working paper

13 10R

Covariates and Causal Effects: The Problem of Context

Dionissi Aliprantis

FEDERAL RESERVE BANK OF CLEVELAND

Working papers of the Federal Reserve Bank of Cleveland are preliminary materials circulated to stimulate discussion and critical comment on research in progress. They may not have been subject to the formal editorial review accorded official Federal Reserve Bank of Cleveland publications. The views stated herein are those of the authors and are not necessarily those of the Federal Reserve Bank of Cleveland or of the Board of Governors of the Federal Reserve System.

Working papers are available on the Cleveland Fed's website at:

www.clevelandfed.org/research.

Working Paper 13-10R March 2015

Covariates and Causal Effects: The Problem of Context

Dionissi Aliprantis

Because future experience is outside the support of the data, any prediction is based on assumptions about how the Data Generating Process (DGP) evolves over time. Investigating these assumptions raises basic questions about inductive inference with causal effects. This paper studies a class of DGPs lending themselves to analysis with potential outcomes. I show that even when treatment is randomized, there is a tradeoff between identification and prediction driven by a fact I call the problem of context: Treatment always influences the outcome variable in combination with covariates. While the response of covariates to variation in treatment impedes identification of direct effects, changes over time to the process generating covariates impedes prediction with total effects. As a result, stronger assumptions are required to identify direct effects than total effects, but direct effects can be used for prediction under weaker assumptions about the evolution of the DGP than total effects. To illustrate implications for applied work, I show that even if successfully identified in past data, total effects of educational attainment require strong assumptions about the behavior of covariates to be used for prediction.

Keywords: Identification, Prediction, Direct Effect, Total Effect, Dynamics, Lucas Critique, Transportability across Time, Exogeneity.

JEL Codes: C31, C36, C40, C51, C53, I21.

Suggested citation: Aliprantis, Dionissi, 2015. “Covariates and Causal Effects: The Problem of Context,” Federal Reserve Bank of Cleveland, working paper no. 13-10R.

Dionissi Aliprantis is at the Federal Reserve Bank of Cleveland; he can be reached at Dionissi. Aliprantis@clev.frb.org. The author thanks Francisca G.-C. Richter, Michela Tincani, Daniel Carroll, Mahmoud Elamin, Martin Huber, Giovanni Mellace, Peter Hinrichs, Nate Baum-Snow, Andrew Butters, Bruce Fallick, JamesWoodward, two anonymous referees, and seminar participants at the Cleveland Fed, GATE/Lyon 2, St. Gallen, and the MEA Conference for helpful comments.

*First version August 2013. First revision August 2014.

1 Introduction

Observing a randomized treatment, whether induced by human agency or natural processes, is the central objective of much empirical research. This is because randomization overcomes selection bias, or unobserved confounders, to identify causal effects within the potential outcomes framework (ie, the Rubin Causal Model). This paper studies a class of Data Generating Processes (DGPs) in which the goal of randomization is achieved and causal effects are identified.

Of what use are such causal effects once identified? Causal effects are thought to be useful for predicting the future state of the world in ways that descriptive statistics are not.[1] Despite this intended use and the careful attention given to the assumptions required for identification, the assumptions required to predict with causal effects have received little formal treatment.

This paper explicitly connects the identification of causal effects in past data with the prediction of future experience. Focusing on direct and total effects, which I define in terms of how corresponding interventions to the DGP affect covariates, I show that while identifying each distinct type of effect requires different assumptions, the same is true for prediction.[2] Unfortunately, the assumptions necessary for identification and prediction do not coincide, and so there is a tradeoff: Stronger assumptions about covariates must be invoked in the identification stage when using direct effects and in the prediction stage when using total effects.[3]

The relative strengths of direct and total effects are generated by a fact I call the problem of context: Treatment always influences the outcome variable in combination with covariates. I show how the relative strength of total effects for identification results from the response of covariates to variation in treatment. This feature of the DGP impedes scientists from identifying direct effects, which requires not only that an intervention would randomize treatment, but that it would also hold covariates at fixed values. In contrast, random variation in treatment is sufficient to identify total effects. When considered separately from the issue of prediction, the relative difficulty of identifying direct effects might actually be seen as a point of agreement in the literature contrasting structural and experimental approaches to econometrics (Deaton (2010),

[1] Zellner (2007) follows Jeffreys (2011) and others to distinguish between two key steps of science as being (1) *Description* of the past, and (2) Generalization/*Prediction* of future (or as of yet unobserved) experience. Angrist (2004) notes that "empirical research is almost always motivated by a belief that estimates for a particular context provide useful information about the likely effects of similar programmes or events in the future" (p C52). Similarly, Angrist and Pischke (2009) "believe that the most interesting research in social science is about questions of cause and effect . . . , " because "A causal relationship is useful for making predictions about the consequences of changing circumstances or policy; it tells us what would happen in alternative (or 'counterfactual') worlds" (p 3). Heckman and Vytlacil (2007) (pp 4787-92) and Manski (2007) (p 6) contrast this common goal in economics with a competing view that understanding causal effects adds to "knowledge" that is useful in some general sense.

[2] I define covariates as observable (but not necessarily observed) variables other than treatment that causally influence the outcome variable. As discussed in the paper, direct effects are useful for predicting outcomes under interventions allocating treatment while holding the values of covariates fixed (Pearl (2014b), Robins et al. (2009)). Total effects predict outcomes under interventions allocating treatment while holding the process generating covariates fixed (Angrist et al. (1996)). Because covariates can respond to treatment as a part of this process, interventions allocating treatment in this way can influence the outcome variable through any number of covariates known as mediators (Imai et al. (2010), VanderWeele (2009), Heckman and Pinto (2015), Sobel and Arminger (1992)).

[3] This tradeoff is under-appreciated because internal validity has received more formal attention than external validity in the literature contrasting the structural and experimental approaches devoted to one effect or the other.

Rosenzweig and Wolpin (2000), Keane (2010), White and Chalak (2013), Heckman (1997), Imbens (2010), Freedman (1987), Holland (1988)).

I also show how the relative strength of direct effects for prediction results from changes to the process generating covariates. Remember the problem of induction: Because future experience is outside the support of the data, any prediction is based on unverifiable assumptions about how the DGP evolves over time. A scientist predicting with total effects must make restrictive assumptions about the evolution of the DGP: *All* features of the DGP will remain the same.[4] In contrast, direct effects allow for prediction when only *some* features of the DGP remain the same.

A causal effect discussed in Freedman (1987) is illustrative: If we spend another million dollars on schools, how much will that affect test scores? In terms of Woodward (2003)'s notion of degrees of invariance, education funding and test scores have a weakly invariant relationship:[5] It matters how the money is spent! While randomized spending on schools might identify the total effect of spending on test scores, prediction based on this total effect will require that the behavior of all covariates remains the same over time.

I also illustrate implications of the identification-prediction tradeoff for the current literature by interpreting estimates of returns to schooling. I cite evidence that a long list of covariates that are determined in response to educational attainment also have large effects on wages, including on-the-job training; participation in job training programs; self-employment; vocational education; criminal behavior; arrest; incarceration; fertility; household formation; geographic location; military service; working while in school; smoking; and neighborhood quality. Selection into these covariates is likely to result in violations of the direct effect exclusion restriction requiring randomized and controlled variation in treatment (Deaton (2010), Rosenzweig and Wolpin (2000), Keane (2010)).

In such cases where we do not expect to observe randomized *and* controlled variation in treatment, randomized variation in treatment, and therefore identification of total effects, may be the best we can hope to observe (Imbens (2010)). But thinking in terms of returns to schooling, since human behavior related to selection into the above covariates is likely to change over time, so too are the total effects of the corresponding DGPs likely to change. If total effects of social systems are unstable over time, why should they be more useful than descriptive statistics for predicting the outcome variable under interventions manipulating the treatment variable?

Explicitly connecting identification and prediction links the literature on causal effects from the Rubin Causal Model to the macroeconomic literature using theory to construct predictions when

[4] When the DGP does not change over time the main obstacle to prediction is linking Local Average Treatment Effects (LATEs, Imbens and Angrist (1994)) relevant to a particular subpopulation to the ATE summarizing information about the DGP for the entire population (Angrist (2004)). A more complicated version of the issues studied in this paper also apply to the DGPs studied in the dynamic treatment effects literature where selection into treatment are dynamic (Frangakis and Rubin (2002), Robins (1986), Lechner and Miquel (2010), Heckman and Navarro (2007)). The distinguishing feature of the issues studied in this paper is that the dynamics under consideration will occur from the present moment into the future instead of occurring entirely in the past.

[5] This recent work in philosophy would characterize specific causal effects neither as exceptionless laws of nature nor as complete accidents, but rather as having a degree of invariance located somewhere between these all-or-nothing extremes (Woodward (2000), Woodward (2003)). This analysis has been strongly influenced by Woodward (2003)'s argument that generalizations invariant under some interventions need not be invariant under all others.

the process generating covariates is dynamic (Lucas (1976), Kydland and Prescott (1977)).[6] For such DGPs, formally studying prediction illustrates that internal validity need not be the *summum bonum* of inductive inference. If we define credibility as the strength of assumptions required to make a claim (Manski (2007)), then explicitly connecting identification and prediction highlights that the credibility revolution in empirical economics has been focused on only one step of inductive inference: Identifying causal effects in past data. Explicitly connecting identification and prediction also helps to extend the formal analysis of transportability (Only recently begun: See Pearl and Bareinboim (2011), Bareinboim and Pearl (2013b), Bareinboim and Pearl (2013a), Angrist (2004).) to populations for which not even passive observations can be collected (ie, populations in the future).

The paper proceeds as follows: Section 2 defines the set of DGPs to be considered in the paper. Section 3 presents three definitions of causal effects, and Section 4 discusses the identification of these causal effects in past data. Section 5 states the assumptions necessary to accurately predict future effects from causal effects identified in past data, and makes clear that researchers' choice between direct effects and total effects represents a tradeoff between the generality of the DGP that can be studied in the past and the generality of the DGP for which the future can be predicted. Section 6 discusses implications for the literature, with a focus on the fact that LATE and MTE estimates of returns to schooling are total effects. Section 7 concludes.

2 Data Generating Processes

Suppose that at time $t \in \mathbb{N}$ data are generated by a Data Generating Process (DGP) $\mathcal{D}_t$ in which the outcome variable (Y_{ti}) for each individual i is causally determined by two observed variables, treatment (D_{ti}) and observed covariates (X_{ti}), as well as unmeasured covariates (U_{ti}^Y), or additional factors not observed by the econometrician. The unmeasured covariates can be broken down into those factors that are unobserved (E_{ti}) and those that are unobservable (ϵ_{ti}) at the given level of measurement. To generalize the typical mediation problem (Pearl (2014a), Pearl (2014b)), we consider a class of four DGPs in which both observed and unobserved covariates might be determined by treatment. Where variables U are unmeasured variables, the four DGPs are characterized by the following structural equations:

Endogenous Variable	$\mathcal{D}_t^I$	$\mathcal{D}_t^{II}$	$\mathcal{D}_t^{III}$	$\mathcal{D}_t^{IV}$
$D_{ti} \stackrel{\leftarrow}{=}$	$f_t^D(U_{ti}^D)$	$f_t^D(U_{ti}^D)$	$f_t^D(U_{ti}^D)$	$f_t^D(U_{ti}^D)$
$X_{ti} \stackrel{\leftarrow}{=}$	$f_t^X(U_{ti}^X)$	$f_t^X(D_{ti}, U_{ti}^X)$	$f_t^X(U_{ti}^X)$	$f_t^X(D_{ti}, U_{ti}^X)$
$E_{ti} \stackrel{\leftarrow}{=}$	$f_t^E(U_{ti}^E)$	$f_t^E(U_{ti}^E)$	$f_t^E(D_{ti}, U_{ti}^E)$	$f_t^E(D_{ti}, U_{ti}^E)$
$Y_{ti} \stackrel{\leftarrow}{=}$	$f_t^Y(X_{ti}, D_{ti}, E_{ti}, \epsilon_{ti})$	$f_t^Y(X_{ti}, D_{ti}, E_{ti}, \epsilon_{ti})$	$f_t^Y(X_{ti}, D_{ti}, E_{ti}, \epsilon_{ti})$	$f_t^Y(X_{ti}, D_{ti}, E_{ti}, \epsilon_{ti})$

[6] A simple version of the Lucas (1976) critique can be restated as X_t being an indicator for guards at Fort Knox, D_t being an indicator for an attack on Fort Knox, and Y_t being the gold stolen from Fort Knox with $Y_t \stackrel{\leftarrow}{=} \delta D_t - \delta D_t X_t$. Total effects will change, and therefore so will decisions about whether to intervene to allocate treatment, when all features of DGP stay the same over time except for the process generating the covariate X_t (ie, f_t^X).

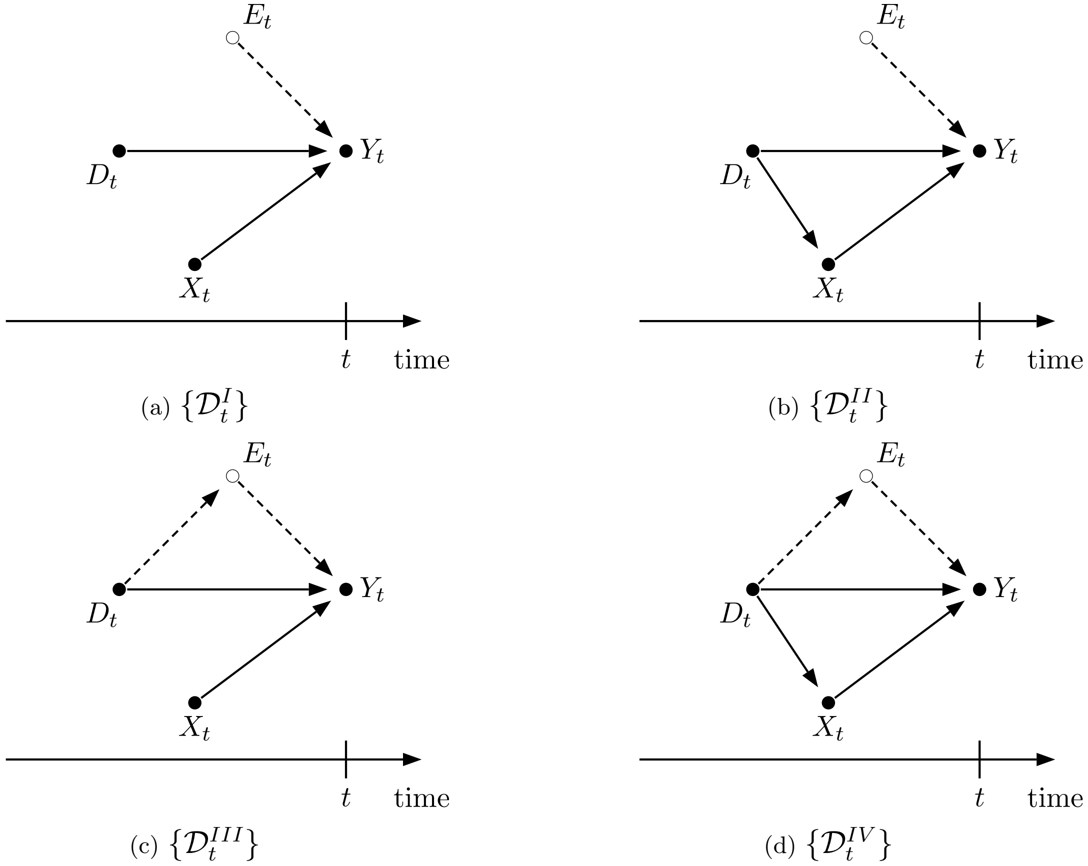
These DGPs are considered because they represent a canonical simple system that lend themselves to analysis with potential outcomes (Pearl et al. (2014)). I study a typical mediation system without confounders in order to focus on the role of unobserved mediators over time.[7] In order to distinguish clearly between functional forms f^V and parameterizations Θ^V, I will define a DGP as

$$\mathcal{D}_t \equiv < U_t, V_t, F_t, \Theta_t > \tag{1}$$

where $U_t \equiv (U_{ti}^D, U_{ti}^X, U_{ti}^E, U_{ti}^Y) = (U_{ti}^D, U_{ti}^X, U_{ti}^E, E_{ti}, \epsilon_{ti})$, $V_t \equiv (D_{ti}, X_{ti}, E_{ti}, Y_{ti})$, $F_t \equiv (f_t^D, f_t^X, f_t^E, f_t^Y)$, and $\Theta_t \equiv (\Theta_t^D, \Theta_t^X, \Theta_t^E, \Theta_t^Y)$.[8] I assume the functional form of f_t^Y is linear, so that the outcome equation is:[9]

$$Y_{ti} \overset{\leftarrow}{=} \theta_t^0 + D_{ti}\theta_t^1 + X_{ti}\theta_t^2 + E_{ti} + \epsilon_{ti}.$$

Figure 1 shows the four sets of DGPs $\{\mathcal{D}_t^I\}$, $\{\mathcal{D}_t^{II}\}$, $\{\mathcal{D}_t^{III}\}$, and $\{\mathcal{D}_t^{IV}\}$ comprised of each of the possible functional form specifications and parameterizations of the structural equations, and omitting the unmeasured factors in $\mathbf{U}$ except the unobserved factors in U^Y:



(a) $\{\mathcal{D}_t^I\}$

(b) $\{\mathcal{D}_t^{II}\}$

(c) $\{\mathcal{D}_t^{III}\}$

(d) $\{\mathcal{D}_t^{IV}\}$

Figure 1: Directed Acyclic Graphs of the Four Sets of Data Generating Processes

[7] Appendix D shows that the key results also apply to DGPs extended to include unobserved confounders.

