

A Unified Framework for Defining and Identifying Causal Effects

Halbert White and Karim Chalak

Department of Economics

University of California, San Diego

La Jolla, CA 92093-0508

USA

September 1, 2006

Abstract This paper unifies three complementary approaches to defining, identifying, and estimating causal effects: the classical structural equations approach of the Cowles Commission; methods of the labor econometrics and related treatment effects literatures; and the structural Directed Acyclic Graph (DAG) approach of the machine learning literature. The settable system framework nests these prior approaches, while affording significant improvements to each. For example, the settable system approach permits identification of causal effects without requiring exogenous instruments; instead, a weaker conditional exogeneity condition suffices. It removes the stable unit treatment value assumption of the treatment effect approach and provides significant insight into the selection of covariates. It generalizes the DAG approach by accommodating mutual causality and attributes. We provide a variety of results ensuring structural identification of general covariate-conditioned average causal effects, laying the foundation for parametric and nonparametric estimation of effects of interest and new tests for structural identification.

Acknowledgments: The authors are indebted to Patrick Bajari, Clive W.J. Granger, Jinyong Hahn, Lars Hansen, James Heckman, Kei Hirano, Kevin Hoover, Meng Huang, Chuck Manski, Massimiliano Marinucci, Robert Marshall, Rosa Matzkin, Judea Pearl, Chris Taber, and Rosalind Wu for helpful discussions and suggestions. Any errors remain the authors' responsibility.

JEL Classification Numbers: C10, C20, C30, C50.

1 Introduction

The introduction by Tinbergen, Frisch, Koopmans, Haavelmo, Marschak, and the other pioneers of the Cowles Commission in the late 1940s and early 1950s of the “simultaneous equations” or “structural equations” approach to econometric modeling revolutionized economics by providing sophisticated methods with which to attempt to understand and measure causal effects in economics. Despite this momentous advance, economists’ focus on causal relationships declined in the following decades, due at least in part to the difficult conceptual and statistical issues raised in the ensuing scholarly debate. Hoover (2004) provides enlightening documentation and discussion of this decline; he further documents the strong resurgence of interest in causal issues in economics over the last decade.

This resurgence has been led by the work of a group prominently populated by labor economists, including that of Heckman and Robb (1985), Heckman, Ichimura, and Todd (1997, 1998), Angrist (1998), Hahn (1998), Hirano and Imbens (2001), Hirano, Imbens, and Ridder (2003), Heckman, Urzua, and Vytlačil (2005) (HUV), and Heckman and Vytlačil (2005) (HV). These authors have developed powerful methods for estimating causal effects, blending the classical structural equations approach with methods developed in the treatment effects literature by Rubin (1974) and Rosenbaum and Rubin (1983).

Another important strand of thinking about causal structures has emerged from the machine learning literature. There, Pearl (1988, 1993a, 1993b, 1995, 1998, 2000) and his colleagues (Verma and Pearl, 1992; Spirtes, Glymour, and Scheines, 1993 (SGS); and Dawid, 2002, among others) have developed insightful techniques for analyzing causal structures using directed acyclic graphs (DAGs).

These different approaches to understanding causality are not fully compatible, as each admits certain features ruled out by the others. So far, there does not exist a formal mathematical framework that encompasses these various approaches to defining, identifying, modeling, and estimating causal effects. Our goal here is to provide such a framework, permitting rigorous definitions of notions of cause and effect; and then, based on this framework, to provide conditions for the identification of causal effects of interest. In a companion paper (White and Chalak 2006), we study parametric and nonparametric estimators of these effects.

The “settable system” framework proposed here requires a non-trivial reconsideration of the foundations of each of the approaches it seeks to unify. This would not be worth the required effort unless it yields significant benefits relative to each prior approach. In fact, settable systems offer important improvements and advantages in each case.

Specifically, relative to the classical structural equations approach, settable systems

permit the identification of well-defined causal effects in the absence of traditional exogenous variables. Instead, weaker conditional exogeneity conditions suffice, yielding an extension of the concept of instrumental variables. Our framework shares the flexibility of modern treatments of structural equations, such as Matzkin (2003, 2004, 2005), Imbens and Newey (2003), HUV, and HV, by not imposing the classical restrictions of linearity or separability on the structural relations. The effects of interest identified here do not, however, require the monotonicity requirements imposed by Matzkin (2004, 2005) and Imbens and Newey (2003). Our framework also accommodates unobserved (“essential”) heterogeneity emphasized by HUV and HV. It thus provides a framework complementing the work of Matzkin (2003, 2004, 2005), Imbens and Newey (2003), HUV, and HV that permits the causal notions either implicit or explicit in their approaches to be fully formalized and that permits generalizations of their methods. Another benefit is that the identification results that emerge provide a significant refinement of the standard *ceteris paribus* interpretation of estimated regression coefficients.

Relative to the treatment effects literature, which assumes that the treated units respond passively to treatment, the settable system framework permits the units to be active optimizing agents and explicitly permits these agents to interact with one another. A significant consequence is that agents may be affected not only by treatments applied to them specifically, but also by treatments applied to others. We thus explicitly remove the Stable Unit Treatment Value Assumption (SUTVA) standard in the treatment effects literature. The settable system framework introduces precise definitions of interventions and counterfactuals and clarifies their relation. It further provides significant insight into the selection of the covariates required to ensure identification of causal effects and into the central role of economic theory and domain knowledge in specifying covariates.

Much of the treatment effects literature analyzes effects of single binary treatments. The settable systems analyzed here admit multiple treatments, each of which may be binary, categorical, or continuous. We thus provide causal foundations and identification results complementary to the treatment effects literature, including those that analyze more general treatments (e.g., Robins, Hernan, and Brumback, 2000; Lechner, 2001; Imbens and Newey, 2003; Hirano and Imbens, 2004; Robins, Hernan, and Siebert, 2004).

Relative to the machine learning literature focused on DAGs, the settable system framework provides a formal framework in which notions of mutual causality, central to economic concerns with optimization and equilibrium, are well defined and non-paradoxical. The settable system framework provides an underlying foundation from which standard directed acyclic graph-based causal structures can emerge as special cases, without having to impose useful properties (e.g., acyclicity) as axiomatic. Settable systems also provide

a complement to work in the machine learning literature studying the causal interpretation of graphs in which the requirements of directedness and acyclicity are removed (e.g. Lauritzen and Richardson, 2002). The settable system framework further accommodates attributes that may act as response and effect modifiers, and it leads naturally to straightforward and general statistical methods for estimating causal effects.

Section 2 gives formal definitions of the settable systems supporting our further analysis. We blend ideas explicit and implicit in the classical structural equations framework with concepts from the treatment effect literature and the structural DAG approach of the machine learning literature. The result is a framework possessing two complementary components, one providing stochastic structure and the other explicit causal structure.

As our causal structure accommodates definitions of causal effects for binary, categorical, and continuous causes, we consider two kinds of effects: marginal *ceteris paribus* effects, involving an infinitesimal change to a continuous cause of interest holding all other causes of interest fixed; and effects of more general interventions, involving non-infinitesimal changes to multiple causes of interest.

Because the variables involved in the responses underlying these effects typically cannot all be observed, we study expectations or averages of these effects, conditioned on observable covariates. In Section 3 we use settable systems to define covariate-conditioned average effects, formalizing concepts of Pearl (1995, 2000) and extending notions of conditional effects introduced by Abadie and Imbens (2002). We introduce the notion of conditional exogeneity and show how this ensures structural identification, that is, the equality of the covariate-conditioned average counterfactual response with the standard conditional expectation of the response given causes of interest and covariates.

Section 4 studies structural identification of general effects on the conditional counterfactual response distribution, based on covariate-conditioned explicit moments, optimizers, and implicit moments. We examine how conditional effects can be used to construct unconditional effects and contrast structural identification with “stochastic identification,” which corresponds to the absence of exact multicollinearity in linear regression.

Section 5 treats structural identification without conditional exogeneity, providing local identification and near identification results. We introduce the “discrepancy” and “marginal discrepancy” scores, which quantify departures from conditional exogeneity.

Section 6 discusses selection of covariates using “predictive proxies,” based on results of Section 3. Section 7 discusses the relation between conditional exogeneity and the treatment effect literature notion of “unconfoundedness.” Section 8 provides a summary and concluding remarks. The Mathematical Appendix contains all proofs.

2 Settable Systems and Causality

2.1 Attribute-Indexed Settable Systems

We begin by defining settable systems, a mathematical framework that supports formal notions of cause and effect.

Definition 2.1 Let (Ω, \mathcal{F}, P) be a complete probability space, and let \mathcal{A} be a non-empty multi-dimensional Borel set. For *agents* $h = 1, 2, \dots$, let *attributes* a_h belong to \mathcal{A} , and put $a \equiv \{a_h\}$. For $h = 0, 1, \dots$, and $j = 1, 2, \dots$, let *settings* $Z_{h,j} : \Omega \rightarrow \mathbb{R}$ be measurable functions. For $h = 1, 2, \dots$ and $j = 1, 2, \dots$, let *attribute-indexed response functions* $r_{h,j} : \mathbb{R}^\infty \times \mathcal{A}^\infty \rightarrow \mathbb{R}$ be measurable functions.

For $h = 1, 2, \dots$ and $j = 1, 2, \dots$, let $Z_{(h,j)}$ be the vector including every setting except $Z_{h,j}$, and define the *attribute-indexed settable variables* $\mathcal{X}_{h,j} : \{0, 1\} \times \Omega \rightarrow \mathbb{R}$ such that

$$\begin{aligned}\mathcal{X}_{h,j}(1, \cdot) &= Z_{h,j} \\ \mathcal{X}_{h,j}(0, \cdot) &= r_{h,j}(Z_{(h,j)}, a).\end{aligned}$$

For $h = 0$ and $j = 1, 2, \dots$, let $\mathcal{X}_{h,j}(0, \cdot) = \mathcal{X}_{h,j}(1, \cdot) = Z_{h,j}$.

Put $Z \equiv \{Z_{h,j} : j = 1, 2, \dots; h = 0, 1, \dots\}$, $r \equiv \{r_{h,j}, j = 1, 2, \dots; h = 1, 2, \dots\}$, and $\mathcal{X} \equiv \{\mathcal{X}_{h,j}, j = 1, 2, \dots; h = 0, 1, \dots\}$.

The pair $\mathcal{S} \equiv \{(\Omega, \mathcal{F}, P), (\mathcal{A}, a, Z, r, \mathcal{X})\}$ is an *attribute-indexed settable system*. ■

In this definition each of a collection of *agents* ($h = 1, 2, \dots$) responds to an environment consisting of the other agents in the collection. The notion of “agent” should be interpreted broadly. This may be an individual economic agent, such as a consumer or firm; a collection of individual agents, such as an industry or market; or a physical, natural, or technological process, such as a production function. This framework thus permits analysis of systems in the social sciences as well as in the clinical or natural sciences.

Our focus is on economic systems. In non-economic contexts, our agents might be viewed as passive responders called “units,” “subjects,” “cases,” or “patients.” We do not rule out passive responders, but it is important for our language to reflect the active and interactive nature of the components of economic systems. Further, our interactive structure maps directly to the primitive structures of economic theory.

Each agent h has an attribute vector a_h belonging to the multi-dimensional *attribute space* \mathcal{A} , with dimensions for each possible characteristic of the agent, including what kind of agent it is (individual, firm, market, technological process). Attributes a_h are

fixed characteristics, such as the race or gender of an individual. If a given attribute is not strictly immutable in some larger context (e.g., gender), we view it as determined in some more comprehensive system that nests the present system. Thus, within a given settable system, attributes can vary across agents but are fixed for a given agent.

The j th *response* of agent h , denoted $Y_{h,j} = \mathcal{X}_{h,j}(0, \cdot)$, depends on settings $Z_{(h,j)} = \mathcal{X}_{(h,j)}(1, \cdot)$ and attributes a . The key idea is that $Y_{h,j}$ is generated by setting all other variables of the system to $Z_{(h,j)}$, leaving only the j th variable for agent h free to respond however the agent may determine, represented as $r_{h,j}(Z_{(h,j)}, a)$. The agent’s response may depend not only on its own attributes, but also others’ characteristics.

