1)$ for those individuals in the strata $\{S(0) = S(1) = s_0\}$. Also, note that Assumption 4 only uses information on $Y(1, 1)$, $S(1)$ and $S(0)$. As mentioned before, alternative assumptions that also use the information on $Y(0, 0)$ for the strata with $\{S(1) = s_1, S(0) = s_0\}$ and $s_1 \neq s_0$ may also be appropriate in particular situations. Any such assumption can use the general framework described in this section.

³¹Based on the current discussion, one way of imputing the unknown $Y_i(1, 0)$ for each individual is to use a matching approach. Under assumption 4, for any individual with $S_i(1) = s_1$ and $S_i(0) = s_0$, we would assign $Y_i(1, 0)$ equal to the value of $Y(1, 1)$ of people in the strata $S(1) = S(0) = s_0$. For simplicity, though, the estimation approach for estimation of *CANTE* described below suggests a regression approach to estimate $f_1(S(1), S(0))$ and then obtain $\widehat{Y}(1, 0)$. Both approaches are feasible and one may be chosen having in mind the application at hand.

As in our previous identification strategy, we impute the unobserved $S(0)$ and $S(1)$ by using their predicted values based on a set of covariates X . Specifically, we can again let $\widehat{S}_i(1)$ and $\widehat{S}_i(0)$ be given, for example, by (9) and (10). Therefore, we need to make use of Assumption 2 in order to guarantee that treatment assignment is random based on the observed strata. Under Assumptions 1, 2 and 4, and following the same steps as in (14), we have that *CANTE* is identified:

$$CANTE = E \left\{ f_1 \left(\widehat{S}(1), \widehat{S}(0) \right) \Big|_{\widehat{S}(1)=\widehat{S}(0), \widehat{S}(0)} \right\} - E \left\{ E \left[Y^{obs} | T = 0, \widehat{S}(0) = s_0, \widehat{S}(1) = s_1 \right] \right\} \quad (15)$$

In practice, this identification strategy can be implemented as follows: (i) specify and estimate a model (based on X) to predict missing $S(1)$ and $S(0)$ (see discussion in section 4.1); (ii) estimate a model for $E[Y^{obs} | \widehat{S}(0) = s_0, \widehat{S}(1) = s_1, T = 1] = f_1(\widehat{S}(1), \widehat{S}(0))$; (iii) compute $E[Y(1,0) | \widehat{S}(0) = s_0, \widehat{S}(1) = s_1] = f_1(\widehat{S}(0), \widehat{S}(0))$ based on the model in (ii); (iv) estimate *CANTE* using (15); (v) estimate *CAMTE* = *ATE* – *CANTE*. For steps (ii) and (iii), a simple way to proceed is to run a linear regression of Y^{obs} on $\widehat{S}(1)$ and $\widehat{S}(0)$ for treated units and evaluate this estimated model on $\widehat{S}(1) = \widehat{S}(0) = \widehat{E}[\widehat{S}_i(0)]$. One can also allow this function to be more flexible, for instance, by approximating it with a polynomial series expansion of $\widehat{S}(1)$ and $\widehat{S}(0)$.

We close this section with a reminder that in some situations it may not be straightforward to use $Y(1,1)$ and $Y(0,0)$ to obtain information about $Y(1,0)$. This only reflects the difficulty of answering our question of interest with the available data. Hence, stating the assumptions explicitly, and carefully justifying their plausibility in particular settings is a necessary step to increase the confidence with which the observed data is used to draw inferences about $Y(1,0)$, and thus net and mechanism effects in general.

4.3 Identification and Estimation under Non-random Assignment

In the previous sections we analyzed the problem of estimating *CANTE* and *CAMTE* when the treatment T is randomly assigned. There, the two main challenges for estimation of *CANTE* is that we only observe $S(1)$ for treated units and $S(0)$ for controls (i.e., principal strata is unobserved), and that a key potential outcome ($Y(1,0)$) is practically missing for all observations. Unfortunately, in economics we usually do not have an experiment in which T is randomly assigned, and this adds a new dimension to our problem. In this case, a

common approach in the literature to estimate the ATE is to assume that selection into treatment is based on a given set of observed covariates, X , and on unobserved components not correlated with the potential outcomes. We extend the results from the previous sections for estimation of our parameters to the case when T is not randomly assigned, but is assumed to be randomly assigned given a set of covariates X . This assumption is usually known in the literature as unconfoundedness, conditional independence, or selection on observables.³² Throughout this section, we keep the following unconfoundedness or selection on observables assumption.

Assumption 5 $Y(1, 1), Y(0, 0), Y(1, 0), S(1), S(0) \perp T | X$

Assumption 5 implies that the treatment received by each individual is independent of her potential outcomes and potential values of the post-treatment variable given the set of covariates X . Similar to Assumption 1, this assumption implies $Y(t, w) \perp T | (X, S(1), S(0))$ for $t = \{0, 1\}$. As with the case of a randomly assigned treatment, we present two estimation strategies: one based on estimation of $CANTE$ for a specific subpopulation (i.e., $LCANTE$), and another based on using the observed values of $Y(1, 1)$ and $Y(0, 0)$ to make inferences about $Y(1, 0)$.

We start by considering the estimation strategy discussed in section 4.1. Remember that the key insight to identify $CANTE$ in that section was to note that for those individuals with $S_i(1) = S_i(0)$, we have that $Y_i(1, 0) = Y_i(1, 1)$, and therefore, $Y_i(1, 0)$ is observed if they are treated. We can use the same idea to identify a local $CANTE$ under Assumption 5. However, we now need to compare treated and control individuals who have $S(1) = S(0) = s$ and, in addition, the same set of covariates X . In this case, define $LCANTE$ as

$$LCANTE = E \{E[Y(1, 0) - Y(0, 0) | S(0) = s, S(1) = s, X = x]\} = E \{\Delta(s, x)\} \quad (16)$$

where $\Delta(s, x) = E[Y(1, 0) - Y(0, 0) | S(0) = s, S(1) = s, X = x]$ is the local $CANTE$ in the strata $S(1) = S(0) = s$ and with $X = x$. As before, $\Delta(s, x)$ is a causal effect because it is defined within a principal stratum. Hence, the $LCANTE$ in (16) represents the expected value of the average effects over all the subpopulations (i.e., principal strata) where causal comparisons can be made and for which $Y(1, 0) = Y(1, 1)$.

When the treatment is not randomly assigned, we need to add an overlap assumption in order to ensure that, for sufficiently large samples, there will be both treated and control

³²See for example Heckman, Lalonde and Smith (1999) and Imbens (2004).

individuals at each value of X for those values of S for which it is possible to have $S(1) = S(0) = s$. Specifically, we make the following assumption.

Assumption 6 $0 < \Pr(T = 1|S(0) = s, S(1) = s, X = x) < 1$, for all x and those values s for which $\Pr(S(1) = S(0) = s) > 0$.