[8] This is a slight deviation from Pearl (2009)'s Definition 7.1.1 of a structural causal model, where the triple $< U, V, F >$ would be defined as $U \equiv (U_{ti}^D, U_{ti}^X, U_{ti}^E, E_{ti}, \epsilon_{ti})$, $V \equiv (D_{ti}, X_{ti}, E_{ti}, Y_{ti})$, $F \equiv (f_t^D, f_t^X, f_t^E, f_t^Y)$.

[9] For the sake of exposition I refer interchangeably to Θ_t^Y and $(\theta_t^0, \theta_t^1, \theta_t^2)$.

The $\overset{\leftarrow}{=}$ notation communicates that an equation is structural in the following two senses (a) The equation represents an asymmetric relationship between the variables on the left and right hand sides of the equation; and (b) All variables at the given level of observation not included on the right hand side satisfy an exclusion restriction. Note that (b) rules out the possibility for finding new mediators at the given level of observation (Imai et al. (2010)), as mediators can only be found at a finer level of observation (A process which, as noted by Holland (1988), always appears to be possible.). Aliprantis (2015) discusses the definition of structural equation and its implications in greater detail.

3 Defining Causal Effects

3.1 Defining Causal Effects as Changes from Interventions to a DGP

One definition of causal effects is as a quantitative characterization of the change in the outcome variable that would result from an intervention to the DGP. Such interventions to the DGP can be characterized by how they would, or would not, impact covariates, especially unmeasured variables. In order to be precise about which features of the DGP are manipulated, and which are not, under specific interventions, I use Pearl (2009)'s *do*-operator throughout the remainder of the analysis.

Direct effects characterize the change in the outcome variable from a specific type of intervention to the DGP. Specifically, the controlled direct effect of D_t on Y_t, $\Delta_t^{CDE}(d', d)$, is the change in Y_t that would result from an intervention setting D_t from d to d' while setting all other variables entering as arguments in f^Y to fixed values:

$$\begin{aligned}\mathbf{\Delta_t^{CDE}(d', d)} &\equiv \mathbb{E}[Y_{ti}|do(D_{ti} = d', X_{ti} = x, E_{ti} = e)] - \mathbb{E}[Y_{ti}|do(D_{ti} = d, X_{ti} = x, E_{ti} = e)] \\ &= \mathbb{E}[f_t^Y(d', x, e, \epsilon_{ti})] - \mathbb{E}[f_t^Y(d, x, e, \epsilon_{ti})].\end{aligned}$$

Following Pearl (2014a), this definition is made at the population level, with individual-level effects given by the expressions under the expectation. Expectations are taken over ϵ_{ti} for $\Delta_t^{CDE}(d', d)$.

A second useful definition of causal effect is the total effect, which is DGP-specific:

DGP	$\mathbf{\Delta_t^{TE}(d', d)} \equiv \mathbb{E}[Y_{ti}|do(D_{ti} = d')] - \mathbb{E}[Y_{ti}|do(D_{ti} = d)]$
$\mathcal{D}_t^I$:	$\Delta_t^{TE}(d', d) = \mathbb{E}[\ f_t^Y(d',\ f_t^X(U_{ti}^X),\ f_t^E(U_{ti}^E),\ \epsilon_{ti})\ ] - \mathbb{E}[\ f_t^Y(d,\ f_t^X(U_{ti}^X),\ f_t^E(U_{ti}^E),\ \epsilon_{ti})\ ]$
$\mathcal{D}_t^{II}$:	$\Delta_t^{TE}(d', d) = \mathbb{E}[\ f_t^Y(d',\ f_t^X(d', U_{ti}^X),\ f_t^E(U_{ti}^E),\ \epsilon_{ti})\ ] - \mathbb{E}[\ f_t^Y(d,\ f_t^X(d, U_{ti}^X),\ f_t^E(U_{ti}^E),\ \epsilon_{ti})\ ]$
$\mathcal{D}_t^{III}$:	$\Delta_t^{TE}(d', d) = \mathbb{E}[\ f_t^Y(d',\ f_t^X(U_{ti}^X),\ f_t^E(d', U_{ti}^E),\ \epsilon_{ti})\ ] - \mathbb{E}[\ f_t^Y(d,\ f_t^X(U_{ti}^X),\ f_t^E(d, U_{ti}^E),\ \epsilon_{ti})\ ]$
$\mathcal{D}_t^{IV}$:	$\Delta_t^{TE}(d', d) = \mathbb{E}[\ f_t^Y(d',\ f_t^X(d', U_{ti}^X),\ f_t^E(d', U_{ti}^E),\ \epsilon_{ti})\ ] - \mathbb{E}[\ f_t^Y(d,\ f_t^X(d, U_{ti}^X),\ f_t^E(d, U_{ti}^E),\ \epsilon_{ti})\ ]$

As with the $\Delta_t^{DE}(d', d)$, this definition is also made at the population level, with individual-level effects given by the expressions under the expectation. Expectations are taken over U_{ti}^X, U_{ti}^E, and ϵ_{ti} for the $\Delta_t^{TE}(d', d)$.[10]

[10] I also refer to the Controlled Direct Effect simply as the Direct Effect, or $\Delta_{ti}^{CDE}(d', d) \equiv \Delta_{ti}^{DE}(d', d)$, since the

Defining the vector $S \equiv (D, X, E, \epsilon)$ allows us to re-write the outcome equation more compactly as $Y_{ti} \overset{\leftarrow}{=} Y_{ti}(S_{ti})$, so that both direct and total effects can be written in terms of the econometric or graphical definitions given in Heckman (2008) and Pearl (2009) (Definition 3.2.1).

Both $\Delta_t^{DE}(d', d)$ and $\Delta_t^{TE}(d', d)$ are defined in terms of interventions to the DGP. While the precise manipulation of variables can be described by the *do*-operator, these interventions can also be mimicked by instrumental variables. In order to be consistent with the DGP's structural equations (ie, Definition 5.4.1 in Pearl (2009)), instrumental variables must be components of the unmeasured variables U_t that become measured, perhaps due to a researcher's interest. Thus, instrumental variables should be added to a DAG as an observed variable (extracted from U_t) when observed, but left off of the DAG when remaining unmeasured (included in U_t).

For an instrumental variable to mimic the external variation generating a total effect, it must be an element of U_{ti}^D and U_{ti}^D alone. For an instrumental variable to mimic the variation generating a direct effect, it must be an element of U_{ti}^D and the U_t^{pa} of all of the parents pa_t of Y_t. When using Z_t to denote an instrumental variable, I will denote those mimicking total effects by Z_t^T and those mimicking direct effects by Z_t^D. I will at times add observed instruments to DAGs, without explicitly stating that this changes the relevant U_t.

3.2 Defining Causal Effects as Changes from Interventions to a Model

The Rubin Causal Model (Rubin (2005), Angrist et al. (1996)) defines causal effects in terms of the counterfactual outcome variable that would be observed under interventions to treatment. These counterfactual outcomes are also known as potential outcomes,

$$Y_{ti}(D_{ti}),$$

where $Y_{ti}(d)$ is the outcome of individual i at time t if treatment were set to $D_{ti} = d$ by an intervention setting D_{ti} but affecting none of the mediators of the total effect of D_{ti} on Y_{ti} (ie, none of the other parents of Y_{ti}). Although potential outcomes are generated by the DGP, they are defined as features of a model $\mathcal{M}_t$ that can describe many DGPs. That is, the average causal effect in the Rubin Causal Model is defined as

$$\boldsymbol{\Delta}_{\mathbf{t}}^{\mathbf{RCM}}(\mathbf{d}', \mathbf{d}) \equiv \mathbb{E}[Y_{ti}(d') - Y_{ti}(d)].$$

The expectation in $\Delta_t^{RCM}(d', d)$ is taken over individuals in the given population, allowing for any number of underlying functional forms and distributions. In contrast, the expectation in $\Delta_t^{TE}(d', d)$ is taken with respect to the single set of functional forms and distributions specified by the DGP. Aliprantis (2015) shows how interventions to a model rather than the DGP generate causal effects

Controlled Direct Effect is equal to the Natural Direct Effect in the DGPs under consideration. I leave generalization of this paper's results to nonlinear systems where $\Delta_{ti}^{CDE}(d', d) \neq \Delta_{ti}^{NDE}(d', d)$ for future work, since such generalizations are non-trivial (Pearl (2012)). Examples of such systems include DGPs similar to those in $\{\mathcal{D}_t^I\}$–$\{\mathcal{D}_t^{IV}\}$ but with a non-parametric structural outcome equation, or a parametric outcome equation with interaction terms like $Y_{ti} \overset{\leftarrow}{=} \theta_t^0 + D_{ti}\theta_t^1 + X_{ti}\theta_t^2 + D_{ti}X_{ti}\theta_t^3 + E_{ti} + \epsilon_{ti}$.

with distinct DAG representations and distinct uses in inductive inference.

4 Description of the Past: Identification of Causal Effects

In standard mediation or causal inference problems resembling those found in DGPs $\{\mathcal{D}_t^I\}$, $\{\mathcal{D}_t^{II}\}$, $\{\mathcal{D}_t^{III}\}$, or $\{\mathcal{D}_t^{IV}\}$, one of the following three regression equations is typically estimated via OLS, where H, K, and L are statistical error terms:

$$Y_{ti} = \alpha_t^0 + D_{ti}\alpha_t^1 + H_{ti}, \tag{2}$$

$$Y_{ti} = \beta_t^0 + D_{ti}\beta_t^1 + X_{ti}\beta_t^2 + K_{ti}, \tag{3}$$

$$Y_{ti} = \gamma_t^0 + D_{ti}\gamma_t^1 + X_{t_0 i}\gamma_t^2 + L_{ti}. \tag{4}$$

Equation 3 can be justified as the classical regression model with a few additional assumptions (Goldberger (1991), Chapters 15 and 16). And when it is suspected that D_{ti} is an argument in f_t^X, Angrist and Pischke (2009) recommend estimating Equation 2 (p 66) and Duflo et al. (2007) (p 3949) and Rosenbaum (1984) recommend estimating Equation 4.

We now investigate whether the parameters obtained from these three OLS regression equations identify any of the causal effects defined in Section 3. The main result is stated below in Proposition 1, but first we state some notation and maintained assumptions. We denote the current moment in time $t^* \in \mathbb{N}$, and consider DGPs indexed to three additional points in time:

$$t_0 < t < t^* < t'.$$

We interpret t_0 as the time at which treatment was assigned, t as the time at which observed variables were measured, and t' being some point in time in the future.[11] Given the dearth of attention paid to selection into covariates, here we assume selection into treatment is not a problem by assuming that treatment is randomly assigned:[12]

Random Assignment (RA) The variable D_{ti} is randomly allocated in the sample. Specifically, D_{ti} is an *iid* random variable that follows the triangle distribution with lower limit -1, upper limit 1, and mode 0.

[11] An implicit assumption in static models of causal effects is that dependent variables occur a short but finite time interval after independent variables (Pearl (1993), Holland (1986)). To be explicit about this assumption, we assume that at time $t - 4\epsilon \in \mathbb{R}^+$ nature evaluates the arguments of f_t^D, applies f_t^D to them under the given parameterization Θ_t, and sets the value of D_{ti} accordingly, where $0 < \epsilon << 1$. At time $t - 3\epsilon$, nature then proceeds to do the same for f_t^E. Nature proceeds similarly until ultimately finishing at time $t - \epsilon$ by evaluating the arguments of f_t^Y, applying f_t^Y to them under the parameterization Θ_t, and setting the value of Y_{ti} accordingly. The DGP $\mathcal{D}_t$ is indexed by $t \in \mathbb{N}$ because all of the observed variables in $\mathcal{D}_t$ are observed at time $t \in \mathbb{N}$, after all Equations have been evaluated by nature.

[12] I focus on OLS estimators applied to a randomized treatment because I want to focus on the issues raised by selection into covariates in isolation from those raised by selection into treatment. I show in Appendix A that the OLS estimators of α_t^1, β_t^1, and γ_t^1 converge in probability to the analogous 2SLS estimators when there is perfect compliance between the instrument and treatment. I also provide simulation results in Appendix C showing that the analysis generalizes immediately to identification schemes using instrumental variables estimators to overcome the distinct issue of selection into treatment.

Under these conditions, I show the following:

Proposition 1 (Identification Tradeoff)**.** *Define $\mathcal{I}_{t^*}(\Delta_t)$ as the subset of DGPs in $\{\mathcal{D}_t^I\} \cup \{\mathcal{D}_t^{II}\} \cup \{\mathcal{D}_t^{III}\} \cup \{\mathcal{D}_t^{IV}\}$ for which causal effects of type Δ_t are identified at the present time t^* by one of the OLS estimators $\widehat{\alpha}_t^{1,OLS}$, $\widehat{\beta}_t^{1,OLS}$, or $\widehat{\gamma}_t^{1,OLS}$. Then $\mathcal{I}_{t^*}(\Delta_t^{CDE}) \subset \mathcal{I}_{t^*}(\Delta_t^{TE})$.*

Proof. In Appendix A it is shown that when (D_{ti}, X_{ti}, E_{ti}) are excluded from f_{ti}^{ϵ},

$$\widehat{\alpha}_t^{1,OLS} \quad \xrightarrow{p} \quad \theta_t^1 + \frac{\mathbb{E}\left[D_t X_t\right]}{\mathbb{E}\left[D_t D_t\right]}\theta_t^2 + \frac{\mathbb{E}\left[D_t E_t\right]}{\mathbb{E}\left[D_t D_t\right]}, \tag{5}$$

$$\widehat{\beta}_t^{1,OLS} \quad \xrightarrow{p} \quad \theta_t^1 + \frac{\mathbb{E}[X_t X_t]\,\mathbb{E}[D_t E_t] \;-\; \mathbb{E}[D_t X_t]\,\mathbb{E}[X_t E_t]}{\mathbb{E}[D_t D_t]\,\mathbb{E}[X_t X_t] \;-\; \mathbb{E}[D_t X_t]\,\mathbb{E}[X_t D_t]}, \quad \text{and} \tag{6}$$

$$\begin{aligned} \widehat{\gamma}_t^{1,OLS} \quad \xrightarrow{p} \quad \theta_t^1 &+ \frac{\mathbb{E}[X_{t_0} X_{t_0}]\,\mathbb{E}[D_t X_t] \;-\; \mathbb{E}[D_t X_{t_0}]\,\mathbb{E}[X_{t_0} X_t]}{\mathbb{E}[D_t D_t]\,\mathbb{E}[X_{t_0} X_{t_0}] \;-\; \mathbb{E}[D_t X_{t_0}]\,\mathbb{E}[X_{t_0} D_t]}\theta_t^2 \\ &+ \frac{\mathbb{E}[X_{t_0} X_{t_0}]\,\mathbb{E}[D_t E_t] \;-\; \mathbb{E}[D_t X_{t_0}]\,\mathbb{E}[X_{t_0} E_t]}{\mathbb{E}[D_t D_t]\,\mathbb{E}[X_{t_0} X_{t_0}] \;-\; \mathbb{E}[D_t X_{t_0}]\,\mathbb{E}[X_{t_0} D_t]}. \end{aligned} \tag{7}$$

Since D_{ti} has mean zero, the following orthogonality conditions result from **RA**:

Orthogonality DX-t_0 $\mathbb{E}[D_t X_{t_0}] = 0$

Orthogonality DE-t_0 $\mathbb{E}[D_t E_{t_0}] = 0$

Remarkably, one of the orthogonality conditions induced by randomization (**DX-t_0**) implies that $\widehat{\alpha}_t^{1,OLS} \xrightarrow{p} \widehat{\gamma}_t^{1,OLS}$ as $N \to \infty$. Thus Equations 5-7 can be rewritten as:

$$\widehat{\alpha}_t^{1,OLS} \quad \xrightarrow{p} \quad \widehat{\gamma}_t^{1,OLS} \quad \xrightarrow{p} \quad \theta_t^1 + \frac{\mathbb{E}\left[D_t X_t\right]}{\mathbb{E}\left[D_t D_t\right]}\theta_t^2 + \frac{\mathbb{E}\left[D_t E_t\right]}{\mathbb{E}\left[D_t D_t\right]}, \quad \text{and} \tag{8}$$

$$\widehat{\beta}_t^{1,OLS} \quad \xrightarrow{p} \quad \theta_t^1 + \frac{\mathbb{E}[X_t X_t]\,\mathbb{E}[D_t E_t] \;-\; \mathbb{E}[D_t X_t]\,\mathbb{E}[X_t E_t]}{\mathbb{E}[D_t D_t]\,\mathbb{E}[X_t X_t] \;-\; \mathbb{E}[D_t X_t]\,\mathbb{E}[X_t D_t]}. \tag{9}$$