For example, $Y_{1,1}$ may represent an individual’s wage as it responds, among other things, to settings of the individual’s education and ability, given that individual’s race and gender attributes. Or $Y_{1,1}$ may represent the market price of a call option as it responds to settings of the risk-free interest rate and underlying asset price, given that option’s attributes of underlying asset, strike price, and expiration date.

Although settings and attributes are conceptually distinct, we often need to refer to them jointly. For convenience, we refer jointly to settings and attributes as *explanatory variables*. As Hoover (2001, pp. 147-148) notes, Wold argued for the “explanatory variables” nomenclature so as to avoid referring directly to “causal” variables, given the lack of a clear meaning of causality. (See also Hoover, 2001, section 7.1.) Our use is distinct, in that within our framework causal meanings will be crisply defined. Nevertheless, the spirit is similar, in that explanatory variables are not necessarily causal: as discussed below, attributes are not causal.

The response relation $Y_{h,j} = r_{h,j}(Z_{(h,j)}, a)$ corresponds directly to a classical structural equation. A key difference between these is that classical structural equations are often “simultaneous,” whereas our response relations by construction do not admit simultaneity. Simultaneity arises when the same variables occur on both the left- and right-hand sides in a system of structural equations. In settable systems, simultaneity is ruled out, because when a settable variable appears on the left-hand side, it is as a response, $\mathcal{X}_{h,j}(0, \cdot)$, whereas when it appears on the right, it is as a setting, $\mathcal{X}_{h,j}(1, \cdot)$. This formalizes a device introduced in Strotz and Wold (1960) and later used by Fisher (1970) that Strotz and Wold describe as “wiping out” equations of the system that otherwise govern the behavior of a given variable, and replacing these with arbitrary values (“settings” here). This device is also at work in Pearl’s DAG approach (e.g., Pearl, 1995, 1998) where it is referred to as “intervention.” We employ “setting” to convey the same idea, reserving “intervention” for formal definition later in a closely related context.

The response relation corresponds to Pearl’s “functional causal equation” (see e.g.,

Pearl, 2000, p. 27). Here, however, we do not (yet) distinguish between observable and unobservable causes and we permit the appearance of attributes.

Our discussion above refers to agents $h = 1, 2, \dots$; the definition also references settings $Z_{h,j}$ for $h = 0$. Nevertheless, $h = 0$ does not refer to an agent. Rather, $Z_0 \equiv \{Z_{0,j} : j = 1, 2, \dots\}$ refers to particular settings that play a unique role in the settable system. For example, some agents' responses may not depend on any other variable of the system. For such cases, for a given index j^* $Z_{0,j}$ specifies the “default” value for such a response: for some indexes h and j , $\mathcal{X}_{h,j}(0, \cdot) = Z_{0,j^*}$. A second role is that of “initial” values: the impact of other variables is to change the response $\mathcal{X}_{h,j}(0, \cdot)$ to something different than its initial value, Z_{0,j^*} . A third role for Z_0 , containing the first two as special cases, is as “fundamental” variables, in that \mathcal{X}_0 has settings $Z_0 = \mathcal{X}_0(1, \cdot)$ that may impact other system variables, but its responses $\mathcal{X}_0(0, \cdot)$ do not depend on any other system variables. This justifies our convention $\mathcal{X}_0(0, \cdot) = \mathcal{X}_0(1, \cdot)$.

We relate Rubin's (1974) treatment effect framework to Definition 2.1 using Holland's (1986) mathematical description. Holland represents Rubin's framework as a quadruple (U, K, Y, S) , with sets U and K , $Y : U \times K \rightarrow \mathbb{R}$, and $S : U \rightarrow K$. U contains the population of “units” whose responses to treatment are of interest. K contains the admissible values of treatment “labels” (e.g., $K = \{t, c\}$, with “ t ” for treated, “ c ” for control). Y is the “potential outcome” function: $Y(u, k)$ is the response of unit u to treatment k . S is the treatment assignment function: unit u receives treatment $S(u) \in K$. Holland also references other measurements $X : U \times K \rightarrow \mathbb{R}$ (e.g., covariates), and a special measurement $X : U \rightarrow \mathbb{R}$ explicitly identified as attributes.

The function $Y : U \times K \rightarrow \mathbb{R}$ corresponds to a response function, say $r_{\cdot,1} : \mathbb{R}^\infty \times \mathcal{A}^\infty \rightarrow \mathbb{R}$ in Definition 2.1. The units u of U correspond to agents $h \in \{1, 2, \dots\}$ of Definition 2.1. Holland's K is the analog of \mathbb{R}^∞ . By restricting $r_{\cdot,1}$ to depend on a single variable (treatment) and not to depend on attributes except through the agent index, we obtain mapping Y . The treatment assignment $S(u)$ corresponds to a setting, say $Z_{h,2}$, in Definition 2.1. The covariates X correspond to other responses or settings in Definition 2.1, and Holland's attributes $X(u)$ correspond to attributes a_h . Thus, Holland's formulation of Rubin's treatment effect framework is contained in Definition 2.1. Definition 2.1 contains much more, however. The stochastic structure (Ω, \mathcal{F}, P) is easily adjoined to Rubin's framework, so this is not an important difference. A major difference is that we allow agent optimizing behavior and interaction with other agents.

As a concrete example, Definition 2.1 permits trade. Consider an experiment in which agents differ by the attribute of preferences and possess varying initial endowments of several goods (fundamental settings). A subset is “treated” by receiving a standardized

additional endowment; a “control” group does not receive this increment. Then the agents are permitted to trade, with the responses of interest being agents’ consumption of each good after trading ceases. In such systems, the treatment of one agent generally affects the responses of others. In the treatment effect literature, this is typically ruled out by the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980, 1986): treatments affect only treated units. We remove this: the response $r_{h,j}$ may depend on the settings of all other variables $Z_{(h,j)}$, not just the settings for that agent.

Another difference between Rubin’s framework and the settable system framework is that responses may differ not only on the basis of the agent’s own attributes, but also the attributes of all other agents a given agent may interact with, either directly or indirectly.

2.2 Heuristics of Cause and Effect

The formal notion of response functions makes possible formal definitions of causality. We begin with some heuristics. First, we emphasize that the notion of “cause” here is not absolute: it is always defined relative to a specified settable system \mathcal{S} . For given \mathcal{S} , the essential idea is that if, other things equal, the response of a given settable variable $\mathcal{X}_{h,j}$ does not depend on the setting of some other settable variable $\mathcal{X}_{i,k}$ ($i \neq h$ or $j \neq k$), then $\mathcal{X}_{i,k}$ does not cause $\mathcal{X}_{h,j}$. Otherwise, $\mathcal{X}_{i,k}$ causes $\mathcal{X}_{h,j}$. We give formal definitions below, but this suffices for now. We write $\mathcal{X}_{i,k} \Rightarrow_{|\mathcal{S}} \mathcal{X}_{h,j}$ if for a given \mathcal{S} $\mathcal{X}_{i,k}$ does not cause $\mathcal{X}_{h,j}$ according to this heuristic definition and $\mathcal{X}_{i,k} \Rightarrow_{\mathcal{S}} \mathcal{X}_{h,j}$ if $\mathcal{X}_{i,k}$ causes $\mathcal{X}_{h,j}$.

We emphasize that causality is properly defined with respect to the *settable variables* $\{ \mathcal{X}_{h,j} \}$, and not with respect to the random variables representing settings or responses or to events involving these random variables. This reflects the explicit distinction provided by the settable system \mathcal{S} between the *stochastic structure* (Ω, \mathcal{F}, P) and the *causal structure* $(\mathcal{A}, a, Z, r, \mathcal{X})$. Accordingly, we also call settable variables “causal variables.” Causal variables $\mathcal{X}_{h,j}$ thus generate random variables $\mathcal{X}_{h,j}(1, \cdot)$ and $\mathcal{X}_{h,j}(0, \cdot)$.

Although the present framework is substantially inspired by the work of Pearl (1988, 1993a, 1993b, 1995, 1998, 2000) and of SGS, our focus on settable variables distinguishes it from that work. There, causality is defined in terms of either events or random variables, and it is formally *axiomatic* that if A causes B , then B does not cause A . (SGS, p. 42.) This axiom rules out the perplexities of “simultaneous causation” in Pearl’s framework and enforces acyclicity of the directed graphs embodying causal relations there.

This axiom is antithetical to economics, given our central interest in processes of optimization and equilibrium. For example, in game theory, each agent’s behavior is related to that of all others through the agent’s “best response” function. Formally, these are response functions, as defined above. In the sense just given, each agent’s behavior is caused by that of the others. Indeed, the lack of a satisfying way to handle this “mutual

causality” may be one reason why the important advances of Pearl and his colleagues have had less impact in economics than elsewhere. (For example, Shipley (2000) exposit Pearl’s framework from a biological viewpoint.) In contrast, settable systems easily handle mutual causality, without creating the perplexities of simultaneous causation.

To see this, consider that causality, as defined heuristically above, admits four possibilities: $\mathcal{X}_{i,k} \Rightarrow_S \mathcal{X}_{h,j}$ and $\mathcal{X}_{h,j} \Rightarrow_S \mathcal{X}_{i,k}$ (mutual non-causality); $\mathcal{X}_{i,k} \Rightarrow_S \mathcal{X}_{h,j}$ and $\mathcal{X}_{h,j} \Rightarrow_S \mathcal{X}_{i,k}$ (directed causality from $\mathcal{X}_{h,j}$ to $\mathcal{X}_{i,k}$); $\mathcal{X}_{h,j} \Rightarrow_S \mathcal{X}_{i,k}$ and $\mathcal{X}_{i,k} \Rightarrow_S \mathcal{X}_{h,j}$ (directed causality from $\mathcal{X}_{i,k}$ to $\mathcal{X}_{h,j}$); and $\mathcal{X}_{i,k} \Rightarrow_S \mathcal{X}_{h,j}$ and $\mathcal{X}_{h,j} \Rightarrow_S \mathcal{X}_{i,k}$ (mutual causality). The latter case creates no paradoxes or perplexities, because it arises from two separate scenarios, one in which one variable is free to respond to the settings of the others, and another in which the other variable is free to respond to the settings of the remaining variables.

Settable systems also admit causal “cycles.” That is, we may have $\mathcal{X}_{i,k} \Rightarrow_S \mathcal{X}_{h,j}$, $\mathcal{X}_{h,j} \Rightarrow_S \mathcal{X}_{g,l}$, and $\mathcal{X}_{g,l} \Rightarrow_S \mathcal{X}_{i,k}$. This is a form of “indirect” mutual causality. Again there is no paradox, due to the separation enforced between settings and responses.

Settable systems are both a refinement and an extension of Pearl’s framework. Dawid’s (2002) “intervention directed acyclic graphs” (IDAGs) extend Pearl’s framework to explicitly accommodate “interventions” (direct changes) to each variable of the system. Settable systems constitute an extension of Dawid’s IDAGs, in that a settable system can be represented as a collection of modified IDAGS, one for each settable variable of the system, each representing the response (causal) dependencies for a given settable variable. Each such IDAG is modified in that it admits interventions only, and not to all system variables, but to all but the settable variable whose response it represents.

Notions of “effect” are straightforwardly defined in terms of the response function. Specifically, when the derivative exists, the *marginal ceteris paribus effect on $\mathcal{X}_{h,j}$ of $\mathcal{X}_{i,k}$ given $z_{(h,j)}$* is $(\partial/\partial z_{i,k}) r_{h,j}(z_{(h,j)}, a)$. The “reciprocal” marginal ceteris paribus effect on $\mathcal{X}_{i,k}$ of $\mathcal{X}_{h,j}$ given $z_{(i,k)}$ is $(\partial/\partial z_{h,j}) r_{i,k}(z_{(i,k)}, a)$. Because of the directionality of causal relations, there is no necessary inverse relation between reciprocal effects. When $\mathcal{X}_{i,k}$ does not cause $\mathcal{X}_{h,j}$, its marginal effects are zero everywhere. Nevertheless, the marginal effect of $\mathcal{X}_{i,k}$ on $\mathcal{X}_{h,j}$ can be zero for certain values of $z_{(h,j)}$, even when $\mathcal{X}_{i,k}$ causes $\mathcal{X}_{h,j}$.

