DISCUSSION PAPER SERIES

IZA DP No. 4237

Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment under Unconfoundedness

Carlos A. Flores
Alfonso Flores-Lagunes

June 2009

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Identification and Estimation of Causal Mechanisms and Net Effects of a Treatment under Unconfoundedness

Carlos A. Flores
University of Miami

Alfonso Flores-Lagunes
University of Florida
and IZA

Discussion Paper No. 4237
June 2009

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 4237
June 2009

ABSTRACT

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

An important goal when analyzing the causal effect of a treatment on an outcome is to understand the mechanisms through which the treatment causally works. We define a causal mechanism effect of a treatment and the causal effect net of that mechanism using the potential outcomes framework. These effects provide an intuitive decomposition of the total effect that is useful for policy purposes. We offer identification conditions based on an unconfoundedness assumption to estimate them, within a heterogeneous effect environment, and for the cases of a randomly assigned treatment and when selection into the treatment is based on observables. Two empirical applications illustrate the concepts and methods.

JEL Classification: C13, C21, C14

Keywords: causal inference, causal mechanisms, post-treatment variables, principal stratification

Corresponding author:

Alfonso Flores-Lagunes
Food and Resource Economics Department
University of Florida
P.O. Box 110240
Gainesville, FL 32611
USA
E-mail: alfonsofl@ufl.edu

* We thank Damon Clark, Jon Guryan, Kei Hirano, Hilary Hoynes, Sabine Kroger, Oscar Mitnik, participants at the 2006 annual meeting of the Society of Labor Economists, the University of Miami Labor Lunch, 2006 Economic Science Association meetings, 2006 Midwest Econometrics Group meeting, 2006 Latin American meetings of the Econometric Society, and seminar participants at the Industrial Relations Section (Princeton), Laval, Purdue, and Virginia Tech for useful comments and discussions. NSF funding under grants SES-0852211 and SES-0852139 is gratefully acknowledged. All errors are our own.

1 Introduction

The main purpose in the estimation of causal effects of a treatment or intervention is to estimate its total impact on a particular outcome. Commonly estimated parameters are the average treatment effect, the average treatment effect on the treated, and the marginal treatment effect.[1] In addition, it is of interest to estimate causal mechanisms through which the treatment or intervention works, and/or causal effects of the treatment on the outcome *net* of these mechanisms. Knowledge of these causal parameters 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, within a heterogeneous effect environment, identification and estimation based on an unconfoundedness or selection-on-observables assumption of the average causal mechanism through which a treatment affects an outcome and the average causal effect of the treatment net of this mechanism. Using the potential outcomes framework, we precisely define our estimands of interest, consider different assumptions that can be employed in their identification and estimation when only information on additional covariates is available, and analyze other related parameters mentioned elsewhere in the literature.

To briefly motivate the importance of understanding the mechanism through which a treatment works, consider the following example that is later employed as empirical illustration of the methods developed in the paper. When analyzing the causal effect of smoking during pregnancy on birth weight, it is of particular interest to determine what part of this causal effect 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 deemed helpful.

Several studies in economics are concerned with estimating mechanism effects and effects net of one or more mechanisms, many of which are based on unconfoundedness-type assumptions. For example, the literature on the effect of school quality on labor market outcomes recognizes that part of this effect may work through increasing years of schooling. To address this, Dearden et al. (2002) present results of the effect of school quality on wages with and without controlling for schooling to measure the total impact of school quality on wages and the effect that works through higher educational attainment. Similarly, Black and Smith (2004) use propensity score matching methods with and without including years of education in the propensity score specification. Another example is Simonsen and Skipper (2006), who estimate the effect of motherhood on wages in Denmark controlling for various mechanisms through which motherhood may affect wages.[2] They use a propensity score matching approach and dis-

[1] See, for instance, Heckman, Lalonde and Smith (1999) for a detailed discussion of these parameters.

[2] 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

cuss assumptions needed to estimate the total effect of motherhood on wages and the effect of motherhood on wages net of the mechanisms. As a final example, Ehrenberg et al. (2007) 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.[3] In general, every time causal effects of a treatment are estimated, it is natural to ask about the relative importance of different potential causal mechanisms through which the treatment works.[4]

A common problem in the existing literature attempting to estimate causal mechanisms of a treatment is that the parameters 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 of the estimates are not always made explicit.[5] To avoid this problem, we use the potential outcomes framework (Neyman, 1923; Rubin, 1974) to clearly define our parameters of interest and decompose the average (total) treatment effect into the average causal mechanism and the average causal effect net of that mechanism. In addition, to give a causal interpretation to our parameters we use the concept of principal stratification introduced in Frangakis and Rubin (2002) for estimating treatment effects controlling for a post-treatment variable (in our case the mechanism variable). The basic idea behind Frangakis and Rubin (2002) is to compare treated and control individuals based on the potential values of the post-treatment variable. As stressed in Rubin (2005), when drawing causal inferences it is very important to keep the distinction between observed values of a variable (e.g., observed gestation) and the potential values it represents (e.g., gestation if smoked during pregnancy or not). In the previously mentioned papers this important distinction is missing.

The following ideal situation provides intuition for the definition of our parameters and the challenges faced in their estimation. 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,

conditions than the private sector. Other channels they consider are working experience and occupation. They consider the use of exclusion restrictions for estimation of net effects, although for estimation of the total average effect they assume motherhood is unconfounded, i.e., random given a set of covariates.

[3] In 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.

[4] Another situation in which it is relevant estimating causal mechanisms is in the evaluation of government programs that are a combination of different services (e.g. they are "bundled treatments"), especially when policymakers aim at reforming them. For example, see the discussions in Meyer (1995) and Currie and Neidell (2007) in the context of unemployment insurance reforms and the Head Start program, both in the U.S. A related study is Card and Hyslop (2005) on the different incentives provided by Canada's Self Sufficiency Project.

[5] Simonsen and Skipper (2006) define their parameter before discussing its estimation. However, as we discuss later, they do not acknowledge explicitly the fact that the mechanism variable represents two different potential variables, and the relationship of their parameter to the total average treatment effect is not discussed.

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. We define the net average treatment effect as the difference in mean potential outcomes of this new experiment and the control treatment. If this counterfactual experiment is available, estimation of the net average treatment effect is straightforward by comparing the average outcomes of the individuals that took this new treatment and those in the control group. Therefore, intuitively, estimation of treatment effects net of a mechanism requires learning about a different treatment from the one we have at hand. This motivates the difficulty in estimating this kind of treatment effects since, unfortunately, the commonly available data may provide limited information about this counterfactual experiment. Given the usual trade-offs between data availability and assumptions, it is not surprising that estimation of causal net treatment effects requires stronger assumptions than estimation of total average effects.

In this paper we focus on identification strategies for our parameters based on an unconfoundedness assumption. Methods for estimating (total) average treatment effects based on unconfoundedness or selection-on-observables assumptions are important in economics and continue to receive considerable attention in the econometrics literature (e.g., Hahn, 1998; Heckman, Ichimura and Todd, 1998; Hirano, Imbens and Ridder, 2003; Abadie and Imbens, 2006). This assumption states that assignment to treatment is independent of the potential outcomes conditional on a set of covariates. For identification of our parameters, our unconfoundedness assumption states that the potential outcomes are independent of the potential values of the mechanism variable conditional on covariates. Contrary to identification of total average effects, this assumption alone does not yield identification of our parameters. We present two approaches for their estimation employing unconfoundedness. The first is based on a functional form assumption relating the potential outcomes of interest; while the second is based on estimation of the causal net average treatment effect for a particular subpopulation: those individuals for which the treatment does not affect the mechanism variable. We present each of these approaches for the case in which the treatment is randomly assigned and when selection into the treatment is based on a set of observable covariates.

For comparison, we also discuss a set of assumptions under which the usual approach of controlling for the observed value of the mechanism variable (e.g., Dearden et al., 2002; Black and Smith, 2004) can be interpreted as a causal net average treatment effect. We also illustrate the practical relevance of the estimation approaches in this paper employing two empirical applications. The first analyzes the importance of the "lock-in" effect of a major training program on participant's earnings using experimental data; while the second analyzes the importance of gestation as a mechanism for the effect of smoking during pregnancy on the incidence of low birth weight, under the assumption that smoking is random conditional on a rich set of covariates.

In cases when the unconfoundedness assumption employed in this paper in not tenable, we develop elsewhere results for nonparametric partial identification for the case of a randomly assigned treatment (Flores and Flores-Lagunes, 2009a), as well as for estimation employing instrumental variables when an instrument for the mechanism variable is available (Flores and Flores-Lagunes, 2009b). The paper is organized as follows. Section 2 reviews 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 the two empirical applications that illustrate the methods discussed in this paper. Concluding remarks are provided in the last section.

2 Related Literature

Our goal is to analyze two related effects: a causal mechanism through which a treatment affects an outcome, and the causal effect of the treatment net of this mechanism. To achieve this goal we employ the potential outcomes framework and, more specifically, build on literature related to the estimation of causal effects adjusting for covariates that are affected by the treatment. This literature relates to our goal since estimating the causal mechanism of a treatment implies accounting for variables that are observed after the treatment and that are affected by it.[6]

Rosenbaum (1984) analyzes the consequences of adjusting for covariates that are affected by the treatment using the potential outcomes framework. He concludes that estimators adjusting for these variables are generally biased, and specifies sufficient conditions under which controlling for such variables yields the average treatment effect (ATE).[7] Trivially, these conditions imply that the ATE can be identified when the post-treatment variables are not affected by the treatment, in which case they can be regarded as pre-treatment variables. Rosenbaum (1984) also defines the "net treatment difference" (NTD), a parameter that is estimated by simply adjusting for the observed value of the post-treatment variable and is argued to "provide insight into the treatment mechanism", even though it lacks causal interpretation.

More recently, Frangakis and Rubin (2002) introduced the concept of principal stratification to define causal effects when controlling for post-treatment variables in a variety of settings. Principal stratification, which will be further discussed in the following section, defines causal effects by comparing individuals with the same potential values of the post-treatment variable

[6] Note that to assess the importance of a potential mechanism through which a treatment works one needs to have a measurement of it. In practice, the mechanisms considered may depend on the availability of a variable measuring them. Since such variables are measured after the treatment, we indistinctly refer to them as "post-treatment variables" or "mechanisms".

[7] More recently, Imbens (2004) also warns about similar pitfalls when controlling for post-treatment variables affected by a treatment, while Lechner (2005) specifies more explicit conditions to assess the endogeneity bias introduced when controlling for variables influenced by the treatment. Both deal with situations in which interest lies on identification of the ATE.

under each of the treatment arms. In this paper, we define our estimands to have causal interpretation based on principal stratification.

Some of the work closer to ours is in Mealli and Rubin (2003) and Rubin (2004). Both papers 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 consider here. A direct effect corresponds to a causal effect of a treatment net of a post-treatment variable, while an indirect effect corresponds to the causal effect of a treatment that is mediated by another variable (i.e., a mechanism). The main goal in both papers is to illustrate that the use of principal stratification clarifies the concepts of causality when controlling for post-treatment variables, and that other methods that ignore potential values of variables influenced by the treatment can potentially lead to misleading causal conclusions.

Even though the concepts of direct and indirect effects in the previous papers are similar to the causal mechanism and causal net effects we define and analyze here, there are important differences between those papers and ours. First, the relationship of the concepts of direct and indirect effects to the (total) ATE is not discussed in those papers, while the parameters to be presented here intuitively decompose the ATE into two effects (a mechanism and a net effect). Second, as we explain later, the concept of direct effect as defined in those papers is a special case of our causal net average treatment effect for a specific subpopulation. Third, we formally discuss identification and estimation under different assumptions and present empirical applications, 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), and Pearl (2001) in artificial intelligence. Robins and Greenland (1992) make a similar distinction of direct and indirect effects and present conditions under which they can be estimated. Pearl (2001) introduces the concepts of "controlled" and "natural" direct effects[8] and discusses their estimation, whereas Petersen, Sinisi and van der Laan (2006) provide conditions for estimation of the natural direct effect. The present paper differs from this literature in important ways. Most notably, those papers do not employ the concept of principal stratification we employ here and do not distinguish the potential values of the post-treatment variable in their assumptions for identification. In our view, this obscures the assessment of the plausibility of the assumptions and, as discussed in Rubin (2005), may lead to invalid causal conclusions.[9]

[8] These two concepts are discussed later in section 3.3.