Not only does $\alpha_t^{1,OLS}$ convey the change in the outcome Y_{ti} for each unit of change in D_{ti} we observe in the data,

$$\begin{aligned} \mathbb{E}[Y_{ti}|D_{ti} = d+1] - \mathbb{E}[Y_{ti}|D_{ti} = d] &= \mathbb{E}[f_{ti}^Y(D_{ti}, X_{ti}, E_{ti}, \epsilon_{ti})|D_{ti} = d+1] \\ &\quad - \mathbb{E}[f_{ti}^Y(D_{ti}, X_{ti}, E_{ti}, \epsilon_{ti})|D_{ti} = d] \\ &= \mathbb{E}[\theta_t^0 + D_{ti}\theta_t^1 + X_{ti}\theta_t^3 + E_{ti} + \epsilon_{ti} \,|\, D_{ti} = d+1] \qquad (10) \\ &\quad - \mathbb{E}[\theta_t^0 + D_{ti}\theta_t^1 + X_{ti}\theta_t^2 + E_{ti} + \epsilon_{ti} \,|\, D_{ti} = d] \\ &= (d+1)\theta_t^1 + \mathbb{E}[X_{ti}\theta_t^2|D_{ti} = d+1] + \mathbb{E}[E_{ti} \,|\, D_{ti} = d+1] \qquad (11) \\ &\quad - d\theta_t^1 - \mathbb{E}[X_{ti}\theta_t^2|D_{ti} = d] + \mathbb{E}[E_{ti}|D_{ti} = d] \\ &= \theta_t^1 + \frac{\mathbb{E}\left[D_t X_t\right]}{\mathbb{E}\left[D_t D_t\right]}\theta_t^2 + \frac{\mathbb{E}\left[D_t E_t\right]}{\mathbb{E}\left[D_t D_t\right]} \qquad (12) \\ &= plim\ \widehat{\alpha}_t^{1,OLS}, \end{aligned}$$

but Random Assignment (**RA**) implies that it also identifies the change in Y_{ti} that would result if we were to counterfactually set treatment D_{ti} one unit higher. Specifically, **RA** implies that the average covariate for the subpopulation with D_{ti} observed to be d is equal to the average covariate when D_{ti} is set to d for the population:[13]

$$\begin{align}
\mathbb{E}[E_{ti}|D_{ti}=d] &= \mathbb{E}[E_{ti}|do(D_{ti}=d)], \text{ and} \tag{13}\\
\mathbb{E}[X_{ti}|D_{ti}=d] &= \mathbb{E}[X_{ti}|do(D_{ti}=d)]. \tag{14}
\end{align}$$

Equations 13 and 14 are a restatement of the independence assumption in Holland (1986) using the *do* operator, and are the link by which conditioning on treatment status identifies the total effect (justifying Equation 16):

$$\begin{align}
\mathbb{E}[Y_{ti}|D_{ti}=d+1]-\mathbb{E}[Y_{ti}|D_{ti}=d] &= \mathbb{E}[f^Y_{ti}(D_{ti},X_{ti},E_{ti},\epsilon_{ti})|D_{ti}=d+1] \notag\\
&\quad -\mathbb{E}[f^Y_{ti}(D_{ti},X_{ti},E_{ti},\epsilon_{ti})|D_{ti}=d] \notag\\
&= \mathbb{E}[\theta^0_t + D_{ti}\theta^1_t + X_{ti}\theta^3_t + E_{ti} + \epsilon_{ti}\,|\,D_{ti}=d+1] \tag{15}\\
&\quad -\mathbb{E}[\theta^0_t + D_{ti}\theta^1_t + X_{ti}\theta^2_t + E_{ti} + \epsilon_{ti}\,|\,D_{ti}=d] \notag\\
&= (d+1)\theta^1_t + \mathbb{E}[X_{ti}\theta^2_t|D_{ti}=d+1] + \mathbb{E}[E_{ti}\,|\,D_{ti}=d+1] \notag\\
&\quad - d\theta^1_t - \mathbb{E}[X_{ti}\theta^2_t|D_{ti}=d] - \mathbb{E}[E_{ti}|D_{ti}=d] \notag\\
&= \theta^1_t + \mathbb{E}[X_{ti}\theta^2_t|do(D_{ti}=d+1)] + \mathbb{E}[E_{ti}|do(D_{ti}=d+1)] \tag{16}\\
&\qquad - \mathbb{E}[X_{ti}\theta^2_t|do(D_{ti}=d)] - \mathbb{E}[E_{ti}|do(D_{ti}=d)] \notag\\
&= \theta^1_t + \mathbb{E}[f^X_t(d+1,U^X_{ti})\theta^2_t] + \mathbb{E}[f^E_t(d+1,U^E_{ti})] \notag\\
&\qquad - \mathbb{E}[f^X_t(d,U^X_{ti})\theta^2_t] - \mathbb{E}[f^E_t(d,U^E_{ti})] \notag\\
&= \mathbb{E}[Y_{ti}|do(D_{ti}=d+1)] - \mathbb{E}[Y_{ti}|do(D_{ti}=0)]. \tag{17}
\end{align}$$

Since D_{ti} is randomly assigned $\mathbb{E}[Y_{ti}|D_{ti}=d]=\mathbb{E}[Y_{ti}|D_{ti}=d,\, X_{t_0 i}=x]$, so that the preceding arguments resulting in Equations 8, 12, and 17 also prove that

$$plim\ \widehat{\gamma}^{1,OLS}_t = plim\ \widehat{\alpha}^{1,OLS}_t = \mathbb{E}[Y_{ti}|do(D_{ti}=d+1)] - \mathbb{E}[Y_{ti}|do(D_{ti}=d)] \equiv \mathbb{E}[\Delta^{TE}_{ti}(d+1,d)].$$

Note that the preceding identification result holds due to the functional form of f^Y_t, regardless of the specifications of f^X_t or f^E_t. Thus $\widehat{\alpha}^{1,OLS}_t$ and $\widehat{\gamma}^{1,OLS}_t$ identify the total effect of treatment on the outcome variable for any of the DGPs in $\{\mathcal{D}^I_t\}$, $\{\mathcal{D}^{II}_t\}$, $\{\mathcal{D}^{III}_t\}$, or $\{\mathcal{D}^{IV}_t\}$. Stated formally,

$$\mathcal{I}_{t^*}(\Delta^{TE}_t) = \{\mathcal{D}^I_t\} \cup \{\mathcal{D}^{II}_t\} \cup \{\mathcal{D}^{III}_t\} \cup \{\mathcal{D}^{IV}_t\}. \tag{18}$$

[13] Equations 13 and 14 do not ensure that the treated and nontreated groups are equal in all aspects apart from the treatment status (Heckman (1996)). The related assumptions in the program evaluation hold because E_{ti} in any of our DGPs is a different random variable than the error terms in the program evaluation literature (Imbens and Wooldridge (2009), Blundell and Dias (2009), Angrist et al. (1996)). A related discussion can be found in Heckman and Navarro-Lozano (2004). E_{ti} in the DGPs is also a different random variable than the error terms in the Conditional Expectation Functions (CEFs) discussed in Goldberger (1991) and Angrist and Pischke (2009).

In contrast, one can immediately see from Equations 8 and 9 that the orthogonality conditions **DX-t_0** and **DE-t_0** resulting from randomization are not sufficient for any of the regression coefficients in Equations 2-4 to identify the direct effect of treatment θ_t^1. This point has received considerable attention in the literature (Deaton (2010), Rosenzweig and Wolpin (2000), Keane (2010), White and Chalak (2013), Heckman (1997), Leamer (2010), Heckman and Smith (1995)).

Consider additional orthogonality conditions at the time of measurement:

Orthogonality DX-t $\mathbb{E}[D_t X_t] = 0$

Orthogonality DE-t $\mathbb{E}[D_t E_t] = 0$

Orthogonality XE-t $\mathbb{E}[X_t E_t] = 0$.

If DE-t and XE-t hold, then $\widehat{\beta}_t^{1,OLS}$ will converge in probability to the direct effect θ_t^1. These conditions will hold for DGPs in $\{\mathcal{D}_t^I\}$ or $\{\mathcal{D}_t^{II}\}$. Similarly, $\widehat{\alpha}_t^{1,OLS}$ and $\widehat{\beta}_t^{1,OLS}$ will converge in probability to the direct effect θ_t^1 when DX-t and DE-t hold. These conditions will only hold for DGPs in $\{\mathcal{D}_t^I\}$, for which the direct effect was already identified by $\widehat{\beta}_t^{1,OLS}$. Thus the direct effect of treatment on the outcome variable is only identified by one of the regression coefficients for DGPs in $\{\mathcal{D}_t^I\}$ or $\{\mathcal{D}_t^{II}\}$:

$$\mathcal{I}_{t^*}(\Delta_t^{CDE}) = \{\mathcal{D}_t^I\} \cup \{\mathcal{D}_t^{II}\}. \tag{19}$$

Thus, for the simple class of DGPs representing canonical mediation problems, traditionally used OLS estimators identify total effects for a broader class of DGPs than they identify direct effects:

$$\mathcal{I}_{t^*}(\Delta_t^{CDE}) \subset \mathcal{I}_{t^*}(\Delta_t^{TE}).$$

□

5 Generalization: Prediction with Causal Effects

Suppose that a social scientist has successfully identified a causal effect of one of the DGPs under consideration. Why would anyone be interested in such information? If we follow Zellner (2007) to distinguish between two key steps of science being (1) *Description* of the past, and (2) Generalization/*Prediction* of future (or as of yet unobserved) experience, social scientists' interest in causal effects is typically justified in terms of their use for *Prediction*.[14]

The problem of induction might be summarized as follows: Because future experience is outside the support of the data, any prediction is based on assumptions about how the DGP evolves over time. That is, the researcher at time t^* must assume the DGP at time t' will be $\mathcal{D}_{t^*,t'}$ in order to make a prediction about the random variable $V_{t'}$ or the parameter $\Theta_{t'}$. Recalling that $t_0 < t < t^* < t'$, I will use the subscript $_{t*,t'}$ to denote predictions at time t^* about features of the

[14] Zellner himself follows Karl Pearson, Harold Jeffreys, and others in this distinction. See Footnote 1 for prominent researchers citing *Prediction* as a justification for interest in causal effects.

future DGP. This includes functional forms, parameterizations, and random variables:

$$f^V_{t^*,t'} \in \mathcal{D}_{t^*,t'}, \qquad \Theta_{t^*,t'} \in \mathcal{D}_{t^*,t'}, \quad \text{or} \quad V_{t^*,t'} \in \mathcal{D}_{t^*,t'}$$

Evidence from the past like causal effects $\triangle^{TE}_t(d+1,d)$ or $\triangle^{DE}_t(d+1,d)$ can be used to construct predictions about the effects from future interventions, $\triangle^{TE}_{t^*,t'}(d+1,d)$. These predictions will be accurate (ie, $\triangle^{TE}_{t^*,t'}(d+1,d) = \triangle^{TE}_{t'}(d+1,d)$), under restrictions about the evolution of the DGP between times t and t'. The first restriction we impose, for all constructions of predictions, is that the DGPs under consideration exhibit a certain level of temporal stability:

Stability of the DGP **(S-DGP)**: $\qquad \mathcal{D}_{t'} \in \{\mathcal{D}^I_t\} \cup \{\mathcal{D}^{II}_t\} \cup \{\mathcal{D}^{III}_t\} \cup \{\mathcal{D}^{IV}_t\}$

Considering the following two ways of constructing predictions:

Prediction 1 $\triangle^{TE}_{t^*,t'}(d+1,d) = \mathbb{E}[Y_{t^*,t'}|do(d+1)] - \mathbb{E}[Y_{t^*,t'}|do(d)]$

Prediction 2 $\triangle^{TE}_{t^*,t'}(d+1,d) = \triangle^{TE}_t(d+1,d)$

allows us to state the following proposition:

Proposition. *Define $\mathcal{P}_{t^*,t'}(p)$ as the subset of DGPs at time t' satisfying S-DGP for which Prediction p made at time t^* could possibly be accurate, or for which it would be possible that $\Delta_{t^*,t'}(d+1,d) = \Delta_{t'}(d+1,d)$. Then*

$$\mathcal{P}_{t^*,t'}(2) \subset \mathcal{P}_{t^*,t'}(1).$$

Since Prediction 1 uses direct effects from past data, and Prediction 2 uses total effects from past data, we could restate this Proposition as

Proposition 2 (Prediction Tradeoff)**.** *Define $\mathcal{P}_{t^*,t'}(\Delta_t)$ as the subset of DGPs at time t' satisfying S-DGP for which causal effects of type $\Delta_t(d+1,d)$ can be used under constructions Prediction 1 or Prediction 2 at the present time t^* to make predictions that could possibly be accurate. Then*

$$\mathcal{P}_{t^*,t'}(\Delta^{TE}_t) \subset \mathcal{P}_{t^*,t'}(\Delta^{DE}_t).$$

Proof. Prediction 1 is accurate under the following assumptions about the researcher's ability to forecast features of the future DGP $\mathcal{D}_{t'}$:

Structural Equations (**SEs**)	$f^Y_{t^*,t'} = f^Y_{t'}$	$f^X_{t^*,t'} = f^X_{t'}$	$f^E_{t^*,t'} = f^E_{t'}$
Parameterizations (**Ps**)	$\Theta^Y_{t^*,t'} = \Theta^Y_{t'}$	$\Theta^X_{t^*,t'} = \Theta^X_{t'}$	$\Theta^E_{t^*,t'} = \Theta^E_{t'}$
Unmeasured Variables (**UVs**)	$\mu_{t^*,t'}(U) = \mu_{t'}(U)$		

Assuming without loss of generality that $\mathcal{D}_{t'} \in \{\mathcal{D}^{IV}_t\}$. Assumptions S-DGP, SEs, Ps, UVs imply that Prediction 1 is accurate as follows:

$$\triangle^{TE}_{t^*,t'}(d+1,d) = \mathbb{E}[Y_{t^*,t'}|do(d+1)] - \mathbb{E}[Y_{t^*,t'}|do(d)] \tag{Prediction 1}$$

$$= \mathbb{E}\Big[\; f^Y_{t^*,t'}\big(d+1,\;\; f^X_{t^*,t'}(d+1, U^X_{t^*,t'i}),\;\; f^E_{t^*,t'}(d+1, U^E_{t^*,t'i}),\;\; \epsilon_{t^*,t'i};\;\; \Theta^Y_{t^*,t'}\big)\;\Big] \tag{Def/}$$

$$-\;\; \mathbb{E}\Big[\; f^Y_{t^*,t'}\big(d,\;\; f^X_{t^*,t'}(d, U^X_{t^*,t'i}),\;\; f^E_{t^*,t'}(d, U^E_{t^*,t'i}),\;\; \epsilon_{t^*,t'i};\;\; \Theta^Y_{t^*,t'}\big)\;\Big] \tag{$\mathcal{D}_{t'} \in \{\mathcal{D}^{IV}_t\}$}$$

$$\begin{aligned} &= (d+1)\,\theta^1_{t^*,t'} + \mathbb{E}[f^X_{t^*,t'}(d+1, U^X_{t^*,t'i})]\theta^2_{t^*,t'} + \mathbb{E}[f^E_{t^*,t'}(d+1, U^E_{t^*,t'i})] \\ &\quad -\;\; d\,\theta^1_{t^*,t'} + \mathbb{E}[f^X_{t^*,t'}(d, U^X_{t^*,t'i})]\theta^2_{t^*,t'} + \mathbb{E}[f^E_{t^*,t'}(d, U^E_{t^*,t'i}))] \end{aligned} \tag{\textbf{S-DGP}}$$

$$\begin{aligned} &= (d+1)\,\theta^1_{t^*,t'} + \mathbb{E}[\boldsymbol{f^X_{t'}}(d+1, U^X_{t^*,t'i})]\theta^2_{t^*,t'} + \mathbb{E}[\boldsymbol{f^E_{t'}}(d+1, U^E_{t^*,t'i})] \\ &\quad -\;\; d\,\theta^1_{t^*,t'} + \mathbb{E}[\boldsymbol{f^X_{t'}}(d, U^X_{t^*,t'i})]\theta^2_{t^*,t'} + \mathbb{E}[\boldsymbol{f^E_{t'}}(d, U^E_{t^*,t'i}))] \end{aligned} \tag{\textbf{SEs}}$$

$$\begin{aligned} &= (d+1)\,\boldsymbol{\theta^1_{t'}} + \mathbb{E}[f^X_{t'}(d+1, U^X_{t^*,t'i})]\boldsymbol{\theta^2_{t'}} + \mathbb{E}[f^E_{t'}(d+1, U^E_{t^*,t'i})] \\ &\quad -\;\; d\,\boldsymbol{\theta^1_{t'}} + \mathbb{E}[f^X_{t'}(d, U^X_{t^*,t'i})]\boldsymbol{\theta^2_{t'}} + \mathbb{E}[f^E_{t'}(d, U^E_{t^*,t'i}))] \end{aligned} \tag{\textbf{Ps}}$$

$$\begin{aligned} &= (d+1)\,\theta^1_{t'} + \mathbb{E}[f^X_{t'}(d+1, \boldsymbol{U^X_{t'i}})]\theta^2_{t'} + \mathbb{E}[f^E_{t'}(d+1, \boldsymbol{U^E_{t'i}})] \\ &\quad -\;\; d\,\theta^1_{t'} + \mathbb{E}[f^X_{t'}(d, \boldsymbol{U^X_{t'i}})]\theta^2_{t'} + \mathbb{E}[f^E_{t'}(d, \boldsymbol{U^E_{t'i}}))] \end{aligned} \tag{\textbf{UVs}}$$

$$= \triangle^{TE}_{t'}(d+1,d). \tag{Definition/$\mathcal{D}_{t'} \in \{\mathcal{D}^{IV}_t\}$ + \textbf{S-DGP}}$$

Identification of the direct effect at time t in the past ($\widehat{\theta^1_t} = \theta^1_t$) gives credibility to assumption Ps with respect to $\Theta^Y_{t^*,t'}$ given one stability assumption:

Stability of Direct Effect (**S-DE**) $\theta^1_t = \theta^1_{t'}$

To prove that these assumptions are necessary, and not only sufficient, suppose by way of contradiction that any of SEs, Ps, or UVs were not true, and the above equations will deliver a contradiction of the prediction's accuracy.