Effects involving several settings can also be readily defined. For example, the *effect on $\mathcal{X}_{h,j}$ of the intervention $z_{(h,j)} \rightarrow z_{(h,j)}^*$ to $\mathcal{X}_{(h,j)}$* , is the difference

$$\Delta r_{h,j}(z_{(h,j)}, z_{(h,j)}^*, a) \equiv r_{h,j}(z_{(h,j)}^*, a) - r_{h,j}(z_{(h,j)}, a).$$

In defining this, we use the notion of intervention, as is standard. At one level, this is purely formal: an effect is a property of the response function involving evaluation at

two different values of its first argument. The intervention $z_{(h,j)} \rightarrow z_{(h,j)}^*$ specifies the two values. At another level, however, “intervention” implies active manipulation of some kind, suggesting a change in the settings of $\mathcal{X}_{(h,j)}$ from $z_{(h,j)}$ to $z_{(h,j)}^*$. Nevertheless, this manipulation is purely conceptual and not empirically observable. The definition compares the response under two possible outcomes of the settings, $z_{(h,j)}^* = Z_{(h,j)}(\omega^*)$ and $z_{(h,j)} = Z_{(h,j)}(\omega)$ for $\omega^*, \omega \in \Omega$. Accordingly, we formally define an *intervention* as a pair (ω, ω^*) of Ω elements. Any effect of interest can be defined using this formalism. For example, a marginal effect is the limit of a sequence of ratios involving such effects.

Empirically, only one realization from Ω operates to generate the observed data; for a given intervention (ω, ω^*) , this realization need not correspond to either ω^* or ω . Put differently, if the realization from Ω generating the data is “fact,” then for a given intervention, either ω^* or ω or *both* must be counterfactual. Consistent with common usage, we will speak of interventions as “changes in settings,” but with the conceptual and counterfactual nature of the manipulations involved clearly in mind.

In contrast to settings, attributes do not depend on ω so are *fixed* in this precise sense. For a given settable system, attributes are thus not subject to even the conceptual manipulations applicable to settings. This justifies the convention that attributes do not have causal properties, consonant with Holland and Rubin’s dictum, “No causation without manipulation” (Holland, 1986, p. 959). In other words, attributes *cannot act as causes and cannot have effects*. Rather, their role is to modify responses and effects.

By formalizing notions of interventions and effects as we have, we define these concepts *within* the settable system. Another way to manipulate realizations of settings of interest is to change the mappings $Z_{h,j}$ defining the settings, while keeping ω fixed. This corresponds to changing what Heckman (2006) calls the “policy regime.” For clarity, we call such alterations *modifications to the settable system*. Modifications are not necessary for defining concepts of effect, so we do not employ them for this purpose.

Modifications play other important roles, however. Certain modifications correspond to the exercise of experimental control. For example, modifying the *stochastic* structure by altering the probability measure P can give different patterns of treatment assignment. In particular, randomized treatment assignments are useful for learning about causal effects, so one may exercise experimental control to enforce a randomizing P .

Modifying the *causal* structure can also involve experimental control. For example, the choice of mapping Z can determine whether treatments are binary, categorical, or continuous. These choices are part of the design of the experiment. Modifying the attribute space and the attribute sequence permits a choice of the population of agents studied: by this means one may focus a given study on women, on collections of neurons,

or on firms in a given industry, pre-merger or post-merger.

Accordingly, one may view a system as an *experiment* and view modifications to the system as the exercise of experimental control implementing experimental design.

Not all modifications of settable systems require physical control, however. An especially important modification may involve only conceptual manipulations. These impact the response functions by partitioning the settable system in useful ways. As we discuss below, partitions that lead to recursive structures can be particularly useful.

Although settable systems admit mutual causality and cycles without paradox, systems in which these are absent are more straightforward to analyze. To keep the analysis here manageable, we focus on identifying causal effects in such systems. We leave analysis of causal effects in systems admitting mutual causality or cycles to other work.

2.3 Partitioned Settable Systems and Formal Causality

To resolve mutual causalities, we begin by defining partitioned systems.

Definition 2.2 Let (Ω, \mathcal{F}, P) , \mathcal{A} , a , and Z be as in Definition 2.1. Suppose that settings of $\mathcal{X}_{h,j}$ are given by $\mathcal{X}_{h,j}(1, \cdot) = Z_{hj}$, $j = 1, 2, \dots$, $h = 0, 1, \dots$, and let $\Pi = \{\Pi_b\}$ be a partition of the ordered pairs $\{(h, j): j = 1, 2, \dots; h = 1, 2, \dots\}$. Suppose there exists a countable sequence of measurable functions $r^\Pi \equiv \{r_{h,j}^\Pi\}$ such that for all (h, j) in Π_b the responses $Y_{h,j} = \mathcal{X}_{h,j}(0, \cdot)$ are jointly determined as

$$Y_{h,j} = r_{h,j}^\Pi(Z_{(b)}, a), \quad b = 1, 2, \dots,$$

where $Z_{(b)}$ is the countable vector containing $Z_{i,k}$, $(i, k) \notin \Pi_b$. Then $\mathcal{S} \equiv \{(\Omega, \mathcal{F}, P), (\mathcal{A}, a, Z, \Pi, r^\Pi, \mathcal{X})\}$ is an *attribute-indexed partitioned settable system*. ■

The response functions $r_{h,j}^\Pi$ give the *joint* response of settable variables within a block to settings *outside* that block. The responses in a given block cannot depend on settings in the same block, eliminating mutual causality within the block.

The response functions of Definition 2.1 correspond to the *elementary partition* with typical block $\Pi_b = \{(h, j)\}$. Other useful partitions group together all responses for a given agent or collection of agents. The former case yields the *agent partition*, where $\Pi_b = \{(b, j), j = 1, 2, \dots\}$ (here $b = h$), permitting the responses to represent the outcome of an agent's (joint) optimization problem. The latter permits the responses to represent, for example, the equilibrium price response of a competitive market or of a cartel.

To illustrate, consider a firm optimizing its factor inputs of capital and labor. The joint response functions for capital and labor for the agent partition specify jointly optimal choices for capital and labor in response to rents, wages, and other relevant factors. The

response functions for the elementary partition embody short run responses separately describing (1) the firm’s optimal choice of labor given settings of capital, wages, and rents, and (2) the firm’s optimal choice of capital given settings of labor, wages, and rents. Mutual causality exists between the capital and labor responses for the elementary partition but are absent within the block for the agent partition.

As another example, consider a static game with complete information. Here each of n agents has attributes defining their playable strategies (finite in number) and the utility (payoff) derived from each configuration of playable strategies (an n -tuple). By employing the agent partition Π (each agent’s responses depend only on settings for the other agents) the partitioned system can represent each player’s mixed strategy best response. This is a vector of probabilities, each element of which specifies the probability that a given player will play one of their available strategies. These best response probabilities are determined by the probabilities other players assign to playing their available strategies. Thus, $\mathcal{X}_{h,j}(0, \cdot)$ gives the probability that agent h will play strategy j , determined as a function $r_{h,j}^{\Pi}$ (the best response function) of (1) the given probabilities of play $\mathcal{X}_{(h)}(1, \cdot)$ for all other agents; and (2) agent h ’s own attributes, a_h .

Continuing this example, consider the partition Π^* that groups together all agents. Now the responses represent each agent’s behavior solely as a function of the attributes of all agents, namely their playable strategies and payoffs. A standard result of game theory for this game (e.g., Gibbons, 1992, p. 45) ensures at least one mixed strategy Nash equilibrium. Suppose this equilibrium is unique. Then $\mathcal{X}_{h,j}(0, \cdot) = r_{h,j}^{\Pi^*}(a)$ can represent the mixed strategy equilibrium response for player h playing strategy j .

If Nash equilibrium is not unique, then an equilibrium selection mechanism will deliver a unique equilibrium response. This can be straightforwardly accommodated by introducing fundamental variables \mathcal{X}_0 , in which case $\mathcal{X}_{h,j}(0, \cdot) = r_{h,j}^{\Pi^*}(\mathcal{X}_0(1, \cdot), a)$ represents the selected mixed strategy equilibrium. More elaborate selection mechanisms are possible (see, e.g. Bajari, Hong, and Ryan, 2004). Another role for \mathcal{X}_0 here is to determine the observable action taken by each agent according to the selected equilibrium.

Although partitioning removes mutual causalities within a block, it still permits mutual causalities between blocks or causal cycles. To make this precise and to define structures in which both direct and indirect mutual causality are absent, we need a formal definition of causality. We now have a sufficiently rich formal framework in which to state the required definition. We add one more item of notation: for a given $\Pi = \{\Pi_b\}$, $z_{(b),(i,k)}$ denotes the vector with elements corresponding to every setting of the system except those indexed by the elements of Π_b and by $(i, k) \notin \Pi_b$.

Definition 2.3 Let $\mathcal{S} \equiv \{(\Omega, \mathcal{F}, P), (\mathcal{A}, a, Z, \Pi, r^{\Pi}, \mathcal{X})\}$ be an attribute-indexed par-

tioned settable system. For given positive integer b , let $(h, j) \in \Pi_b$. (i) If for given $(i, k) \notin \Pi_b$ the function $z_{i,k} \rightarrow r_{h,j}^\Pi(z^{(b)}, a)$ is constant in $z_{i,k}$ for every $z^{(b),(i,k)}$, then we say $\mathcal{X}_{i,k}$ *does not cause* $\mathcal{X}_{h,j}$ in \mathcal{S} and write $\mathcal{X}_{i,k} \not\Rightarrow_S \mathcal{X}_{h,j}$. Otherwise, we say $\mathcal{X}_{i,k}$ *causes* $\mathcal{X}_{h,j}$ in \mathcal{S} and write $\mathcal{X}_{i,k} \Rightarrow_S \mathcal{X}_{h,j}$. (ii) For $(i, k), (h, j) \in \Pi_b$, $\mathcal{X}_{i,k} \Rightarrow_S \mathcal{X}_{h,j}$. ■

In defining non-causality, this definition formalizes the heuristic given above that the response of $\mathcal{X}_{h,j}$ does not depend on the setting of $\mathcal{X}_{i,k}$. For brevity, we let the settings take real values. Binary variables require only the admissible values $\{0, 1\}$, and similarly for categorical or bounded continuous variables. Only the admissible values of $z_{i,k}$ and $z^{(b),(i,k)}$ are relevant. Extensions where settable variables take values in a function space (e.g. conditional probability densities) accommodate a causal role for beliefs.

Our definition corresponds to that of *direct* cause in Pearl’s structural DAG framework. In other work (Chalak and White, 2006b) a carefully defined notion of indirect causality plays a key role, but for present purposes direct causality suffices.

By definition, responses in a given block can depend only on settings in other blocks. Part (ii) explicitly states the ensuing necessary within-block absence of causality. This includes the convention that causality is non-reflexive: $\mathcal{X}_{i,k} \Rightarrow_S \mathcal{X}_{i,k}$. Generally, however, when we write $\mathcal{X}_{i,k} \Rightarrow_S \mathcal{X}_{h,j}$, it will be understood that the referenced variables belong to different blocks, unless explicitly stated to the contrary.

The notations \Rightarrow_S and \Rightarrow_S extend to accommodate disjoint sets of variables on the left and right by the convention that the indicated relation holds between each pair of left-hand and right-hand variables. Thus, $\mathcal{X}_i \Rightarrow_S \mathcal{X}_h$ means $\mathcal{X}_{i,k} \Rightarrow_S \mathcal{X}_{h,j}$ for all k and j .

2.4 Recursive and Reduced Settable Systems

In the standard DAG framework, causality is anti-symmetric: an axiom is that “if A causes B, then B does not cause A” (SGS, p. 42). So far we do not require this. We now define a system with anti-symmetry.