[9] Robins and Greenland (1992) is actually an application of a more general literature on the estimation of dynamic causal effects (e.g. Robins (1986) in epidemiology and more recently Lechner and Miquel (2005) in economics). In this literature, 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. We abstract from modeling dynamics explicitly, so we concentrate on a static model of causal effects as in Robins and Greenland (1992).

Finally, two recent papers in economics–Lee (2009) and Zhang, Rubin and Mealli (2006)–focus on the problem of estimating the effect of a randomly assigned training program on wages taking into consideration the fact that wages are only observed for those individuals who are employed, which leads to a sample selection problem because employment status may also be affected by the training program. This problem is related to ours since employment status may be regarded as a mechanism through which training affects wages. Zhang, Rubin and Mealli (2006) use a principal stratification approach to analyze this problem and argue that the relevant average treatment effect of training on wages is for the subpopulation of individuals who would be employed whether they received training or not. They derive bounds for this effect and propose a Bayesian approach for its estimation. Similarly, Lee (2009) proposes a trimming procedure that yields sharp bounds for the average treatment effect for those individuals who would be employed whether trained or not.

An important distinction of our work with those two papers is that they keep the focus on estimation of the average treatment effect of training on wages controlling for employment status (an intermediate variable), while we focus on the more general problem of decomposing the part of the effect of a treatment on an outcome that works through a mechanism or intermediate variable. Naturally, in some cases these two objectives will coincide. In fact, as will be discussed later in section 4.3, the average treatment effect for those who would be employed whether trained or not equals our definition of the causal net average treatment effect for this subpopulation. Other differences are that our framework allows the mechanism variable to be polychotomous or continuous (in their case, employment status is binary) and that in the problem analyzed by Lee (2009) and Zhang, Rubin and Mealli (2006) the observability of the outcome (wages) depends on an intermediate variable (employment status), while in our framework the outcome is always observed. Lastly, while those papers focus mainly on partial identification (especially Lee, 2009), the focus in this paper is on point identification of our parameters. Partial identification results for the parameters discussed in the present paper can be found in Flores and Flores-Lagunes (2009a).

Finally, note that most of the work described in this section considers the case when the treatment is randomly assigned. In this paper, our framework starts under that same assumption but is then extended to the case when selection into the treatment is based on observable variables. Flores and Flores-Lagunes (2009b) considers the use of exclusion restrictions in the form of instrumental variables to identify the parameters defined in the next section.

3 The Estimands of Interest

3.1 Definition of Estimands

We employ the potential outcomes framework (Neyman, 1923; Rubin, 1974). 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 ($T_i = 0$). We are interested on the effect of the treatment T on an outcome Y. Let $Y_i(1)$ denote the potential outcome for individual i under treatment and $Y_i(0)$ denote the potential outcome under the control treatment. The (population) average treatment effect is hence given by $ATE = E[Y(1) - Y(0)]$.[10] We are interested on analyzing the part of the ATE that works through a mechanism variable S, and the causal effect of T on Y net of the effect through S. Since S is affected by the treatment, we must 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 exposed to the treatment, and $S_i(0)$ represents the value she would get if exposed to the control treatment.[11] For each unit i, we observe the vector $\left(T_i, Y_i^{obs}, S_i^{obs}\right)$, where $Y_i^{obs} = T_i Y_i(1) + (1 - T_i) Y_i(0)$ and $S_i^{obs} = T_i S_i(1) + (1 - T_i) S_i(0)$. It is important to stress the fact that S^{obs} represents two different potential variables: $S(1)$ for treated units and $S(0)$ for controls.[12]

In our case, it is convenient to let the potential outcomes be a function of the mechanism variable S. For each individual i, define the "composite" potential outcomes $Y_i(\tau, \zeta)$, where the first argument refers to one of the treatment arms ($\tau \in \{0,1\}$) and the second argument represents one of the potential values of the post-treatment variable S ($\zeta \in \{S_i(0), S_i(1)\}$). Using this notation, we can consider the following composite potential outcomes for any given individual:

1. $Y_i(1, S_i(1))$: this is the potential outcome the individual would obtain if she received treatment and post-treatment variable level $S_i(1)$. It includes the total effect of receiving treatment on Y (i.e., through S or not). This is exactly the potential outcome $Y_i(1)$ under the treatment.

2. $Y_i(0, S_i(0))$: this is the potential outcome when no treatment is received and the post-treatment variable value is $S_i(0)$. It is the outcome an individual would obtain if the

[10] Another 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 and other parameters, as is the case in section 5.1.

[11] Note that S is not restricted to be binary.

[12] We 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.

treatment is not given to her and if the value of her post-treatment variable is not altered either. This is exactly the potential outcome $Y_i(0)$ under the control treatment.

3. $Y_i(1, S_i(0))$: this is the potential outcome the individual would receive if she were exposed to the treatment but kept the level of S she would obtain had not been treated. In other words, it is the outcome the individual would get if we were to give her the treatment but held the value of her post-treatment variable at $S_i(0)$. As a result, 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.[13]

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

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

(b) $Y_i(1, S_i(1)) - Y_i(1, S_i(0))$: 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 $Y_i(1, S_i(0))$ 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*.

(c) $Y_i(1, S_i(0)) - Y_i(0, S_i(0))$: 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*.

Given these comparisons, we can decompose the individual total treatment effect in (a) into the part of the effect due to a change in S because of a change in T (mechanism effect) and the part of the effect holding S fixed at $S(0)$ (net effect):

$$ITTE = [Y_i(1, S_i(1)) - Y_i(1, S_i(0))] + [Y_i(1, S_i(0)) - Y_i(0, S_i(0))]. \quad (1)$$

The population (total) average treatment effect (ATE) can be decomposed in a similar way as:

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

[13] For completeness, note that $Y_i(0, S_i(1))$ is the potential outcome the individual would obtain when the treatment is not given to her but she receives a value of the post-treatment variable equal to $S_i(1)$.

As 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 term shows the part of the average effect holding S fixed at $S(0)$.[14] A decomposition similar to (2) appears in Pearl (2001) based on what he calls the Natural Direct Effect, discussed in Section 3.3.

It is clear from the decomposition in (2) that we need to make treatment comparisons adjusting for the post-treatment variable S that is affected by the treatment. In order to causally interpret our parameters of interest, we employ the concept of principal stratification developed in Frangakis and Rubin (2002) (hereafter FR). The idea is to define the set of "comparable individuals" based on the potential values of the post-treatment variable.[15] 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 $\{S(0) = s_0, S(1) = s_1\}$, where s_0 and s_1 are generic values of $S(0)$ and $S(1)$, respectively. A principal effect with respect to a principal strata is defined as 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.[16]

Based on FR, we condition on the principal strata $\{S(0) = s_0, S(1) = s_1\}$ in order to give a causal interpretation to our parameters. Write the ATE controlling for $S(0)$ and $S(1)$ as

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

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

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

We define the (causal) net average treatment effect or $NATE$ as:

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

and the (causal) mechanism average treatment effect or $MATE$ as:[17]

$$MATE = E\{E[Y(1, S(1)) - Y(1, S(0))|S(0) = s_0, S(1) = s_1]\}. \tag{6}$$

[14] While we consider a decomposition of the total effect based on one mechanism of interest, it is possible to extend the decomposition to accommodate more than one mechanism. See Flores and Flores-Lagunes (2009a).

[15] 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.

[16] FR's idea of principal stratification is closely related to the local average treatment effect interpretation of instrumental variables in Imbens and Angrist (1994). For example, in their terminology, 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 this group $\{S(0) = 0, S(1) = 1\}$, where S is an indicator of actual treatment reception.

[17] Although we could have used the terms "direct effect" and "indiret effect" to define our parameters, we

3.2 Discussion of Estimands

An intuitive way to think about our estimands is to consider $Y(1, S(0))$ as the potential outcome of an alternative counterfactual experiment in which the treatment is the same as the original one but blocks the effect of T on S by holding S fixed at $S_i(0)$ for each individual i. The $NATE$ for individual i is the difference between the outcome of this alternative treatment, $Y_i(1, S_i(0))$, and $Y_i(0, S_i(0))$ from the original control treatment. Similarly, her $MATE$ is given by the difference in the potential outcomes of the original treatment and the alternative one.

An important property of $NATE$ in (5) is that 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 $S(0)$, the treatment may still affect the way in which S affects the outcome, and this is counted as part of $NATE$. To illustrate this point, consider one of our empirical applications in which we analyze the lock-in effect as a causal mechanism of a job training program, i.e. the labor market experience lost due to participation in the program. If participants lose substantial labor market experience due to the program, and this negatively affects their future earnings, a policy maker may want to change the program to be as the original one but without affecting labor market experience (i.e., holding experience fixed at $S(0)$). In this case the policy maker would like to know the average effect of this alternative training program on future earnings. This effect 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 program on how experience affects wages, i.e., the program's effect on the returns to experience. $NATE$ takes this into account, correctly measuring what the effect of this alternative treatment (training program) would be.

We argue that including the effect of T on how S affects Y (i.e., returns to S) in $NATE$ is more relevant from a policy perspective, compared to a different parameter that holds constant the way S affects Y. The reason is that a policy maker typically has some degree of control over S, while very rarely over how S affects Y. 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. by offering training while on the job or shortening the time to completion of the program), but it seems unlikely that they could influence the (potentially) different returns to experience that the market awards to trained versus non-trained individuals.[18] Our argument is consistent with the notion of a "treatment"

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 3.3. Second, our names make clear that these effects are considered with respect to a particular mechanism S. Strictly speaking, a "pure" direct effect would have to net out all possible mechanisms through which the treatment may affect the outcome.

[18] One potentially interesting case where the policymaker might have some degree of influence on how S affects Y is when general equilibrium effects due to the treatment are present.

being an intervention that can be potentially applied to each individual (e.g., Holland, 1986).

As a final remark about $NATE$ and $MATE$, we note that their definitions conform to intuition in the following two extreme cases. First, consider the situation in which all the effect of T on Y works exclusively through S for the entire population. In this case $Y(1, S(0)) = Y(0, S(0))$ and, as expected, $NATE = 0$ and $MATE = ATE$ from equations (5) and (6), respectively. Second, consider the situation in which none of the effect of T on Y is through S, in which case $NATE$ should equal ATE and $MATE$ should be zero. This can arise due to two reasons: either S does not affect Y (even though S may be affected by T) and thus $\{S(1), S(0)\}$ is independent of $\{Y(1), Y(0)\}$; or T simply does not affect S and thus $S(1) = S(0)$. Regardless of the reason, the consequence is that $Y(1, S(1)) = Y(1, S(0))$ and thus (5) and (6) imply $NATE = ATE$ and $MATE = 0$, respectively. This desirable property is not shared by some of the parameters available in the literature.

3.3 Relation of the Estimands to Other Parameters in the Literature

As discussed in section 2, Rosenbaum (1984) defines the NTD. This parameter is characterized by conditioning on the observed post-treatment variable and, without further assumptions, has no causal interpretation when the post-treatment variable is affected by the treatment. It can be written as $NTD = E\{E[Y(1) - Y(0) | S^{obs}]\}$. The reason for NTD's lack of causal interpretation is that it compares individuals with the same values of S^{obs}. Since S^{obs} represents two different potential variables, $S(1)$ and $S(0)$, units with the same value of S^{obs} are generally not comparable. This point is further discussed and illustrated in Mealli and Rubin (2003), Rubin (2004), and Rubin (2005).[19] In contrast, by conditioning on principal strata, $NATE$ explicitly accounts for the possibility that the post-treatment variable is affected by the treatment. Furthermore, our parameters effectively decompose the ATE into causal mechanism and net effects (see (4)).

Although both our estimands and the concepts of direct and indirect effects in Mealli and Rubin (2003) and Rubin (2004) rely on the idea of principal stratification and thus can be interpreted as causal effects, they differ in other aspects. Mealli and Rubin (2003) define a direct effect as a comparison of $Y(1)$ and $Y(0)$ within the stratum for which $S(0) = S(1) = s$, which implies $Y(1, S(1)) = Y(1, S(0))$. Using our notation in (3), we can write their direct effect as $DE(s) = \tau(s, s)$, which corresponds to $NATE$ in (5) defined *for this particular subpopulation or strata*. More generally, we can define the direct average effect as $DAE = E[\tau(s, s)]$, which is the average of the direct effects over the possible values s of S. Note that, unless $NATE$ is constant in the population, DAE will differ from $NATE$. Moreover, DAE does not decompose the

[19] Another 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 $E[Y(1) | S^{obs}] = E\left[Y^{obs} | T = 1, S^{obs}\right]$. However, in general, we can infer something about the treatment assignment T based on S^{obs} and hence the assumption fails. See Rubin (2005) for further discussion.