Prediction 2 is accurate under the following assumptions about the stability of the DGP:

Stability of Structural Equations **(S-SEs)** $\quad f^Y_t = f^Y_{t'} \quad f^X_t = f^X_{t'} \quad f^E_t = f^E_{t'}$
Stability of Parameterizations **(S-Ps)** $\quad \Theta^Y_t = \Theta^Y_{t'} \quad \Theta^X_t = \Theta^X_{t'} \quad \Theta^E_t = \Theta^E_{t'}$
Stability of Unmeasured Variables **(S-UVs)** $\quad \mu_t(U) = \mu_{t'}(U)$

Assumptions S-DGP, S-SEs, S-Ps, S-UVs imply that Prediction 2 is accurate as follows:

$$
\begin{aligned}
\triangle^{TE}_{t^*,t'}(d+1,d) &= \triangle^{TE}_{t}(d+1,d) && \text{(Prediction 2)}\\
&= \mathbb{E}\Big[\ f^Y_t\big(d+1,\ \ f^X_t(d+1,U^X_{ti}),\ \ f^E_t(d+1,U^E_{ti}),\ \ \epsilon_{ti};\ \ \Theta^Y_t\big)\ \Big] && \text{(Definition)}\\
&\quad -\ \mathbb{E}\Big[\ f^Y_t\big(d,\ \ f^X_t(d,U^X_{ti}),\ \ f^E_t(d,U^E_{ti}),\ \ \epsilon_{ti};\ \ \Theta^Y_t\big)\ \Big]\\[2ex]
&= \mathbb{E}\Big[\ \boldsymbol{f^Y_{t'}}\big(d+1,\ \ \boldsymbol{f^X_{t'}}(d+1,U^X_{ti}),\ \ \boldsymbol{f^E_{t'}}(d+1,U^E_{ti}),\ \ \epsilon_{ti};\ \ \Theta^Y_t\big)\ \Big] && \textbf{(S-SEs)}\\
&\quad -\ \mathbb{E}\Big[\ \boldsymbol{f^Y_{t'}}\big(d,\ \ \boldsymbol{f^X_{t'}}(d,U^X_{ti}),\ \ \boldsymbol{f^E_{t'}}(d,U^E_{ti}),\ \ \epsilon_{ti};\ \ \Theta^Y_t\big)\ \Big]\\[2ex]
&= \mathbb{E}\Big[\ f^Y_{t'}\big(d+1,\ \ f^X_{t'}(d+1,U^X_{ti}),\ \ f^E_{t'}(d+1,U^E_{ti}),\ \ \epsilon_{ti};\ \ \boldsymbol{\Theta^Y_{t'}}\big)\ \Big] && \textbf{(S-Ps)}\\
&\quad -\ \mathbb{E}\Big[\ f^Y_{t'}\big(d,\ \ f^X_{t'}(d,U^X_{ti}),\ \ f^E_{t'}(d,U^E_{ti}),\ \ \epsilon_{ti};\ \ \boldsymbol{\Theta^Y_{t'}}\big)\ \Big]\\[2ex]
&= \mathbb{E}\Big[\ f^Y_{t'}\big(d+1,\ \ f^X_{t'}(d+1,\boldsymbol{U^X_{t'i}}),\ \ f^E_{t'}(d+1,\boldsymbol{U^E_{t'i}}),\ \ \boldsymbol{\epsilon_{t'i}};\ \ \Theta^Y_{t'}\big)\ \Big] && \textbf{(S-UVs)}\\
&\quad -\ \mathbb{E}\Big[\ f^Y_{t'}\big(d,\ \ f^X_{t'}(d,\boldsymbol{U^X_{t'i}}),\ \ f^E_{t'}(d,\boldsymbol{U^E_{t'i}}),\ \ \boldsymbol{\epsilon_{t'i}};\ \ \Theta^Y_{t'}\big)\ \Big]\\[2ex]
&= \triangle^{TE}_{t'}(d+1,d). && \text{(Definition)}
\end{aligned}
$$

Note that the first line, $\triangle^{TE}_{t^*,t'}(d+1,d) = \triangle^{TE}_{t}(d+1,d)$, implies that

$$
\begin{aligned}
&\mathbb{E}\Big[\ f^Y_{t^*,t'}\big(d+1,\ \ f^X_{t^*,t'}(d+1,U^X_{t^*,t'i}),\ \ f^E_{t^*,t'}(d+1,U^E_{t^*,t'i}),\ \ \epsilon_{t^*,t'i};\ \ \Theta^Y_{t^*,t'}\big)\ \Big]\\
&\quad -\ \mathbb{E}\Big[\ f^Y_{t^*,t'}\big(d,\ \ f^X_{t^*,t'}(d,U^X_{t^*,t'i}),\ \ f^E_{t^*,t'}(d,U^E_{t^*,t'i}),\ \ \epsilon_{t^*,t'i};\ \ \Theta^Y_{t^*,t'}\big)\ \Big]\\
=&\mathbb{E}\Big[\ f^Y_t\big(d+1,\ \ f^X_t(d+1,U^X_{ti}),\ \ f^E_t(d+1,U^E_{ti}),\ \ \epsilon_{ti};\ \ \Theta^Y_t\big)\ \Big]\\
&\quad -\ \mathbb{E}\Big[\ f^Y_t\big(d,\ \ f^X_t(d,U^X_{ti}),\ \ f^E_t(d,U^E_{ti}),\ \ \epsilon_{ti};\ \ \Theta^Y_t\big)\ \Big],
\end{aligned}
$$

which under the stability assumptions S-SEs, S-Ps, and S-UVs implies the original forecastability assumptions SEs, Ps, and UVs.[15] To prove that Prediction 2's assumptions are necessary like Prediction 1's, and not only sufficient, suppose analogously by way of contradiction that any of S-SEs, S-Ps, or S-UVs were not true, and the above equations will deliver a contradiction of the prediction's accuracy.

The set of DGPs at future time t' for which assumptions S-DGP, S-DE, SEs, Ps, and UVs could possibly be true is

$$\mathcal{P}_{t^*,t'}(\Delta_t^{DE}) = \left\{ \mathcal{D}_{t'} \;\middle|\; \theta_{t'}^1 = \theta_t^1 \;\;,\;\; \mathcal{D}_{t'} \in \{\mathcal{D}_t^I\} \cup \{\mathcal{D}_t^{II}\} \cup \{\mathcal{D}_t^{III}\} \cup \{\mathcal{D}_t^{IV}\} \right\}.$$

The set of DGPs at future time t' for which assumptions S-DGP, S-SEs, S-Ps, and S-UVs could possibly be true is

$$\mathcal{P}_{t^*,t'}(\Delta_t^{TE}) = \left\{ \mathcal{D}_{t'} \;\middle|\; \Theta_{t'}^Y = \Theta_t^Y, \Theta_{t'}^X = \Theta_t^X, \Theta_{t'}^E = \Theta_t^E; \quad f_{t'}^Y = f_t^Y, f_{t'}^X = f_t^X, f_{t'}^E = f_t^E; \quad \mu_{t'}(U) = \mu_t(U); \right.$$
$$\left. \mathcal{D}_{t'} \in \{\mathcal{D}_t^I\} \cup \{\mathcal{D}_t^{II}\} \cup \{\mathcal{D}_t^{III}\} \cup \{\mathcal{D}_t^{IV}\} \right\}.$$

Thus, it is clearly the case that

$$\mathcal{P}_{t^*,t'}(\Delta_t^{TE}) \subset \mathcal{P}_{t^*,t'}(\Delta_t^{DE}).$$

$\square$

6 Implications for the Literature

6.1 One Example: The Effect of Education Spending on Test Scores

Proposition 2 showed that for a class of DGPs resembling standard mediation problems, accurate prediction with total effects requires strong stability restrictions on the DGP. In contrast, accurate prediction with direct effects is possible for DGPs that change over time. The temporal stability required of the DGP to predict with total effects is more likely to be violated in social systems than in physical or biological systems.

A causal effect discussed in Freedman (1987) can help to illustrate: If we spend another million dollars on schools, how much will that affect test scores? This might be seen as an ill-posed problem: It matters how the money is spent! One can imagine many different mechanisms through which test scores might be affected, and these mechanisms need not be stable over time (Fruehwirth (2014), Carrell et al. (2013)).

Consider one such mechanism: The quantity supplied of teachers with content knowledge at a

[15] Also note that this first line asserts an analogue to direct transportability (Pearl and Bareinboim (2014)).

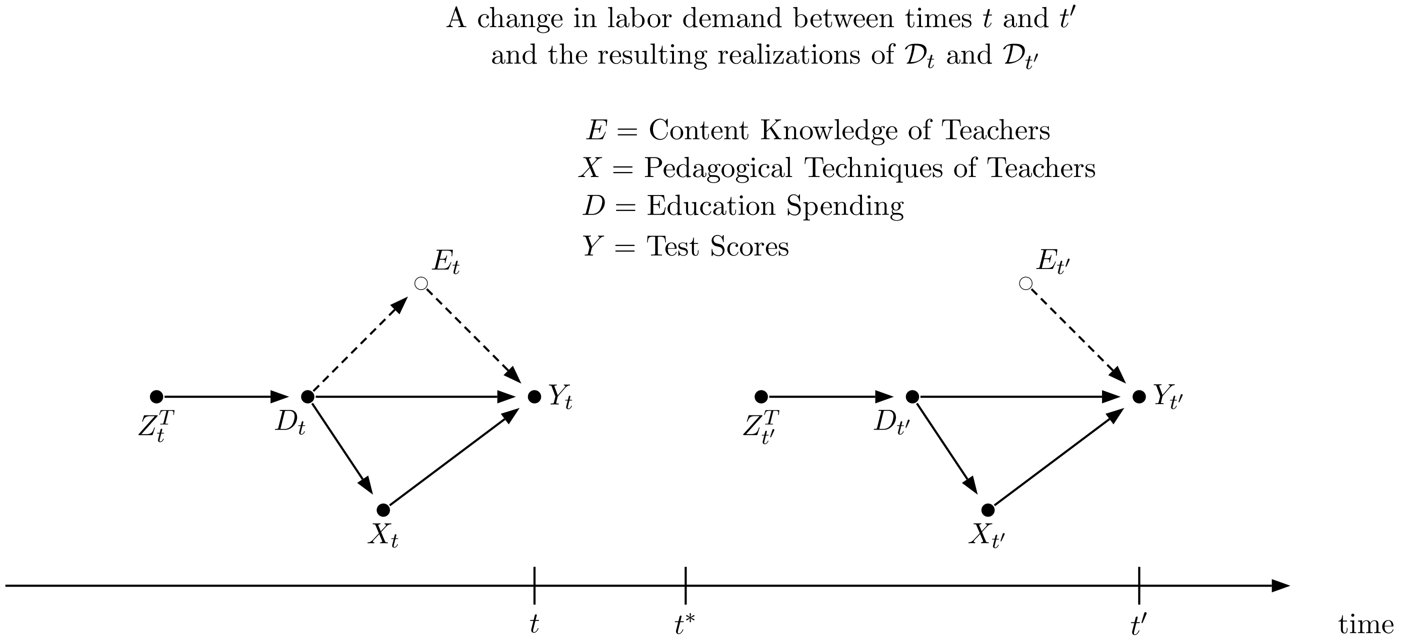

A change in labor demand between times t and t'
and the resulting realizations of $\mathcal{D}_t$ and $\mathcal{D}_{t'}$
E = Content Knowledge of Teachers
X = Pedagogical Techniques of Teachers
D = Education Spending
Y = Test Scores
E_t
Z_t^T
D_t
Y_t
X_t
$E_{t'}$
$Z_{t'}^T$
$D_{t'}$
$Y_{t'}$
$X_{t'}$
t
t^*
t'
time

Figure 2: An example in which predictions of the future effect of $D_{t'}$ on $Y_{t'}$ (ie, of changes to $Y_{t'}$ from the intervention $Z_{t'}^T$) will be biased when constructed at current time t^* using total effects identified from past data collected at time t

given salary. Figure 2 shows an example DGP in which math test scores Y are determined in part by education spending D at time t ($\mathcal{D}_t$), along with its successor DGP determining Y at time $t' > t$ ($\mathcal{D}_{t'}$). Even if a researcher at the current moment in time (t^*) knew the value of the total effect identified from data in the past (at time t), the DGP might change in many ways between times t and t', rendering predictions with total effects inaccurate. In the example shown in Figure 2, a change in labor demand in the broader economy makes the previously responsive quantity supplied of teachers with mathematics knowledge unchanged over teachers' salary range.

6.2 Another Example: Returns to Schooling

A large literature is devoted to estimating the causal effects of educational attainment. The key reason for this focus is that policy makers and citizens might intervene to the DGP to encourage or discourage students from finishing high school and/or college. If we knew the changes from an intervention manipulating educational attainment D, it would help us to decide how much to spend as a society to implement that intervention.

Identifying causal effects of educational attainment is complicated by selection into treatment in response to the unobserved covariate ability (Card (2001), Belzil (2007)). Although overcoming selection into treatment is a non-trivial task (Angrist and Krueger (1991), Aliprantis (2012), Barua and Lang (2009)), suppose for the moment that social scientists had found empirical methods overcoming the unobserved nature of ability, and had identified the returns to schooling based on samples in which attainment were randomly assigned.

Would the total effects of education on earnings identified in past data be useful for predicting how wages would change in the future under interventions to increase educational attainment? Marginal Treatment Effects (MTEs) and Local Average Treatment Effects (LATEs) of educational attainment estimated in the literature like in Carneiro et al. (2011), Oreopoulos (2006a), and Oreopoulos (2006b) are non-parametric parameters, so it is not clear that Propositions 1 and 2 apply to them. Nevertheless, many mediators, or covariates like X_t or E_t, could themselves respond to the assignment of treatment. MTEs and LATEs of educational attainment may reflect the direct effect of educational attainment, but due to selection into covariates may also reflect direct effects from many causal variables like:

Table 1: Covariates Whose Behavior Determines the Total Effect

Covariate	Evidence of Selection into Covariate in Response to Ed Attainment	Evidence of Effect of Covariate on Wages
On-the-Job Training	Altonji and Blank (1999)	Brown (1989)
Job Training Program	–	Lee (2009)
Self-Employment	Blanchflower (2000)	Hamilton (2000)
Vocational Education	Bishop and Mane (2004)	Meer (2007)
Criminal Behavior	Jacob and Lefgren (2003)	Nagin and Waldfogel (1998)
Arrest	Grogger (1995)	Bushway (2004)
Incarceration	Lochner and Moretti (2004)	Kling (2006), Western et al. (2001)
Fertility	McCrary and Royer (2011)	Simonsen and Skipper (2006)
Household Formation	Nielsen and Svarer (2009)	Gemici (2011)
Geographic Location	Costa and Kahn (2000)	Baum-Snow and Pavan (2013), Black et al. (2009)
Military Service	Small and Rosenbaum (2008)	Angrist (1990)
Health (smoking)	Currie and Moretti (2003)	Auld (2005)
Working While in School	–	Light (2001)
Neighborhood Quality	–	Rosenbaum (1995), Aliprantis and Richter (2013)

This example helps to illustrate why total effects can be so weakly invariant in social settings. If the process generating the covariate were to change over time for any of these covariates, the total effects of attainment estimated in the literature would give biased predictions. Figure 3 displays DAGs of the total effects of educational attainment on wages at times t and t'. Suppose that at time t companies only provided on-the-job training to employees with certain levels of educational attainment. If this policy were to change between times t and t' so that at time t' companies provided training to all employees, regardless of their educational attainment level, then total effects would change.[16]

Accurate prediction with total effects requires there are not changes over time to the DGP related to *any* of the direct effects contributing to the total effect. It is difficult to imagine that the social processes related to *each* of the above mediators do not change in important ways over time.

[16] Similar examples can be found in Cartwright and Hardie (2012), who distinguish the evidence necessary to make accurate statements about the past from the evidence necessary to make accurate statements about the future. This can also be thought of as an example of how super exogeneity assumptions need not follow from weak exogeneity (Engle et al. (1983), Hendry and Richard (1982)).