Definition 2.4 Let $\mathcal{S} \equiv \{(\Omega, \mathcal{F}, P), (\mathcal{A}, a, Z, \Pi, r^\Pi, \mathcal{X})\}$ be an attribute-indexed partitioned settable system. For $b = 1, 2, \dots$, let $\mathcal{X}_{[b]}$ denote the vector containing the settable variables $\mathcal{X}_{h,j}$ for $(h, j) \in \Pi_b$ and let $\mathcal{X}_{[0]} = \mathcal{X}_0$. If Π is such that $\mathcal{X}_{[b]} \Rightarrow_S \mathcal{X}_{[0], \dots, \mathcal{X}_{[b-1]}}$, $b = 1, 2, \dots$, then \mathcal{S} is an *attribute-indexed recursive settable system*. ■

This block recursive structure eliminates mutual causalities and cycles. It is an analog of classical recursive systems analyzed by Simon (1953), Wold (1954, 1956), and Fisher (1961, 1966), among others. The analogy is loose at best, however. Specifically, in classical block recursive systems, the responses in a given block may be simultaneously determined (e.g., Fisher, 1966, pp 99-100). Here, however, the responses in a given block

are explicitly *not* mutually causal. Further, the classical unobservable disturbances must obey a block diagonal covariance structure. Here, this need not hold, as unobservables in a given block may have effects in succeeding blocks.

Every attribute-indexed settable system can be partitioned to form at least one recursive system. Specifically, form the two block *fundamental partition* containing \mathcal{X}_0 in block $b = 0$ and $\mathcal{X}_{(0)}$ (the remaining settable variables) in block $b = 1$.

The block ordering enables us to refer to “levels” b of the partition, and to define “successors” (higher level blocks or their elements) and “predecessors” (lower level blocks or their elements). If $\mathcal{X}_{h,j}$ belongs to a block succeeding that containing $\mathcal{X}_{i,k}$, we say $\mathcal{X}_{h,j}$ \mathcal{S} -*succeeds* $\mathcal{X}_{i,k}$ or $\mathcal{X}_{i,k}$ \mathcal{S} -*precedes* $\mathcal{X}_{h,j}$ and write $\mathcal{X}_{h,j} \leftarrow_{\mathcal{S}} \mathcal{X}_{i,k}$. Then $\mathcal{X}_{h,j} \leftarrow_{\mathcal{S}} \mathcal{X}_{i,k}$ implies $\mathcal{X}_{h,j} \Rightarrow_{\mathcal{S}} \mathcal{X}_{i,k}$, but the converse does not hold, nor does $\mathcal{X}_{h,j} \leftarrow_{\mathcal{S}} \mathcal{X}_{i,k}$ imply $\mathcal{X}_{i,k} \Rightarrow_{\mathcal{S}} \mathcal{X}_{h,j}$. When $\mathcal{X}_{h,j}$ succeeds each element of a vector $\tilde{\mathcal{X}}$, we write $\mathcal{X}_{h,j} \leftarrow_{\mathcal{S}} \tilde{\mathcal{X}}$.

So far, time has played no role here, as there has been no formal need for it. Nevertheless, time can play an important role by imposing natural recursive structure. To specify this role, we introduce a special identifier, “time,” such that settings or responses for different time indexes are necessarily distinct. For agent h , this means that the indexes j implicitly depend on time, such that settings or responses for different times, say $t \neq t'$, necessarily have $j \neq j'$. With this dependence explicit, settable variables are indexed by (h, j, t) ; (j, t) now plays the role previously played by j . (Attributes may also be time-indexed; as such, they are deterministic functions of time, such as age.)

With time explicit, the settable system obeys *weak time structure* if for all (i, k) and (j, h) $\mathcal{X}_{i,k,t} \Rightarrow_{\mathcal{S}} \mathcal{X}_{h,j,\tau}$ for all $t > \tau$, $t = 1, 2, \dots$. It obeys *strong time structure* if for all (i, k) and (j, h) $\mathcal{X}_{i,k,t} \Rightarrow_{\mathcal{S}} \mathcal{X}_{h,j,\tau}$ for all $t \geq \tau$, $t = 1, 2, \dots$. Strong time structure thus rules out “contemporaneous” causality. In the classical framework, this eliminates simultaneity (cf. Wold, 1954, 1956; Cartwright, 1989). Here this eliminates mutual causality and cycles. Weak time structure permits these. In practice, empirical researchers often use weak time structures. Economic theory then provides further structure, specifying which causal variables may or may not cause which others.

It may seem counterintuitive that weak time structure is preferred to strong in practice, as contemporaneous causation may seem both physically and metaphysically unappealing. This preference may be a response to the fact that the *operational* time scale at which responses occur may be much smaller than the *observational* time scale on which responses are measured. As the researcher can only work with measured responses, contemporaneous causation may afford a useful approximation to the observed phenomena.

Although settable systems permit explicit consideration of time and dynamics, to keep the analysis manageable, we revert to our previous implicit treatment, but recognize that

recursive structures may arise in significant part from temporal relationships. In game theoretic terms, the system can be viewed as representing repeated or dynamic games in extensive form.

Recursive settable systems are especially convenient because, as we now show, the settings of predecessor variables driving a given response can be viewed as responses to further predecessor settings. Nor is it necessary to be fully explicit about such predecessor responses. This is especially useful in economics, where both the ability to directly set potential causes and the full understanding of their genesis otherwise may be absent.

To justify this claim, we introduce convenient notation. Let $r_{[b]}^{\Pi}$ denote the joint response functions for levels $b = 1, 2, \dots$ of a recursive settable system such that

$$\mathcal{X}_{[b]}(0, \cdot) = r_{[b]}^{\Pi}(\mathcal{X}_{[0]}(1, \cdot), \dots, \mathcal{X}_{[b-1]}(1, \cdot)).$$

For simplicity, we leave attributes implicit. Let $Z_{[b]} = \mathcal{X}_{[b]}(1, \cdot)$ denote the settings of $\mathcal{X}_{[b]}$, $b = 0, 1, \dots$. We define *canonical settings* recursively as

$$Z_{[b]} = r_{[b]}^{\Pi}(\mathcal{X}_{[0]}(1, \cdot), \dots, \mathcal{X}_{[b-1]}(1, \cdot)), \quad b = 1, 2, \dots$$

These are valid settings, because $Z_{[b]}$ is a composition of measurable functions, which are themselves measurable functions. They are canonical in that by construction $Z_{[b]} = \mathcal{X}_{[b]}(0, \cdot)$, $b = 0, 1, \dots$, so that

$$Z_{[b]} = r_{[b]}^{\Pi}(\mathcal{X}_{[0]}(0, \cdot), \mathcal{X}_{[1]}(0, \cdot), \dots, \mathcal{X}_{[b-1]}(0, \cdot)).$$

Consequently, *these are the settings generated in the absence of experimental control.*

When the settings are canonical, we can view the settings driving the response at any level as responses to predecessor settings. Although this may seem to blur the distinction between settings and responses, it will continue to be helpful to recognize this distinction, if only implicitly. We formalize the discussion above as follows.

Definition 2.5 Let $\mathcal{S} \equiv \{(\Omega, \mathcal{F}, P), (\mathcal{A}, a, Z, \Pi, r^{\Pi}, \mathcal{X})\}$ be an attribute-indexed recursive settable system. Suppose the settings are the *canonical settings* such that

$$\mathcal{X}_{[b]}(1, \cdot) = Z_{[b]} = r_{[b]}^{\Pi}(\mathcal{X}_{[0]}(1, \cdot), \dots, \mathcal{X}_{[b-1]}(1, \cdot)), \quad b = 1, 2, \dots$$

Then \mathcal{S} is an *attribute-indexed canonical recursive settable system*. ■

In the classical framework, recursivity eliminates simultaneity but does not by itself eliminate “endogeneity,” i.e., the correlation of observable right-hand side variables with unob-

servable causes. The classical reduced form eliminates endogeneity as well as simultaneity by solving a system of structural equations to express the endogenous variables as functions of observable exogenous variables, independent of unobservable “error terms.” In the recursive case, simple substitution suffices. In the simultaneous case, the reduced form equations represent the fixed point or equilibrium (assuming it exists) of the underlying simultaneous equations (see e.g., Fisher, 1970 and Matzkin, 2005).

The settable system analog of the reduced form is given by the fundamental partition Π^0 defined above. Here

$$\mathcal{X}_{[1]}(0, \cdot) = r_{[1]}^{\Pi^0}(\mathcal{X}_{[0]}(1, \cdot)),$$

where $\mathcal{X}_{[0]} = \mathcal{X}_0$ and $\mathcal{X}_{[1]} = \mathcal{X}_{(0)}$, the non-fundamental variables. We call $r_{[1]}^{\Pi^0}$ the “reduced response function.” There are several important differences between the classical reduced form and the reduced response. First, although $\mathcal{X}_{[0]}(1, \cdot)$ is determined outside the system, it is not exogenous in the standard sense. Some elements of the fundamental variables may be observed, whereas others may not. Further, even if some are observable, we do not assume independence between observables and unobservables. Nor do we assume the separability of the reduced response between its observable and unobservable arguments, as does the classical reduced form.

Moreover, our goal is to identify and ultimately estimate effects of non-fundamental variables on particular successors. As we show, we can identify these effects without involving the reduced response. Accordingly, we do not consider it further.

2.5 Constructing Recursive Settable Systems

It is not always obvious how to construct a recursive settable system from a given system admitting mutual causality or cycles; this is often true even in their absence. Fortunately, powerful methods for obtaining recursive systems are provided by graph theory.

Specifically, for finite settable systems (realistic in practice), proposition 1.4.3 of Bang-Jensen and Gutin (2001) (BJG) establishes the existence of a recursive settable system whenever a partitioned system is acyclic. The DFSA algorithm of BJG (chapter 4.2) delivers the required ordering of the variables.

For systems with mutual causalities or cycles, notions of joint optimization and equilibrium can play a central role in suggesting partitions yielding recursive structures, as our examples of Section 2.3 should suggest. Lauritzen and Richardson (2002, Section 6) discuss several physically motivated equilibrium processes. Equilibrium concepts and processes relevant to economics (e.g., those of dynamic games) can be applied analogously. This is an extensive topic that must necessarily be taken up elsewhere. Once a suitable

acyclic structure has been found, the DFSA algorithm then delivers the required recursive ordering of the settable variables.

2.6 Settable Systems Generating Samples

In applications, interest often focuses on comparable responses for a class of comparable agents, responding to settings with common meanings. For example, interest might focus on retail store price responses at various locations to settings of wages, rents, pre-existing density of nearby competitors, nearby consumer demographics, etc.

In such circumstances, one identifies a subset \mathbf{A} of attribute space \mathcal{A} designating the comparable agents whose responses are of interest and then maps the relevant comparable settable variables to a standardized collection of settable variables, $(\mathcal{Y}_h, \mathcal{Z}_h)$ say, permitting representation of the responses as

$$Y_h \stackrel{c}{=} r(Z_h, a_h, a_{(h)}) \quad a_h \in \mathbf{A}, \quad a_{(h)} \in \mathcal{A}_{(h)}.$$

In this representation the notation $\stackrel{c}{=}$ emphasizes the causal nature of the relation. This makes causal roles clear even when settings and responses are not explicitly identified. For our applications, we require that \mathcal{Y}_h \mathcal{S} -succeeds \mathcal{Z}_h ($\mathcal{Y}_h \leftarrow_{\mathcal{S}} \mathcal{Z}_h$). For convenience, when this holds for each $a_h \in \mathbf{A}$, we write $\mathcal{Y} \leftarrow_{\mathcal{S}} \mathcal{Z}$.

We split the dependence of the responses on the attributes into two parts, the “own attributes” a_h of the agent, which by construction belong to \mathbf{A} , and the “other agent attributes” $a_{(h)}$. When other agent attributes enter the response, the relevant other agents should be interpreted in relation to the given agent (e.g., as siblings, customers, or potential competitors). We let $\mathcal{A}_{(h)}$ designate the space in which $a_{(h)}$ takes its values.

So far, we have focused on the full population of agents, $h = 1, 2, \dots$. In economics, we usually observe only a sample from the population. Here, sampling corresponds to generating random positive integers, say $\{H_i\}$, governed by the probability measure P , such that realizations of a_{H_i} belong to \mathbf{A} . This yields sample responses

$$Y_{H_i} \stackrel{c}{=} r(Z_{H_i}, a_{H_i}, a_{(H_i)}), \quad i = 1, 2, \dots$$

To simplify notation, let $A_i \equiv (a_{H_i}, a_{(H_i)})$, and, engaging in a mild abuse of notation, write $Z_i = Z_{H_i}$ and $Y_i = Y_{H_i}$, so that sample responses can be represented as

$$Y_i \stackrel{c}{=} r(Z_i, A_i), \quad i = 1, 2, \dots$$

We refer to this standard sampling situation by saying that \mathcal{S} *generates a sample from \mathbf{A} involving settable variables $(\mathcal{Y}, \mathcal{Z})$* .