ATE in the way $NATE$ does because DAE ignores 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 subpopulations for which there is no mechanism effect. For these reasons, $NATE$ and $MATE$ are more general and, in our view, more relevant for policy purposes.

There are other parameters related to $NATE$ that have been used in the epidemiology literature: the controlled direct effect (CDE) and the natural direct effect (NDE).[20] The CDE at a specific value $\overline{s}$ of S can be written as $CDE = E[Y(1, S(1) = \overline{s}) - Y(0, S(0) = \overline{s})]$. The CDE gives the average difference between the counterfactual outcome under the two treatment arms controlling for the value of the mechanism variable at $\overline{s}$. While this parameter may be informative in some applications, in our view has some undesirable features for the estimation of net effects. First, it does not decompose the ATE into a net and a mechanism effect in the way $NATE$ and $MATE$ do.[21] Second, since neither of the two potential outcomes used in the definition of CDE necessarily correspond to the observed outcome (Y^{obs}) for any particular individual, its estimation requires stronger assumptions than the ones used for estimation of $NATE$, where at least one of the potential outcomes ($Y(0, S(0))$) is observed for some individuals.[22] Lastly, using the CDE to estimate net effects has the undesirable property that, even if in fact the treatment does not affect the mechanism variable S, the ATE may be different from the CDE if there is heterogeneity in the effect of T on Y along the values of S. Conversely, as previously discussed for our parameters, $NATE = ATE$ and $MATE = 0$ in this case.

The NDE used in epidemiology can be written as $E[Y(1, S(0)) - Y(0, S(0))]$. Hence, this parameter is similar to $NATE$ in (5) with the subtle but important difference that $NATE$ conditions on principal strata in order to retain causal interpretation. This distinction becomes crucial when stating and evaluating the assumptions needed for estimation.[23]

4 Identification and Estimation of the Parameters of Interest

In this section we discuss identification and estimation of the parameters $NATE$ and $MATE$ defined in section 3.1. We focus our attention on $NATE$ since, by definition, we can obtain $MATE = ATE - NATE$. We start by discussing in section 4.1 the type of assumptions needed to interpret the standard approach of directly controlling for S^{obs} as an estimate of $NATE$. Unfortunately, these assumptions are too strong to be useful in practice. Next, we

[20] See, for instance, Pearl (2000) and Petersen, Sinisi and van der Laan (2006).

[21] For example, we could write the ATE as: $ATE = E[Y(1, S(1)) - Y(1, S(1) = \overline{s})] + CDE + E[Y(0, S(0) = \overline{s}) - Y(0, S(0))]$. The first term gives the average effect of giving the treatment to the individuals and moving the value of the post-treatment variable from $\overline{s}$ to $S(1)$. The third term represents the average effect of giving the control treatment to the individuals and moving the value of the post-treatment variable from $S(0)$ to $\overline{s}$. These two effects are hard to interpret as mechanism effects of T on Y through S.

[22] See following section for details.

[23] For further discussion on the importance of conditioning on principal strata see Rubin (2004, 2005) and Mealli and Rubin (2003).

present two different estimation strategies based on an unconfoundedness assumption for each of two treatment-assignment mechanisms. We first consider the situation 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. We then discuss the case in which the treatment is assumed to be random given a set of observed covariates.

Regardless of the mechanism used to assign the treatment, identification and estimation of $NATE$ 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: S^{obs} represents $S(1)$ for treated units and $S(0)$ for control units. This implies that the principal strata $\{S(0) = s_0, S(1) = s_1\}$, which is necessary for a causal interpretation of $NATE$, is unobserved. Note that S can be regarded as an outcome, and thus the distribution of the principal strata equals the joint distribution of the potential outcomes $\{S(1), S(0)\}$, which is not easily identifiable (e.g., Heckman, Smith and Clements, 1997). The second challenge is that a key potential outcome needed for estimation of $NATE$, $Y_i(1, S_i(0))$, is generally not observed–this is in contrast to the case of estimation of the ATE, where only one of the relevant potential outcomes is missing for every unit. In an ideal situation in which we could perform the alternative counterfactual experiment and observe $Y(1, S_i(0))$ for some units, none of these two challenges would arise and estimation of $NATE$ would be straightforward.[24] Despite the missing data challenges that result from the unavailability of the alternative counterfactual experiment, we can still impose assumptions under which $NATE$ can be identified from the available data. We present in sections 4.2 and 4.3 two strategies for its estimation.

4.1 Assumptions under which controlling directly for S^{obs} yields $NATE$

It is important to state conditions under which the standard approach of controlling for the observed value of the post-treatment variable (S^{obs}), and possibly a set of covariates X, yields $NATE$. Examples of the use of this approach are Dearden et al. (2002) and Black and Smith (2004). It turns out that the kind of assumptions needed are very strong–certainly stronger than those we present in the following sections to identify our parameters.

Consider the following parameter that is representative of the standard approach, where the second line uses the fact that S^{obs} represents $S(0)$ or $S(1)$ depending on the treatment received:

$$\begin{aligned} \gamma &= E\{E[Y^{obs}|T=1, S^{obs}=s, X=x] - E[Y^{obs}|T=0, S^{obs}=s, X=x]\} \\ &= E\{E[Y^{obs}|T=1, S(1)=s, X=x] - E[Y^{obs}|T=0, S(0)=s, X=x]\}. \end{aligned} \quad (7)$$

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

A set of sufficient conditions under which $\gamma = ATE$ are (Rosenbaum, 1984): (i) $S(1) = S(0)$ for all subjects in the population ("unaffected post-treatment variable"), and (ii) the treatment assignment is ignorable in the sense that $\{Y(1), Y(0)\} \perp T|X$ and $0 < \Pr(T = 1|X) < 1$ for all X.[25] Intuitively, the issue when estimating the ATE based on (7) is that the outer expectation should be taken with respect to the distribution $\Pr(S(1)|X)$ for the first term and with respect to $\Pr(S(0)|X)$ for the second. As a result, if S is affected by T, bias will arise from averaging both terms over $\Pr\left(S^{obs}|X\right)$ instead. In other words, looking at units with the same values of S^{obs} in fact compares treated units with $S(1) = s$ to control units with $S(0) = s$, which are in general not comparable.[26] Condition (i) implies that $S^{obs} = S(1) = S(0)$, ensuring that the averaging is over the correct distribution; nonetheless, this condition is too strong since it rules out an effect of T on S.

Unfortunately, the same conditions are needed to have $\gamma = NATE$. Even if we were to assume that people in different strata are comparable conditional on X,[27] we still need to assume that $S_i(1) = S_i(0) = s$ for all units in order to have $Y_i(1, S_i(0)) = Y_i(1, S_i(1))$ and thus deal with the problem that $Y_i(1, S_i(0))$ is unobserved. Only then could we have $\gamma = NATE$. In a linear regression context, the fact that $Y_i(1, S_i(0))$ is generally unobserved implies that even if all explanatory variables in the regression $Y^{obs} = a + bT + cS^{obs} + d'X + u$ are uncorrelated to the error term u (e.g., if T and S are randomly assigned), b equals $NATE$ only if $S_i(1) = S_i(0) = s$ for all units. However, this condition rules out the role of S as a mechanism by assumption. In the following sub-sections we present weaker assumptions that allow us to estimate $NATE$ and $MATE$.

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

We first consider the case in which individuals are randomly assigned to the treatment. We keep the following assumption until our discussion of non-random treatment assignment in section 4.4.

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

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 Assumption 1 implies $Y(1, S(1)), Y(0, S(0)), Y(1, S(0)) \perp T|\{S(1), S(0)\}$, so that potential outcomes are independent of the treatment given the principal strata.[28]

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

[26] Yet another 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).

[27] In which case the groups with $\{T = 1, S(1) = s, X = x\}$ and $\{T = 0, S(0) = s, X = x\}$ would be comparable.

[28] See, for instance, Lemma 4 in Dawid (1979).

Let us start by considering the challenge that the principal strata $\{S(0) = s_0, S(1) = s_1\}$ is not observed. Note that identification of the principal strata is difficult since it entails determining the effect of the treatment T on the intermediate outcome S for every individual using only the marginal distributions of $S(1)$ and $S(0)$ for treated and controls, respectively. In this paper we follow an approach analogous to the commonly-used selection on observables framework in program evaluation (e.g., Imbens, 2004) and assume that the principal strata is independent of the potential outcomes given a rich set of covariates X (unconfounded principal strata).

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

Assumption 2 implies that individuals in different strata are comparable once we condition on a set of covariates X, ruling out the existence of variables not included in X that simultaneously affect the principal strata an individual belongs to and her potential outcomes (i.e., confounders). Assumption 2 further implies that by conditioning on X we rule out confounders of the relationship between (i) each of the potential values of S ($S(1)$ and $S(0)$) and the potential outcomes, and (ii) any function of $S(1)$ and $S(0)$ and the potential outcomes. Finally, Assumptions 1 and 2 imply that $Y(1, S(1)), Y(0, S(0)), Y(1, S(0)) \perp \{T, S(1), S(0)\} | X$, so that control and treated units in different strata, but with the same values of covariates, are comparable.[29]

The other challenge in the estimation of $NATE$ is making inferences about a potential outcome that is usually not observed, $Y(1, S(0))$. One approach is to use the information in $Y(1, S(1))$, and possibly $Y(0, S(0))$, to learn about $Y(1, S(0))$. This can be done in many different ways, with the specific assumption to be made depending on what is judged plausible in the particular application at hand. We present here one assumption to illustrate the approach. Suppose that the conditional expectations of the potential outcomes $Y(1, S(0))$ and $Y(1, S(1))$ have the same functional form in terms of $\{X, S(0)\}$ and $\{X, S(1)\}$, respectively, but the former sets $S(1) = S(0)$. As a simple example, let $E[Y(1, S(1))]$ be of the form $E[Y(1, S(1)) | S(1), X] = a_1 + b_1 S(1) + c_1 X$. Then, this assumption implies that $E[Y(1, S(0)) | S(0), X] = a_1 + b_1 S(0) + c_1 X$. We can state this assumption more generally as follows:

Assumption 3 Suppose we can write $E[Y(1, S(1)) | S(1) = s_1, X = x] = f_1(S(1), X)$. Then, assume

$$E[Y(1, S(0)) | S(0) = s_0, X = x] = f_1(S(0), X).$$

[29] Note the importance of stating the assumptions used in terms of principal strata as opposed to using simply S, as commonly done in the literature (e.g., Petersen, Sinisi and van der Laan, 2006; Simonsen and Skipper, 2006). Principal strata is not affected by T, and it acknowledges the fact that S represents two potential variables: $S(1)$ and $S(0)$. If we were to use S instead of the principal strata in Assumption 2, its interpretation (which is needed to gauge its plausibility in practice) would be obscured by the fact that S is affected by the treatment.

We offer a few comments on this assumption. First, Assumption 3 directly acknowledges that we are trying to learn about a counterfactual treatment based on the information available on the original treatment. Second, we clarify that regarding $Y(1, S(0))$ as the outcome of the counterfactual treatment does not imply Assumption 3. The definition of $Y_i(1, S_i(0))$ implies that $Y_i(1, S_i(0)) = Y_i(1, S_i(1))$ for those units with $S_i(1) = S_i(0)$; however, for those with $S_i(1) \neq S_i(0)$ it is not necessarily the case that $Y_i(1, S_i(0))$ has the same functional form as $Y_i(1, S_i(1))$ but setting $S_i(1) = S_i(0)$. Finally, note that Assumption 3 implies that the covariates X and the mechanism variable S affect the outcome in the same way in both the original and counterfactual treatments.

Under Assumptions 1-3 we can identify $NATE$ by writing it as a function of observed variables as:

$$\begin{aligned} NATE &= E\{E[Y(1,S(0)) - Y(0,S(0))|S(0)=s_0, S(1)=s_1, X=x]\} \\ &= E\{E[Y(1,S(0))|S(0)=s_0, S(1)=s_1, X=x]\} \\ &\quad -E\{E[Y(0,S(0))|S(0)=s_0, S(1)=s_1, X=x]\} \\ &= E\{E[Y(1,S(0))|S(0)=s_0, X=x]\} - E\{E[Y(0,S(0))|S(0)=s_0, X=x]\} \\ &= E\{f_1(S(0), X)\} - E\left\{E\left[Y^{obs}|T=0, S^{obs}=s_0, X=x\right]\right\} \end{aligned} \tag{8}$$

where we have used Assumption 2 in the third equality, Assumptions 1 and 3 in the last equality, and we have that $E[Y(1,S(1))|S(1)=s_1, X=x] = f_1(S(1), X) = E[Y^{obs}|T=1, S^{obs}=s_1, X=x]$.