Overlap assumptions are common in the program evaluation literature.³³ For estimation of the ATE under selection on observables, this assumption is usually stated as $0 < \Pr(T = 1|X = x) < 1$ for all X ; which ensures that in infinite samples there will be both treated and control units at each value of X so that it is possible to estimate $E[Y|T = t, X = x]$ for all values of t and x . Assumption 5 is similar to the usual overlap assumption including $S(1)$ and $S(0)$ as additional covariates, except that it does not have to hold for all values of S , but only for those values for which it is possible to have $S(1) = S(0)$. This is the case because we want to estimate a local average effect that is defined precisely for that subpopulation with $S(1) = S(0)$.³⁴

For the moment, suppose we know $S(1)$ and $S(0)$ and thus the individuals that belong to the strata $S(1) = S(0) = s$. Then, under Assumptions 5 and 6 we can write (16) as:

$$\begin{aligned}
 LCANTE &= E\{\Delta(s, x)\} \\
 &= E\{E[Y(1, 0) - Y(0, 0)|S(0) = s, S(1) = s, X = x]\} \\
 &= E\{E[Y^{obs}|T = 1, S(0) = s, S(1) = s, X = x] \\
 &\quad - E[Y^{obs}|T = 0, S(0) = s, S(1) = s, X = x]\}
 \end{aligned} \tag{17}$$

which can be readily estimated if indeed $S(1)$ and $S(0)$ are known.

As in the previous sections, we further have to deal with the problem that the principal strata is unobserved. Hence, we use instead observed strata, that is, the predicted strata based on a set of covariates. As before, let $\widehat{S}(1)$ and $\widehat{S}(0)$ be the predicted values of the potential post-treatment values based on X . Then, we add the following assumption, which is an extension of Assumption 2 to this selection on observables framework.

Assumption 7 $Y(1, 1), Y(0, 0), Y(1, 0) \perp T | \{\widehat{S}(1), \widehat{S}(0), X\}$

³³See for example Heckman, Lalonde and Smith (1999) and Imbens (2004).

³⁴It may be possible that for some values of S we cannot have that $S(1) = S(0) = s$. Hence, we do not require the overlap assumption to hold for all possible values of S . In the random assignment case the overlap assumption is not needed since by definition the probability of receiving treatment is positive for everyone.

Assumption 7 implies that, once we are within a group of individuals that share the same values of X , $\widehat{S}(1)$ and $\widehat{S}(0)$, there are no systematic differences between treated and control groups. In other words, we assume that those individuals with the same observed strata and values of X are comparable, so that simple differences in means between treated and control individuals yield causal effects for those in this principal strata.

Note that when T is randomly assigned, the role played by the covariates in the estimation of the parameters is that of predicting the principal strata each individual belongs to and also the potential outcome $Y(1,0)$. Due to the non-random assignment of the treatment and under the unconfoundedness assumption, the covariates now play the additional role of making the treatment random once we condition on them.³⁵ Hence, the assumptions in this section are stronger than those employed under a randomly assigned treatment, and the role played by the covariates is even more important.

Finally, we need one additional assumption in order to identify *LCANTE* with the observed data. From Assumption 5 it is implied that $S(t) \perp T|X$ for $t = \{0,1\}$; but now, since the treatment is not randomly assigned, we need to add the following overlap assumption to be able to predict $S(1)$ and $S(0)$ based on X .

Assumption 8 $0 < \Pr(T = 1|X = x) < 1$ for all x .

This assumption is indeed the same overlap assumption used for estimation of the *ATE*, which allows here predicting $S(1)$ and $S(0)$ under Assumptions 5 and 8 using any of the approaches discussed in section 4.1 (e.g., (9) and (10)).

Under Assumptions 5 to 8, following the same steps as in (17), we can write *LCANTE* as a function of observed data:

$$\begin{aligned}
 LCANTE = E \left\{ E[Y^{obs}|T = 1, \widehat{S}(0) = s, \widehat{S}(1) = s, X = x] \right. \\
 \left. - E[Y^{obs}|T = 0, \widehat{S}(0) = s, \widehat{S}(1) = s, X = x] \right\}. \tag{18}
 \end{aligned}$$

³⁵In principle, the set of covariates used for the selection on observables assumption and the ones used to predict $S(1)$ and $S(0)$ may differ. Without loss of generality, and given that in any particular application it may be hard to justify why a particular covariate may be used in only one of the assumptions, we maintain the same set of covariates in both assumptions. On the other hand, using a different set of covariates for predicting $S(1)$ and $S(0)$ will reduce the dependence of the estimation on the covariates used for the selection on observables assumption. Note that for the selection on observables assumption we use pre-treatment covariates. For predicting $S(1)$ and $S(0)$ we can use other post-treatment variables that are not affected by the treatment (otherwise are mechanisms) but can help predict $S(1)$ and $S(0)$. For instance, local labor market characteristics as used in Hotz, Imbens and Klerman (2006) and Flores-Lagunes, Gonzalez and Neumann (2006) for a different purpose.

Finally, note that under Assumption 3 in section 4.1 (i.e., constant $CANTE$) we have that the $LCANTE$ defined in (16) equals $CANTE$ and $CAMTE = ATE - CANTE$.

One way to implement this approach is the following: (i) specify and estimate a model (based on X) to predict $S(1)$ and $S(0)$; (ii) identify the subpopulation for which $S(1) = S(0)$ based on the predictions of the model; (iii) obtain for this subpopulation $LCANTE$ using (18); (iv) estimate $CAMTE = ATE - CANTE$. Note that the same general comments about this approach outlined in section 4.1 apply here.

We now turn our attention to the second approach for estimating $CANTE$ discussed in section 4.2 in the context of a randomly assigned treatment. This approach is based on using the observed outcomes, $Y(0,0)$ for controls and $Y(1,1)$ for treated, to learn about $E[Y(1,0)]$. As discussed in section 4.2, one can use the observed outcomes in different ways to make inferences about $E[Y(1,0)]$. Here we focus on the natural extension of the particular assumption used in section 4.2 to illustrate the general approach. Suppose that the conditional expectations of the potential outcomes $Y(1,0)$ and $Y(1,1)$ have the same functional form in terms of $S(1)$, $S(0)$ and X , but the former sets $S(1) = S(0)$. In general,

Assumption 9 Let $E[Y(1,1)|S(0) = s_0, S(1) = s_1, X = x] = f_1(S(1), S(0), X)$. Then,

$$E[Y(1,0)|S(0) = s_0, S(1) = s_1, X = x] = f_1(S(1), S(0), X)|_{S(1)=S(0), S(0), X}$$

Assumption 9 is a straightforward extension of Assumption 4, and thus the same notes given for that assumption apply here. Furthermore, due to the nonrandom treatment assignment, to estimate our parameters we need an overlap assumption to guarantee that in sufficiently large samples we are able to find treated and control units who are comparable in terms of $S(1)$, $S(0)$ and X .

Assumption 10 $0 < \Pr(T = 1|S(0) = s_0, S(1) = s_1, X = x) < 1$, for all s_0 , s_1 and x .