A change in the direct effect of educational attainment on on-the-job training
between times t and t' and the resulting realizations of $\mathcal{D}_t$ and $\mathcal{D}_{t'}$

M_1 = On-the-Job Training
M_2 = Job Training Program
M_3 = Self-Employment
M_4 = Vocational Education
M_5 = Criminal Behavior
M_6 = Arrest
M_7 = Incarceration
M_8 = Fertility
M_9 = Household Formation
M_{10} = Geographic Location
M_{11} = Military Service
M_{12} = Health (smoking)
M_{13} = Working While in School
M_{14} = Neighborhood Quality
Y = Wages

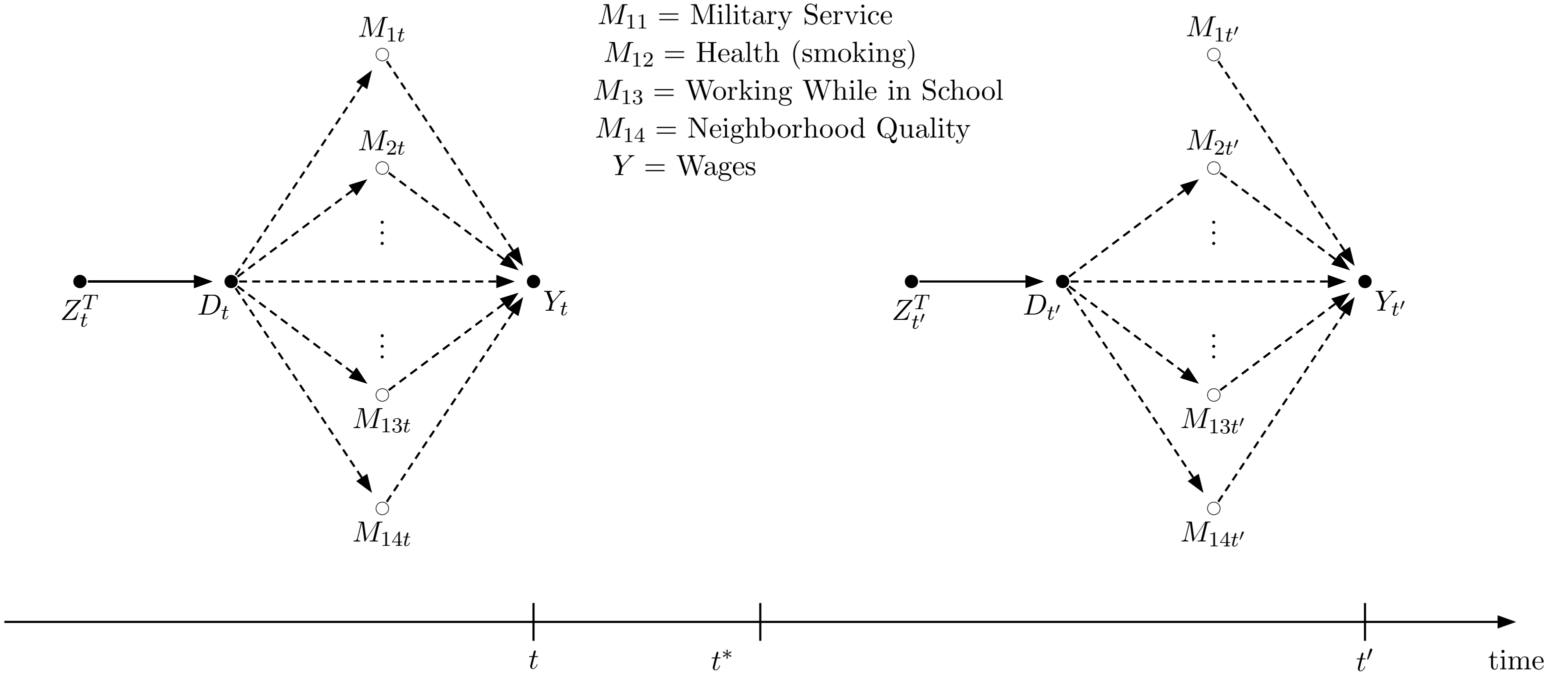


Figure 3: An example in which predictions of the future effect of $D_{t'}$ on $Y_{t'}$ (ie, of changes to $Y_{t'}$ from the intervention $Z_{t'}^T$) will be biased when constructed at current time t^* using total effects identified from past data (ie, from data collected at time t)

7 Conclusion

This paper studied a simple dynamic extension to the canonical static treatment effect framework. Treatment always influences the outcome variable in combination with other variables, which I refer to as covariates. I showed that for a class of DGPs representing a standard mediation problem, there is a tradeoff between how easy it is to identify a causal effect in past data and its usefulness for predicting the future. This tradeoff arises because covariates can respond even to a randomized treatment, and the behavior of covariates can change over time. I used the effects of education spending on test scores and of schooling on wages as examples to discuss why human agency is likely to change the behavior of covariates over time in many social systems.

References

Aliprantis, D. (2012). Redshirting, compulsory schooling laws, and educational attainment. *Journal of Educational and Behavioral Statistics 37*(2), 316–338.

Aliprantis, D. (2015). A distinction between causal effects in Structural and Rubin Causal Models. *Federal Reserve Bank of Cleveland Working Paper*.

Aliprantis, D. and F. G.-C. Richter (2013). Evidence of neighborhood effects from MTO: LATEs of neighborhood quality. *Mimeo., Federal Reserve Bank of Cleveland.*

Altonji, J. G. and R. M. Blank (1999). Race and gender in the labor market. In O. Ashenfeher and D. Card (Eds.), *Handbook of Labor Economics*, Volume 3, pp. 3143–3259. North-Holland.

Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from Social Security Administration records. *American Economic Review 80*(3), 313–335.

Angrist, J. D. (2004). Treatment effect heterogeneity in theory and practice. *The Economic Journal 114*(494), pp. C52–C83.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using Instrumental Variables. *Journal of the American Statistical Association 91*(434), 444–455.

Angrist, J. D. and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics 106*(4), 979–1014.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Auld, M. C. (2005). Smoking, drinking, and income. *Journal of Human Resources 40*(2), 505–518.

Bareinboim, E. and J. Pearl (2013a). A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference 1*(1), 107–133.

Bareinboim, E. and J. Pearl (2013b, Forthcoming). Meta-transportability of causal effects: A formal approach. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS).*

Barua, R. and K. Lang (2009). School entry, educational attainment and quarter of birth: A cautionary tale of LATE. *NBER Working Paper 15236*.

Baum-Snow, N. and R. Pavan (2013, Forthcoming). Inequality and city size. *Review of Economics and Statistics*.

Belzil, C. (2007). The return to schooling in structural dynamic models: A survey. *European Economic Review 51*(5), 1059–1105.

Bishop, J. H. and F. Mane (2004). The impacts of career-technical education on high school labor market success. *Economics of Education Review 23*(4), 381–402.

Black, D., N. Kolesnikova, and L. Taylor (2009). Earnings functions when wages and prices vary by location. *Journal of Labor Economics 27*(1), 21–47.

Blanchflower, D. G. (2000). Self-employment in OECD countries. *Labour Economics 7*(5), 471 – 505.

Blundell, R. and M. C. Dias (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources 44*(3), 565–640.

Brown, J. N. (1989). Why do wages increase with tenure? on-the-job training and life-cycle wage growth observed within firms. *The American Economic Review*, 971–991.

Bushway, S. D. (2004). Labor market effects of permitting employer access to criminal history records. *Journal of Contemporary Criminal Justice 20*(3), 276–291.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica 69*(5), 1127–1160.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011). Estimating marginal returns to education. *American Economic Review 101*(6), 2754–2781.

Carrell, S. E., B. I. Sacerdote, and J. E. West (2013). From natural variation to optimal policy? The importance of endogenous peer group formation. *Econometrica 81*(3), 855–882.

Cartwright, N. and J. Hardie (2012). *Evidence-Based Policy: A Practical Guide to Doing It Better.* Oxford University Press.

Chalak, K. and H. White (2011). An extended class of instrumental variables for the estimation of causal effects. *Canadian Journal of Economics 44*(1), 1–51.

Costa, D. L. and M. E. Kahn (2000). Power couples: Changes in the locational choice of the college educated, 1940-1990. *The Quarterly Journal of Economics 115*(4), 1287–1315.

Currie, J. and E. Moretti (2003). Mother's education and the intergenerational transmission of human capital: Evidence from college openings. *The Quarterly Journal of Economics 118*(4), 1495–1532.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature 48*(2), 424–455.

Duflo, E., R. Glennerster, and M. Kremer (2007). Using randomization in development economics research: A toolkit. In T. P. Schultz and J. A. Strauss (Eds.), *Handbook of Development Economics*, Volume 4, pp. 3895–3962. Elsevier.

Engle, R. F., D. F. Hendry, and J.-F. Richard (1983). Exogeneity. *Econometrica 51*(2), pp. 277–304.

Frangakis, C. E. and D. B. Rubin (2002). Principal stratification in causal inference. *Biometrics 58*(1), 21–29.

Freedman, D. A. (1987). As others see us: A case study in path analysis. *Journal of Educational Statistics 12*(2), 101–128.

Fruehwirth, J. C. (2014). Can achievement peer effect estimates inform policy? A view from inside the black box. *The Review of Economics and Statistics 96*(3), 514–523.

Gemici, A. (2011). Family migration and labor market outcomes. *Mimeo., New York University.*

Goldberger, A. S. (1991). *A Course in Econometrics.* Harvard University Press.

Grogger, J. (1995). The effect of arrests on the employment and earnings of young men. *The Quarterly Journal of Economics 110*(1), 51–71.

Hamilton, B. H. (2000). Does entrepreneurship pay? An empirical analysis of the returns to self-employment. *Journal of Political Economy 108*(3), 604–631.

Heckman, J. J. (1996). Randomization as an Instrumental Variable. *The Review of Economics and Statistics 78*(2), pp. 336–341.

Heckman, J. J. (1997). Instrumental Variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources 32*(3), 441–462.

Heckman, J. J. (2008). Econometric causality. *International Statistical Review 76*(1), 1–27.

Heckman, J. J. and S. Navarro (2007). Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics 136*(2), 341–396.

Heckman, J. J. and S. Navarro-Lozano (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *The Review of Economics and Statistics 86*(1), 30–57.

Heckman, J. J. and R. Pinto (2014). Causal analysis after Haavelmo. *Econometric Theory 31*(1), 115–151.

Heckman, J. J. and R. Pinto (2015). Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs. *Econometrics Reviews 34*(1-2), 6–31.

Heckman, J. J. and J. A. Smith (1995). Assessing the case for social experiments. *The Journal of Economic Perspectives 9*(2), 85–110.

Heckman, J. J. and E. Vytlacil (2007). Econometric evaluation of social programs, Part I: Causal models, structural models and econometric policy evaluation. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, Chapter 70, pp. 4779 – 4874. Elsevier.

Hendry, D. F. and J.-F. Richard (1982). On the formulation of empirical models in dynamic econometrics. *Journal of Econometrics 20*(1), 3 – 33.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology 18*(1), 449–484.

Imai, K., L. Keele, and D. Tingley (2010). A general approach to causal mediation analysis. *Psychological Methods 15*(4), 309–334.

Imbens, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of the Economic Literature 48*(2), 399–423.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of Local Average Treatment Effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature 47*(1), 586.

Jacob, B. A. and L. Lefgren (2003). Are idle hands the devil's workshop? Incapacitation, concentration, and juvenile crime. *The American Economic Review 93*(5), pp. 1560–1577.

Jeffreys, H. (2011). *Scientific Inference.* Cambridge University Press.

Keane, M. P. (2010). Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics 156*(1), 3–20.

Kling, J. R. (2006). Incarceration length, employment, and earnings. *The American Economic Review 96*(3), 863–876.

Kydland, F. E. and E. C. Prescott (1977). Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy 85*(3), pp. 473–492.

Leamer, E. E. (2010). Tantalus on the road to Asymptopia. *The Journal of Economic Perspectives 24*(2), 31–46.

Lechner, M. and R. Miquel (2010). Identification of the effects of dynamic treatments by sequential conditional independence assumptions. *Empirical Economics 39*(1), 111–137.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies 76*(3), pp. 1071–1102.

Light, A. (2001). In-school work experience and the returns to schooling. *Journal of Labor Economics 19*(1), 65–93.

Lochner, L. and E. Moretti (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *The American Economic Review 94*(1), 155–189.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. In K. Brunner and A. Meltzer (Eds.), *The Phillips Curve and Labor Markets*, Volume 1, pp. 19–46. Carnegie-Rochester Conference Series on Public Policy.

Manski, C. F. (2007). *Identification for Prediction and Decision.* Harvard University Press.

McCrary, J. and H. Royer (2011). The effect of female education on fertility and infant health: Evidence from school entry laws using exact date of birth. *American Economic Review 101*(1), 158195.

Meer, J. (2007). Evidence on the returns to secondary vocational education. *Economics of Education Review 26*(5), 559–573.

Nagin, D. and J. Waldfogel (1998). The effect of conviction on income through the life cycle. *International Review of Law and Economics 18*(1), 25 – 40.

Nielsen, H. S. and M. Svarer (2009). Educational homogamy: How much is opportunities? *Journal of Human Resources 44*(4), 1066–1086.

Oreopoulos, P. (2006a). The compelling effects of compulsory schooling: Evidence from Canada. *The Canadian Journal of Economics / Revue canadienne d'Economique 39*(1), pp. 22–52.

Oreopoulos, P. (2006b). Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *The American Economic Review*, 152–175.

Pearl, J. (1993). On the statistical interpretation of structural equations. *Mimeo., UCLA Cognitive Systems Laboratory*.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press.

Pearl, J. (2012). The causal mediation formula - a guide to the assessment of pathways and mechanisms. *Prevention Science 13*, 426–436.

Pearl, J. (2014a). The deductive approach to causal inference. *Journal of Causal Inference 2*(2), 115–129.

Pearl, J. (2014b, Forthcoming). Interpretation and identification of causal mediation. *Psychological Methods*.

Pearl, J. and E. Bareinboim (2011). Transportability of causal and statistical relations: A formal approach. *Proceedings of the Twenty-Fifth National Conference on Artificial Intelligence*, 247–254.

Pearl, J. and E. Bareinboim (2014). External validity: From *do*-calculus to transportability across populations. *Statistical Science 29*(4), 579–595.

Pearl, J., G. Imbens, B. Chen, and E. Bareinboim (2014, October-November). Are economists smarter than epidemiologists? (Comments on Imbens' recent paper). *UCLA Causality Blog*. http://www.mii.ucla.edu/causality/?p=1241.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling 7*(9), 1393–1512.

Robins, J., T. Richardson, and P. Spirtes (2009). On identification and inference for direct effects. *Mimeo., Carnegie-Mellon University*.

Rosenbaum, J. E. (1995). Changing the geography of opportunity by expanding residential choice: Lessons from the Gautreaux program. *Housing Policy Debate 6*(1), 231–269.

Rosenbaum, P. R. (1984). The consquences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series A (General) 147*(5), pp. 656–666.

Rosenzweig, M. R. and K. I. Wolpin (2000). Natural "Natural Experiments" in Economics. *Journal of Economic Literature 38*, 827–874.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association 100*(469), 322–331.

Simonsen, M. and L. Skipper (2006). The costs of motherhood: an analysis using matching estimators. *Journal of Applied Econometrics 21*(7), 919–934.

Small, D. S. and P. R. Rosenbaum (2008). War and wages: The strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association 103*(483), 924–933.

Sobel, M. E. and G. Arminger (1992). Modeling household fertility decisions: A nonlinear simultaneous probit model. *Journal of the American Statistical Association 87*(417), 38–47.

VanderWeele, T. J. (2009). Mediation and mechanism. *European Journal of Epidemiology 24*(5), 217–224.

Western, B., J. R. Kling, and D. F. Weiman (2001). The labor market consequences of incarceration. *Crime & delinquency 47*(3), 410–427.

White, H. and K. Chalak (2013). Identification and identification failure for treatment effects using structural systems. *Econometric Reviews 32*(3), 273–317.

Woodward, J. (2000). Explanation and invariance in the special sciences. *The British Journal for the Philosophy of Science 51*, 197–254.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation.* New York: Oxford University Press.

Wooldridge, J. (2005). Violating ignorability of treatment by controlling for too many factors. *Econometric Theory 21*(05), 1026–1028.

Zellner, A. (2007). Philosophy and objectives of econometrics. *Journal of Econometrics 136*, 331–339.

A Derivation of OLS and 2SLS Estimators

A.1 Notation for Matrix Algebra

For the sake of exposition, assume for these derivations that the constant term in the structural outcome equation $\theta_0 = 0$, and that the regressions are specified without constants. Additionally for the sake of exposition, recall that at the given level of measurement $\epsilon_{ti} \sim$ iid f^{ϵ}_{ti}, where f^{ϵ}_{ti} is simply the distribution of a random variable with finite variance, and no observable variables enter as argument of f^{ϵ}_{ti} at the given level of measurement. Since ϵ_{ti} is mean zero, has finite variance, and is independent of all observable variables, we can ignore it when taking expectations and constructing estimators. The ensuing analysis therefore considers DGPs omitting this variable.