Although the dependence of agent responses on settings and attributes of other agents introduces stochastic dependence among agents in the population, independent sampling results in independence of sample observations, conditional on population values. Other forms of sampling may preserve aspects of the population dependence. These may be assessed in applications and dealt with appropriately for estimation and inference using suitable laws of large numbers and central limit theorems. Our identification results will not require independence.

3 Structural Identification with Conditional Exogeneity

We build on the foundations of Section 2 and add structure enabling us to define and identify causal effects when not all relevant explanatory variables are observed.

3.1 A Causally Structured Data Generating Process

We first specify how the data are generated. Our conditions permit but do not require randomization, as randomization is often not plausible in observational studies. In such cases, suitable covariates can be exploited to identify effects of interest. The availability of such covariates is typically assumed, but the literature provides little guidance as to their construction. Our first assumption accommodates covariates in a way that yields significant insight into their choice. Here, \mathbb{N} denotes the natural numbers, including zero by convention; \mathbb{N}^+ denotes the positive integers; and $\bar{\mathbb{N}} \equiv \mathbb{N} \cup \{\infty\}$.

Assumption A.1 Let an attribute-indexed recursive settable system \mathcal{S} generate a sample from $\mathbf{A} \subset \mathcal{A}$ involving settable variables $(\mathcal{Y}, \mathcal{D}, \mathcal{W}, \mathcal{Z})$ such that $\mathcal{Y} \leftarrow_S (\mathcal{D}, \mathcal{W}, \mathcal{Z})$. In addition: (a) Let attributes $\{(A_i, \tilde{B}_i) \equiv (\tilde{A}_i, \ddot{A}_i, \tilde{B}_i)\}$ be a sequence of random vectors, and let $(\mathcal{D}, \mathcal{W}, \mathcal{Z})$ generate $\{(D_i, W_i, Z_i) \equiv (D_i, W_i, \tilde{Z}_i, \ddot{Z}_i)\}$ such that the joint distribution of $(D_i, X_i) \equiv (D_i, W_i, \tilde{Z}_i, \tilde{A}_i, \tilde{B}_i)$ is H and the conditional distribution of $\ddot{X}_i \equiv (\ddot{Z}_i, \ddot{A}_i)$ given $(D_i, X_i) = (d, x)$ is $G(\cdot \mid d, x)$ for all $i = 1, 2, \dots$, where D_i is \mathbb{R}^{k_1} -valued, $k_1 \in \mathbb{N}^+$, W_i is \mathbb{R}^{k_2} -valued, $k_2 \in \mathbb{N}$, \tilde{Z}_i is \mathbb{R}^{k_3} -valued, $k_3 \in \mathbb{N}$, \ddot{Z}_i is \mathbb{R}^{k_4} -valued, $k_4 \in \bar{\mathbb{N}}$, \tilde{A}_i is ℓ_1 -valued, $\ell_1 \in \mathbb{N}$, \ddot{A}_i is \mathbb{R}^{ℓ_2} -valued, $\ell_2 \in \bar{\mathbb{N}}$, and \tilde{B}_i is ℓ_3 -valued, $\ell_3 \in \mathbb{N}$; (b) The responses $\{Y_i\}$ of \mathcal{Y} are

$$Y_i \stackrel{c}{=} r(D_i, Z_i, A_i), \quad i = 1, 2, \dots,$$

where r is an unknown measurable scalar-valued function; (c) (i) $\mathcal{D} \leftarrow_S (\mathcal{W}, \mathcal{Z})$; (ii) $\mathcal{W} \leftarrow_S \mathcal{Z}$; (d) The realizations of $Y_i, D_i, W_i, \tilde{Z}_i, \tilde{A}_i$, and \tilde{B}_i are observed; those of \ddot{Z}_i and \ddot{A}_i are not. ■

To see what this assumption entails, we begin with (b). This focuses attention on the response of \mathcal{Y}_i and the effects embodied in r , as modified by relevant attributes, A_i . We consider a single response without essential loss of generality. The response does not

depend on W_i or \tilde{B}_i . We thus call these “structurally irrelevant.” They nevertheless play an instrumental role in identifying causal effects. We defer discussion of this role until we have examined the response function more closely.

Because \mathcal{Y}_i \mathcal{S} -succeeds $(\mathcal{D}_i, \mathcal{Z}_i)$ for each i , \mathcal{Y}_i does not cause $(\mathcal{D}_i, \mathcal{Z}_i)$, either directly or indirectly. We view \mathcal{D}_i and \mathcal{Z}_i as potential causes of \mathcal{Y}_i , as we may not know *a priori* which elements of \mathcal{D}_i and \mathcal{Z}_i are truly causal. Permitting $(\mathcal{D}_i, \mathcal{Z}_i)$ to be countably dimensioned (A.1(a)), ensures that we reference every variable truly causal for \mathcal{Y}_i .

We reference \mathcal{D} and \mathcal{Z} separately to single out \mathcal{D}_i as containing the causal variables whose effects on \mathcal{Y}_i are of primary interest. Interest in effects of \mathcal{Z}_i is secondary. We thus call \mathcal{D}_i “causal variables of interest” and \mathcal{D}_i “causes of interest.” We call \mathcal{Z}_i “ancillary causal variables” and \mathcal{Z}_i “ancillary causes.” Rosenbaum (1984) calls these “concomitants”; our nomenclature emphasizes their causal role. The causes \mathcal{D}_i correspond to “treatments” in the treatment effect literature. There, treatments are often a binary scalar. Here, \mathcal{D}_i may be a vector with binary, categorical, or continuous components.

According to A.1(d), Y_i and \mathcal{D}_i are observable, as are sub-vectors \tilde{Z}_i and \tilde{A}_i of \mathcal{Z}_i and \mathcal{A}_i , respectively. Realizations of sub-vectors \tilde{Z}_i and \tilde{A}_i are not observed. We call $\tilde{X}_i \equiv (\tilde{Z}_i, \tilde{A}_i)$ “observed” explanatory variables or “observables” and $\ddot{X}_i \equiv (\ddot{Z}_i, \ddot{A}_i)$ “unobserved” explanatory variables or “unobservables.”

Typically, economic theory does not provide strong guidance about r . Accordingly, A.1(b) imposes the mildest possible structure on r . Classical assumptions of linearity or of separability between observables and unobservables are not imposed; nor are monotonicity, convexity, or other specific structures. This provides a framework where such restrictions can be tested. Correctly imposing such structure may not only help to identify features of interest not otherwise identifiable (as in Matzkin, 2003) but also enhance estimation efficiency. For brevity, we leave these possibilities aside here.

The response function r is assumed identical for all observations, but because r depends explicitly on agent attributes this is without loss of generality. The unobserved heterogeneity admitted here corresponds to the “essential heterogeneity” of HV and HUV. When \mathcal{D}_i is itself a response, essential heterogeneity is also permitted in determining \mathcal{D}_i .

In A.1(b) the response incorporates both observed and unobserved “fixed effects.” These are the attributes \mathcal{A}_i . Given the preceding discussion, the “fixed effect” nomenclature appears singularly unhelpful. Attributes are indeed fixed, but they are not effects (i.e., changes in a response with respect to an intervention) nor are they even potentially causal, given their immutability. They are rather “fixed effect/response modifiers,” that is, effect and response modifiers that are fixed.

The role of the observable but structurally irrelevant W_i and \tilde{B}_i is to serve as proxies

for the unobservables \ddot{X} . We call \mathcal{W} and \tilde{B} *predictive proxies*, as their role is to provide predictive information about \ddot{X} . We call W “proxy settings” and \tilde{B} “proxy attributes.” The imposed recursive structure ensures that \mathcal{Y}_i does not cause \mathcal{W}_i , either directly or indirectly. We call $X \equiv (W, \tilde{Z}, \tilde{A}, \tilde{B})$ *covariates*. Below we examine in detail both the instrumental role of the covariates in identifying effects of interest and their selection.

Assumption A.1(c.i) requires that \mathcal{D}_i \mathcal{S} -succeeds both \mathcal{W}_i and \mathcal{Z}_i : both \mathcal{W}_i and \mathcal{Z}_i may cause \mathcal{D}_i , but \mathcal{D}_i does not cause either \mathcal{W}_i or \mathcal{Z}_i either directly or indirectly. Often, one can ensure this by measuring W_i and Z_i prior to assigning or measuring D_i (e.g., see Rosenbaum and Rubin, 1983). But this is not always convenient or appropriate; in such cases A.1(c.i) is essential to a proper accounting of the effects of \mathcal{D}_i on \mathcal{Y}_i .

Specifically, if, contrary to A.1(c.i), elements of \mathcal{Z}_i respond either directly or indirectly to \mathcal{D}_i , then these elements generally mediate some of the effects of \mathcal{D}_i . Rosenbaum (1984), Angrist and Krueger (1999), Heckman and Vytlacil (2005), and Wooldridge (2005) among others, discuss what happens when mediating causes are present, in violation of A.1(c.i). Specifically, any statistical analysis of such a system will fail to recover the indirect effects of \mathcal{D}_i , unless the analysis explicitly accounts for the mediating responses. The requirement that \mathcal{D}_i \mathcal{S} -succeed \mathcal{Z}_i is a weak analog of Assumption A-6 of HUV and HV.

As \mathcal{W}_i is structurally irrelevant, it cannot mediate causal effects of \mathcal{D}_i . Nevertheless, W_i will be used to predict Y_i , so A.1(c.i) precludes W_i from absorbing some of the predictive impact otherwise attributable to D_i . We examine the motivation for A.1(c.ii) below. Because the A_i ’s are attributes, they cannot be causally affected by any settable variable of the system, and thus cannot mediate effects of \mathcal{D}_i .

Thus, we focus on the total effect of the causes of interest, encompassing both direct and indirect effects, by requiring the covariates not to respond to the causes of interest.

In A.1(a), the probability measure P generates explanatory variables and predictive proxies identically across observations. This can be weakened substantially without changing the character of the results, but at great notational expense. We do not impose independence, as this is not needed to analyze structural identification.

Assumption A.1(a) permits (D_i, Z_i, W_i) to be arbitrary settings, canonical settings, or a mix. For example, in a clinical study, D_i may be a dosage set by the researcher. In non-experimental situations, D_i may be a response to variables outside researcher control. Assumption A.1(a) explicitly permits dependence between observables and unobservables. Thus, the observed causes and covariates (D_i, X_i) are generally endogenous.

3.2 Structural Identification Via Conditional Expectations

Because we generally do not observe realizations of all structurally relevant causes and attributes, we cannot estimate r itself, making the effects defined in Section 2 inaccessible.

Heckman (2005) forcefully argues that notions of “effect” are inherently counterfactual. As Dawid (2000) discusses, provided that the counterfactuals involved relate to certain properly behaved averages, they can be assigned empirical meaning. We now show how this works here by defining average counterfactual responses and providing conditions that identify these with corresponding standard conditional expectations. This then supports empirically meaningful definitions of effects.

We start with a representation of the conditional expectation of the response given covariates. Let $\text{supp}(D, X)$ denote the support of (D, X) , the smallest measurable set on which the density dH of (D, X) integrates to one. (We exploit identical distribution (A.1(a, b)) to drop the subscript i). Corresponding to the notation \tilde{X}, \ddot{X} , and X , we write $\tilde{x} \equiv (\tilde{z}, \tilde{a})$, $\ddot{x} \equiv (\ddot{z}, \ddot{a})$, and $x \equiv (w, \tilde{z}, \tilde{a}, \tilde{b})$. We also write $r(d, \tilde{x}, \ddot{x}) = r(d, z, a)$.