Note that Assumption 10 will be typically stronger than Assumption 6, as the latter refers to a particular subpopulation. As before, if $S(1)$ and $S(0)$ were both observed, under Assumptions 5, 9, and 10, we can write $CANTE$ as:

$$\begin{aligned} CANTE &= E\{E[Y(1,0) - Y(0,0)|S(0) = s_0, S(1) = s_1, X = x]\} \\ &= E\{f_1(S(1), S(0), X)|_{S(1)=S(0), S(0), X}\} \\ &\quad - E\{E[Y^{obs}|T = 0, S(0) = s_0, S(1) = s_1, X = x]\} \end{aligned} \tag{19}$$

where we have that $E[Y(1,1)|S(0) = s_0, S(1) = s_1, X = x] = f_1(S(1), S(0), X) = E[Y^{obs}|T = 1, S(0) = s_0, S(1) = s_1, X = x]$. That is, we estimate the model for $Y(1,1)$ based on the treated units. Finally, given that $S(1)$ and $S(0)$ are never observed simultaneously for the same unit, we have to work with their predicted values based on the set of covariates X , and add Assumption 7. Therefore, under Assumptions 7 through 10, we can estimate $CANTE$ using this approach and the observed data:

$$CANTE = E\{f_1(\widehat{S}(1), \widehat{S}(0), X)|_{\widehat{S}(1)=\widehat{S}(0), \widehat{S}(0), X}\} - E\{E[Y^{obs}|T = 0, \widehat{S}(0) = s_0, \widehat{S}(1) = s_1, X = x]\} \quad (20)$$

The implementation of this strategy is similar to the one discussed in section 4.2: (i) specify and estimate a model (based on X) to predict missing $S(1)$ and $S(0)$; (ii) estimate a model for $E[Y^{obs}|T = 1, S(0) = s_0, S(1) = s_1, X = x] = f_1(S(1), S(0), X)$; (iii) compute $E[Y(1,0)|\widehat{S}(0) = s_0, \widehat{S}(1) = s_1, X = x] = f_1(\widehat{S}(0), \widehat{S}(0), X)$ based on the model in (ii); (iv) estimate $CANTE$ using (20); (v) estimate $CAMTE = ATE - CANTE$. For the third step, a simple way to proceed is to run a regression of Y^{obs} on $\widehat{S}(1)$, $\widehat{S}(0)$ and X for treated units, and evaluate the estimated model on $\widehat{S}(1) = \widehat{S}(0) = \widehat{E}[\widehat{S}_i(0)]$.³⁶ It is also possible to allow the function to be more flexible by approximating $f_1(\widehat{S}(1), \widehat{S}(0), X)$ using a series expansion on $\widehat{S}(1)$ and $\widehat{S}(0)$.

4.4 Assumptions under which controlling directly for S^{obs} yields $CANTE$

It is important to specify conditions under which the usual approach of controlling for the observed value of the post-treatment variable (S^{obs}), and possibly a set of covariates, yields $CANTE$ as defined in (5). It turns out that these assumptions are stronger than those presented in the previous sections.

Rosenbaum (1984) states a set of sufficient conditions under which controlling for S^{obs} and a set of covariates X yields the ATE . Specifically, he shows that (i) if $S(1) = S(0)$ for all subjects in the population ("unaffected concomitant variable"), and, (ii) if $\{Y(1,1), Y(0,0), S(1), S(0)\} \perp T|X$ and $0 < \Pr(T = 1|X) < 1$ for all X ("ignorable

³⁶It is important to note that, because of nonrandom assignment of treatment, we need to evaluate $f_1(S(1), S(0), X)|_{S(1)=S(0), S(0), X}$ in step (iii) for treated and control units to estimate $E[Y(1,0)]$. This is different from the case when the treatment is randomly assigned, since in that case one can calculate $E[Y(1,0)]$ based only on the treated units because by random assignment treatment and control groups are comparable.

treatment assignment"); then

$$ATE = E \{ E [Y (1, 1) | T = 1, S^{obs} = s, X = x] - E [Y (0, 0) | T = 0, S^{obs} = s, X = x] \} \quad (21)$$

The main problem when using equation (21) to estimate the *ATE* is that the outer expectation over the first term inside curly brackets should be taken with respect to the distribution of $\Pr (S (1) | X)$, and the expectation over the second term should be taken with respect to $\Pr (S (0) | X)$. As a result, it is likely that bias will arise from averaging both terms over $\Pr (S^{obs} | X)$ instead.³⁷ Note that condition (i) above implies that $S^{obs} = S (1) = S (0)$, guaranteeing that the averaging is over the correct distribution. Regarding estimation of *CANTE*, note that condition (i) above implies $Y (1, 1) = Y (1, 0)$, and therefore $ATE = CANTE$. Hence, under conditions (i) and (ii) by Rosenbaum (1984), equation (21) estimates both the *ATE* and the *CANTE*. Unfortunately, this result is of little help since those conditions will only be satisfied when S is not affected by the treatment, hence ruling out the existence of a mechanism effect through S in the population.³⁸

We now take a closer look at the approach of controlling for S^{obs} using the concept of principal stratification. Similar to equation (21), we can write this estimator as

$$\begin{aligned} \gamma &= E\{E[Y^{obs}|T = 1, S^{obs} = s, X = x] - E[Y^{obs}|T = 0, S^{obs} = s, X = x]\} \\ &= E\{E[Y^{obs}|T = 1, S(1) = s, X = x] - E[Y^{obs}|T = 0, S(0) = s, X = x]\} \end{aligned} \quad (22)$$

From (22), we see that in comparing units with the same values of S^{obs} we in fact compare treated units with $S(1) = s$ to control units with $S(0) = s$. Unless those units belong to the strata for which $\{S(1) = S(0) = s, X = x\}$, this comparison will not yield a causal effect. Following Rosenbaum (1984), if the treatment does not affect S , then $S(1) = S(0)$ for all units and, as discussed above, (22) yields *ATE* and *CANTE*. However, in general, we will have some units for which $S(1) \neq S(0)$ and thus equation (22) compares units from different strata. Without further assumptions, (24) does not yield a causal net effect.

Recall that in estimating *CANTE* in the previous sections we dealt with two challenges:

³⁷Another way to see the problem of estimating *ATE* controlling for S^{obs} is to regard S^{obs} as an endogenous control variable since it is affected by the treatment. See Lechner (2005).

³⁸Note that conditions (i) and (ii) are different from our approach to estimate *CANTE* discussed in section 4.1. There, we use a set of people for which $\widehat{S}(1) = \widehat{S}(0)$ to identify the local *CANTE* for this principal strata or subpopulation. Under Rosenbaum's (1984) conditions, *CANTE* is implicitly assumed to be zero in the population.

only $S(1)$ or $S(0)$ is observed for each unit, and $Y(1,0)$ is never observed.³⁹ In dealing with the first challenge, we proposed to use the predicted values of $S(1)$ and $S(0)$ based on a set of covariates X . Note that under the assumptions for estimation of *LCANTE* discussed in section 4.3 (employing the selection-on-observables assumption) we know that, within principal strata with $\{\widehat{S}(1) = \widehat{S}(0) = s\}$, the estimator in (22) yields *LCANTE* as defined in (16). We can use this observation to present a set of sufficient assumptions under which (22) estimates *CANTE* using the predicted $\widehat{S}(1)$ and $\widehat{S}(0)$. These assumptions are slightly weaker than those in Rosenbaum (1984), but they are stronger than our assumptions outlined above. We start with the following assumption:

Assumption 11 For all units in the sample we have: $\widehat{S}_i(1) = \widehat{S}_i(0) = s$, for some s .

This assumption implies that the treatment T does not affect the predicted values of S based on X , *for every unit*. In other words, it implies that the treatment does not affect the part of the mechanism S that depends on X , so that any difference in the true values of $S(1)$ and $S(0)$ for every unit must come from unobserved variables. For instance, in the context of our smoking during pregnancy application, it implies that, for all mothers, the predicted gestation if a mother had smoked is equal to the predicted gestation if she had not smoked, where the prediction is based on a set of observed covariates.