Remember that $\boldsymbol{D}_t$ represents the $N \times 1$ vector of observations of D_{ti}. We also have N observations of $\boldsymbol{X}$ at both the time of assignment and the time of measurement, which were labeled as $\boldsymbol{X}_{t_0}$ and $\boldsymbol{X}_t$. Define the following $N \times 2$ vectors

$$\begin{aligned} \boldsymbol{W}_{t_0} &\equiv [\boldsymbol{D}_t, \boldsymbol{X}_{t_0}]\,, \qquad & \boldsymbol{J}_{t_0} &\equiv [\boldsymbol{Z}_t, \boldsymbol{X}_{t_0}] \\ \boldsymbol{W}_t &\equiv [\boldsymbol{D}_t, \boldsymbol{X}_t]\,, \text{ and } & \boldsymbol{J}_t &\equiv [\boldsymbol{Z}_t, \boldsymbol{X}_t]\,. \end{aligned}$$

Defining the $N \times 1$ and 2×1 vectors

$$\boldsymbol{Y}_t \equiv \begin{bmatrix} Y_{t1} \\ \vdots \\ Y_{tN} \end{bmatrix}, \qquad \boldsymbol{E}_t \equiv \begin{bmatrix} E_{t1} \\ \vdots \\ E_{tN} \end{bmatrix}, \qquad \boldsymbol{\theta}_t \equiv \begin{bmatrix} \theta^1_t \\ \theta^2_t \end{bmatrix},$$

it is possible to write the structural potential outcome Equation **??** in Section 3,

$$Y_{ti} \overset{\leftarrow}{=} D_{ti}\theta^1_t + X_{ti}\theta^2_t + E_{ti}, \tag{1}$$

as

$$\boldsymbol{Y}_t \overset{\leftarrow}{=} \boldsymbol{W}_t\boldsymbol{\theta}_t + \boldsymbol{E}_t.$$

Recall the regression Equations 2-4:

$$\boldsymbol{Y}_t = \boldsymbol{D}_t\alpha^1_t + \boldsymbol{H}_t \tag{5}$$

$$\boldsymbol{Y}_t = \boldsymbol{W}_t\boldsymbol{\beta}_t + \boldsymbol{K}_t \tag{6}$$

$$\boldsymbol{Y}_t = \boldsymbol{W}_{t_0}\gamma_t + \boldsymbol{L}_t \tag{7}$$

A.2 Derivation of OLS Estimators

A little matrix algebra shows that:

$$\begin{aligned}
\widehat{\alpha}_t^{1,OLS} &= (\boldsymbol{D}_t'\boldsymbol{D}_t)^{-1}[\boldsymbol{D}_t'\boldsymbol{Y}_t] \\
&= (\boldsymbol{D}_t'\boldsymbol{D}_t)^{-1}[\boldsymbol{D}_t'(\boldsymbol{D}_t\theta_t^1 + \boldsymbol{X}_t\theta_t^2 + \boldsymbol{E}_t)] \\
\\
&= \theta_t^1 + \frac{\boldsymbol{D}_t'\boldsymbol{X}_t}{\boldsymbol{D}_t'\boldsymbol{D}_t}\theta_t^2 + \frac{\boldsymbol{D}_t'\boldsymbol{E}_t}{\boldsymbol{D}_t'\boldsymbol{D}_t}.
\end{aligned}$$

Rewriting a ratio of the dot products of two $N \times 1$ vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ as

$$\frac{\boldsymbol{A}'\boldsymbol{B}}{\boldsymbol{B}'\boldsymbol{B}} = \frac{\frac{1}{N}\sum_{i=1}^{N} A_iB_i}{\frac{1}{N}\sum_{i=1}^{N} B_iB_i} \tag{20}$$

the Weak Law of Large Numbers implies that as N goes to infinity,

$$\widehat{\alpha}_t^{1,OLS} \quad \xrightarrow{p} \quad \theta_t^1 + \frac{\mathbb{E}\left[D_tX_t\right]}{\mathbb{E}\left[D_tD_t\right]}\theta_t^2 + \frac{\mathbb{E}\left[D_tE_t\right]}{\mathbb{E}\left[D_tD_t\right]}$$

as long as the above means are all finite.

Similarly,

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_t^{OLS} &= (\boldsymbol{W}_t'\boldsymbol{W}_t)^{-1}[\boldsymbol{W}_t'\boldsymbol{Y}_t] \\
&= (\boldsymbol{W}_t'\boldsymbol{W}_t)^{-1}[\boldsymbol{W}_t'(\boldsymbol{W}_t\boldsymbol{\theta}_t + \boldsymbol{E}_t)] \\
&= \boldsymbol{\theta}_t + (\boldsymbol{W}_t'\boldsymbol{W}_t)^{-1}[\boldsymbol{W}_t'\boldsymbol{E}_t] \\
&= \boldsymbol{\theta}_t + \begin{bmatrix} \boldsymbol{D}_t'\boldsymbol{D}_t & \boldsymbol{D}_t'\boldsymbol{X}_t \\ \boldsymbol{X}_t'\boldsymbol{D}_t & \boldsymbol{X}_t'\boldsymbol{X}_t \end{bmatrix}^{-1} \left(\begin{bmatrix} D_{t1} & \cdots & D_{tN} \\ X_{t1} & \cdots & X_{tN} \end{bmatrix} \begin{bmatrix} E_{t1} \\ \vdots \\ E_{tN} \end{bmatrix} \right) \\
\\
&= \boldsymbol{\theta}_t + \frac{1}{(\boldsymbol{D}_t'\boldsymbol{D}_t)(\boldsymbol{X}_t'\boldsymbol{X}_t) - (\boldsymbol{D}_t'\boldsymbol{X}_t)(\boldsymbol{X}_t'\boldsymbol{D}_t)} \begin{bmatrix} \boldsymbol{X}_t'\boldsymbol{X}_t & -\boldsymbol{D}_t'\boldsymbol{X}_t \\ -\boldsymbol{X}_t'\boldsymbol{D}_t & \boldsymbol{D}_t'\boldsymbol{D}_t \end{bmatrix} \begin{bmatrix} \boldsymbol{D}_t'\boldsymbol{E}_t \\ \boldsymbol{X}_t'\boldsymbol{E}_t \end{bmatrix} \\
\\
&= \begin{bmatrix} \theta_t^1 \\ \\ \theta_t^2 \end{bmatrix} + \begin{bmatrix} \frac{(\boldsymbol{X}_t'\boldsymbol{X}_t)(\boldsymbol{D}_t'\boldsymbol{E}_t) - (\boldsymbol{D}_t'\boldsymbol{X}_t)(\boldsymbol{X}_t'\boldsymbol{E}_t)}{(\boldsymbol{D}_t'\boldsymbol{D}_t)(\boldsymbol{X}_t'\boldsymbol{X}_t) - (\boldsymbol{D}_t'\boldsymbol{X}_t)(\boldsymbol{X}_t'\boldsymbol{D}_t)} \\ \\ \frac{-(\boldsymbol{X}_t'\boldsymbol{D}_t)(\boldsymbol{D}_t'\boldsymbol{E}_t) + (\boldsymbol{D}_t'\boldsymbol{D}_t)(\boldsymbol{X}_t'\boldsymbol{E}_t)}{(\boldsymbol{D}_t'\boldsymbol{D}_t)(\boldsymbol{X}_t'\boldsymbol{X}_t) - (\boldsymbol{D}_t'\boldsymbol{X}_t)(\boldsymbol{X}_t'\boldsymbol{D}_t)} \end{bmatrix}.
\end{aligned}$$

Recalling Equation 20, as N goes to infinity

$$\widehat{\beta}_t^{1,OLS} \quad \xrightarrow{p} \quad \theta_t^1 + \frac{\mathbb{E}[X_tX_t]\,\mathbb{E}[D_tE_t] \; - \; \mathbb{E}[D_tX_t]\,\mathbb{E}[X_tE_t]}{\mathbb{E}[D_tD_t]\,\mathbb{E}[X_tX_t] \; - \; \mathbb{E}[D_tX_t]\,\mathbb{E}[X_tD_t]}$$

if the above means are finite.

And finally,

$$
\begin{aligned}
\widehat{\gamma}_t^{OLS} &= (\boldsymbol{W}'_{t_0}\boldsymbol{W}_{t_0})^{-1}[\boldsymbol{W}'_{t_0}\boldsymbol{Y}_t] \\
&= (\boldsymbol{W}'_{t_0}\boldsymbol{W}_{t_0})^{-1}[\boldsymbol{W}'_{t_0}(\boldsymbol{W}_t\boldsymbol{\theta}_t + \boldsymbol{E}_t)] \\
&= \frac{1}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0}) - (\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)} \left[\begin{array}{cc} \boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0} & -\boldsymbol{D}'_t\boldsymbol{X}_{t_0} \\ -\boldsymbol{X}'_{t_0}\boldsymbol{D}_t & \boldsymbol{D}'_t\boldsymbol{D}_t \end{array}\right] \left[\begin{array}{c} \boldsymbol{D}'_t\boldsymbol{D}_t\theta_t^1 + \boldsymbol{D}'_t\boldsymbol{X}_t\theta_t^2 + \boldsymbol{D}'_t\boldsymbol{E}_t \\ \boldsymbol{X}'_{t_0}\boldsymbol{D}_t\theta_t^1 + \boldsymbol{X}'_{t_0}\boldsymbol{X}_t\theta_t^2 + \boldsymbol{X}'_{t_0}\boldsymbol{E}_t \end{array}\right]
\end{aligned}
$$

$$
= \left[\begin{array}{c}
\frac{(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})(\boldsymbol{D}'_t\boldsymbol{D}_t)\theta_t^1+(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})(\boldsymbol{D}'_t\boldsymbol{X}_t)\theta_t^2+(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})(\boldsymbol{D}'_t\boldsymbol{E}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)} \\
\qquad - \frac{(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)\theta_t^1+(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{X}_t)\theta_t^2+(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{E}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)} \\
\\
\frac{-(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)(\boldsymbol{D}'_t\boldsymbol{D}_t)\theta_t^1-(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)(\boldsymbol{D}'_t\boldsymbol{X}_t)\theta_t^2-(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)(\boldsymbol{D}'_t\boldsymbol{E}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)} \\
\qquad + \frac{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)\theta_t^1+(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_t)\theta_t^2+(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{E}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)}
\end{array}\right]
$$

$$
= \left[\begin{array}{c}
\theta_t^1 + \frac{(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})(\boldsymbol{D}'_t\boldsymbol{X}_t)-(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{X}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)}\theta_t^2 + \frac{(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})(\boldsymbol{D}'_t\boldsymbol{E}_t)-(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{E}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)} \\
\\
\frac{-(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)(\boldsymbol{D}'_t\boldsymbol{D}_t)+(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)}\theta_t^1 + \frac{-(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)(\boldsymbol{D}'_t\boldsymbol{X}_t)+(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_A\boldsymbol{D}_t)}\theta_t^2 + \frac{-(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)(\boldsymbol{D}'_t\boldsymbol{E}_t)+(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{E}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D}'_t\boldsymbol{X}_{t_0})(\boldsymbol{X}'_{t_0}\boldsymbol{D}_t)}
\end{array}\right],
$$

so

$$
\begin{aligned}
\widehat{\gamma}_t^{1,OLS} \quad \xrightarrow{p} \quad & \theta_t^1 + \frac{\mathbb{E}[X_{t_0}X_{t_0}]\,\mathbb{E}[D_tX_t] \;-\; \mathbb{E}[D_tX_{t_0}]\,\mathbb{E}[X_{t_0}X_t]}{\mathbb{E}[D_tD_t]\,\mathbb{E}[X_{t_0}X_{t_0}] \;-\; \mathbb{E}[D_tX_{t_0}]\,\mathbb{E}[X_{t_0}D_t]}\theta_t^2 \\
& \qquad + \frac{\mathbb{E}[X_{t_0}X_{t_0}]\,\mathbb{E}[D_tE_t] \;-\; \mathbb{E}[D_tX_{t_0}]\,\mathbb{E}[X_{t_0}E_t]}{\mathbb{E}[D_tD_t]\,\mathbb{E}[X_{t_0}X_{t_0}] \;-\; \mathbb{E}[D_tX_{t_0}]\,\mathbb{E}[X_{t_0}D_t]}.
\end{aligned}
$$

A.3 Derivation of 2SLS Estimators

Similarly, we can perform some matrix algebra to see that

$$
\begin{aligned}
\widehat{\alpha}_t^{1,2SLS} &= (\boldsymbol{Z}'_t\boldsymbol{D}_t)^{-1}[\boldsymbol{Z}'_t\boldsymbol{Y}_t] \\
&= (\boldsymbol{Z}'_t\boldsymbol{D}_t)^{-1}[\boldsymbol{Z}'_t(\boldsymbol{D}_t\theta_t^1 + \boldsymbol{X}_t\theta_t^2 + \boldsymbol{E}_t)] \\
\\
&= \theta_t^1 + \frac{\boldsymbol{Z}'_t\boldsymbol{X}_t}{\boldsymbol{Z}'_t\boldsymbol{D}_t}\theta_t^2 + \frac{\boldsymbol{Z}'_t\boldsymbol{E}_t}{\boldsymbol{Z}'_t\boldsymbol{D}_t}, \qquad (21)
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}_t^{2SLS} &= \Big[(\boldsymbol{W}'_t\boldsymbol{J}_t)(\boldsymbol{J}'_t\boldsymbol{J}_t)^{-1}(\boldsymbol{J}'_t\boldsymbol{W}_t)\Big]^{-1}\Big[(\boldsymbol{W}'_t\boldsymbol{J}_t)(\boldsymbol{J}'_t\boldsymbol{J}_t)^{-1}(\boldsymbol{J}'_t\boldsymbol{Y}_t)\Big] \\
&= \Big[(\boldsymbol{W}'_t\boldsymbol{J}_t)(\boldsymbol{J}'_t\boldsymbol{J}_t)^{-1}(\boldsymbol{J}'_t\boldsymbol{W}_t)\Big]^{-1}\Big[(\boldsymbol{W}'_t\boldsymbol{J}_t)(\boldsymbol{J}'_t\boldsymbol{J}_t)^{-1}(\boldsymbol{J}'_t\boldsymbol{W}_t\boldsymbol{\theta}_t)\Big] \\
&\quad + \Big[(\boldsymbol{W}'_t\boldsymbol{J}_t)(\boldsymbol{J}'_t\boldsymbol{J}_t)^{-1}(\boldsymbol{J}'_t\boldsymbol{W}_t)\Big]^{-1}\Big[(\boldsymbol{W}'_t\boldsymbol{J}_t)(\boldsymbol{J}'_t\boldsymbol{J}_t)^{-1}(\boldsymbol{J}'_t\boldsymbol{E}_t)\Big] \\
&= \boldsymbol{\theta}_t + \Big[(\boldsymbol{W}'_t\boldsymbol{J}_t)(\boldsymbol{J}'_t\boldsymbol{J}_t)^{-1}(\boldsymbol{J}'_t\boldsymbol{W}_t)\Big]^{-1}\Big[(\boldsymbol{W}'_t\boldsymbol{J}_t)(\boldsymbol{J}'_t\boldsymbol{J}_t)^{-1}(\boldsymbol{J}'_t\boldsymbol{E}_t)\Big],
\end{aligned}
\tag{22}
$$

and

$$
\begin{aligned}
\widehat{\boldsymbol{\gamma}}_t^{2SLS} &= \Big[(\boldsymbol{W}'_{t_0}\boldsymbol{J}_{t_0})(\boldsymbol{J}'_{t_0}\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J}'_{t_0}\boldsymbol{W}_{t_0})\Big]^{-1}\Big[(\boldsymbol{W}'_{t_0}\boldsymbol{J}_{t_0})(\boldsymbol{J}'_{t_0}\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J}'_{t_0}\boldsymbol{Y}_t)\Big] \\
&= \Big[(\boldsymbol{W}'_{t_0}\boldsymbol{J}_{t_0})(\boldsymbol{J}'_{t_0}\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J}'_{t_0}\boldsymbol{W}_{t_0})\Big]^{-1}\Big[(\boldsymbol{W}'_{t_0}\boldsymbol{J}_{t_0})(\boldsymbol{J}'_{t_0}\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J}'_{t_0}\boldsymbol{W}_t\boldsymbol{\theta}_t)\Big] \\
&\quad + \Big[(\boldsymbol{W}'_{t_0}\boldsymbol{J}_{t_0})(\boldsymbol{J}'_{t_0}\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J}'_{t_0}\boldsymbol{W}_{t_0})\Big]^{-1}\Big[(\boldsymbol{W}'_{t_0}\boldsymbol{J}_{t_0})(\boldsymbol{J}'_{t_0}\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J}'_{t_0}\boldsymbol{E}_M)\Big].
\end{aligned}
\tag{23}
$$