Proposition 3.1 Suppose Assumptions A.1(a, b) hold such that $E(Y) < \infty$. Then (i) $\mu(D, X) \equiv E(Y \mid D, X)$ exists and is finite; and (ii) for each (d, x) in $\text{supp}(D, X)$

$$\mu(d, x) = \int r(d, \tilde{x}, \ddot{x}) dG(\ddot{x} \mid d, x). \quad \blacksquare$$

This represents $\mu(d, x)$ as the *average response given* $(D, X) = (d, x)$, or just the “average response,” leaving conditioning implicit. We call μ an *average response function*. It tells us the expected response given realizations (d, x) of (D, X) generated according to A.1(a, b) but does not say what to expect under interventions to causes of interest.

3.2.1 Covariate-Conditioned Average Effects

Given A.1(c.i), we can define a particular conditional expectation with a clear counterfactual interpretation. Specifically, when $E(r(d, Z, A))$ exists and is finite for each d in $\text{supp}(D)$, we define the *average counterfactual response at d given $X = x$* as

$$\rho(d, x) \equiv E(r(d, Z, A) \mid X = x) = \int r(d, \tilde{x}, \ddot{x}) dG(\ddot{x} \mid x),$$

where $dG(\cdot \mid x)$ is the conditional density of \ddot{X} given $X = x$. A.1(c.i) ($\mathcal{D} \leftarrow_S (\mathcal{W}, \mathcal{Z})$) ensures that different values for d do not necessitate different realizations of (W, Z) .

That is, we view $\rho(d, x)$ as representing $\rho(d, X(\omega))$ for $X(\omega) = x$, where X explicitly does not depend on d . Otherwise, one must consider

$$\rho(d, X(d, \cdot)) = E(r(d, Z(d, \cdot), A) \mid X(d, \cdot))$$

making the dependence of (W, Z) (hence X) on d explicit. This permits analysis of

mediated effects, but we leave this aside here. Given A.1(c.i), we thus interpret $\rho(d, x)$ as a representation in which d and x are variation-free without further explicit indication.

Pearl (1995, 2000) introduced the “do” operator to express counterfactual settings of causes. If \mathcal{D} is set to d , Pearl writes the expected response given $X = x$ as $E(Y \mid \text{do}(d), X = x)$. In our framework, $E(Y \mid \text{do}(d), X = x) = \rho(d, x)$.

The function ρ is a conditional analog of Blundell and Powell’s (2003) “average structural function.” We call ρ a *covariate-conditioned counterfactual average response function* or a “counterfactual average response function,” leaving conditioning implicit.

Comparing $\mu(d, x)$ and $\rho(d, x)$, we see that $dG(\ddot{x} \mid d, x)$ appears in $\mu(d, x)$, whereas $dG(\ddot{x} \mid x)$ appears in $\rho(d, x)$. If $dG(\cdot \mid d, x) = dG(\cdot \mid x)$, then $\mu(d, x) = \rho(d, x)$, so μ does provide counterfactual information. A sufficient condition for this is:

Assumption A.2 \ddot{X} is independent of D given X , written $\ddot{X} \perp D \mid X$. ■

In A.2 we use Dawid’s (1979) conditional independence notation. This ensures that for all relevant (d, x, \ddot{x}) , $dG(\ddot{x} \mid d, x) = dG(\ddot{x} \mid x)$. By analogy with the use of “exogeneity” to describe regressors independent of unobservable “disturbances” (e.g., Wooldridge, 2002, p. 50), when A.2 holds we say D is *conditionally exogenous*. As is readily verified, this concept is distinct from weak, strong, or super exogeneity (Engle, Hendry, and Richard, 1983). It contains strict exogeneity ($\ddot{X} \perp D$) as a special case (when $X \equiv 1$). When the covariates suffice to ensure conditional exogeneity for D , we call them *sufficient covariates*, following Dawid (1979). Conditional exogeneity ensures Rosenbaum and Rubin’s (1983) “unconfoundedness” condition. Section 7 provides discussion.

Conditional exogeneity holds under a generalization of *randomization*. Randomization, in which \mathcal{D} generates settings D_i randomly and independently of all observed and unobserved attributes and potential causes, was introduced by Fisher (1935, ch.2) as a powerful means of identifying causal effects of interest. Our next result shows how conditional exogeneity can be ensured by *conditional randomization*.

Assumption A.2' \mathcal{S} involves settable variables \mathcal{U} generating random vectors U_i , $i = 1, \dots, n$ such that (X_i, \ddot{X}_i, U_i) is identically distributed, $D_i \stackrel{c}{=} c(X_i, U_i)$, where c is a measurable vector-valued function, and $\ddot{X}_i \perp U_i \mid X_i$. ■

For compatibility with A.1, $\mathcal{Y}, \mathcal{D} \leftarrow_{\mathcal{S}} \mathcal{U}$ must hold. Pure randomization occurs when $D = U$, with $\ddot{X} \perp U$ (put $X \equiv 1$).

Proposition 3.2 Suppose A.1(a) and A.2' hold. Then $\ddot{X} \perp D \mid X$. ■

In experimental studies, this result justifies treatment assignment (and receipt) conditional on covariates X . In observational studies, this result provides key insight into the selection of covariates. We explore this in depth in Section 6.

Our first formal identification result identifies ρ with μ .

Theorem 3.3 Suppose A.1(a, b, c.i) and A.2 or A.2' hold with $E(Y) < \infty$. (i) Then for all $(d, x) \in \text{supp}(D, X)$, $\rho(d, x) \equiv \int r(d, \tilde{x}, \ddot{x}) dG(\ddot{x} | x)$ exists and $\rho(d, x) = \mu(d, x)$. ■

We call this a *structural identification* result because it identifies an aspect of the causal structure, ρ , with μ , a standard stochastic object. Because μ is empirically meaningful and empirically accessible, so is ρ , under the conditions given.

Using ρ , we can define the expected effect of any intervention to \mathcal{D} . Specifically, the *average effect on \mathcal{Y} of the intervention $d \rightarrow d^*$ to \mathcal{D} given $X = x$* is

$$\Delta\rho(d, d^*, x) \equiv \rho(d^*, x) - \rho(d, x).$$

We call this a “covariate-conditioned average effect of intervention,” or, leaving conditioning implicit, an “average effect of intervention.” As a special case, this includes the conditional average treatment effect of a binary treatment introduced by Abadie and Imbens (2002). Formally, the interventions (ω, ω^*) underlying $d \rightarrow d^*$ given $X = x$ are those satisfying $d = D(\omega)$, $x = X(\omega)$ and $d^* = D(\omega^*)$, $x = X(\omega^*)$. With structural identification, when $(d^*, x), (d, x) \in \text{supp}(D, X)$, we have

$$\Delta\rho(d, d^*, x) = \Delta\mu(d, d^*, x) \equiv \mu(d^*, x) - \mu(d, x),$$

so if $\Delta\mu(d, d^*, x)$ can be consistently estimated, so can $\Delta\rho(d, d^*, x)$.

Replacing x with X yields a random version of the average effect, $\Delta\rho(d, d^*, X)$, with an optimal prediction property. Specifically, by the mean-square optimality of conditional expectation, $\Delta\rho(d, d^*, X)$ is the mean-squared-error best predictor of the random effect $\Delta r(d, d^*, Z, A) \equiv r(d^*, Z, A) - r(d, Z, A)$ among all predictors based on X .

Structural identification has fundamental implications for the interpretation of regression coefficients. Consider the familiar linear regression, $E(Y | D, X) = D\beta^* + X'\gamma^*$, with D a scalar binary indicator (dummy). Without structural identification, β^* and γ^* have only a predictive interpretation: they deliver mean squared error-optimal predictions of Y given D and X (e.g., White, 1980). Given structural identification, however, β^* acquires causal meaning. Specifically, let $d = 0$ and $d^* = 1$. Then for all x , $\Delta\mu(d, d^*, x) = \beta^*$; so structural identification ensures $\Delta\rho(d, d^*, x) = \beta^*$.

This significantly refines the usual *ceteris paribus* interpretation for regression coefficients. Specifically, β^* is the covariate-conditioned average effect on the response of an intervention $d = 0 \rightarrow d^* = 1$, *averaging over* unobserved \ddot{X} , *conditional on* observed X . The unobserved \ddot{X} is not “held constant,” but is averaged over; the observed covariates are not “held constant,” but are conditioned on. These distinctions are important: “holding constant” is a counterfactual operation associated with interventions, whereas averaging and conditioning are stochastic operations. Further, the remaining coefficients γ^* do not have any necessary causal interpretation; they possess only their predictive interpretation.

Next, let $D = (D_1, D_2)$, and suppose $E(Y | D, X) = D_1\beta_1^* + D_2\beta_2^* + X'\gamma^*$. Now let $d = (0, 0)$ and $d^* = (1, 0)$, so that intervention changes only d_1 , holding d_2 constant at 0. Under structural identification, for all x , $\Delta\rho(d, d^*, x) = \Delta\mu(d, d^*, x) = \beta_1^*$. This is the expected effect of an intervention to $d_1(0 \rightarrow 1)$ holding d_2 constant (at 0), averaging over the unobserved explanatory variables \ddot{X} conditional on the observed covariates. Now one of the causes of interest (D_2) is held constant. Here β_1^* and β_2^* (corresponding to conditionally exogenous variables) have causal interpretations, but γ^* (corresponding to covariates) has only a predictive interpretation.

These interpretive considerations hold generally. In linear regression, this means some coefficients may have causal meaning and others may not. Recognizing this could have far-reaching implications for the way researchers assess the plausibility of empirical results. Similarly, in more general contexts, certain differences or derivatives may have causal meaning and others may not. Which these are is determined by structural identification.

3.2.2 Covariate-Conditioned Average Marginal Effects

The *average marginal ceteris paribus effect on \mathcal{Y} of \mathcal{D}_j given $(D, X) = (d, x)$* , is

$$\xi_j(d, x) \equiv \int \mathbf{D}_j r(d, \tilde{x}, \ddot{x}) dG(\ddot{x} | x),$$

where $\mathbf{D}_j \equiv (\partial/\partial d_j)$, provided the indicated derivative and integral exist. This is related to the average derivatives of Stoker (1986) and Powell, Stock, and Stoker (1989). It is a weighted average of the unobservable marginal effect $\mathbf{D}_j r(d, z, a)$, averaging over unobserved causes and attributes, given observed covariates. We call $\xi_j(d, x)$ a “covariate-conditioned average marginal effect,” or just an “average marginal effect.”

The next condition permits interchange of derivative and integral for ρ to structurally identify average marginal effects. By $d_{(j)}$ we denote the $(k_1 - 1) \times 1$ sub-vector of d containing all but d_j , $j \in \{1, \dots, k_1\}$.

Assumption A.3 For given $(d, x) \in \text{supp}(D, X)$, suppose the function $\ddot{x} \rightarrow r(d, \tilde{x}, \ddot{x})$

is integrable with respect to $G(\cdot | x)$, that is, $\int r(d, \tilde{x}, \ddot{x}) dG(\ddot{x} | x) < \infty$, and suppose that for the given $(d_{(j)}, \tilde{x})$, $(d_j, \ddot{x}) \rightarrow \mathbf{D}_j r(d, \tilde{x}, \ddot{x})$ exists on $C_j \times \text{supp}(\ddot{X} | \tilde{x})$, where C_j is a convex compact neighborhood of the given d_j , and $\text{supp}(\ddot{X} | \tilde{x})$ is the support of \ddot{X} given $\tilde{X} = \tilde{x}$. Suppose further that for the given $(d_{(j)}, \tilde{x})$ and for each \ddot{x} in $\text{supp}(\ddot{X} | \tilde{x})$,

$$\sup_{d_j \in C_j} |\mathbf{D}_j r(d, \tilde{x}, \ddot{x})| \leq q(d_{(j)}, \tilde{x}, \ddot{x}),$$

where q is a measurable function such that $E(q(D_{(j)}, \tilde{X}, \ddot{X})) < \infty$ ■

When A.3 holds for r and $\mathbf{D}_j r$, we say “ $\mathbf{D}_j r(d, \tilde{x}, \ddot{x})$ is dominated on C_j by an integrable function.” Our next structural identification result is a continuation of Theorem 3.3:

Theorem 3.3(ii) If, in addition to the conditions of Theorem 3.3(i), Assumption A.3 holds, then the functions $d_j \rightarrow \rho(d, x)$ and $d_j \rightarrow \mu(d, x)$ are differentiable on C_j , and $\mathbf{D}_j \mu(d, x) = \mathbf{D}_j \rho(d, x) = \xi_j(d, x) = \int \mathbf{D}_j r(d, \tilde{x}, \ddot{x}) dG(\ddot{x} | d, x)$. ■

This identifies an aspect of the causal structure, ξ_j , with an aspect of the stochastic structure, $\mathbf{D}_j \mu$. To consistently estimate ξ_j , it thus suffices to consistently estimate $\mathbf{D}_j \mu$. White and Chalak (2006) discuss this estimation.