Under Assumption 11, we know each and every unit belongs to a principal strata for which $\{\widehat{S}(1) = \widehat{S}(0) = s\}$. Therefore, under Assumptions 5 to 8 discussed in section 4.3, plus Assumption 11, we have that (22) estimates *LCANTE* in (16). Moreover, Assumption 11 implies that *LCANTE* = *CANTE*. While the additional Assumption 11 allows employing the approach of section 4.3 to estimate *LCANTE* controlling for S^{obs} , it is a strong assumption. In particular, note that our approach for identification and estimation of *LCANTE* in sections 4.1 and 4.3 is based on finding a subpopulation for which $\{\widehat{S}(1) = \widehat{S}(0) = s\}$; while Assumption 11 assumes that such subpopulation is in fact the complete population under study. In this respect, the latter is a stronger assumption. Finally, note that we can easily get some sense on the plausibility of Assumption 11 in our data by developing different models to predict $S(1)$ and $S(0)$ based on observed covariates and checking whether the complete sample satisfies $\{\widehat{S}(1) = \widehat{S}(0) = s\}$.

³⁹Rosenbaum's (1984) assumptions implicitly solve these challenges by assuming $S(1) = S(0)$ for all units, in which case $Y(1,0) = Y(1,1)$, so $Y(1,0)$ is observed for treated units; and $S^{obs} = S(1) = S(0)$.

Rosenbaum’s (1984) conditions and Assumption 11 exemplify the type of strong assumptions that are needed for estimation of causal net effects when controlling directly for S^{obs} and X . At the end of the day, any set of assumptions used to justify (22) as an estimator of a causal net effect needs to take into account the two challenges mentioned above: only $S(1)$ or $S(0)$ is observed for each unit; and $Y(1,0)$ is never observed. Hence, any such set of assumptions is likely to be as strong as the ones discussed here. We end by remarking again that it is important to make any assumptions explicit in order to evaluate their plausibility in a particular application.

5 Two Empirical Applications

In this section, we present two empirical applications that illustrate the implementation of our strategies to identify and estimate the causal average net treatment effect (*CANTE*) and the causal average mechanism treatment effect (*CAMTE*). The first application illustrates the case of a randomly assigned treatment, using data from the social experiment undertaken in the National Job Corps Study (NJCS), while the second application implements our estimators to observational data from Natality Data Sets from Pennsylvania (1989-1991).

5.1 Random Assignment

We first consider the case of random assignment of the treatment. If reception of treatment is randomly assigned, as it would be in a randomized experiment, then we can focus on the estimation of *CANTE* and *CAMTE* without the additional complication of controlling for self-selection into the treatment. The data comes from the National Job Corps Study (NJCS), a randomized experiment to evaluate the effectiveness and social value of the Job Corps training program. Job Corps (JC) provides low-skilled and less-educated young people (ages 16-24) with marketable skills to enhance their labor market outcomes. It does this by offering academic, vocational, and social skills training at 122 centers throughout the United States, where nearly all students reside during their enrollment period.⁴⁰

An important finding of the NJCS was that, 48 months after randomization, individuals in the treatment group earned a statistically significant 12% more per week than individuals in the control group (Burghardt et al., 2001). However, upon looking at different race and

⁴⁰For more information on Job Corps and the NJCS see Burghardt et al. (2001).

ethnic groups, it was found that Hispanics in the treatment group earned 10% less (not statistically significant) than those in the control group during the same period of time. In contrast, black and white treatment-group members experienced a statistically significant earnings increase of 14% and 24% over their control group members, respectively (Schochet, Burghardt and Glazer, 2001).⁴¹

The bold differential impact on Hispanics was labeled the most prominent "failure" of JC and it could not be explained by individual and institutional variables (Burghardt et al., 2001). In a recent paper, Flores-Lagunes, Gonzalez and Neumann (2006) (hereafter FGN) document that Hispanics in the control group earned a significant amount of labor market experience during the study compared to treated Hispanics and also to control-group blacks and whites. They show, using several pieces of evidence, that such accumulated experience resulted in an earnings advantage that treated Hispanics were not able to overcome by the end of the study (i.e. 48 months after randomization). In other words, this post-treatment variable potentially accounts for much of the lack of earnings gain of Hispanics in JC. Nevertheless, the methods employed in FGN (2006) come up short of having clean causal interpretations, which the authors recognize explicitly.

This setup offers an important situation in which the estimation of our causal parameters are policy-relevant: if lost labor market experience (i.e. the lock-in effect) is a relevant causal mechanism through which JC fails to increase the earnings of Hispanics, policies that reduce the lock-in effect of JC on Hispanics will be beneficial to consider. At the same time, focusing on subgroups that seem to differ in terms of their lock-in effect, this application provides an interesting setting in which our parameters should result in different inferences for the different demographic groups, at least in principle.

The first row in Table 1 reports, for comparison, the original NJCS estimates for the full sample of individuals that participated in the NJCS, and also for the three subgroups of interest: Hispanics, whites and blacks. The original NJCS estimates were computed using differences-in-means adjusted for non-compliance (see fn. 41), thus identifying a *LATE* on those who comply with their treatment assignment (Imbens and Angrist, 1994). Given that these estimates do not control for any pre-treatment variables, the sample used is one that

⁴¹These estimated effects reported by the NJCS were computed using differences-in-means estimates adjusted for non-compliance, thus identifying a *LATE* on those who comply with their treatment assignment (Imbens and Angrist, 1994). The adjustment is performed by dividing the differences-in-means by the difference in the proportion of those in the treatment group who enrolled in Job Corps (about 73%) minus the proportion of those in the control group that managed to enroll in Job Corps (about 1.4%).

contains individuals with earnings information at the 48-month interview. These estimates imply an overall gain from Job Corps of \$25.2 per week, although it is not uniform across demographic groups: whites and blacks gain \$46.2 and \$22.8 per week, respectively, both statistically significant, while Hispanics show a statistically insignificant loss of \$15.1.

Given that one needs to make use of observed covariates plus the post-treatment variable "average hours worked per week during the study", the remaining rows in Table 1 restrict the original sample used by the NJCS to those individuals without missing values in all of those variables.⁴² The same sample was used by FGN (2006), where it was compared with the sample used in the original NJCS estimates and was found to be consistent with the overall profile of the total JC population.⁴³ Rows 2 through 4 in Table 1 report estimates of the average "intention-to-treat" (*ITT*) parameter using this sample.⁴⁴ Row 2 presents unadjusted differences in means between treatment- and control-group individuals. Note that, despite smaller effects (even for Hispanics) relative to the original NJCS estimates, the main conclusions as in the NJCS hold. Even though controlling for covariates is not necessary since the treatment is randomly assigned, covariance adjustment should result in estimates with higher precision. Thus, our estimates will control for observed pre-treatment covariates, as shown in the *ITT* estimates in rows 3 and 4. We control for pre-treatment variables using a propensity score (pscore) approach (e.g., Dehejia and Wahba, 1999).⁴⁵ These estimates are fairly comparable to the unadjusted estimates in row 2, except for Hispanics (and consequently the full sample). This might be due to the smaller sample size of this group and the fact that the group shows some imbalances in pre-treatment covariates.⁴⁶ Again, the same conclusions as in the NJCS hold in rows 3 and 4, although the

⁴²The observed pre-treatment variables include: indicators for a high school diploma or GED, speaks English as a native language, married or cohabitating, household head, one or more children, gender, vocational degree, ever been convicted, employed, unemployed, not in the labor force, resides in a PMSA, MSA, pre-treatment weekly earnings, age, and indicators for race and ethnicity in the full sample.

⁴³The original NJCS sample contains 11,313 individuals. Out of these, 219 individuals do not complete the baseline interview, 1,295 more have missing values in any of the variables we control for, and 694 are individuals whose race or ethnicity is not white, black or Hispanic. The resulting sample contains 9,105 individuals.