In practice, this identification strategy can be implemented as follows: (i) estimate a model for $E[Y^{obs}|T=1, S^{obs}=s_1, X=x] = f_1(S(1), X)$; (ii) compute $E[Y(1,S(0))|S(0)=s_0, X=x] = f_1(S(0), X)$ based on the model in (i); (iii) estimate $NATE$ based on (8) and $MATE = ATE - NATE$. For steps (i) and (ii) a simple way to proceed is to run a linear regression of Y^{obs} on $S(1)$ and X for treated units and evaluate this estimated model on $S(1) = \widehat{E}[S_i(0)]$. One may allow this function to be more flexible by employing a polynomial series expansion of $S(1)$ and interactions with the covariates, for instance.

4.3 Estimation of $NATE$ Based on a Specific Subpopulation

In this section we present the second approach to estimate $NATE$ by focusing on a particular subpopulation or principal strata: those for which T does not affect S, so that $S_i(1) = S_i(0)$. For them we have that $Y_i(1, S_i(0)) = Y_i(1, S_i(1))$ and hence $Y_i(1, S_i(0))$ is in fact observed for those receiving treatment. Therefore, any non-zero causal effect $Y_i(1, S_i(1)) - Y_i(0, S_i(0))$ in this subpopulation is due to factors different from the change in S caused by T. For this particular subpopulation with $S_i(1) = S_i(0)$, we define its local $NATE$ (hereafter $LNATE$) as:

$$LNATE = E\{E[Y(1,S(0)) - Y(0,S(0))|S(0)=s, S(1)=s]\} = E\{\Delta(s)\} \tag{9}$$

where $\Delta(s) = E[Y(1, S(0)) - Y(0, S(0))|S(0) = s, S(1) = s]$ is the local $NATE$ in the stratum $S(1) = S(0) = s$.[30] The key to regard $LNATE$ 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). $LNATE$ is similar to the "direct effect" discussed in Mealli and Rubin (2003) and Rubin (2004). More precisely, $LNATE$ equals the direct average effect (DAE) discussed in section 3.3, which is simply the average of the direct effects for all the stratum for which $S(1) = S(0) = s$. Hence, the direct effect is a local $NATE$ since it is defined for a specific subpopulation.

The $LNATE$ equals the local average treatment effect for the subpopulation for which the treatment does not affect the mechanism variable ($LATE^{sub}$), since $Y_i(1, S_i(0)) = Y_i(1, S_i(1))$ and there is no mechanism effect by definition. There is precedent in the literature on the estimation and practical importance of local average treatment effects. In particular, Imbens and Angrist (1994) interpret instrumental variables (IV) estimators as estimators of a local average treatment effect ($LATE$). The importance of $LATE$ in economics is discussed, for instance, in Imbens (2009). In addition, note that the parameter of interest in Lee (2009) and Zhang, Rubin and Mealli (2006) is a special case of $LNATE$. They focus on the (local) average treatment effect of training on wages for those individuals who would be employed whether trained or not. This is a subset of the subpopulation for which training does not affect employment status, since there may be unemployed individuals for which training does not affect their employment status. Hence, the local average treatment effect considered in those papers equals the $LNATE$ for those individuals employed whether trained or not. We can write this parameter as $LNATE_{S=1} = E[Y(1) - Y(0)|S(0) = 1, S(1) = 1]$, where S stands for employment status.[31]

In practice, knowledge about a subpopulation with $S_i(1) = S_i(0)$ may or may not be available. In some situations, the nature of the treatment and post-treatment variables conveys knowledge about a subpopulation for which T does not affect S. An illustration of this is when a law or regulation (i.e., a "natural experiment") restricts the effect of the treatment on the post-treatment variable and results on $S(1)$ being equal to $S(0)$ for a known group. For example, suppose interest lies on estimating the importance of trans-fat consumption (S) as a mechanism through which consuming fast-food (T) affects overweight incidence (Y). The subpopulation of interest for estimation of $LNATE$ is that for which T does not affect S. Given that the city of New York banned the use of trans fats in restaurant cooking in 2006, individuals in New York city represent a group that can be employed to estimate $LNATE$.[32] In such cases, one can restrict attention to the subpopulation with $S_i(1) = S_i(0)$ for estimation of $LNATE$.

[30] Note that the outer expectation in $LNATE$ is taken over all strata with $S(1) = S(0) = s$. For instance, in the context of a binary post-treatment variable, $LNATE$ would be the average of the net effects for the stratum with $S(0) = S(1) = 0$ and $S(0) = S(1) = 1$.

[31] Recall that, due to the nature of the problem analyzed in those papers, they do not observe wages for those individuals who would be unemployed whether trained or not, as discussed in section 2.

[32] We thank Jinyong Hahn for suggesting this example.

Under Assumption 1, and since in this subpopulation $LNATE = LATE^{sub}$, we can identify $LNATE$ in this natutal-experiment setting as the simple difference in mean outcomes between treated and controls:

$$LNATE = LATE^{sub} = E[Y^{obs}|T = 1, S(0) = S(1)] - E[Y^{obs}|T = 0, S(0) = S(1)] \quad (10)$$

It is important to note that in this natural-experiment setting only Assumption 1 (or Assumption 5 below if T is not randomly assigned) is needed. Indeed, in this setting we can estimate $LNATE$ under the weakest assumptions since the subpopulation with $S_i(1) = S_i(0)$–the only one for which $Y_i\left(1, S_i\left(0\right)\right)$ is observed in the data–is known.[33] Finally, note that in this case a variable measuring S is not even necessary to estimate $LNATE$.

If knowledge about a subpopulation with $S_i(1) = S_i(0)$ is not available, one possibility is to use the covariates X to find a subpopulation for which there is no effect of T on S–or for which such effect is close to zero–and then estimate the corresponding local $NATE$ for this subpopulation. To find such subpopulation one can rely on predicted values of the potential values of the post-treatment variable based on the covariates X. Let $\widehat{S}\left(1\right)$ and $\widehat{S}\left(0\right)$ be the estimators of the potential values of S based on X and S^{obs}.[34] Then, in this case the focus is on estimation of the local $NATE$ for the subpopulation with $\widehat{S}_i\left(1\right) = \widehat{S}_i\left(0\right)$:

$$LNATE_{\widehat{S}} = E\{E[Y(1, S\left(0\right)) - Y(0, S\left(0\right))|S(0) = S(1), \widehat{S}\left(1\right) = \widehat{S}\left(0\right)]\} \quad (11)$$

The conditioning on principal strata in the definition of $LNATE_{\widehat{S}}$ is necessary in order to interpret it as a causal effect. Estimating (11) as the difference in mean outcomes between treated and control units with $\widehat{S}_i\left(1\right) = \widehat{S}_i\left(0\right)$ as in (10) introduces two sources of bias. First, since both $\widehat{S}_i\left(1\right)$ and $\widehat{S}_i\left(0\right)$ are functions of S^{obs}, conditioning on them brakes the independence between T and the potential outcomes $Y\left(1\right)$ and $Y\left(0\right)$–i.e., it is not true that $Y\left(1\right), Y\left(0\right) \perp T|\{\widehat{S}\left(1\right), \widehat{S}\left(0\right)\}$. Hence, the difference $E[Y^{obs}|T = 1, \widehat{S}(0) = \widehat{S}(1)] - E[Y^{obs}|T = 0, \widehat{S}(0) = \widehat{S}(1)]$ is not equal to the local average treatment effect for the subpopulation with $\widehat{S}_i\left(1\right) = \widehat{S}_i\left(0\right)$, say, $LATE^{sub}_{\widehat{S}}$. Second, unless $\widehat{S}_i\left(1\right)$ and $\widehat{S}_i\left(0\right)$ are perfect predictors, the subpopulation for which $\widehat{S}_i\left(1\right) = \widehat{S}_i\left(0\right)$ will likely have units for which $S_i\left(1\right) \neq S_i\left(0\right)$, in which case $LNATE_{\widehat{S}} \neq LATE^{sub}_{\widehat{S}}$. Therefore, in analogy to equation (10), the two biases imply: $E[Y^{obs}|T = 1, \widehat{S}_i\left(1\right) = \widehat{S}_i\left(0\right)] - E[Y^{obs}|T = 0, \widehat{S}_i\left(1\right) = \widehat{S}_i\left(0\right)] \neq LATE^{sub}_{\widehat{S}} \neq LNATE_{\widehat{S}}$.

[33] Flores and Flores-Lagunes (2009a) impose monotonicity assumptions on the effect of T on S to identify subpopulations with $S_i(1) = S_i(0)$. They use these subpopulations to create nonparametric bounds and point identify $NATE$ for various subpopulations, including the overall population.

[34] One can construct estimators $\widehat{S}\left(1\right)$ and $\widehat{S}\left(0\right)$ in different ways. For example, we could use a single matching approach and let $\widehat{S}_i\left(0\right) = S^{obs}$ and $\widehat{S}_i\left(1\right) = S^{obs}_k$ if unit i is a control, and $\widehat{S}_i\left(1\right) = S^{obs}$ and $\widehat{S}_i\left(0\right) = S^{obs}_k$ if unit i is treated, where S^{obs}_k is the observed value of S for the closest unit to i in terms of a given distance measure $||X_i - X_j||$, with $T_i \neq T_j$. Alternatively, we could use a regression function approach to predict $S\left(1\right)$ and $S\left(0\right)$. Let $\mu_t\left(x\right) = E\left[S\left(t\right)|X = x\right]$ for $t = \{0, 1\}$ be the regression functions of the post-treatment potential values on X. Then, given the estimators $\widehat{\mu}_t\left(x\right)$ of these regression functions, we would define $\widehat{S}\left(1\right)$ and $\widehat{S}\left(0\right)$ for each unit i as $\widehat{\mu}_1\left(x\right)$ and $\widehat{\mu}_0\left(x\right)$, respectively.

Adding Assumption 2 (unconfounded strata) allows estimation of $LATE^{sub}_{\widehat{S}}$ without bias. Together, assumption 1 and 2 imply: $Y(1), Y(0) \perp T | \{X, , \widehat{S}(1), \widehat{S}(0)\}$.[35] Hence, $LATE^{sub}_{\widehat{S}}$ is identified as

$$\begin{aligned} & LATE^{sub}_{\widehat{S}} \\ = \; & E[Y(1) - Y(0) | \widehat{S}(0) = \widehat{S}(1)] = E\{E[Y(1) - Y(0) | X, \widehat{S}(0) = \widehat{S}(1)] | \widehat{S}(0) = \widehat{S}(1)\} \\ = \; & E\{E[Y(1) | T = 1, X, \widehat{S}(0) = \widehat{S}(1)] - E[Y(0) | T = 0, X, \widehat{S}(0) = \widehat{S}(1)] | \widehat{S}(0) = \widehat{S}(1)\} \\ = \; & E\{E[Y^{obs} | T = 1, X, \widehat{S}(0) = \widehat{S}(1)] - E[Y^{obs} | T = 0, X, \widehat{S}(0) = \widehat{S}(1)] | \widehat{S}(0) = \widehat{S}(1)\} \end{aligned} \tag{12}$$

where the fact that $Y(1), Y(0) \perp T | \{X, , \widehat{S}(1), \widehat{S}(0)\}$ is used in the third equality. Based on (12), $LATE^{sub}_{\widehat{S}}$ can be estimated by identifying those units with $\widehat{S}(0) = \widehat{S}(1)$ and then employing on this subsample any of the available methods for estimating the ATE of T on Y under an unconfounded treatment.[36] The role the covariates play in (12) is to control for any bias arising from comparing treated and control outcomes in the subpopulation with $\widehat{S}(0) = \widehat{S}(1)$.