Average marginal effects are identified only for elements of the conditionally exogenous causes D . Derivatives of μ with respect to elements of X have no causal interpretation.

4 Structural Identification For General Measures of Effect

Interest also attaches to effects of interventions on aspects of the conditional response distribution other than the mean. Heckman, Smith, and Clements (1997) draw attention to this issue in the context of programme evaluation. Imbens and Newey (2003) discuss a variety of such effects. For wage determination, Firpo, Fortin, and Lemieux (2005) study effects of binary treatments on aspects of the unconditional response distribution, such as the variance, median, or density. Here we discuss identifying causal effects of general interventions on general aspects of the conditional response distribution.

4.1 Three Ways to Define General Effects

One approach uses the *covariate-conditioned counterfactual moment*

$$\rho_0(d, x) \equiv \tau_0(\rho_1(d, x), \rho_2(d, x), \dots),$$

where τ_0 is a known function, and for known scalar-valued functions τ_k ,

$$\rho_k(d, x) \equiv \int \tau_k(r(d, \tilde{x}, \ddot{x})) dG(\ddot{x} | x), \quad k = 1, 2, \dots$$

The *moment effect* on \mathcal{Y} of the intervention $d \rightarrow d^*$ to \mathcal{D} given $X = x$ is

$$\Delta\rho_0(d, d^*, x) \equiv \rho_0(d^*, x) - \rho_0(d, x),$$

and the *marginal moment effect* on \mathcal{Y} of \mathcal{D}_j given $X = x$ is $D_j\rho_0(d, x)$.

For example, let $\tau_1(r) = 1[r \leq y]$ for $y \in \mathbb{R}$ (cf. Imbens, 2004, p.9), and let $\tau_0(\rho_1) = \rho_1$. Then effects on the conditional response distribution are defined from the counterfactual conditional distribution function

$$\rho_0(d, x) = \int 1[r(d, \tilde{x}, \ddot{x}) \leq y] dG(\ddot{x} | x).$$

Or let $\tau_1(r) = r$, $\tau_2(r) = r^2$, and put $\rho_0(d, x) = \rho_2(d, x) - \rho_1(d, x)^2$. This defines the covariate-conditioned counterfactual variance, yielding conditional variance effects.

When conditional exogeneity holds, the counterfactual moment function and the corresponding effects are structurally identified.

Theorem 4.1 Suppose Assumptions A.1(a, b) hold. For $k = 1, 2, \dots$, let $\tau_k : \mathbb{R} \rightarrow \mathbb{R}$ be a known measurable function such that $E(\tau_k(Y)) < \infty$. (i) Then $\mu_k(D, X) \equiv E(\tau_k(Y) | D, X)$ exists and is finite, and for each (d, x) in $\text{supp}(D, X)$

$$\mu_k(d, x) = \int \tau_k(r(d, \tilde{x}, \ddot{x})) dG(\ddot{x} | d, x), \quad k = 1, 2, \dots \quad .$$

If $\tau_0 : \mathbb{R}^\infty \rightarrow \mathbb{R}$ is a known measurable function, then the function μ_0 defined by $\mu_0(d, x) \equiv \tau_0(\mu_1(d, x), \mu_2(d, x), \dots)$ is also measurable. (ii) If A.1(c.i) and A.2 or A.2' also hold, then for each (d, x) in $\text{supp}(D, X)$

$$\rho_k(d, x) = \int \tau_k(r(d, \tilde{x}, \ddot{x})) dG(\ddot{x} | x)$$

exists and is finite, and $\rho_k = \mu_k$, $k = 1, 2, \dots$; the function ρ_0 defined by $\rho_0(d, x) = \tau_0(\rho_1(d, x), \rho_2(d, x), \dots)$ is measurable; and $\rho_0 = \mu_0$. ■

General effects also arise from the covariate-conditioned counterfactual optimizer

$$\rho_0(d, x) \equiv \arg \max_m \int \tau(r(d, \tilde{x}, \ddot{x}), m) dG(\ddot{x} | x),$$

where $\tau : \mathbb{R} \times \mathbb{R}^\lambda \rightarrow \mathbb{R}$ is known, so that $\rho_0(d, x)$ is a $\lambda \times 1$ vector of aspects of the counterfactual conditional distribution. For example, effects on the conditional α -quantiles

of the response arise from

$$\tau(r, m) = -|r - m|(\alpha 1[r \geq m] + (1 - \alpha)1[r < m]).$$

Now $\rho_0(d, x)$ defines the covariate-conditioned counterfactual α -quantile function, a conditional analog of the “quantile structural function” of Imbens and Newey (2003). The associated effect is the covariate-conditioned analog of the quantile treatment effect of Lehmann (1974) and Abadie, Angrist, and Imbens (2002).

Taking m to be a vector and $\tau(r, m)$ to define a quasi-log-likelihood function permits the optimization approach to focus attention simultaneously on multiple aspects of the counterfactual conditional response distribution, such as location and scale. Taking $\tau(r, m)$ to define a utility function yields effects on optimal actions of interventions to d conditional on $X = x$.

Theorem 4.2 Suppose Assumptions A.1(a, b) hold. (i) For $\lambda \in \mathbb{N}$, let $\tau : \mathbb{R} \times \mathbb{R}^\lambda \rightarrow \mathbb{R}$ be a known measurable function such that $E(\tau(Y, m)) < \infty$ for each m in \mathbb{R}^λ . Then for each (d, x, m) in $\text{supp}(D, X) \times \mathbb{R}^\lambda$ the conditional expectation

$$\varphi_{\tau,d}(d, x, m) \equiv \int \tau(r(d, \tilde{x}, \ddot{x}), m) dG(\ddot{x} | d, x)$$

exists and is finite. (ii) Further, let τ, r , and $(d, x) \rightarrow G(\cdot | d, x)$ be such that $\varphi_{\tau,d}(d, x, m)$ defines a continuous real-valued function on $\text{supp}(D, X) \times \mathbb{R}^\lambda$, and let $M: \text{supp}(D, X) \rightarrow \mathbb{R}^\lambda$ be a non-empty and compact-valued continuous correspondence. Then for each (d, x) in $\text{supp}(D, X)$ the correspondence

$$\mu_0(d, x) = \arg \max_{m \in M(d, x)} \varphi_{\tau,d}(d, x, m)$$

is non-empty, compact-valued, and upper hemi-continuous. (iii) If A.1(c.i) and A.2 or A.2' hold also, then

$$\varphi_\tau(d, x, m) \equiv \int \tau(r(d, \tilde{x}, \ddot{x}), m) dG(\ddot{x} | x)$$

defines a continuous real-valued function on $\text{supp}(D, X) \times \mathbb{R}^\lambda$ such that for each (d, x, m) in $\text{supp}(D, X) \times \mathbb{R}^\lambda$ we have $\varphi_\tau(d, x, m) = \varphi_{\tau,d}(d, x, m)$; the correspondence

$$\rho_0(d, x) = \arg \max_{m \in M(d, x)} \varphi_\tau(d, x, m)$$

is non-empty, compact-valued, and upper hemi-continuous; and $\rho_0 = \mu_0$. ■

A third way to define effects uses implicitly defined moments $\rho_0(d, x)$ such that

$$\int \tau(r(d, \tilde{x}, \ddot{x}), \rho_0(d, x)) dG(\ddot{x} | x) = 0,$$

where $\tau : \mathbb{R} \times \mathbb{R}^\lambda \rightarrow \mathbb{R}^\lambda$ is known. This method has not been previously studied to the best of our knowledge, although it contains many instances of the moment or optimizer approaches as special cases. For example, this $\rho_0(d, x)$ can represent first order conditions defining the interior optimizer of some objective function. The implicit moment approach generalizes and complements the optimizer approach in the same way that method of moments estimation generalizes and complements maximum likelihood estimation.

Theorem 4.3 Suppose Assumptions A.1(a, b) hold. (i) For $\lambda \in \mathbb{N}$, let $\tau : \mathbb{R} \times \mathbb{R}^\lambda \rightarrow \mathbb{R}^\lambda$ be a measurable function such that $E(\tau(Y, m)) < \infty$ for each $m \in M \subset \mathbb{R}^\lambda$. Then for each (d, x, m) in $\text{supp}(D, X) \times M$ the conditional expectation

$$\psi_{\tau, d}(d, x, m) \equiv \int \tau(r(d, \tilde{x}, \ddot{x}), m) dG(\ddot{x} | d, x)$$

exists and is finite. (ii) Further, let τ, r , and $(d, x) \rightarrow G(\cdot | d, x)$ be such that for each (d, x, m) in $\text{supp}(D, X) \times M$, $\psi_{\tau, d}$ is differentiable on a neighborhood of (d, x, m) , the $\lambda \times \lambda$ matrix $\nabla_m \psi_{\tau, d}(d, x, m)$ is non-singular, and $\psi_{\tau, d}(d, x, m) = 0$. Then there exists a unique function μ_0 such that for each $(d, x) \in \text{supp}(D, X)$, μ_0 is differentiable at (d, x) , and

$$\int \tau(r(d, \tilde{x}, \ddot{x}), \mu_0(d, x)) dG(\ddot{x} | d, x) = 0.$$

(iii) If A.1(c.i) and A.2 or A.2' hold also, then there exists a unique function ρ_0 such that for each $(d, x) \in \text{supp}(D, X)$, ρ_0 is differentiable at (d, x) ;

$$\int \tau(r(d, \tilde{x}, \ddot{x}), \rho_0(d, x)) dG(\ddot{x} | x) = 0;$$

and $\rho_0 = \mu_0$. ■

In each case, for (d, x) and (d^*, x) in $\text{supp}(D, X)$ and given the covariate-conditioned counterfactual distribution aspect ρ , the ρ -effect of the intervention $d^* \rightarrow d$ to \mathcal{D} given $X = x$ is

$$\Delta\rho(d, d^*, x) \equiv \rho(d^*, x) - \rho(d, x).$$

Our results identify these in terms of the corresponding μ as

$$\Delta\rho(d, d^*, x) = \Delta\mu(d, d^*, x) \equiv \mu(d^*, x) - \mu(d, x).$$

The value x need not be factual, so there is an allowed counterfactual aspect to conditioning. Comparing the effects of an intervention $d^* \rightarrow d$ for different values x and x^* gives an *average ρ -effect difference* $\Delta\rho(d, d^*, x^*) - \Delta\rho(d, d^*, x)$, measuring how the expected effect changes with a “change” in the covariate outcome. The covariates are thus accurately called “effect modifiers” in the epidemiological literature.

In Section 3.2.2, we studied average marginal effects. We defer treating general marginal effects to Section 5, where we study weaker conditions for identification.

4.2 Unconditional Effects

Covariate-conditioned effects can be used to construct unconditional effects. Let F be any distribution with support on a subset of $\text{supp}(X)$. The unconditional($-F$) ρ -effect mean is

$$\mu_1(d, d^*; \Delta\rho, F) \equiv \int \Delta\rho(d, d^*, x) dF(x).$$

When conditional exogeneity structurally identifies $\Delta\rho$, it also identifies μ_1 . For example, consider a single binary treatment, with $\Delta\rho(0, 1, x)$ the covariate-conditioned average effect of treatment. Let $F = F_1$, the covariate distribution for the treated observations ($D = 1$). Then $\mu_1(0, 1; \Delta\rho, F_1)$ is the familiar average effect of treatment on the treated (e.g., Rubin, 1974). Similarly, the unconditional($-F$) marginal ρ -effect mean is

$$\mu_1(d; D_j\rho, F) \equiv \int D_j\rho(d, x) dF(x).$$

Other descriptors of the distribution of ρ -effects can be straightforwardly defined using other unconditional($-F$) ρ -effect moments, such as

$$\mu_k(d, d^*; \Delta\rho, F) \equiv \int \Delta\rho(d, d^*, x)^k dF(x) \quad \text{or}$$

$$\mu_k(d; D_j\rho, F) \equiv \int D_j\rho(d, x)^k dF(x).$$

For brevity, we leave aside further discussion of these possibilities.