⁴⁴Given the presence of random assignment non-compliance in the sample, the *ATE* can not be estimated using the randomization exclusively. To exploit the random assignment, we estimate the "intention-to-treat" (*ITT*) parameter. This parameter is commonly estimated in the program evaluation literature and allows us to focus on the estimation of our parameters under a randomly assigned treatment. Consequently, in this application our parameters are decomposing the *ITT* and not the *ATE*.

⁴⁵The propensity score (pscore) is estimated using all pre-treatment variables, their squares, and interactions in a logit model. Then, the pscore is included linearly in a regression, while in another specification a square and cubic terms are included as well.

⁴⁶The misalignment of observed pre-treatment variables for Hispanics is documented in Flores-Lagunes,

negative effects on Hispanics are less dramatic.⁴⁷ These "total effect" estimates will be the benchmark to compare our estimated effects net of the lock-in mechanism effect.

Rows 5 through 8 of Table 1 present estimates of Rosenbaum's (1984) *NTD* parameter. All of them are obtained by controlling for the observed value of post-treatment labor market experience, and differ in the way this is done. Recall that *NTD* estimates typically lack causal interpretation as estimates of the total effect (*ITT*), and correspond to a causal net effect under very stringent conditions (see section 4.4). We present these estimates here for comparison to estimates of *CANTE* using our identification strategies. Consistent with the findings in FGN (2006), the *NTD* estimates are less than 10% larger than the *ITT* estimates for whites and blacks (35.5 v. 32 and 20.2 v. 19.8, respectively, comparing averages of rows 5-8 and 3-4); while they are starkly different (180% larger) for Hispanics (6.1 v. -7.2, comparing same averages). Despite the statistically insignificant effects for Hispanics and the lack of causal interpretation of *NTD*, the point estimates are consistent with a relevant lock-in effect that explains an important part of the lack of effects of JC on them.

Some heterogeneity in the estimates in rows 5 to 8 arise as a result of the specific way in which the estimates are obtained. In addition to controlling for observed labor market experience (S^{obs}), rows 5 and 6 adjust for covariates in the same way rows 3 and 4 do. Not surprisingly, the heterogeneity is highest for Hispanics (the smallest sample). The specifications in rows 7 and 8 are similarly obtained adding S^{obs} to the estimated propensity score, as opposed to including it separately. This way of controlling for a post-treatment variable is followed by Black and Smith (2004). These alternative specifications result in estimates which are close to each other and also to that in row 6, which controls for experience in a more flexible way. The implied *NTD* estimate is about 6 dollars per week for Hispanics.

We present three estimates of *LCANTE* that differ in the way they are implemented. In all of them, the potential values of post-treatment labor market experience (i.e. $S(0)$ and $S(1)$) have to be estimated based on covariates X . To accomplish this, we implement equations (9)-(10) using a single match on the propensity score estimated before.⁴⁸ In order to estimate *LCANTE* we need to identify the population for which $\{\widehat{S}(1) = \widehat{S}(0)\}$. Since in this case S is defined as the average number of hours worked per week during the study, it is very difficult to find individuals for which $\widehat{S}(1) = \widehat{S}(0)$. We approach this problem in two

Gonzalez and Neumann (2006). We note that randomization was applied to the full sample and not explicitly for any of the subgroups.

⁴⁷This is also documented in Flores-Lagunes, Gonzalez and Neumann (2006).

⁴⁸Alternatively one could use, for example, an OLS approach to predict $S(1)$ and $S(0)$ for all units.

different ways. First, we define a window around $\widehat{S}(1) - \widehat{S}(0) = 0$ using a Silverman-type bandwidth to define the subpopulation with $\{\widehat{S}(1) = \widehat{S}(0)\}$. Since for those units falling in this window we have that $\widehat{S}(1) \approx \widehat{S}(0)$, then we can think of S^{obs} as any other covariate since it is (approximately) not affected by the treatment. Hence, in rows 9 and 10, we include S^{obs} in the propensity score to estimate *LCANTE* because S^{obs} can be regarded as another covariate within this subpopulation. The second way in which we estimate *LCANTE* is by using weighted least squares on the full data of the outcome on the treatment indicator, the usual propensity score and observed experience. The weights in this case are given by an Epanechnikov kernel function of the distance between $\widehat{S}(1)$ and $\widehat{S}(0)$, so that the weight for each unit declines as $\widehat{S}(1)$ and $\widehat{S}(0)$ are farther apart. The estimate of *LCANTE* obtained this way is shown in row 11.⁴⁹ Interestingly, the proportional size of the resulting *LCANTE* subpopulation is very similar across samples, ranging from 17% for blacks to 20% for Hispanics.

All three *LCANTE* estimates are larger relative to the estimated total effect (rows 3 and 4) and the *NTD* estimates (rows 5-8), although some heterogeneity is evident in the way the estimates are implemented, as we discuss below. Recall that these estimates correspond to a particular subpopulation unless we assume a constant *CANTE* over the population (Assumption 3). In contrast to the *NTD* estimates, the *LCANTE* estimates imply a more significant role of the lock-in effect across all the different samples, and not just for Hispanics. Looking at the average estimates from rows 9-10, they are 85%, 25% and 225% larger than the average estimates of the total effect for whites, blacks and Hispanics, respectively. In turn, the estimates of *LCANTE* in row 11 are even larger for whites and Hispanics. Taken together and embracing Assumption 3, they imply that the *NTD* underestimates the causal average net effect (*CANTE*), especially for whites and Hispanics.

Finally, row 12 presents an estimate of *CANTE* using the second identification and estimation strategy (section 4.2). To implement it, we model (under Assumption 4) the first term in (15) as a linear function of $\widehat{S}(0)$ and $\widehat{S}(1)$, adding, as before, the estimated propensity score (up to a cubic term) to gain efficiency.⁵⁰ Similarly, the second term in

⁴⁹In both instances we use a Silverman-type bandwidth equal to $0.79 * IQR * N^{-1/5}$, where *IQR* is the interquartile range and *N* is the sample size. This bandwidth has the advantage of being more robust to outliers than the most usual one based on the standard deviation (see, e.g., Pagan and Ullah, 1999).

⁵⁰Note that this term estimates the missing counterfactual $E[Y(1,0)]$. In practice, we first run a regression of the outcome on $\widehat{S}(0)$, $\widehat{S}(1)$ and the pscore terms for individuals in the treatment group ($T = 1$), and then use the estimated coefficients to predict $E[Y(1,0)]$ using all individuals and setting $\widehat{S}_i(1) = \widehat{S}_i(0)$. Finally,

(15) is predicted with the same specification. Compared to the *NTD* estimates, they imply a slightly larger lock-in effect for whites; while for blacks they are of essentially the same magnitude. For Hispanics, the magnitude is larger than the *NTD* estimates by about 16%. Compared to the *LCANTE* estimates, the *CANTE* estimate is smaller for blacks and Hispanics, and considerably so for whites. These differences across samples might as well reflect the heterogeneity among them, which in turn has implications for the plausibility of the different assumptions employed to estimate the parameters of interest. It is important to remember that the estimates of *LCANTE* and *CANTE* are based on different assumptions and that, without assuming a constant *CANTE* in the population, they estimate different parameters. Hence, the difference in the point estimates for *LCANTE* and *CANTE* may raise some doubts on the constant-*CANTE* assumption in this particular application.