Removing the second source of bias is more difficult since $LATE^{sub}_{\widehat{S}} = LNATE_{\widehat{S}}$ only if $S(0) = S(1)$ for all units with $\widehat{S}(0) = \widehat{S}(1)$. However, it is possible to derive an expression for this bias by noting that the bias associated with estimating the local $NATE$ for any given subpopulation using an unbiased estimator of its local ATE equals the difference between ATE and the $NATE$ (i.e., $MATE$) for that subpopulation. Therefore, the bias from using an estimator based on (12), call it $\widehat{LATE^{sub}_{\widehat{S}}}$, to estimate $LNATE_{\widehat{S}}$ equals the "local $MATE$" for the subpopulation with $\widehat{S}_i(1) = \widehat{S}_i(0)$. More precisely, letting $Z_i = 1$ if indeed $S_i(1) = S_i(0)$ and zero otherwise, the bias associated with estimating $LNATE_{\widehat{S}}$ using the unbiased estimator $\widehat{LATE^{sub}_{\widehat{S}}}$ can be written as:[37]

$$\begin{aligned} Bias(\widehat{LATE^{sub}_{\widehat{S}}}) &= LATE^{sub}_{\widehat{S}} - LNATE_{\widehat{S}} \\ &= E\left[Y(1, S(1)) - Y(1, S(0)) | \widehat{S}(1) = \widehat{S}(0)\right] \\ &= E\left\{E\left[Y(1, S(1)) - Y(1, S(0)) | Z = z, \widehat{S}(1) = \widehat{S}(0)\right] | \widehat{S}(1) = \widehat{S}(0)\right\} \\ &= \Pr(Z = 0 | \widehat{S}(1) = \widehat{S}(0)) E[Y(1, S(1)) - Y(1, S(0)) | Z = 0, \widehat{S}(1) = \widehat{S}(0)] \end{aligned} \tag{13}$$

The first term states that the closer $\Pr(S(1) \neq S(0) | \widehat{S}(1) = \widehat{S}(0))$ is to zero, the smaller the bias associated with the estimation of $LNATE_{\widehat{S}}$. Consequently, the better we predict $S(1)$ and

[35] By Assumption 1 we have that $Y(1), Y(0) \perp T | \{X, , S(1), S(0)\}$, which along with Assumption 2 implies (Lemma 4.3 in Dawid, 1979): $Y(1), Y(0) \perp \{T, , S(1), S(0)\} | X$. Since both $\widehat{S}(1)$ and $\widehat{S}(0)$ are functions of $S(1)$, $S(0)$, X and T we have: $Y(1), Y(0) \perp \{T, \widehat{S}(1), \widehat{S}(0)\} | X$. The result then follows by employing again Lemma 4.3 in Dawid (1979).

[36] See, for instance, Heckman, LaLonde and Smith (1999), Imbens (2004) or Imbens and Wooldridge (2009) for reviews on methods for estimation of average effects based on unconfoundedness.

[37] To simplify notation we omit the conditioning on the principal strata in the expression below.

$S(0)$, the smaller the bias will be. In the limit, if we perfectly predict $S(1)$ and $S(0)$, $LNATE_{\widehat{S}}$ is estimated without bias. The second term equals the local average mechanism effect for those units with $S(0) \neq S(1)$ and $\widehat{S}(1) = \widehat{S}(0)$. In principle, one would expect this term to be small to the extent that $S(0) \approx S(1)$ for this subpopulation with $\widehat{S}(1) = \widehat{S}(0)$. Importantly, note that the sign of the bias is given by the second term in (13). This is useful in determining the direction of the bias if information is available about the sign of the mechanism effect for this subpopulation, or if an assumption about its sign is tenable. In this case, $\widehat{LATE}^{sub}_{\widehat{S}}$ provides either an upper or a lower bound for $LNATE_{\widehat{S}}$.

In practice, one can perform some checks on the subpopulation with $\widehat{S}_i(1) = \widehat{S}_i(0)$ to gauge the extent to which it satisfies that T does not affect S. One check is to employ a Fisher randomization test (Fisher, 1935) for the sharp null hypothesis that the treatment effect of T on S is zero for all units in the subpopulation of interest. While failure to reject this null hypothesis does not necessarily mean that the treatment effect is zero for all units, rejecting it is a clear indication that the subpopulation characterized by $\widehat{S}_i(1) = \widehat{S}_i(0)$ is not appropriate to estimate $LNATE_{\widehat{S}}$. To describe a second check, note that under the ideal situation in which all the individuals in the subpopulations with $\widehat{S}_i(1) = \widehat{S}_i(0)$ have $S_i(1) = S_i(0)$, an estimator based on (12) will be unbiased for $LNATE_{\widehat{S}}$ whether it includes S^{obs} in the conditioning set or not (since $S_i(1) = S_i(0) = S_i^{obs}$). Therefore, the extent to which $S_i(1) \neq S_i(0)$ for those with $\widehat{S}_i(1) = \widehat{S}_i(0)$ can be gauged comparing $\widehat{LATE}^{sub}_{\widehat{S}}$ with and without controlling for S^{obs}. A statistically significant difference between the two can be regarded as evidence that the corresponding subpopulation is not appropriate to estimate $LNATE_{\widehat{S}}$.

So far we have discussed estimation of $LNATE$ or $LNATE_{\widehat{S}}$. The following assumption can be employed, when considered plausible, to interpret this estimator as an estimator of $NATE$ as well.

Assumption 4 *$NATE$ is constant over the population.*

Under this assumption, $NATE = LNATE_{\widehat{S}} = LNATE$. In addition, the part of the ATE that is due to the mechanism S is given by $MATE = ATE - LNATE_{\widehat{S}}$.

A few observations about Assumption 4 are in order. First, note that this assumption is analogous to the necessary assumption of a constant average treatment effect when estimating ATE using instrumental variables. In that case we can only identify $LATE$ 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 $LATE = ATE$. Second, we point out that Assumption 4 is weaker than assuming a constant ATE, which is a relatively common assumption in the literature (see, e.g., Heckman, LaLonde and Smith, 1999). Assumption 4 allows for heterogeneous effects of the treatment on the outcome variable, but such heterogeneity is restricted to work through the mechanism or post-treatment variable S (i.e.

through $MATE$). The plausibility of this assumption can be gauged in light of this observation. Third, note that the standard approach of controlling directly for S^{obs} implicitly assumes a stronger condition than Assumption 4, since it imposes a zero mechanism effect ($MATE$) for the population. Finally, when Assumption 4 is judged to be untenable in a particular application, $LNATE$ or $LNATE_{\widehat{S}}$ can still be an informative parameter for policymakers, just as $LATE$ commonly is (e.g., Imbens, 2009). We exemplify this last point in section 5.2.

For the case of a randomly assigned treatment, one way to implement this estimation strategy–in the absence of a natural experiment–is as follows: (i) specify a model (based on X) to estimate $S(1)$ and $S(0)$; (ii) identify the subpopulation for which $\widehat{S}(1) = \widehat{S}(0)$;[38,39] (iii) For that subpopulation, estimate $LATE^{sub}_{\widehat{S}}$ based on (12); (iv) if Assumption 4 is tenable, estimate $MATE = ATE - LNATE_{\widehat{S}}$.

In sum, if there is knowledge of a subpopulation for which T does not affect S, one can estimate $LNATE$ based on (12) under Assumption 1 only. If finding that subpopulation requires prediction of the potential values of S, Assumption 2 needs to be added to estimate $LNATE_{\widehat{S}}$ based on (12). This estimator will be biased in general to the extent that $S(1) \neq S(0)$ in the subpopulation with $\widehat{S}(1) = \widehat{S}(0)$, and one can learn about the sign of this bias based on (13). Although these parameters ($LNATE$ and $LNATE_{\widehat{S}}$) are informative on their own right, Assumption 4 can be used (when tenable) to estimate $NATE$. Finally, note that under the approach discussed in this section the functional-form assumption (Assumption 3) is not needed.

4.4 Identification and Estimation under Non-random Assignment

In the previous sections we analyzed the problem of estimating $NATE$ and $MATE$ when the treatment T is randomly assigned. In the absence of an experiment, a common approach in the literature is to assume that selection into treatment is based on a set of observed covariates (X) and on unobserved components not correlated with the potential outcomes (i.e., unconfounded treatment). We extend the framework discussed in sections 4.2 and 4.3 to the case when T is not randomly assigned using the following unconfoundedness assumption:

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

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

[38] If the post-treatment variable under consideration is continuous or if the procedure used 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), one could 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 grows to infinity. Alternatively, one could use a kernel function to give higher weight to observations for which $\widehat{S}(1)$ is closer to $\widehat{S}(0)$.

[39] At this point, one can use a test like the Fisher randomization test to gauge whether the subpopulation meets the minimum requirements for the estimation of $LNATE_{\widehat{S}}$.

X. Hence, the covariates will now have the additional role of controlling for selection into the treatment.[40]

We also add the following overlap assumption:

Assumption 6 $0 < \Pr(T = 1|X = x) < 1$, for all x.

Assumption 6 ensures that in infinite samples we are able to compare treated and control units for all values of X. When Assumptions 5 and 6 hold, the treatment assignment is said to be strongly ignorable (Rosenbaum and Rubin, 1983).

We start by discussing the identification strategy in section 4.2. In this case, as before, we need Assumptions 2 and 3. Assumptions 2 and 5 imply that $Y(1,1), Y(0,0), Y(1,0) \perp \{T, S(1), S(0)\}|X$. Thus, the covariates correct for selection not only into the treatment, but also into the principal strata.[41] Finally, Assumption 3 allows using $Y(1, S(1))$ to learn about $Y(1, S(0))$. Then, under Assumptions 2, 3, 5, and 6 we identify $NATE$ as in (8). This estimator can be implemented using the same approach outlined in section 4.2.

Now consider estimating $NATE$ with a strongly ignorable treatment assignment by focusing on the subpopulation for which T does not affect S. The main difference from the random assignment case is that now focus is on the subpopulation for which $S(1) = S(0) = s$ and that also has the same values of X. Therefore, $LNATE$ can be defined as in (9) including X in the conditioning set. If there is knowledge about a subpopulation for which T does not affect S, $LNATE$ can be estimated as in the previous section (with the additional conditioning on X). If we do not have knowledge of such subpopulation, we could follow the same approach as in section 4.3 and focus on the subpopulation for which $\widehat{S}_i(1) = \widehat{S}_i(0)$. Specifically, estimation focuses on:

$$LNATE_{\widehat{S}} = E\{E[Y(1, S(0)) - Y(0, S(0))|S(0) = s_0, S(1) = s_1, \widehat{S}(1) = \widehat{S}(0), X = x]\} \quad (14)$$

where the term inside the outer expectation is the local $NATE$ for the strata with $S(0) = s_0, S(1) = s_1, \widehat{S}(1) = \widehat{S}(0)$ and $X = x$. As before, (14) is a causal effect because is an average over effects defined within a principal stratum.

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 individuals at each value of X, $S(0)$, and $S(1)$, for those units with $\widehat{S}(1) = \widehat{S}(0)$. Specifically:

Assumption 7 $0 < \Pr(T = 1|S(0) = s_0, S(1) = s_1, X = x, \widehat{S}(1) = \widehat{S}(0)) < 1$, for all s_0, s_1 and x.

[40] Similar to Assumption 1, this assumption implies $Y(1, S(1)), Y(0, S(0)), Y(1, S(0)) \perp T|(X, S(1), S(0))$.

[41] Assumptions 2 and 5 also imply (again using Lemma 4 in Dawid, 1979) that $Y(1,1), Y(0,0), Y(1,0) \perp \{S(1), S(0)\}|\{T, X\}$, so that individuals in different strata but with the same values of the treatment and covariates are comparable.

Assumption 7 is similar to the common overlap condition in Assumption 6, except that it includes $S(1)$ and $S(0)$ as additional covariates and only has to hold for the subpopulation of interest.[42] Finally, under Assumption 4, $NATE = LNATE_{\widehat{S}}$ and $MATE = ATE - LNATE_{\widehat{S}}$. The implementation of this approach follows closely the one outlined when T is randomly assigned.

5 Empirical Applications

In this section, we present two empirical applications that illustrate the implementation of our strategies to estimate $NATE$. 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 implements our estimators to observational data from the Natality Data Sets of Pennsylvania (1989-1991).