Assuming for the sake of exposition that there is perfect compliance, so that $\boldsymbol{D}_t = \boldsymbol{Z}_t$, we can replace $\boldsymbol{J}_t = \boldsymbol{W}_t$. In this case, each of these 2SLS estimators reduces to their OLS counterpart, as:

$$
\begin{aligned}
\widehat{\alpha}_t^{1,2SLS} &= \theta_t^1 + \frac{\boldsymbol{Z}'_t\boldsymbol{X}_t}{\boldsymbol{Z}'_t\boldsymbol{D}_t}\theta_t^2 + \frac{\boldsymbol{Z}'_t\boldsymbol{E}_t}{\boldsymbol{Z}'_t\boldsymbol{D}_t} \\
&= \theta_t^1 + \frac{\boldsymbol{D}'_t\boldsymbol{X}_t}{\boldsymbol{D}'_t\boldsymbol{D}_t}\theta_t^2 + \frac{\boldsymbol{D}'_t\boldsymbol{E}_t}{\boldsymbol{D}'_t\boldsymbol{D}_t} \\
&= \widehat{\alpha}_t^{1,OLS},
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}_t^{2SLS} &= \boldsymbol{\theta}_t + \Big[(\boldsymbol{W}'_t\boldsymbol{J}_t)(\boldsymbol{J}'_t\boldsymbol{J}_t)^{-1}(\boldsymbol{J}'_t\boldsymbol{W}_t)\Big]^{-1}\Big[(\boldsymbol{W}'_t\boldsymbol{J}_t)(\boldsymbol{J}'_t\boldsymbol{J}_t)^{-1}(\boldsymbol{J}'_t\boldsymbol{E}_t)\Big], \\
&= \boldsymbol{\theta}_t + (\boldsymbol{W}'_t\boldsymbol{W}_t)^{-1}(\boldsymbol{W}'_t\boldsymbol{E}_t) \\
&= \begin{bmatrix} \theta_t^1 \\ \\ \theta_t^2 \end{bmatrix} + \begin{bmatrix} \frac{(\boldsymbol{X}'_t\boldsymbol{X}_t)(\boldsymbol{D}'_t\boldsymbol{E}_t)-(\boldsymbol{D}'_M\boldsymbol{X}_t)(\boldsymbol{X}'_t\boldsymbol{E}_t)}{(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_t\boldsymbol{X}_t)-(\boldsymbol{D}'_t\boldsymbol{X}_t)(\boldsymbol{X}'_t\boldsymbol{D}_t)} \\ \\ \frac{-(\boldsymbol{X}'_M\boldsymbol{D}_t)(\boldsymbol{D}'_t\boldsymbol{E}_t)+(\boldsymbol{D}'_t\boldsymbol{D}_t)(\boldsymbol{X}'_t\boldsymbol{E}_t)}{(\boldsymbol{D}'_M\boldsymbol{D}_t)(\boldsymbol{X}'_t\boldsymbol{X}_t)-(\boldsymbol{D}'_t\boldsymbol{X}_t)(\boldsymbol{X}'_t\boldsymbol{D}_t)} \end{bmatrix} \\
&= \widehat{\boldsymbol{\beta}}_t^{OLS},
\end{aligned}
$$

and

$$
\begin{aligned}
\widehat{\boldsymbol{\gamma}}_t^{2SLS} &= \left[(\boldsymbol{W'}_{t_0}\boldsymbol{J}_{t_0})(\boldsymbol{J'}_{t_0}\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J'}_{t_0}\boldsymbol{W}_{t_0})\right]^{-1}\left[(\boldsymbol{W'}_{t_0}\boldsymbol{J}_{t_0})(\boldsymbol{J'}_{t_0}\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J'}_{t_0}\boldsymbol{W}_M\boldsymbol{\theta}_t)\right] \\
&\quad + \left[(\boldsymbol{W'}_{t_0}\boldsymbol{J}_{t_0})(\boldsymbol{J'}_{t_0}\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J'}_{t_0}\boldsymbol{W}_{t_0})\right]^{-1}\left[(\boldsymbol{W'}_{t_0}\boldsymbol{J}_{t_0})(\boldsymbol{J'}_{t_0}\boldsymbol{J}_{t_0})^{-1}(\boldsymbol{J'}_{t_0}\boldsymbol{E}_t)\right] \\
&= \left[(\boldsymbol{W'}_{t_0}\boldsymbol{W}_{t_0})(\boldsymbol{W'}_{t_0}\boldsymbol{W}_{t_0})^{-1}(\boldsymbol{W'}_{t_0}\boldsymbol{W}_{t_0})\right]^{-1}\left[(\boldsymbol{W'}_{t_0}\boldsymbol{W}_{t_0})(\boldsymbol{W'}_{t_0}\boldsymbol{W}_{t_0})^{-1}(\boldsymbol{W'}_{t_0}\boldsymbol{W}_t\boldsymbol{\theta}_t)\right] \\
&\quad + \left[(\boldsymbol{W'}_{t_0}\boldsymbol{W}_{t_0})(\boldsymbol{W'}_{t_0}\boldsymbol{W}_{t_0})^{-1}(\boldsymbol{W'}_{t_0}\boldsymbol{W}_{t_0})\right]^{-1}\left[(\boldsymbol{W'}_{t_0}\boldsymbol{W}_{t_0})(\boldsymbol{W'}_{t_0}\boldsymbol{W}_{t_0})^{-1}(\boldsymbol{W'}_{t_0}\boldsymbol{E}_t)\right]
\end{aligned}
$$

$$
= \begin{bmatrix}
\theta_t^1 + \frac{(\boldsymbol{X'}_{t_0}\boldsymbol{X}_{t_0})(\boldsymbol{D'}_t\boldsymbol{X}_t)-(\boldsymbol{D'}_t\boldsymbol{X}_{t_0})(\boldsymbol{X'}_{t_0}\boldsymbol{X}_t)}{(\boldsymbol{D'}_t\boldsymbol{D}_t)(\boldsymbol{X'}_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D'}_t\boldsymbol{X}_{t_0})(\boldsymbol{X'}_{t_0}\boldsymbol{D}_t)}\theta_t^2 + \frac{(\boldsymbol{X'}_{t_0}\boldsymbol{X}_{t_0})(\boldsymbol{D'}_t\boldsymbol{E}_t)-(\boldsymbol{D'}_M\boldsymbol{X}_{t_0})(\boldsymbol{X'}_{t_0}\boldsymbol{E}_t)}{(\boldsymbol{D'}_t\boldsymbol{D}_t)(\boldsymbol{X'}_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D'}_t\boldsymbol{X}_{t_0})(\boldsymbol{X'}_{t_0}\boldsymbol{D}_t)} \\
\frac{-(\boldsymbol{X'}_{t_0}\boldsymbol{D}_M)(\boldsymbol{D'}_t\boldsymbol{D}_t)+(\boldsymbol{D'}_t\boldsymbol{D}_t)(\boldsymbol{X'}_{t_0}\boldsymbol{D}_t)}{(\boldsymbol{D'}_t\boldsymbol{D}_t)(\boldsymbol{X'}_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D'}_t\boldsymbol{X}_{t_0})(\boldsymbol{X'}_{t_0}\boldsymbol{D}_t)}\theta_t^1 + \frac{-(\boldsymbol{X'}_{t_0}\boldsymbol{D}_t)(\boldsymbol{D'}_t\boldsymbol{X}_t)+(\boldsymbol{D'}_t\boldsymbol{D}_t)(\boldsymbol{X'}_{t_0}\boldsymbol{X}_t)}{(\boldsymbol{D'}_t\boldsymbol{D}_t)(\boldsymbol{X'}_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D'}_t\boldsymbol{X}_{t_0})(\boldsymbol{X'}_{t_0}\boldsymbol{D}_t)}\theta_t^2 + \frac{-(\boldsymbol{X'}_{t_0}\boldsymbol{D}_t)(\boldsymbol{D'}_t\boldsymbol{E}_t)+(\boldsymbol{D'}_M\boldsymbol{D}_t)(\boldsymbol{X'}_{t_0}\boldsymbol{E}_t)}{(\boldsymbol{D'}_t\boldsymbol{D}_t)(\boldsymbol{X'}_{t_0}\boldsymbol{X}_{t_0})-(\boldsymbol{D'}_t\boldsymbol{X}_{t_0})(\boldsymbol{X'}_{t_0}\boldsymbol{D}_{ti})}
\end{bmatrix}
$$

$$
= \widehat{\boldsymbol{\gamma}}_t^{OLS}.
$$

B Monte Carlo Evidence on Identification

Table 2 presents Monte Carlo results showing the performance of the OLS estimators from Equations 2-4 when estimated on 100,000 simulated data points generated by parameterized DGPs from $\{\mathcal{D}_t^I\} - \{\mathcal{D}_t^{IV}\}$, and Figure 5 displays causal effects from these DGPs. The precise parameterizations of the DGPs are as follows: The structural outcome equation is the same across all simulated DGPs:

$$\begin{aligned} Y_{ti} &\overset{\leftarrow}{=} \theta_t^0 + D_{ti}\theta_t^1 + X_{ti}\theta_t^2 + E_{ti} \\ &\overset{\leftarrow}{=} 2.0 + D_{ti} \cdot 1.0 + X_{ti} \cdot 1.0 + E_{ti}. \end{aligned}$$

As well, in all simulated DGPs treatment is randomized with

$$\begin{aligned} D_{ti} \overset{\leftarrow}{=} 0.5Z_{ti}^T + 0.5U_{ti}^D \quad &\text{where} \\ Z_{ti}^T \sim iidU[-1,1] \;\; &\text{and} \\ U_{ti}^D \sim iidU[-1,1], & \end{aligned} \tag{2*}$$

stated equivalently as D_{ti} being an *iid* random variable that follows the triangle distribution with lower limit -1, upper limit 1, and mode 0.

In $\mathcal{D}_t^I$ the remaining selection equations are such that:

$$X_{ti} \sim U[-\tfrac{1}{2}, \tfrac{1}{2}], \qquad \text{and} \qquad E_{ti} \sim U[-\tfrac{1}{2}, \tfrac{1}{2}]. \tag{DGP I}$$

Let $U_{ti}^X \sim iidU[0,1]$. In $\mathcal{D}_t^{II}$ all features of the model are the same as in $\mathcal{D}_t^I$ except that observed covariates are selected in response to treatment:

$$X_{ti} \overset{\leftarrow}{=} \begin{cases} 1 - D_{ti} & \text{if } U_{ti}^X \leq 0.5 \\ A_{ti} & \text{otherwise, where } A_{ti} \sim U[-\tfrac{1}{2}, \tfrac{1}{2}] \end{cases} \tag{DGP II}$$

Similarly, $\mathcal{D}_t^{III}$ is the same as $\mathcal{D}_t^I$ except that now unobserved covariates are selected in response to treatment. Letting $U_{ti}^E \sim iidU[0,1]$, the unobserved factors are determined in response to treatment as:

$$E_{ti} \overset{\leftarrow}{=} \begin{cases} 1 - D_{ti} & \text{if } U_{ti}^E \leq 0.75 \\ B_{ti} & \text{otherwise, where } B_{ti} \sim iidU[-\tfrac{1}{2}, \tfrac{1}{2}] \end{cases} \tag{DGP III}$$

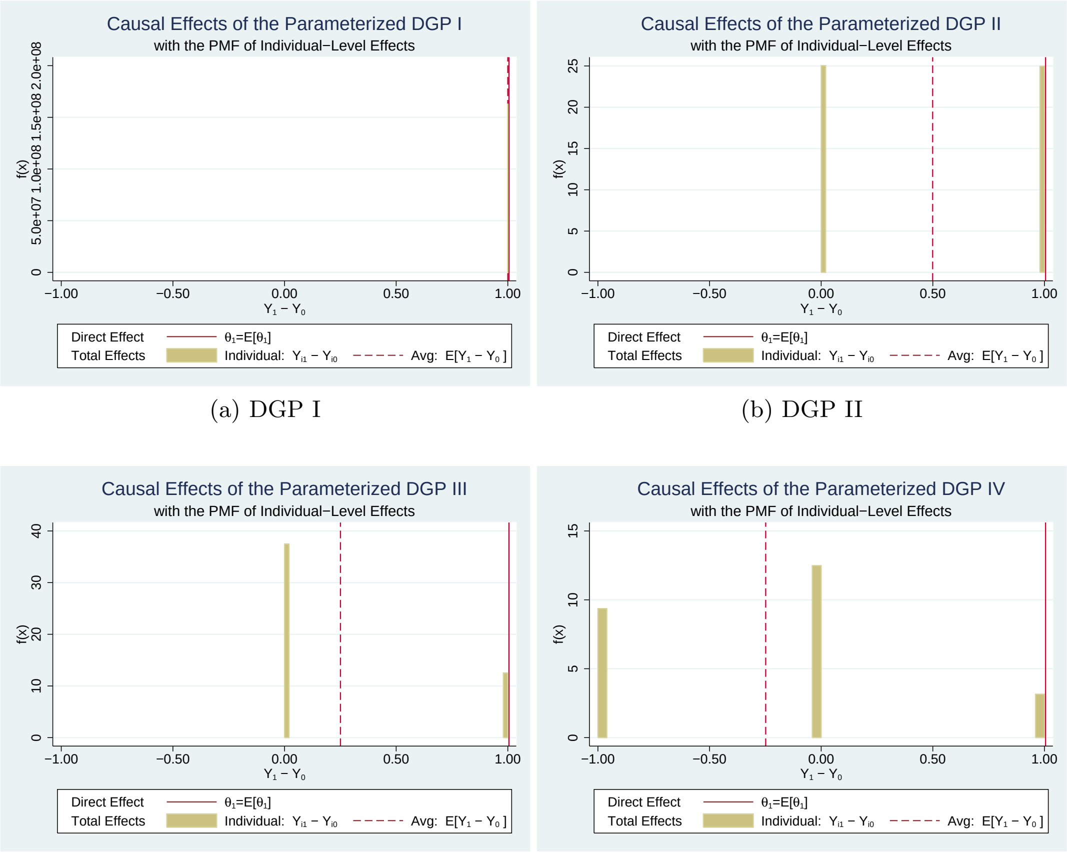
Finally, $\mathcal{D}_t^{IV}$ is the same as $\mathcal{D}_t^I$ except that observed covariates are selected in response to treatment as in DGP $\mathcal{D}_t^{II}$ and unobserved covariates are selected in response to treatment as in $\mathcal{D}_t^{III}$.

Table 2: Estimation Results on Data Simulated from Data Generating Processes with Various Selection Rules

A DGP $\mathcal{D}_t$ is Fully Specified by:
-The Potential Outcome Equation
-Selection Equations f_t^D, f_t^X, f_t^P, f_t^M, and f_t^E

		DGP $\mathcal{D}_t^I$	DGP $\mathcal{D}_t^{II}$	DGP $\mathcal{D}_t^{III}$	DGP $\mathcal{D}_t^{IV}$
	Potential Outcomes:	$Y_{ti} \overset{\leftarrow}{=} \theta_t^0 + \theta_t^1 D_{ti} + \theta_t^2 X_{ti} + E_{ti}$			
	Selection Rule f_t^X:	$X_{ti} \sim U[-\frac{1}{2}, \frac{1}{2}]$	$X_{ti} \overset{\leftarrow}{=} f_t^X(D_{ti})$	$X_{ti} \sim U[-\frac{1}{2}, \frac{1}{2}]$	$X_{ti} \overset{\leftarrow}{=} f_t^X(D_{ti})$
	Selection Rule f_t^E:	$E_{ti} \sim U[-\frac{1}{2}, \frac{1}{2}]$	$E_{ti} \sim U[-\frac{1}{2}, \frac{1}{2}]$	$E_{ti} \overset{\leftarrow}{=} f_t^E(D_{ti})$	$E_{ti} \overset{\leftarrow}{=} f_t^E(D_{ti})$
	Causal Effects				
	DE: θ_t^1	1.00	1.00	1.00	1.00
Randomized D_t, f_t^D:	TE: $\mathbb{E}[Y_{ti(D_{ti}=1)} - Y_{ti(D_{ti}=0)}]$	1.00	0.50	0.25	–0.25
$D_{ti} \overset{\leftarrow}{=} 0.5Z_{ti}^T + 0.5U_{ti}^D$					
	Estimate				
$Z_{ti}^T, U_{ti}^D \sim iidU[-1, 1]$	$\widehat{\alpha}_t^{1,OLS}$ $(\xrightarrow{p} \widehat{\alpha}_t^{1,2SLS})$	1.00	0.49	0.25	–0.23
	$\widehat{\beta}_t^{1,OLS}$ $(\xrightarrow{p} \widehat{\beta}_t^{1,2SLS})$	1.00	1.00	0.25	0.26
	$\widehat{\gamma}_t^{1,OLS}$ $(\xrightarrow{p} \widehat{\gamma}_t^{1,2SLS})$	1.00	0.49	0.25	–0.23
	Selection into Covariates				
	$\mathbb{E}[X_{ti} \mid D_{ti} > 0]$	0.00	0.33	0.00	0.34
	$\mathbb{E}[X_{ti} \mid D_{ti} < 0]$	0.00	0.67	0.00	0.66
	$\mathbb{E}[E_{ti} \mid D_{ti} > 0]$	0.00	0.00	0.50	0.50
	$\mathbb{E}[E_{ti} \mid D_{ti} < 0]$	0.00	0.00	1.00	1.00

Note: The specified DGPs were used to generate 100,000 observations. The precise functions f_t^X and f_t^E used in each simulated DGP are specified in Section 4 in the text.