Some general conclusions can be gathered from this empirical illustration of our methods. First, our estimates of *LCANTE* and *CANTE* suggest that the lock-in effect results in a negative causal mechanism for the effect of JC training on earnings 48-months after randomization, especially for Hispanics whose estimated net effects become positive, although still statistically insignificant. Second, the full set of estimates corroborate the high degree of heterogeneity that exist among whites, blacks and Hispanics in terms of their estimated total, net and mechanism effects from JC training. Lastly, it is useful to experiment with several specifications and assumptions when estimating net and mechanism effects, as one can learn about the plausibility of the assumptions made in order to arrive at causal net effects.

5.2 Non-random Assignment

We now move on to consider an application in which the treatment is not randomly assigned. In this case, we face the additional issue of controlling for self-selection into the treatment. As a result, we employ a selection on observables assumption such that we regard the treatment as randomly assigned conditional on a rich set of observed covariates. For this application, the data is a 10% random sample extracted from Pennsylvania’s Natality Data Sets from 1989 to 1991. The complete data, which includes all births, has been previously

$CANTE = E[\widehat{Y}(1,0)] - E[\widehat{Y}(0,0)]$, where the second term is also a prediction of the same model for controls ($T = 0$).

used and documented by Chay, Flores and Torelli (2005). We concentrate on single births.⁵¹ The large number of observations and the availability of a wide range of observable characteristics, including characteristics of both parents and previous birth history, makes the necessary assumption of selection on observables more plausible.

Our focus is on evaluating the extent to which smoking during pregnancy (the treatment) affects birth weight (the outcome) through a shorter gestation time (a mechanism). The consensus in the literature (e.g., Stein et al., 1983; Center for Disease Control and Prevention, 2001) is that smoking during pregnancy causally reduces birth weight, but the importance of specific mechanisms are not completely understood. In general, there might be two ways in which smoking during pregnancy affects birth weight: a shorter gestation time and intrauterine growth retardation (IGR). The importance of determining the causal relative importance of a channel is that particular policies that can minimize the negative effects of smoking during pregnancy may be considered. For instance, if gestation time is an important causal mechanism, drugs that lengthen gestation time may be deemed useful.

Table 2 presents the results for this application. Given the importance of satisfying the support condition in observational studies using the selection-on-observables assumption (e.g. Heckman, Ichimura and Todd, 1997; Dehejia and Wahba, 1999), we concentrate on a sample in the overlap region of the estimated propensity score (pscore) between the 1 percentile of the pscore values for the treated and 99 percentile of the pscore values for controls.⁵² Rows 1 through 4 present estimates of the total treatment effect (*ATE*). The first row shows, for reference, the unadjusted difference in means between the birth weight of individuals in the treatment versus control group. This unadjusted figure is consistent with a loss of -244.7 grams in birth weight resulting from smoking during pregnancy. To control for self-selection, a propensity score is estimated, and rows 2-4 includes it in different ways.⁵³ Rows 2 and 3 use simple OLS controlling for the estimated pscore (row 2) and its square and cube (row 3), resulting in essentially identical estimates of -202 grams, while row 4 employs a bias-adjusted simple matching approach (Abadie and Imbens, 2005) with one

⁵¹While we have access to the full data set that consists of 496,212 single births, we focus on a 10% random sample to allow obtaining all the estimates in a reasonable amount of time. Similar results as the ones presented in Table 2 are obtained with an alternative 10% random sample.

⁵²The 10% random sample consists of 49,524 individuals, of which 41,335 are contained within the overlap region.

⁵³The propensity score is estimated with a logit model that includes a large list of observed covariates such as mother's and father's demographic characteristics, previous birth history and some medical history. Squares and interactions of these variables are also included. This specification is similar to the one used in Chay, Flores and Torelli (2005).

match on the estimated pscore, resulting in a slightly larger effect of -204 grams. Taken together, the estimated total effects (ATE) that control for self-selection are consistent with a birth weight loss of about 200 grams, with an implied selection bias of about -45 grams when compared to the unadjusted difference (row 1).⁵⁴

The second panel of Table 2 (rows 5-10) presents estimates of the NTD parameter that controls directly for gestation time (S^{obs}) using different specifications. Recall that these estimates are presented for comparison, as they typically lack causal interpretation as estimates of the ATE , and correspond to a causal net effect under very stringent conditions. Rows 5-7 control separately for the estimated pscore and observed gestation, while rows 8-10 include observed gestation in the estimation of the pscore which controls for both pre-treatment variables and S^{obs} . These two alternative specifications result in essentially identical estimates of -196 grams, with the exception of the simple matching estimate of row 7 which is smaller at -192. Taking -196 grams as the benchmark NTD estimate, it suggests that approximately less than 6 grams can potentially be attributed to gestation time.

The next panel in Table 2 presents estimates of $LCANTE$. Similar to the previous application, the potential values of gestation time (i.e. $S(0)$ and $S(1)$) have to be estimated based on covariates X . We implement equations (9)-(10) using a single match on the estimated pscore. Rows 11-13 are based on the subpopulation with $\{\widehat{S}(1) = \widehat{S}(0)\}$, which results in a sample of 5,945 (14.4%).⁵⁵ Since within this subpopulation we have that $S^{obs} = \widehat{S}(1) = \widehat{S}(0)$, S^{obs} can be regarded as any other covariate, so we use a propensity score approach that includes observed gestation time in its estimation to control for S^{obs} and X . As before, we present estimates obtained using OLS (rows 11-12) and simple matching (row 13). All three estimates result in fairly similar $LCANTE$ estimates of about -192.5 grams. Compared to the ATE estimates, the $LCANTE$ estimates are consistent with a causal role of gestation of about 10 grams (5%), and also with a difference with NTD of about 4 grams.

Finally, row 14 presents an estimate of $CANTE$ using the second identification and estimation strategy. We implement it in a similar way to the previous application: we model (under Assumption 9) the first term in (20) as a linear function of $\widehat{S}(0)$, $\widehat{S}(1)$, and the esti-

⁵⁴We note that a birth weight loss of 200 grams due to smoking during pregnancy is a fairly consistent finding in the literature across years and states (Chay, Flores and Torelli, 2005).

⁵⁵Note that, contrary to the previous application, the post-treatment variable gestation time, measured in weeks, is sufficiently discrete to allow identifying a population for which $\{\widehat{S}(1) = \widehat{S}(0)\}$ exactly. If we were to obtain $\widehat{S}(1)$ and $\widehat{S}(0)$ using OLS, the resulting predictions are more continuous.

mated propensity score up to a cubic term.⁵⁶ Similarly, the second term in (20) is predicted with the same specification. The results from employing an Epanechnikov kernel are just slightly larger compared to the previously reported *LCANTE* estimates: the estimate in row 14 is -194 grams. The similarities between this estimate of *CANTE* and the *LCANTE* estimates suggest that the two parameters are close to each other, thus suggesting the validity of Assumption 3.

In sum, the results of this empirical application are consistent with a small causal role of gestation time as a channel through which smoking during pregnancy reduces birth weight. While the total effect is robust at about -200 grams, our results indicate that only about 8-10 grams (4-5%) works causally through a shorter gestation time. Therefore, other channels, such as IGR, are responsible for the bulk of the total effect of smoking on birth weight. Finally, we find that the *NTD* estimates overstate the causal net effect by about 2-4 grams in this application.

Summarizing the implementation of our methods in the two empirical applications, some patterns emerge. First, we do find that our methods are useful in estimating the causal average net effect and the causal average mechanism effect introduced in this paper. Second, the empirical applications illustrate different ways in which we can implement our estimators. Given that different implementations may yield useful information about the particular data at hand, it is desirable to try several such specifications in practice. Third, in both our applications, the *NTD* estimates are different from our causal estimates, although they still appear to be informative about the mechanism, as suggested by Rosenbaum (1984).