5.1 Random Assignment

Our data comes from the National Job Corps Study (NJCS), a randomized experiment to evaluate the effectiveness and social value of the Job Corps (JC) training program. JC provides low-skilled young people (ages 16-24) with marketable skills to enhance their labor market outcomes by offering academic, vocational, and social skills training at JC centers throughout the United States, where most students reside during their enrollment period.[43]

An important finding of the NJCS was that, 16 quarters after randomization, individuals in the treatment group earned a statistically significant 12% more per week ($25.2) than individuals in the control group (Burghardt et al., 2001). However, upon looking at different race and ethnic groups, it was found that Hispanics in the treatment group earned 10% less (a statistically insignificant -$15.1) 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% ($22.8) and 24% ($46.2) over their control group members, respectively (Schochet, Burghardt and Glazerman, 2001).[44]

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 (2009) (hereafter FGN) document that Hispanics in the control group earned a significant amount of labor market experience during

[42] In practice, since we expect $S_i(0) \approx S_i(1)$ within the subpopulation with $\widehat{S}_i(1) = \widehat{S}_i(0)$, one way to check the overlap condition is to look at the overlap in the distribution of the propensity scores for treated and control units within this subpopulation, where the propensity score includes S^{obs} as an additional covariate.

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

[44] These estimated effects reported by the NJCS were computed using differences-in-means estimates adjusted for non-compliance, identifying a $LATE$ on those who comply with their treatment assignment (Imbens and Angrist, 1994). The proportion of those in the treatment group who enrolled in Job Corps was 73%, and the proportion of those in the control group that managed to enroll in Job Corps was 1.4%.

the study compared to treated Hispanics and also to control-group blacks and whites. Thus, if accumulated experience resulted in an earnings advantage that treated Hispanics were not able to overcome by the end of the study, this post-treatment variable (experience) potentially accounts for part of the lack of earnings gains for Hispanics in JC. This setup illustrates the policy relevance of our parameters: 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 can be judged beneficial. At the same time, by 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 distinct inferences for these groups.

Table 1 presents estimates of different parameters of interest using a subsample from the NJCS that includes individuals with information on pre-treatment covariates plus the post-treatment variable "average hours worked per week during the study", and that report being Hispanic, white or black.[45] We pool the samples of blacks and whites for simplicity, since for both of these groups it is found that post-treatment experience is not a relevant mechanism (we present further evidence of this below), making unnecessary to present a separate analysis for them.[46]

Rows 1 through 4 in Table 1 report estimates of the average intention-to-treat (ITT) parameter.[47] Row 1 presents unadjusted differences in means between treatment- and control-group individuals. These estimates are qualitatively similar to the originally reported NJCS estimates: for the full sample and white/black samples the estimates are positive and statistically significant ($15.6 and $23.8, respectively), while for Hispanics the effects show a loss of $19.7 that is marginally statistically significant.

Next, we present ITT estimates that control for pre-treatment variables through weighting by the estimated propensity score (pscore) in order to improve precision. In particular, in this paper we employ the estimator due to Robins and Rotnitzky (1995) that combines weighting by the pscore and regression (i.e., the "double-robust" estimator), as described in Imbens (2004) and Imbens and Wooldridge (2009). Its implementation amounts to applying weighted least squares (WLS) to a regression of the outcome on the treatment indicator and additional covariates or functions of them (e.g., the pscore), with weights given by $\lambda_i = \sqrt{\frac{T_i}{p(X_i)} + \frac{1-T_i}{1-p(X_i)}}$

[45] The 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.

[46] We have estimated all parameters for the black and white samples separately as well, corroborating that this is indeed the case.

[47] Given the presence of non-compliance in the sample, we estimate the intention-to-treat (ITT) parameter. This parameter is commonly estimated in the program evaluation literature and allows relying on the random assignment as much as possible. Consequently, in this application, our parameters decompose the ITT and not the ATE.

where $p(X_i)$ is the estimated pscore.[48] Rows 2, 3, and 4 differ in the way the WLS regression is specified: using no additional covariates, the pscore, and up to a cubic pscore as additional regressors, respectively.

The estimates in rows 2-4 are fairly comparable to the unadjusted estimates in row 1, except for Hispanics. This might be due to the smaller sample size of this group and the fact that the group shows some pre-treatment imbalances in the covariates.[49] For this reason the remaining estimates adjust for covariates. These estimates are in line with the conclusions in the NJCS although the negative effects on Hispanics are less dramatic once covariates are controlled for. These "total effect" estimates will be the benchmark to compare our estimated effects net of the lock-in mechanism effect.

The next set of estimates in Table 1 are of Rosenbaum's (1984) NTD parameter. All of them are obtained controlling for the observed value of post-treatment labor market experience employing the WLS approach described above with a pscore that includes a flexible specification of experience (S^{obs}) in its estimation.[50] This way of controlling for a post-treatment variable by including it in the estimation of the propensity score is followed by Black and Smith (2004), although they use a matching approach to control for the estimated pscore, as opposed to weighting.

Recall that the NTD estimates typically lack causal interpretation as estimates of the total effect, and correspond to $NATE$ under very stringent conditions (see section 4.1). We report them for comparison to our estimates below. The NTD estimates for the full sample are less than 20% larger compared to the ITT estimates, while for whites/blacks they are less than 10% larger. For Hispanics, the two sets of estimates are starkly different: more than 150% larger. In sum, despite the fact that all these effects are statistically insignificant for Hispanics and the lack of causal interpretation of NTD, the point estimates are suggestive of a relevant lock-in effect for Hispanics (contrary to whites/blacks) that would seem to explain an important portion of the lack of effects of JC on them.

Two sets of estimates of $NATE$ appear in rows 8-9, obtained using the estimation strategy outlined in section 4.2. To implement it, we model (under Assumption 3) the first term in (8) as a linear function of $S(1)$ and all available pre-treatment covariates. The second term in (8) is similarly predicted, but using $S(0)$ instead. The two $NATE$ estimates differ on the specification of the experience variable and the covariates included: row 8 includes experience up to a cubic term and all covariates, while row 9 adds interactions between the experience variable and the covariates to this specification. The $NATE$ estimates for whites/blacks and for the full sample

[48] The propensity score (pscore) is estimated using all pre-treatment variables, their squares, and interactions in a logit model.

[49] The misalignment of pre-treatment variables for Hispanics is documented and discussed in FGN (2009).

[50] In particular, we use the same specification of the pscore as in ITT, but include experience up to a cubic term, and interactions of this variable with the pre-treatment covariates.

are closer to the ITT estimates than the NTD estimates, which is consistent with a non-existent lock-in effect for them. Among the two $NATE$ estimates, the richer specification (row 9) is closer to the ITT estimates. For Hispanics, however, the $NATE$ estimates are very different from the ITT estimates (as was the case with NTD), strengthening the notion of a relevant lock-in effect for them. Unfortunately, as before, the estimates are imprecisely estimated.

In the last panel of Table 1, we present estimates of $LNATE_{\widehat{S}}$ that differ in the way they are implemented. In all of them, the potential values of post-treatment labor market experience ($S(0)$ and $S(1)$) are estimated based on covariates X employing the matching approach described in footnote 34, using a single match on the estimated pscore that does not include experience.[51] Given that S is a fairly continuous variable–average number of hours worked per week during the study–, it is difficult to find individuals for which $\widehat{S}(1) = \widehat{S}(0)$. We approach this feature by defining a window around $\widehat{S}(1) - \widehat{S}(0) = 0$ using a Silverman-type bandwidth to characterize the subpopulation with $\{\widehat{S}(1) = \widehat{S}(0)\}$.[52] The proportional size of the resulting $LNATE_{\widehat{S}}$ subpopulation is similar across the three samples.

As mentioned in section 4.3, we can assess the plausibility that the subpopulation found for the estimation of $LNATE_{\widehat{S}}$ satisfies the requirement that experience (S) is not affected by the treatment (or that this effect is small). To do this, we implement several versions of the Fisher randomization test. This test provides evidence on the sharp null hypothesis $H_0 : S_i(1) = S_i(0)$ for every i. In its simplest form, the implementation consists of simulating the distribution under H_0 for the observed test statistic $\frac{\Sigma T_i S_i(1)}{\Sigma T_i} - \frac{\Sigma(1-T_i)S_i(0)}{\Sigma(1-T_i)}$ (or other quantity measuring the effect of T on S) by randomizing the treatment indicator to the units in the sample and computing the test statistic in each repetition. Then, an approximate p-value is constructed by comparing the observed test statistic with the simulated distribution. A rejection of the test indicates that the post-treatment variable S is affected by the treatment.

Panel A of Table 3 presents results of Fisher randomization tests for the three groups under analysis. For each group, tests are applied to the population of the group (for comparison) and to the subpopulation characterized by $\{\widehat{S}(1) = \widehat{S}(0)\}$. We present five versions of the test, which turn out to yield the same conclusion. The first three tests are based on comparing the coefficient from an OLS regression of S on the treatment indicator (T) in row 1, adding the pscore in row 2, and further adding the square and cube of the pscore in row 3.[53] The last two rows are based on applying the simple form of the test described above to the residuals from OLS regressions of S on the pscore and then adding the pscore square and cube, respectively.

The results of the tests are in line with expectations. Whites/blacks, the group that shows

[51] We estimated $S(1)$ and $S(0)$ based on X separately for the full sample, whites/blacks, and Hispanics.

[52] The Silverman-type bandwidth employed is equal to $0.79 * IQR * N\hat{}(-1/5)$, where IQR is the interquartile range and N is the sample size. This bandiwdth has the advantage of being more robust to outliers than the usual one based on the standard deviation (see, e.g., Pagan and Ullah, 1999).

[53] Given that in this case S is the outcome of interest, in all cases the pscore is the specification that does not include experience on it. We re-estimate the pscore within the corresponding subpopulation.

the least amount of lock-in effect, have p-values for their population that range from 0.31 to 0.5, not rejecting the null hypothesis of no effect of the treatment on post-treatment experience. For the full sample, the null hypothesis is rejected at the 10% level in all cases, and at the 5% level in one. However, the subpopulation cannot reject the test with a p-value of 0.77 or higher. Finally, as expected, Hispanics show the strongest rejections of the null hypothesis for their population, consistent with the hypothesis of an important effect of training on experience for them. Importantly, the characterized subsample substantially decreases the strong relationship between S and T, with p-values that range from 0.57 to 0.99. In sum, for the three groups under analysis, the statistical evidence cannot reject the notion that the corresponding subpopulations characterized by $\{\widehat{S}(1) = \widehat{S}(0)\}$ have a zero effect of T on S for all units.

We now discuss the $LNATE_{\widehat{S}}$ estimates based on the subpopulations described above. We start by estimating $LNATE_{\widehat{S}}$ in rows 10-12 as the local average treatment effect ($LATE^{sub}_{\widehat{S}}$) for this subpopulation based on (12). For each group, we estimate a propensity score without including experience in this subpopulation and use WLS with similar specifications as those used in the estimation of ITT and NTD. For the full sample, the $LNATE_{\widehat{S}}$ estimates average \$29.9, substantially higher than ITT (\$19.2), NTD (\$23), and also $NATE$ (\$21.7). In contrast, for whites/blacks the $LNATE_{\widehat{S}}$ estimates are around \$24, which is about the same magnitude as ITT. Under Assumption 4, this result reinforces the observation that for whites/blacks experience is not a mechanism through which JC affects wages. For Hispanics, however, the $LNATE_{\widehat{S}}$ estimates (about \$10.9 on average) are larger than the NTD and $NATE$, and substantially larger than the ITT. Unfortunately, it is also estimated very imprecisely and these differences are not statistically significant. Note that if we were to assume that the average mechanism effect is negative for all three subpopulations, then the estimates in rows 10-12 would be downward biased according to equation (13).

As a robustness check, and following our discussion in section 4.3, rows 13-15 present estimates of $LNATE_{\widehat{S}}$ further controlling for S^{obs}. These estimates differ from those in rows 10-12 in that they include experience in the specification of the pscore.[54] Overall, the estimates for the full population are close to the ones presented in rows 10-12, which give us confidence in our $LNATE_{\widehat{S}}$ results for the full sample. For whites and blacks the results are not as robust as for the full sample; however, they remain below the NTD estimates in rows 5-7. Note that for this group there is a considerable decrease in the precision of our estimates by introducing experience as an additional control, since now the $LNATE_{\widehat{S}}$ estimates in rows 13-15 are not statistically different from zero. Something similar occurs for Hispanics, for which the $LNATE_{\widehat{S}}$ estimates fall when including experience in the pscore specification, although none of their estimates are statistically significant.

[54] Given that we are within the subpopulation with $\widehat{S}_i(1) = \widehat{S}_i(0)$, for which $S(1) \approx S(0)$, we include S^{obs} in the estimation of the propensity score as an additional pre-treatment variable.