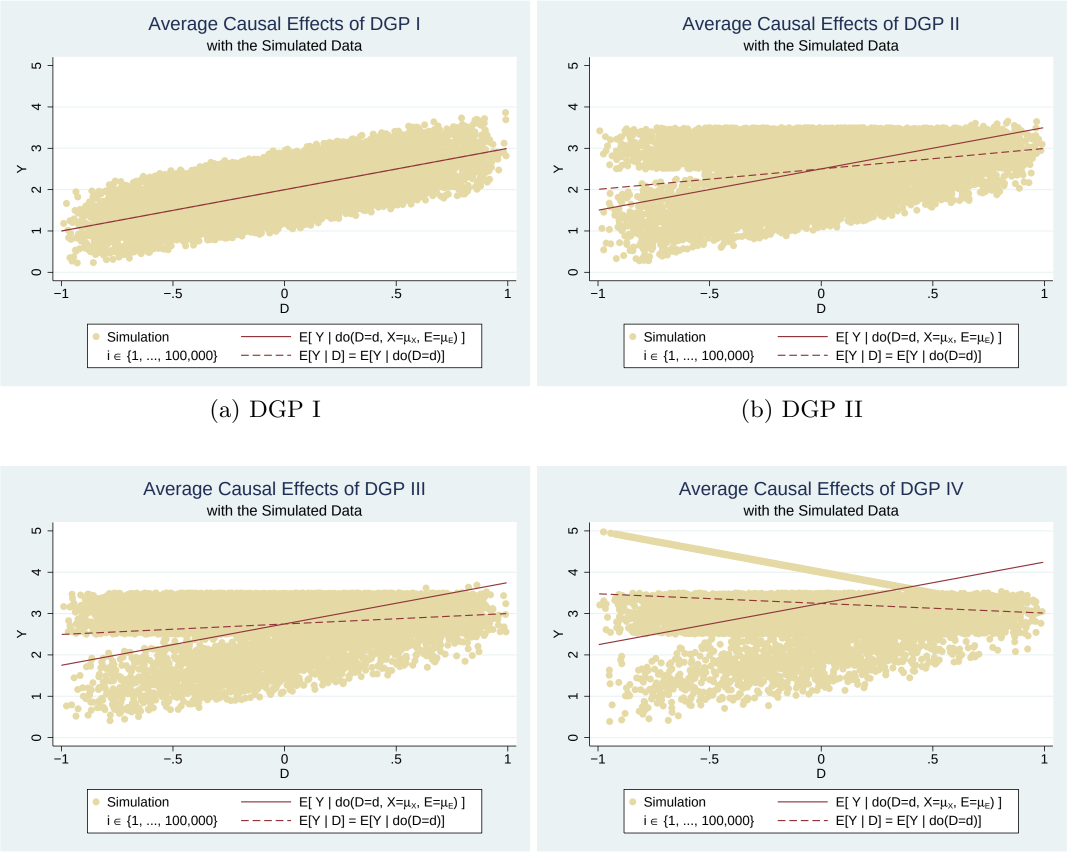
(a) DGP I

(b) DGP II

(c) DGP III

(d) DGP IV

Figure 4: Causal Effects of Data Generating Processes $\mathcal{D}_t^I$-$\mathcal{D}_t^{IV}$



(a) DGP I

(b) DGP II

(c) DGP III

(d) DGP IV

Figure 5: Coefficients of Data Generating Processes $\mathcal{D}_t^I$-$\mathcal{D}_t^{IV}$

B.1 Identification: One Parameter's Bias is Another Parameter's Identification

DGPs $\mathcal{D}_t^{II}$-$\mathcal{D}_t^{IV}$ illustrate that one parameter's bias is another parameter's identification, due to the fact that the exclusion restriction identifying direct effects is distinct from the exclusion restriction identifying total effects.

In DGP $\mathcal{D}_t^{II}$ what represents bias for the researcher trying to identify the total effect represents identification of the direct effect. Should the researcher control for the observed covariates determined in response to treatment and estimate Equation 3, $\widehat{\beta}_t^{1,OLS}$ will be a biased estimator of the average total effect $\mathbb{E}[\triangle_t^{TE}]$ (Wooldridge (2005), Heckman and Navarro-Lozano (2004)). Chalak and White (2011) refer to this as "included variable bias." At the same time, though, these need not be "bad controls:" $\widehat{\beta}_t^{1,OLS}$ will identify the direct causal effect $\triangle_t^{DE} = \theta_t^1$.

In DGPs $\mathcal{D}_t^{III}$ and $\mathcal{D}_t^{IV}$ what represents bias for the researcher trying to identify the direct effect represents identification of the total effect. In light of DGPs $\mathcal{D}_t^{III}$ and $\mathcal{D}_t^{IV}$, previous criticisms of the experimentalist approach can be seen as discussions of identification using DGPs with random variation in treatment that impacts outcomes through covariates. One of Heckman (1997)'s concerns about the total effects identified in Angrist (1990) is that they cannot distinguish between the direct effect of treatment and the direct effect of unobserved covariates selected in response to the quasi-randomly assigned treatment. The concerns raised in Rosenzweig and Wolpin (2000) and Keane (2010) about total effects identified by the quasi-random assignment of treatment generated by natural experiments are likewise related to selection into covariates, creating a difference between the total effect and direct effect identified by IV estimators. Finally, the distinction between exogeneity and orthogonality made in Deaton (2010) can be seen as a distinction between orthogonality conditions made at two points in time, the time of assignment (t_0) and the time of measurement (t). Deaton's concern is that even if orthogonality conditions hold for a given DGP at time t_0, the later ones at t necessary for identification can be violated due to selection into covariates.[17] Further discussion of the cases when conditioning and setting/fixing variables coincide can be found in Heckman and Pinto (2014).

[17] A similar point about DGPs in which the direct effect is not identified is made in White and Chalak (2013). Further discussions on the limitations of effects identified by randomized treatments are provided in Leamer (2010) and Heckman and Smith (1995).

C Extension to DGPs with an Unobserved Confounder

Table 3 presents Monte Carlo results showing the performance of the OLS estimators from Equations 2-4 as well as their 2SLS analogues when DGPs $\mathcal{D}_t^I$-$\mathcal{D}_t^{IV}$ also exhibit selection into treatment. These DGPs are characterized by the following structural equations:

$$E_{ti}^P \Lleftarrow f^P(U_{ti}^P;\ \Theta^P) \tag{24}$$

$$D_{ti} \Lleftarrow f^D(Z_{ti}^T, E_{ti}^P, U_{ti}^D\ ;\ \Theta^D) \tag{25}$$

$$X_{ti} \Lleftarrow f^X(D_{ti}, E_{ti}^P, U_{ti}^X\ ;\ \Theta^X) \tag{26}$$

$$E_{ti}^M \Lleftarrow f^M(D_{ti}, X_{ti}, U_{ti}^M;\ \Theta^M) \tag{27}$$

$$E_{ti} \Lleftarrow f^E(E_{ti}^P, E_{ti}^M\ ;\ \Theta^E) \tag{28}$$

$$Y_{ti} \Lleftarrow \theta_t^0 + D_{ti}\theta_t^1 + X_{ti}\theta_t^2 + E_{ti}. \tag{29}$$

In terms of specification, the potential outcome equation is still the same across all DGPs:

$$Y_{ti} \Lleftarrow 2.0 + D_{ti} \cdot 1.0 + X_{ti} \cdot 1.0 + E_{ti}.$$

The difference is that now, in all simulated DGPs treatment is selected according to

$$D_{ti} \Lleftarrow 0.5Z_{ti}^T + 0.25U_{ti}^D + 0.25E_{ti}^P$$

where both $Z_{ti}^T, U_{ti}^D \sim iidU[-1,1]$. E_{ti}^P represents a permanent component of the unobserved covariate and E_{ti}^M represents a malleable component of the unobserved covariate as follows:

$$E_{ti} \Lleftarrow 0.25E_{ti}^P + 0.75E_{ti}^M.$$

In DGP $\mathcal{D}_t^I$ both $E_{ti}^P, E_{ti}^M \sim iidU[-\frac{1}{2}, \frac{1}{2}]$ and the remaining selection equation is specified to be:

$$X_{ti} \sim U[-\tfrac{1}{2}, \tfrac{1}{2}].$$

In DGP $\mathcal{D}_t^{II}$ all features of the model are the same as in DGP $\mathcal{D}_t^{II}$ except that observed covariates are selected in response to treatment:

$$X_{ti} \Lleftarrow \begin{cases} 1 - D_{ti} & \text{if } U_{ti}^X \leq 0.5 \\ A_{ti} & \text{otherwise, where } A_{ti} \sim U[-\frac{1}{2}, \frac{1}{2}] \end{cases}$$

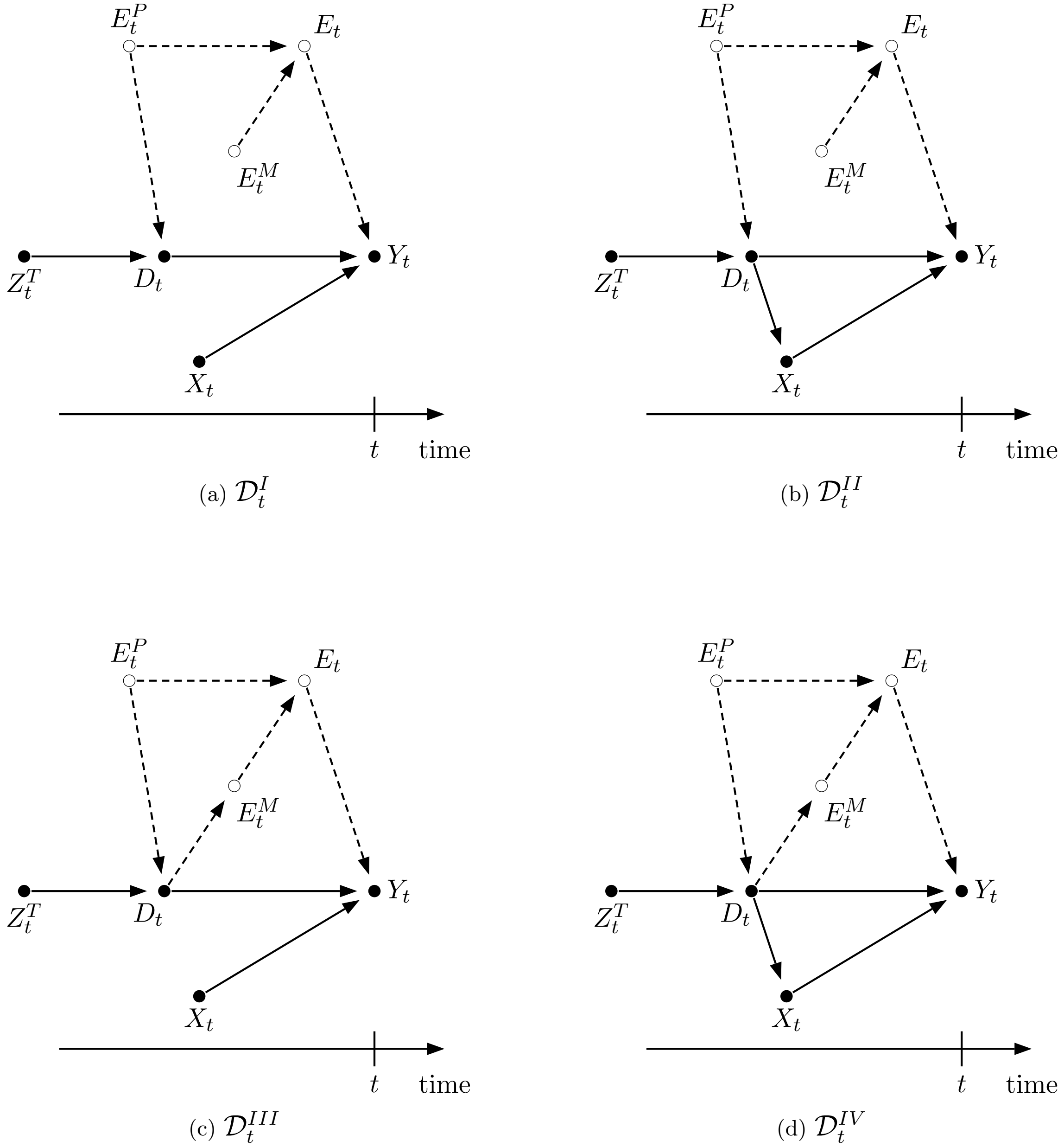
where $U_{ti}^X \sim iidU[0,1]$.

Similarly, DGP $\mathcal{D}_t^{III}$ is the same as DGP $\mathcal{D}_t^I$ except that now unobserved covariates are selected in response to treatment as

$$E_{ti}^{M} \overset{\leftarrow}{=} \begin{cases} 1 - D_{ti} & \text{if } U_{ti}^{E} \leq 0.75; \\ B_{ti} \sim U[-\frac{1}{2}, \frac{1}{2}] & \text{if } U_{ti}^{E} > 0.75, \end{cases}$$

with $U_{ti}^{E} \sim U[0,1]$.

Finally, DGP $\mathcal{D}_t^{IV}$ is the same as DGP $\mathcal{D}_t^{I}$ except that observed covariates are selected in response to treatment as in DGP $\mathcal{D}_t^{II}$ and unobserved covariates are selected in response to treatment as in DGP $\mathcal{D}_t^{III}$.



(a) $\mathcal{D}_t^{I}$

(b) $\mathcal{D}_t^{II}$

(c) $\mathcal{D}_t^{III}$

(d) $\mathcal{D}_t^{IV}$

Figure 6: Directed Acyclic Graphs of Four Data Generating Processes from $\{\mathcal{D}_t\}$ Including the Total Effect Intervention Z_t^T

Table 3: Estimation Results on Data Simulated from Data Generating Processes with Selection into Treatment

A DGP $\mathcal{D}_t$ is Fully Specified by:
-The Potential Outcome Equation
-Selection Equations f^D, f^X, and f^E

	DGP $\mathcal{D}_t^I$	DGP $\mathcal{D}_t^{II}$	DGP $\mathcal{D}_t^{III}$	DGP $\mathcal{D}_t^{IV}$
Potential Outcomes:	$Y_{ti} \Lleftarrow \theta_t^0 + \theta_t^1 D_{ti} + \theta_t^2 X_{ti} + E_{ti}$			
Selection Rule f_t^X:	$X_{ti} \sim U[-\frac{1}{2}, \frac{1}{2}]$	$X_{ti} \Lleftarrow f_t^X(D_{ti})$	$X_{ti} \sim U[-\frac{1}{2}, \frac{1}{2}]$	$X_{ti} \Lleftarrow f_t^X(D_{ti})$
Selection Rule f_t^E:	$E_{ti}^M \sim U[-\frac{1}{2}, \frac{1}{2}]$	$E_{ti}^M \sim U[-\frac{1}{2}, \frac{1}{2}]$	$E_{ti}^M \Lleftarrow f_t^E(D_{ti})$	$E_{ti}^M \Lleftarrow f_t^E(D_{ti})$
Causal Effects				
D1: θ_t^1	1.00	1.00	1.00	1.00
D2: $\mathbb{E}[Y_{ti(D_{ti}=1)} - Y_{ti(D_{ti}=0)}]$	1.00	0.50	0.44	–0.06
Estimate				
$\widehat{\alpha}_t^{1,OLS}$	1.04	0.54	0.49	–0.02
$\widehat{\beta}_t^{1,OLS}$	1.05	1.05	0.49	0.49
$\widehat{\gamma}_t^{1,OLS}$	1.04	0.54	0.49	–0.02
$\widehat{\alpha}_t^{1,2SLS}$	0.99	0.49	0.44	–0.06
$\widehat{\beta}_t^{1,2SLS}$	1.00	1.00	0.45	0.44
$\widehat{\gamma}_t^{1,2SLS}$	0.99	0.49	0.44	–0.06
Exclusion Restrictions				
D1: $\mathbb{E}[Z_{ti}^T E_{ti}]$	0.00	0.00	–0.09	–0.09
D2: $CORR(Y_{ti}, Z_{ti}^T)\vert D_{ti}, E_{ti}^P$	0.00	0.00	0.00	0.00
Selection into Covariates				
$\mathbb{E}[X_{ti}\vert D_{ti} > 0]$	0.00	0.36	0.00	0.36
$\mathbb{E}[X_{ti}\vert D_{ti} < 0]$	0.00	0.64	0.00	0.64
$\mathbb{E}[E_{ti}\vert D_{ti} > 0]$	0.01	0.01	0.42	0.42
$\mathbb{E}[E_{ti}\vert D_{ti} < 0]$	–0.01	–0.01	0.71	0.71

Selection into E:

E_{ti}^P is Permanent, E_{ti}^M is Malleable

$E_{ti} \Lleftarrow 0.25E_{ti}^P + 0.75E_{ti}^M$

$E_{ti}^P \sim iidU[-\frac{1}{2}, \frac{1}{2}]$

$E_{ti}^M \Lleftarrow 1 - D_{ti}$ if $U_{ti}^E \leq 0.75$

$E_{ti}^M \sim iidU[-\frac{1}{2}, \frac{1}{2}]$ if $U_{ti}^E > 0.75$

$U_{ti}^E \sim U[0, 1]$

Selection into D:

$D_{ti} \Lleftarrow 0.5Z_{ti}^T + 0.25U_{ti}^D + 0.25E_{ti}^P$

$Z_{ti}^T, U_{ti}^D \sim iidU[-1, 1]$

Note: The specified DGPs were used to generate 100,000 observations for the previous and current time periods. The precise functions f_t^X and f_t^E used in each simulated DGP are specified in the text.