6 Conclusion

In this paper we discuss identification and estimation of causal average net and mechanism effects, which results from a natural step beyond the evaluation of the effects of a particular treatment or intervention. Using the concept of principal stratification (Frangakis and Rubin, 2002), our definitions of these parameters intuitively decompose the "total effect" of a treatment into the part that is causally due to a particular mechanism (or post-treatment variable) and that part that is net of such mechanism. These causal parameter definitions help clarify several discussions in the literature (mainly applied) that attempt to estimate net effects controlling for a particular post-treatment variable of interest.

⁵⁶See fn. 50 for more specific details.

In general, estimation of causal net effects is a difficult task given the data typically available to a researcher. Because of this, it is very important to make all assumptions necessary for a causal interpretation explicit in order to evaluate their plausibility in any particular application. We develop two identification strategies for the estimation of our parameters, both for the case of a randomly assigned treatment and the more common case of non-random assignment. In particular, the kind of assumptions we employ are in the spirit of the popular selection on observables assumptions (Rosenbaum and Rubin, 1983; Imbens, 2004). In addition, we present two different empirical applications that are used to illustrate the implementation of our methods.

Several natural extensions are left for future work. First, given that most of our assumptions are based on selection on observables, it is of interest to develop a set of alternative assumptions that lead to an identification and estimation strategy that allows for unobservables. A couple of possibilities come to mind, such as the construction of bounds for our parameters in the spirit of Manski (1990). Similarly, an analysis of the way in which additional information can be used to estimate our parameters, such as the availability of instrumental variables, is of interest but is beyond the scope of the current paper. We do plan to explore these avenues in future research.

References

- [1] Abadie, A., and Imbens, G. (2005) "Large Sample Properties of Matching Estimators for Average Treatment Effects" *Econometrica*, 74, 235-267.
- [2] Adams, P., Hurd, M., McFadden, D., Merrill, A. and Ribeiro, T. (2003) "Healthy, Wealthy and Wise? Tests for Direct Causal Paths Between Health and Socioeconomic Status" *Journal of Econometrics*, 112, 3-56.
- [3] Black, D. and Smith, J. (2004), "How Robust is the Evidence on the Effects of College Quality? Evidence from Matching." *Journal of Econometrics*, 121, 99-124.
- [4] Burghardt, J., Schochet, P., McConnell, S., Johnson, T., Gritz, R., et. al. (2001) "Does Job Corps Work? Summary of the National Job Corps Study" 8140-530. Mathematica Policy Research, Inc., Princeton, NJ.
- [5] Center for Disease Control and Prevention (2001) *Women and Smoking: A Report of the Surgeon General*.
- [6] Chay, K.; Flores, C. A. and Torelli, P. (2005) "The Association between Maternal Smoking during Pregnancy and Fetal and Infant health: New Evidence from United States Birth Records", mimeo, University of California, Berkeley.
- [7] Dawid, A. (1979) "Conditional Independence in Statistical Theory (with Discussion)" *Journal of the Royal Statistical Society, Series B*, 41, 1-31.
- [8] Dearden, L. Ferri, J. and Meguir, C. (2002), "The Effect of School Quality on Educational Attainment and Wages." *Review of Economics and Statistics*, 84, 1-20.
- [9] Dehejia, R. and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs", *Journal of the American Statistical Association*, 94, 1053-1062.
- [10] Ehrenberg, R., Jakubson, G., Groen, J., So, E., and Price, J. (2006), "Inside the Black Box of Doctoral Education: What Program Characteristics Influence Doctoral Students' Attrition and Graduation Probabilities?" NBER Working Paper 12065.
- [11] Flores-Lagunes, A., Gonzalez, A., and Neumann, T. (2006) "Learning But Not Earning? The Impact of Job Corps Training on Hispanic Youths", mimeo, University of Arizona.
- [12] Frangakis, C.E. and Rubin D. (2002) "Principal Stratification in Causal Inference" *Biometrics*, 58, 21-29.
- [13] Heckman, J. (1979) "Sample Selection Bias as a Specification Error" *Econometrica*, 47, 153-61.
- [14] Heckman, J.; Ichimura, H. and Todd, P. (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme", *Review of Economic Studies*, 64(4), 605-654.
- [15] Heckman, J., LaLonde, R. and Smith, J. (1999) "The Economics and Econometrics of Active Labor Market Programs" in O. Ashenfelter and D. Card (eds.) *Handbook of Labor Economics*. Elsevier Science North Holland, 1865-2097.
- [16] Hirano, K., Imbens, G. and Ridder, G. (2003) "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score" *Econometrica*, 71, 1161-89.

- [17] Holland, P. (1986) "Statistics and Causal Inference" *Journal of the American Statistical Association*, 81, 945-70.
- [18] Hotz, J., Imbens, G. and Klerman, J. (2006), "Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Reanalysis of the California Gain Program." *Journal of Labor Economics*, 24(3), 521-565.
- [19] Imbens, G. (2004) "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review" *Review of Economics and Statistics*, 84, 4-29.
- [20] Imbens, G. and Angrist, J. (1994) "Identification and Estimation of Local Average Treatment Effects" *Econometrica*, 62, 467-75.
- [21] Lechner, M. (2005) "A Note on Endogenous Control Variables in Evaluation Studies" Discussion paper 2005-16, University of St. Gallen.
- [22] Lechner, M. and R. Miquel (2005) "Identification of the Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions" Discussion paper, University of St. Gallen.
- [23] Manski, C. (1990) "Nonparametric Bounds on Treatment Effects" *American Economic Review Papers and Proceedings*, 80, 319-23.
- [24] Mealli, F. and Rubin, D. (2003) "Assumptions Allowing the Estimation of Direct Causal Effects" *Journal of Econometrics*, 112, 79-87.
- [25] Neyman, J. (1923) "On the Application of Probability Theory to Agricultural Experiments: Essays on Principles" Translated in *Statistical Science*, 5, 465-80.
- [26] Pagan, A. and Ullah A. (1999) *Nonparametric Econometrics*. Cambridge university Press.
- [27] Petersen, M., Sinisi, S., and van der Laan, M. (2006) "Estimation of Direct Causal Effects" *Epidemiology*, 17, 276-284.
- [28] Robins, J. (1986) "A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods—Application to Control of the Healthy Worker Survivor Effect" *Mathematical Modeling*, 7, 1393-1512.
- [29] Robins, J. and Greenland, S. (1992) "Identifiability and Exchangeability for Direct and Indirect Effects" *Epidemiology*, 3, 143-155.
- [30] Rosenbaum, P. (1984) "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment" *Journal of the Royal Statistical Society, Series A*, 147, 656-66.
- [31] Rosenbaum, P. and Rubin, D. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.
- [32] Rubin, D. (1974) "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies" *Journal of Educational Psychology*, 66, 688-701.
- [33] Rubin, D. (1980) "Discussion of 'Randomization Analysis of Experimental Data in the Fisher Randomization Test' by Basu" *Journal of the American Statistical Association*, 75, 591-93.