We gather the following conclusions from this empirical illustration. First, our estimates of $NATE$ and $LNATE_{\widehat{S}}$ suggest that the lock-in effect results in a negative causal mechanism for the effect of JC training on Hispanic's earnings, although the estimates remain statistically insignificant. Second, the full set of estimates corroborate the high degree of heterogeneity that exists among whites/blacks and Hispanics, which results in very different inferences in terms of their estimated total, net and mechanism effects from JC training. Lastly, and unfortunately, in this application many of the differences in the estimates are not statistically significant. This may be due to the absence of differences among the true parameters in this application, or the need for larger sample sizes to increase precision.

5.2 Non-random Assignment

When the treatment is not randomly assigned we face the additional issue of controlling for self-selection, for which we employ a selection on observables assumption and regard the treatment as randomly assigned conditional on a rich set of observed covariates. For this application, the data comes from Pennsylvania's Natality Data Sets from 1989 to 1991, which includes all births (although we focus on single births) and has been previously used and documented by Chay, Flores and Torelli (2005). The availability of a wide range of observable characteristics, including characteristics of both parents and previous birth history, makes the assumption of selection on observables more plausible.

The focus is on evaluating the extent to which smoking during pregnancy (treatment) affects the incidence on low birth weight (outcome) through a shorter gestation time (a mechanism). The outcome "low birth weight" (LBW) has the standard definition in the medical literature of birth weight below 2,500 grams, and is widely associated with a myriad of health, behavioral and socioeconomic problems in later stages of individual development (e.g. UNICEF and WHO, 2004). For instance, LBW has been negatively associated to educational attainment, self-reported health status, and employment (Currie and Hyson, 1999). 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 and thus increases the probability of incidence on LBW, but the importance of specific mechanisms is 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 aimed at minimizing 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 the sample in the overlap region of the estimated pscore between the 1 percentile of the pscore values for the treated and 99 percentile of the pscore values for controls.[55] For reference, the average LBW incidence in the sample employed is 58.3 per 1,000 births, and 20.8% of women smoked during pregnancy. The incidence on LBW is 48.4 per 1,000 births for non-smokers (control) and 95.7 per 1,000 births for smokers (treatment), yielding the unadjusted difference of 47.3 shown in the first row of Table 2. Thus, mothers who smoked during pregnancy were about twice as likely to deliver a LBW baby than those mothers who did not.

Rows 2 through 4 present estimates of the total effect (ATE) of smoking on LBW incidence, controlling for self-selection using an estimated pscore.[56] Rows 2-4 employ the same WLS approach and specifications as in the previous empirical application. The three estimates are close to each other, reflecting an effect of smoking on the incidence on LBW of 33 per 1,000 births. The fact that this figure is smaller than the unadjusted difference in row 1 implies a selection bias of about 14 in the unadjusted figure. Still, the ATE estimate suggests a sizable effect of smoking during pregnancy on LBW incidence, as the probability increases 68%.

The second panel of Table 2 (rows 5-7) presents estimates of the NTD parameter that control directly for gestation time in weeks (S^{obs}) using specifications similar to those used in the previous application. For these estimates the pscore includes observed gestation (in a flexible way) in its estimation. The NTD is precisely estimated at about 28. This suggests that 15.2% of the effect of smoking during pregnancy on the incidence on LBW (5 of 33 per 1,000 births) can potentially be attributed to gestation time, although these estimates correspond to $NATE$ under very stringent conditions.

Rows 8 and 9 present estimates of $NATE$, implemented in the same way as in the previous application. Both estimates are essentially identical to each other at 26.5 per 1,000 births, and slightly smaller than NTD. Based on the $NATE$ estimates and under the assumptions discussed in section 4, about 20% of the effect of smoking during pregnancy on the incidence on LBW can be causally attributed to gestation time. We note that the difference between the $NATE$ and the ATE estimates is statistically significant, whereas the difference between the $NATE$ and the NTD estimates is not.

[55] The sample consists of 496,212 individuals, of which 425,219 are contained within the overlap region.

[56] The propensity score is estimated with a logit model. The covariates used are mother's age, education, race, ethnicity, marital status, foreign-born status; father's age, education, race and ethnicity; dummies for trimester of first prenatal care visits, adequacy of care, number of prenatal visits, number of drinks per week, alcohol use, live birth order, number of previous births were newborn died, parity indicator, interval since last birth, indicators for previous birth over 4000 grams and previous birth preterm or small for gestational age; maternal medical risk factors that are not believed to be affected by smoking during pregnancy: anemia, cardiac disease, lung disease, diabetes, genital herpes, hydramnios/oligohydramnios, hemoglobinopathy, chronic hypertension, eclampsia, incompetent cervix, renal disease, Rh sensitization, uterine bleeding; indicators for: month of birth, county of residence at birth, state of occurrence and residence different, each variable that is missing for some mothers. The particular specification used includes nonlinear functions and interactions, and is similar to the one used in Chay, Flores and Torelli (2005) and Almond, Chay and Lee (2005).

Finally, the last panel in Table 2 presents estimates of $LNATE_{\widehat{S}}$. The subpopulation of interest is obtained, as in the previous application, by estimating the potential values of gestation time using a single match on the estimated pscore and selecting those units with $\{\widehat{S}(1) = \widehat{S}(0)\}$, resulting in a sample of about 15% of the one used for estimation of the ATE.[57] We test whether the individual treatment effect of T on S is zero for all mothers in this subpopulation using different versions of the Fisher randomization test. Panel B of Table 3 shows that for the population the null hypothesis of no effect of T on S is soundly rejected, while in the subpopulation it is not, with p-values ranging from 0.35 to 0.92. Rows 10-12 estimate $LNATE_{\widehat{S}}$ by estimating the ATE within the relevant subpopulation ($LATE^{sub}_{\widehat{S}}$) based on (12) employing a pscore without gestation in its specification; whereas rows 13-15 estimate $LNATE_{\widehat{S}}$ by including gestation in the pscore specification for comparison. For this application, all six $LNATE_{\widehat{S}}$ estimates are essentially the same at about 22.8 per 1,000 births. This supports the notion that within this subpopulation $LNATE_{\widehat{S}} \approx LATE^{sub}_{\widehat{S}}$ and thus the bias in (13) is close to zero.

The estimated value of 22.9 for $LATE^{sub}_{\widehat{S}}$ implies that even for a subpopulation for which smoking has a negligible effect on gestation, smoking during pregnancy has a significant and large effect on LBW incidence. Therefore, even if the effect of smoking on gestation were eliminated, there would remain considerable negative effects of smoking during pregnancy through other mechanisms, at least for the individuals in this subpopulation (15% of the total population), although this general result would likely apply to the entire population. This is the type of conclusions that can be learned from $LNATE_{\widehat{S}}$ (*or* $LATE^{sub}_{\widehat{S}}$) without the need of Assumption 4 (constant $NATE$). Under Assumption 4 these estimates imply that 31% of the effect of smoking during pregnancy on the incidence on LBW can be causally attributed to gestation time.[58]. Remarkably, the difference between ATE and each of NTD, $NATE$ and $LNATE_{\widehat{S}}$ is highly statistically significant, speaking to the relevance of the mechanism. In addition, the difference between $LNATE_{\widehat{S}}$ and NTD is statistically significant at the 7% level, while the difference between $LNATE_{\widehat{S}}$ and $NATE$ is not.

In sum, the results of this empirical application are consistent with a causal role of gestation time as a channel through which smoking during pregnancy increases the incidence on LBW. While the total effect is 33 per 1,000 births (or about 70% higher than non-smokers), our results indicate that between 20 to 30 percent of this effect works causally through a shorter gestation time. Importantly, we also find that the NTD understates the importance of gestation time as a causal mechanism by between 5 to 15 percentage points. Clearly, an advantage of

[57] Note 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.

[58] If we assume that the average mechanism effect is positive for the subpopulation where $LNATE_{\widehat{S}}$ is estimated, then the estimates in rows 10-12 are upward biased accoring to equation (13) and thus the estimated mechanism effect of 31% is downward biased.

this empirical application is that the sample sizes allow estimation of the parameters with considerable precision.

Some patterns emerge in the implementation of our methods in the two empirical applications. First, our methods are feasible to implement in estimating the $NATE$ and the $MATE$. Second, the estimation of these parameters yields new insights about the treatment at hand; although we remark that a careful evaluation of the plausibility of the assumptions made by each estimation strategy in particular applications cannot be overstated. Finally, in both applications, the (non-causal) estimates for NTD, which is the parameter commonly used in the literature to estimate net effects, differ from our estimates (especially in the second empirical application). This underscores the potentially misleading conclusions that can be reached by directly controlling for the observed values of the post-treatment variable.

6 Conclusion

This paper analyzes identification and estimation of an average causal mechanism through which a treatment or intervention affects an outcome, and the average causal effect of the treatment net of this mechanism. These causal effects are of interest since they allow a better understanding of the treatment and, as a result, can be used for policy purposes in the design, development, and evaluation of interventions. Not surprisingly, it is common in the literature to informally analyze potential mechanisms of a treatment as a natural step after estimating the "total" effect of the treatment. Unfortunately, the parameter of interest is usually not clearly defined, and these analyses are typically based on a standard approach that directly controls for observed values of a variable representing a mechanism, resulting in estimates that generally lack causal interpretation.

We avoid this pitfall by using the concept of principal stratification (Frangakis and Rubin, 2002) to define causal parameters of interest. These parameters intuitively decompose the total effect of a treatment into the part that is causally due to a particular mechanism (mechanism average treatment effect or $MATE$) and the part that is net of such mechanism (net average treatment effect or $NATE$). Estimation of these effects is a difficult task given the data typically available to researchers. In addition, we show that to interpret the standard approach as $NATE$ we need to rely on assumptions that are typically too strong to be useful in practice.

We develop two strategies for estimation of our parameters under an unconfoundedness assumption in the spirit of the familiar selection on observables approach (Rosenbaum and Rubin, 1983; Imbens, 2004). The first strategy is based on a functional form assumption relating the partially observable potential outcome $Y(1, S(1))$ to the unobserved $Y(1, S(0))$ that is necessary for the estimation of $NATE$. The second approach estimates $NATE$ by first estimating a local $NATE$ for the subpopulation for which T does not affect S, where

the covariates can be used to find such population if it is not available through other means (e.g., a natural experiment). We present each of these approaches for the case in which the treatment is randomly assigned and when selection into the treatment is based on a set of observable covariates. Finally, we present two different empirical applications that illustrate the implementation of our methods.

Several natural extensions are of interest, such as: (i) an analysis of the way in which additional information can be used to estimate our parameters (e.g., the availability of instrumental variables); (ii) the construction of bounds for our parameters in the spirit of Manski (1990); (iii) the development of a set of alternative assumptions leading to identification and estimation strategies that allow for selection into the treatment to be based on unobservables. Work along these lines is being pursued in Flores and Flores-Lagunes (2009a and 2009b).

References

[1] Abadie, A. and Imbens, G. (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects", *Econometrica*, 74(1), 235-267.

[2] Almond, D., Chay, K. Y. and Lee, D. S. (2005) "The Cost of Low Birth Weight". *Quarterly Journal of Economics*, 120 (3), 1031-1083.

[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] Card, D. and Hyslop, D. (2005) "Estimating the Effects of a Time-Limited Earnings Subsidy for Welfare-Leavers" *Econometrica*, 73, 1723-70.

[6] Center for Disease Control and Prevention (2001) *Women and Smoking: A Report of the Surgeon General.*

[7] 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.

[8] Currie, J. and Hyson, R. (1999) "Is the Impact of Health Shocks Cushioned by Socioeconomic Status? The Case of Low Birthweight " *American Economic Review*, 89, 245-250.

[9] Currie, J. and Neidell, M. (2007) "Getting Inside the 'Black Box' of Head Start Quality: What Matters and What Doesn't" *Economics of Education Review*, 26, 83-99.

[10] Dawid, A. (1979) "Conditional Independence in Statistical Theory (with Discussion)" *Journal of the Royal Statistical Society, Series B*, 41, 1-31.

[11] 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.

[12] 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.

[13] Ehrenberg, R., Jakubson, G., Groen, J., So, E., and Price, J. (2007), "Inside the Black Box of Doctoral Education: What Program Characteristics Influence Doctoral Students' Attrition and Graduation Probabilities?" *Educational Evaluation and Policy Analysis*, 29, 134-150.

[14] Fisher, R.A. (1935). *The Design of Experiments*. Edingburgh: Oliver and Boyd.

[15] Flores, C. and Flores-Lagunes, A. (2009a) "Nonparametric Partial and Point Identification of Net or Direct Causal Effects", mimeo, University of Miami.