- [34] Rubin, D. (2004) "Direct and Indirect Causal Effects via Potential Outcomes" *Scandinavian Journal of Statistics*, 31, 161-70.
- [35] Rubin, D. (2005) "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions" *Journal of the American Statistical Association*, 100, 322-331.
- [36] Stein et al. (1983). "Smoking, Alcohol and Reproduction" *American Journal of Public Health*, 73 (10), 1154-1156.
- [37] Schochet, P., Burghardt, J. and Glazerman, S. (2001) "National Job Corps Study: The Impacts of Job Corps on Participants' Employment and Related Outcomes." 8140-530. Mathematica Policy Research, Inc., Princeton, NJ.
- [38] Simonsen, M. and Skipper, L. (2006), "The Costs of Motherhood: An Analysis Using Matching Estimators", *Journal of Applied Econometrics*, 21, 919-934.

Table 1. Random Assignment Application: Estimation of effect of the Job Corps training program on weekly earnings during quarter 16 after randomization and the effect of the program on same outcome controlling for observed post-treatment labor market experience¹

	Full Sample			Hispanic			White			Black		
	(N=9,105)	t-stat	N	(N=1,693)	t-stat	N	(N=2,533)	t-stat	N	(N=4,879)	t-stat	N
1 Original NCJS Estimates (LATE) ²	25.2	(0.00)		-15.1	(0.19)		46.2	(0.01)		22.8	(0.01)	
<i>Estimation of Intention to Treat Effects, ITT</i>												
2 Unadjusted Difference	15.6	(3.40)		-19.7	(-1.79)		31.3	(3.42)		20.3	(3.42)	
3 OLS using linear pscore	19.3	(4.18)		-7.3	(-0.67)		32.2	(3.48)		20.2	(3.38)	
4 OLS using up to cubic pscore	19.3	(4.18)		-7.0	(-0.64)		32.0	(3.46)		20.2	(3.40)	
<i>Estimation of "Net Treatment Difference" controlling for observed post-treatment experience, NTD</i>												
5 OLS using linear pscore and experience	24.1	(6.26)		8.1	(0.85)		36.0	(4.56)		22.1	(4.45)	
6 OLS using up to cubic pscore and experience	23.5	(6.10)		5.7	(0.61)		35.5	(4.50)		21.8	(4.40)	
7 OLS using linear pscore that includes experience in its estimation	23.0	(5.02)		5.9	(0.56)		35.4	(3.83)		21.6	(3.63)	
8 OLS using up to cubic pscore that includes experience in its estimation	22.7	(4.97)		5.0	(0.47)		35.1	(3.80)		21.9	(3.68)	
<i>Estimation of LCANTE using predicted S(0) and S(1) based on matching on the pscore. Focus on subpopulation with (predicted) S(0)=S(1) obtained using a Silverman-type bandwidth based on IQR: $h=0.79*IQR*N^{(-1/5)}$.</i>												
9 OLS using linear pscore that includes experience in its estimation	34.8	(3.31)	1273	9.2	(0.33)	344	59.5	(2.77)	446	25.5	(2.08)	834
10 OLS using up to cubic pscore that includes experience in its estimation	35.5	(3.38)	1273	10.2	(0.37)	344	59.7	(2.77)	446	24.8	(2.01)	834
<i>Estimation of LCANTE using predicted S(0) and S(1) based on matching on the pscore. We perform weighted least squares, where the weight is determined by a kernel function of the distance between (predicted) S(0) and S(1). Bandwidth is chosen by using a Silverman-type bandwidth based on IQR: $h=0.79*IQR*N^{(-1/5)}$.</i>												
11 Epanechnikov kernel, pscore, experience and their interaction	30.7	(3.18)	1273	15.2	(0.81)	344	70.1	(3.56)	446	24.1	(2.06)	834
<i>Estimation of CANTE using $E[Y(1,1) S(0), S(1), X]$ to predict $E[Y(1,0) S(0), S(1), X]$. $E[Y(0,0) S(0), S(1), X]$ is similarly predicted.</i>												
12 OLS using pscore-predicted S(0), S(1) plus up to cubic pscore	24.1	(6.33)		7.2	(0.76)		37.2	(4.84)		21.7	(4.45)	

¹ All estimates except the first row use the restricted sample that contains those who completed both a 48-month and baseline interview, and with non-missing information on the covariates used in the estimators. The sample sizes are indicated at the top of each column, unless otherwise indicated for a particular estimator.

² Based on differences in means adjusted for non-compliers. The NJCS did not report estimates based on weekly earnings during quarter 16 for Hispanics, whites and blacks; the figures correspond to a different measure of earnings: average weekly earnings during year 4 after randomization. Taken from Schochet, Burghardt, and Glazerman (2001, Tables VI.1 and D.14), which only reports p-values, shown in the "t-stat" column. For these estimates, the sample contains all individuals with earnings information at the 48-month interview, and does not correspond to the sample sizes at the top of each column. The sample sizes are 11,313 (full), 1,948 (Hispanics), 2,953 (whites) and 5,517 (blacks). See Flores-Lagunes, Gonzalez and Neuman (2006) for details on how the samples differ.

Table 2. Non-Random Assignment Application: Estimation of effect of smoking during pregnancy on birth weight and the effect of smoking during pregnancy on birth weight controlling for observed gestation (10% random sample of single births in Pennsylvania from 1989 to 1991)

	Estimate	t-statistic
<i>Estimation of Average Treatment Effects, ATE. Focus on subpopulation with (predicted) $S(0)=S(1)$ and overlap region of pscore between the 1 percentile of pscore for treated and 99 percentile of pscore for controls (N=41,335).</i>		
1 Unadjusted Difference	-244.7	(-35.32)
2 OLS using linear pscore	-201.8	(-26.96)
3 OLS using up to cubic pscore	-201.7	(-26.95)
4 Simple matching on pscore, 1 match	-203.8	(-19.73)
<i>Estimation of "Net Treatment Difference" controlling for observed gestation, NTD. Focus on subpopulation with (predicted) $S(0)=S(1)$ and overlap region of pscore between the 1 percentile of pscore for treated and 99 percentile of pscore for controls (N=41,335).</i>		
5 OLS using linear pscore and gestation	-196.5	(-30.75)
6 OLS using up to cubic pscore and gestation	-195.6	(-31.61)
7 Simple matching on pscore and gestation, 1 match	-191.6	(-22.34)
8 OLS using linear pscore that includes gestation in its estimation (N=41,276)	-196.0	(-26.55)
9 OLS using up to cubic pscore that includes gestation in its estimation (N=41,276)	-196.0	(-26.54)
10 Simple matching on pscore that includes gestation in its estimation, 1 match (N=41,276)	-196.0	(-19.36)
<i>Estimation of LCANTE using predicted $S(0)$ and $S(1)$ based on matching on the pscore. Focus on subpopulation with (predicted) $S(0)=S(1)$ and overlap region of pscore between the 1 percentile of pscore for treated and 99 percentile of pscore for controls (N=5,945)</i>		
11 OLS using linear pscore that includes gestation in its estimation	-192.7	(-11.82)
12 OLS using up to cubic pscore that includes gestation in its estimation	-192.8	(-11.82)
13 Simple matching on pscore that includes gestation in its estimation, 1 match	-192.4	(-7.49)
<i>Estimation of CANTE using $E[Y(1,1) S(0), S(1), X]$ to predict $E[Y(1,0) S(0), S(1), X]$. $E[Y(0,0) S(0), S(1), X]$ is similarly predicted. Focus on subpopulation with overlap region of pscore between the 1 percentile of pscore for treated and 99 percentile of pscore for controls. (N=41,335)</i>		
14 OLS using pscore-predicted $S(0)$, $S(1)$ plus up to cubic pscore	-194.3	(-26.71)