[16] Flores, C. and Flores-Lagunes, A. (2009b) "Partial and Point Identification of Net Effects using Instrumental Variables", mimeo, University of Miami.

[17] Flores-Lagunes, A., Gonzalez, A., and Neumann, T. (2009) "Learning But Not Earning? The Impact of Job Corps Training on Hispanic Youths", *Economic Inquiry*, forthcoming.

[18] Frangakis,C.E. and Rubin D. (2002) "Principal Stratification in Causal Inference", *Biometrics*, 58, 21–29.

[19] Hahn, J. (1998) "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects", *Econometrica*, 66, 315-331.

[20] 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.

[21] Heckman, J.; Ichimura, H. and Todd, P. (1998), "Matching as an Econometric Evaluation Estimator", *Review of Economic Studies,* 65(2), 231-294.

[22] Heckman, J.; Smith, J. and Clements, N. (1997), "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts", *Review of Economic Studies,* 64(3), 487-535.

[23] 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.

[24] Hirano, K.; Imbens, G. and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score", *Econometrica*, 71(4), 1161-1189.

[25] Holland, P. (1986) "Statistics and Causal Inference" *Journal of the American Statistical Association,* 81, 945-70.

[26] Imbens, G. (2004) "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review" *Review of Economics and Statistics*, 84, 4-29.

[27] Imbens, G. (2009) "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)", Working Paper, Harvard University.

[28] Imbens, G. and Angrist, J. (1994) "Identification and Estimation of Local Average Treatment Effects" *Econometrica*, 62, 467-75.

[29] Imbens, G. and Wooldridge, J. (2009) "Recent Developments in the Econometrics of Program Evaluation", *Journal of Economic Literature*, 47(1), 5-86.

[30] Lechner, M. (2005) "A Note on Endogenous Control Variables in Evaluation Studies" Discussion paper 2005-16, University of St. Gallen.

[31] Lechner, M. and R. Miquel (2005) "Identification of the Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions" Discussion paper, University of St. Gallen.

[32] Lee, D.S. (2009) "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects", *Review of Economic Studies*, forthcoming.

[33] Manski, C. (1990) "Nonparametric Bounds on Treatment Effects" *American Economic Review Papers and Proceedings*, 80, 319-23.

[34] Meyer, B. (1995) "Lessons from the U.S. Unemployment Insurance Experiments" *Journal of Economic Literature*, XXXIII, 91-131.

[35] Mealli, F. and Rubin, D. (2003) "Assumptions Allowing the Estimation of Direct Causal Effects" *Journal of Econometrics*, 112, 79-87.

[36] Neyman, J. (1923) "On the Application of Probability Theory to Agricultural Experiments: Essays on Principles" Translated in *Statistical Science*, 5, 465-80.

[37] Pagan, A. and Ullah A. (1999) *Nonparametric Econometrics*. Cambridge University Press.

[38] Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

[39] Pearl, J. (2001) "Direct and Indirect Effects" In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, M. Kaufmann ed., San Francisco, 411-20.

[40] Petersen, M., Sinisi, S., and van der Laan, M. (2006) "Estimation of Direct Causal Effects" *Epidemiology*, 17, 276-284.

[41] 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.

[42] Robins, J. and Greenland, S. (1992) "Identifiability and Exchangeability for Direct and Indirect Effects" *Epidemiology*, 3, 143-155.

[43] Robins, J. and Rotnitzky, A. (1995) "Semiparametric Efficiency in Multivariate Regression Models with Missing Data" *Journal of the American Statistical Association*, 90, 122-129.

[44] 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.

[45] Rosenbaum, P. and Rubin, D. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.

[46] Rubin, D. (1974) "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies" *Journal of Educational Psychology*, 66, 688-701.

[47] 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.

[48] Rubin, D. (2004) "Direct and Indirect Causal Effects via Potential Outcomes" *Scandinavian Journal of Statistics*, 31, 161-70.

[49] Rubin, D. (2005) "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions" *Journal of the American Statistical Association*, 100, 322-331.

[50] Stein et al. (1983). "Smoking, Alcohol and Reproduction" *American Journal of Public Health*, 73 (10), 1154-1156.

[51] 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.

[52] Simonsen, M. and Skipper, L. (2006), "The Costs of Motherhood: An Analysis Using Matching Estimators", *Journal of Applied Econometrics*, 21, 919-934.

[53] Zhang, J. L.; Rubin, D. and Mealli, F. (2006) "Evaluating the Effects of Job Training Programs on Wages through Principal Stratification" In *Advances in Econometrics: Modelling and Evaluating Treatment Effects in Econometrics*, Millimet, D.; Smith, J. and Vytlacil, E. eds., Vol. 21. Amsterdam. Elsevier.

[54] UNICEF and WHO (2004), *Low Birthweight: Country, Regional and Global Estimates*, New York.

Table 1. Random Assignment Application: Estimation of the effect of the Job Corps training program on weekly earnings during quarter 16 after randomization. Mechanism analyzed: post-treatment labor market experience[1]

		Full Sample			White and Black			Hispanic		
		(N=9,105)	*t-stat*	*N*	*(N=7,412)*	*t-stat*	*N*	*(N=1,693)*	*t-stat*	*N*
	Estimation of Intention to Treat Effects, ITT									
1	Unadjusted Difference	15.6	(3.40)		23.8	(4.70)		-19.7	(-1.79)	
2	WLS using pscore in weights. No pscore in regression	19.2	(4.21)		24.8	(4.90)		-7.6	(-0.69)	
3	WLS using pscore in weights. Linear pscore as regressor	19.3	(4.24)		24.8	(4.90)		-7.5	(-0.69)	
4	WLS using pscore in weights. Up to cubic pscore as regressor	19.2	(4.23)		25.0	(4.96)		-7.3	(-0.67)	
	Estimation of "Net Treatment Difference" controlling for observed post-treatment esperience, NTD									
5	WLS using pscore in weights. No pscore in regression	23.1	(5.14)		27.0	(5.42)		5.2	(0.47)	
6	WLS using pscore in weights. Linear pscore as regressor	23.1	(5.21)		27.0	(5.43)		5.8	(0.55)	
7	WLS using pscore in weights. Up to cubic pscore as regressor	22.7	(5.13)		27.0	(5.46)		4.5	(0.44)	
	Estimation of NATE using E[Y(1,S(1)) | S(1), X] to predict E[Y(1,S(0)) | S(0), X]. E[Y(0,0) | S(0), X] is similarly predicted.									
8	OLS outcome on experience and its polynomials up to degree 3 plus linear covariates	22.7	(6.33)		26.8	(6.68)		3.6	(0.71)	
9	OLS outcome on experience and its polynomials up to degree 3, linear covariates, and interactions	20.7	(5.29)		23.6	(5.55)		6.7	(0.69)	
	Estimation of the local NATE for the subpopulation with (predicted) S(0)=S(1), where predicted S(0) and S(1) are based on matching on the pscore. [2]									
	Using pscore that does not include experience in its estimation									
10	WLS using pscore in weights. No pscore in regression	28.9	(2.45)	*1273*	23.9	(2.04)	*1072*	10.5	(0.36)	*344*
11	WLS using pscore in weights. Linear pscore as regressor	28.9	(2.49)	*1273*	24.0	(2.05)	*1072*	9.6	(0.34)	*344*
12	WLS using pscore in weights. Up to cubic pscore as regressor	31.9	(3.08)	*1273*	24.1	(2.07)	*1072*	12.7	(0.44)	*344*
	Using pscore that includes experience in its estimation									
13	WLS using pscore in weights. No pscore in regression	30.1	(2.48)	*1273*	18.2	(1.52)	*1072*	1.6	(0.05)	*344*
14	WLS using pscore in weights. Linear pscore as regressor	30.1	(2.52)	*1273*	18.2	(1.52)	*1072*	2.9	(0.09)	*344*
15	WLS using pscore in weights. Up to cubic pscore as regressor	34.4	(3.28)	*1273*	18.0	(1.49)	*1072*	3.6	(0.12)	*344*

[1] All estimates use a sample that contains those who completed both a 48-month and baseline interviews, and with non-missing information on the covariates employed by the estimators. The sample sizes are indicated at the top of each column, unless otherwise indicated for particular estimators. Standard errors do not take into account the estimation of the propensity score.

[2] Given that S is defined as the average number of hours worked during the study, the predicted values S(1) and S(0) are continuous. The subpopulation with predicted S(1)=S(0) is obtained employing a window around (predicted) S(1)-S(0)=0 using a Silverman-type bandwidth based on the inter-quantile range (IQR): h=0.79*IQR*N^(-1/5).

Table 2. Non-Random Assignment Application: Estimation of the effect of smoking during pregnancy on the incidence of low birth weight (less than 2,500 grams) per 1,000 births. Mechanism analyzed: weeks of gestation (single births in Pennsylvania from 1989 to 1991)

		Estimate	*t-statistic*
Estimation of Average Treatment Effects, ATE. Focus on a population with overlap region of pscore between the 1 percentile of pscore for treated and 99 percentile of pscore for controls (N=425,219).			
1	Unadjusted Difference	47.3	(44.82)
2	WLS using pscore in weights. No pscore in regression	33.1	(26.85)
3	WLS using pscore in weights. Linear pscore as regressor	32.8	(26.57)
4	WLS using pscore in weights. Up to cubic pscore as regressor	32.8	(26.56)
Estimation of "Net Treatment Difference" controlling for observed gestation, NTD. Focus on a population with an overlap region of corresponding pscore (that includes gestation) between the 1 percentile of pscore for treated and 99 percentile of pscore for controls (N=424,677).			
5	WLS using pscore in weights. No pscore in regression	28.2	(23.56)
6	WLS using pscore in weights. Linear pscore as regressor	27.9	(23.29)
7	WLS using pscore in weights. Up to cubic pscore as regressor	27.9	(23.28)
Estimation of NATE using E[Y(1,S(1)) | S(1), X] to predict E[Y(1,S(0)) | S(0), X]. E[Y(0,0) | S(0), 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=425,219)			
8	OLS outcome on gestation and its polynomials up to degree 3 plus covariates	26.6	(23.46)
9	OLS outcome on gestation and its polynomials up to degree 3, covariates, and interactions	26.5	(23.36)
Estimation of the local NATE for the 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. Predicted values of S(0) and S(1) are based on matching on the pscore.			
Using pscore that does not include experience in its estimation. N=63,666			
10	WLS using pscore in weights. No pscore in regression	22.9	(9.51)
11	WLS using pscore in weights. Linear pscore as regressor	22.9	(9.47)
12	WLS using pscore in weights. Up to cubic pscore as regressor	22.9	(9.52)
Using pscore that includes experience in its estimation. N=63,748			
13	WLS using pscore in weights. No pscore in regression	22.8	(9.50)
14	WLS using pscore in weights. Linear pscore as regressor	22.7	(9.45)
15	WLS using pscore in weights. Up to cubic pscore as regressor	22.8	(9.49)

Note: The standard errors do not take into account the estimation of the propensity score. For comparison, the corresponding sample averages of incidence of low birth weight are 58.3 for all mothers, 48.4 for those that do not smoke during pregnancy, and 95.7 for those that smoke during pregnancy.

Table 3. Simulated p-values for Fisher's Randomization Test for the presence of individual effects of T on the post-treatment variable S. Based on 10,000 repetitions

		PANEL A						PANEL B	
		Random Assignment Application: testing the effect of Job Corps training on post-treatment experience						***Non-Random Assignment Application: testing the effect of smoking on gestation***	
		Full Sample		*Hispanics*		*Whites and Blacks*			
	Test based on:	*Population*	*Subpopulation*	*Population*	*Subpopulation*	*Population*	*Subpopulation*	*Population*	*Subpopulation*
1	OLS coefficient, no covariates	0.01	0.96	0.00	0.57	0.31	0.25	--	--
2	OLS coefficient, including pscore	0.06	0.78	0.00	0.99	0.38	0.90	0.00	0.35
3	OLS coefficient, including up to cubic pscore	0.06	0.77	0.01	0.94	0.39	0.91	0.00	0.45
4	OLS residuals, including pscore	0.06	0.79	0.01	0.99	0.50	0.84	0.00	0.77
5	OLS residuals, including up to pscore	0.06	0.79	0.01	0.95	0.39	0.91	0.00	0.92

Note: "Subpopulation" refers to that subpopulation for which (predicted) S(0)=S(1), which is used in the estimation of the local NATE.