4.3 Implications for Estimation

Given structural identification of any counterfactual aspect of the conditional response distribution it suffices to estimate the corresponding standard aspect, as estimates of effects follow by taking suitable differences or derivatives of the estimated aspect. White (2006) and White and Chalak (2006) study estimation for settable systems. Nevertheless, to contrast structural identification with *stochastic identification* we now sketch the construction of estimators under structural identification.

First, consider the covariate-conditioned optimizer of Theorem 4.2,

$$\mu_0(d, x) \equiv \arg \max_m \int \tau(r(d, \tilde{x}, \ddot{x}), m) \, dG(\ddot{x} \mid d, x).$$

Given structural identification, $\rho_0 = \mu_0$ so it suffices to estimate μ_0 . For this, parameterize the argument m by specifying a function of parameters θ , say $\theta \rightarrow m(d, x, \theta)$, such that $m(d, x, \theta^*) = \mu_0(d, x)$ for all (d, x) in $\text{supp}(D, X)$, so that θ^* solves

$$\max_{\theta \in \Theta} \int \tau(r(d, \tilde{x}, \ddot{x}), m(d, x, \theta)) \, dG(\ddot{x} \mid d, x),$$

where Θ is an appropriate finite or infinite dimensional parameter space. When θ^* exists and is unique, we say that θ^* is *stochastically identified*. Structural identification is neither necessary nor sufficient for this. For example, in linear regression stochastic identification fails in the presence of exact multicollinearity.

Given a sample of n observations (Y_i, D_i, X_i) , an estimator $\hat{\theta}_n$ of θ^* solves

$$\max_{\theta \in \Theta_n} n^{-1} \sum_{i=1}^n \tau(Y_i, m(D_i, X_i, \theta)),$$

where $\{\Theta_n\}$ is a suitable sequence of subsets of Θ . If $\Theta_n = \Theta$ and Θ is finite dimensional, the method is parametric. Nonparametric methods arise when Θ is infinite dimensional. For example, apply the method of sieves (Grenander, 1981; Chen, 2005).

Next, consider the covariate-conditioned implicit moment μ_0 such that

$$\int \tau(r(d, \tilde{x}, \ddot{x}), \mu_0(d, x)) \, dG(\ddot{x} \mid d, x) = 0.$$

Given structural identification, $\rho_0 = \mu_0$. Now parameterize the solution μ_0 , specifying a parameter space Θ and a function $\theta \rightarrow m(d, x, \theta)$, such that $m(d, x, \theta^*) = \mu_0(d, x)$ for all (d, x) in $\text{supp}(D, X)$, so θ^* satisfies

$$\int \tau(r(d, \tilde{x}, \ddot{x}), m(d, x, \theta^*)) \, dG(\ddot{x} \mid d, x) = 0.$$

That is, θ^* solves the implicit moment conditions $E(\tau(Y, m(D, X, \theta^*)) \mid D, X) = 0$. Stochastic identification holds when θ^* uniquely solves these. To estimate θ^* , apply parametric or non-parametric versions of the methods of moments (Hansen, 1982; Ai and Chen, 2003) or empirical likelihood methods (e.g., Shennach, 2004; Ragusa, 2005).

5 Structural Identification Without Conditional Exogeneity

We now study identification without conditional exogeneity. This yields conditions ensuring identification at specific values (d, x) , i.e., locally. In some cases, we obtain necessary and sufficient conditions. We also obtain “near identification” results.

5.1 Explicit Moment Effects

First consider the relationship between ρ_k and μ_k of Theorem 4.1 without A.2.

Theorem 5.1 Suppose that A.1(a) holds and let

$$s(d, x, \ddot{x}) \equiv 1 - dG(\ddot{x} | x) / dG(\ddot{x} | d, x) = 1 - dG(d | x) / dG(d | \ddot{x}, x).$$

(i) Then for all $(d, x) \in \text{supp}(D, X)$, $\int s(d, x, \ddot{x}) dG(\ddot{x} | d, x) = 0$; (ii) Further, let A.1(c.i) and the remaining conditions of Theorem 4.1(i) hold, and suppose that $E(s(D, X, \ddot{X})^2) < \infty$ and $E(\tau_k(Y)^2) < \infty$, $k = 1, 2, \dots$. Then for all $(d, x) \in \text{supp}(D, X)$, $\rho_k(d, x)$ as defined in Theorem 4.1(ii) exists and is finite, and

$$\mu_k(d, x) = \rho_k(d, x) + \gamma_k(d, x),$$

where

$$\gamma_k(d, x) \equiv \int \tau_k(r(d, \tilde{x}, \ddot{x})) s(d, x, \ddot{x}) dG(\ddot{x} | d, x), \quad k = 1, 2, \dots$$

(iii) Further, for all $(d, x) \in \text{supp}(D, X)$,

$$|\gamma_k(d, x)| \leq \sigma(d, x; \tau_k) \sigma(d, x; s), \quad k = 1, 2, \dots,$$

where $\sigma(d, x; \tau_k) \equiv [\text{var}(\tau_k(Y) | (D, X) = (d, x))]^{1/2}$ and $\sigma(d, x; s) \equiv [\text{var}(s(D, X, \ddot{X}) | (D, X) = (d, x))]^{1/2}$. ■

The *discrepancy score* $s(d, x, \ddot{x})$ measures the relative departure from conditional independence at (d, x, \ddot{x}) . By (i), the discrepancy score has conditional mean zero.

By (ii), $\mu_k(d, x)$ differs from $\rho_k(d, x)$ by the *moment discrepancy* $\gamma_k(d, x)$, which, given (i), is the conditional covariance of $\tau_k(Y)$ and $s(D, X, \ddot{X})$. Thus, $\gamma_k(d, x) = 0$ is necessary and sufficient for structural identification of $\rho_k(d, x)$. This is a *local identification* result, specific to a particular (d, x) . Conditional exogeneity is sufficient for this, but not necessary. It suffices that $s(d, x, \tilde{x}) = 0$ for all $\tilde{x} \in \text{supp}(\ddot{x} | (d, x))$. Matzkin (2004) gives related results involving local identification under some monotonicity assumptions.

The Cauchy-Schwarz inequality gives (iii), bounding the moment discrepancy and establishing continuity with respect to (i) local dependence of $\tau_k(Y)$ on unobservables,

measured by $\sigma(d, x; \tau_k)$; and (ii) local conditional dependence, measured by $\sigma(d, x; s)$. If either is small, then so is the moment discrepancy. Theorem 5.1(iii) is thus a near identification result. The bound is best possible in that equality is attained when $|\tau_k(r(d, \tilde{x}, \ddot{x}))| = |s(d, x, \ddot{x})|$ for given (d, x) and all \ddot{x} in $\text{supp}(\ddot{X} | (D, X) = (d, x))$. Similar results follow by applying the Hölder inequality in place of Cauchy-Schwarz.

Theorem 4.1 treats $\mu_0(d, x) = \tau_0(\mu_1(d, x), \mu_2(d, x), \dots)$ and $\rho_0(d, x) = \tau_0(\rho_1(d, x), \rho_2(d, x), \dots)$. The *general moment discrepancy* is

$$\gamma_0(d, x) \equiv \mu_0(d, x) - \rho_0(d, x).$$

If τ_0 is affine in its arguments (e.g., the covariate-conditioned average response),

$$\gamma_0(d, x) = \tau_0(\gamma_1(d, x), \gamma_2(d, x), \dots).$$

Theorem 5.1(ii) then implies that the “apparent effect” $\Delta\mu_0(d, d^*, x)$ is contaminated by the *effect discrepancy*

$$\Delta\gamma_0(d, d^*, x) = \tau_0(\Delta\gamma_1(d, d^*, x), \Delta\gamma_2(d, d^*, x), \dots),$$

where $\Delta\gamma_k(d, d^*, x) \equiv \gamma_k(d^*, x) - \gamma_k(d, x)$, $k = 1, 2, \dots$

Even if τ_0 is not affine, γ_0 depends globally and smoothly on the γ_k 's, under plausible conditions. For brevity, let τ_0 depend continuously on $\mu \equiv (\mu_1, \dots, \mu_\kappa)'$ taking values in a compact set, K . Then

$$\tau_0(\mu) = \sum_{i=1}^{\infty} a_i \cos(\mu' \theta_i) + \sum_{i=1}^{\infty} b_i \sin(\mu' \theta_i),$$

gives the Fourier series representation, where a_i 's and b_i 's are Fourier coefficients, θ_i 's are appropriate multi-frequencies, and equality is in the sense of uniform convergence. Then

$$\tau_0(\mu) - \tau_0(\rho) = \left[\sum_{i=1}^{\infty} a_i \cos(\mu' \theta_i) - \sum_{i=1}^{\infty} a_i \cos(\rho' \theta_i) \right] + \left[\sum_{i=1}^{\infty} b_i \sin(\mu' \theta_i) - \sum_{i=1}^{\infty} b_i \sin(\rho' \theta_i) \right]$$

for any $\mu, \rho \in K$. Standard trigonometric identities give

$$\begin{aligned} \cos(u) - \cos(v) &= 2 \sin(u) \cos([v - u]/2) \sin([v - u]/2) + 2 \cos(u) \sin^2([v - u]/2) \\ \sin(u) - \sin(v) &= -2 \cos(u) \cos([v - u]/2) \sin([v - u]/2) + 2 \sin(u) \sin^2([v - u]/2). \end{aligned}$$

Letting $\gamma \equiv \mu - \rho$ and substituting into $\tau_0(\mu) - \tau_0(\rho)$ then gives

$$\begin{aligned}\gamma_0(\mu, \gamma) &= 2 \sum_{i=1}^{\infty} [a_i \sin(\mu' \theta_i) - b_i \cos(\mu' \theta_i)] \cos(\gamma' \theta_i / 2) \sin(\gamma' \theta_i / 2) \\ &\quad + 2 \sum_{i=1}^{\infty} [a_i \cos(\mu' \theta_i) + b_i \sin(\mu' \theta_i)] \sin^2(\gamma' \theta_i / 2).\end{aligned}$$

The general moment discrepancy γ_0 thus depends globally and smoothly on γ . Theorem 5.1(iii) then ensures that the effect discrepancy $\Delta\gamma_0$ inherits continuity with respect to both conditional dependence and local dependence of the response on unobservables.

Significantly, this implies that *neglecting proxies for unobservables that have minor relevance for determining either the response or the cause of interest leads to correspondingly minor distortions in the apparent effect*. Thus, priority should be given to including proxies for variables most relevant to determining the response and causes of interest.

Marginal effects are similarly affected when A.2 fails. We add further structure, gaining analytic convenience without losing much generality. We now require that possible values for \ddot{X} do not depend on the realization of D , though they may depend on that of X . This permits conditional dependence, as the probabilities associated with these possible values can depend on the realization of D .

Recall that a σ -finite measure η is absolutely continuous with respect to a σ -finite measure ν , written $\eta \ll \nu$, if $\eta(\mathbf{B}) = 0$ for every measurable set \mathbf{B} such that $\nu(\mathbf{B}) = 0$. If $\eta \ll \nu$, we say ν *dominates* η and call ν a “dominating measure.” The Radon-Nikodym theorem states that if $\eta \ll \nu$, then there exists a positive measurable function $f = d\eta/d\nu$, the *Radon-Nikodym density*, such that $\eta(\mathbf{A}) = \int_{\mathbf{A}} f d\nu$ for every measurable set \mathbf{A} .

Assumption A.4 For each $x \in \text{supp } X$, there exists a σ -finite measure $\nu(\cdot | x)$ such that (a) for each $(d, x) \in \text{supp}(D, X)$, the measure $G(\mathbf{B} | d, x) = \int_{\mathbf{B}} dG(\ddot{x} | d, x)$ is absolutely continuous with respect to $\nu(\cdot | x)$; (b) for each $x \in \text{supp } (X)$, the measure $G(\mathbf{B} | x) = \int_{\mathbf{B}} dG(\ddot{x} | x)$ is absolutely continuous with respect to $\nu(\cdot | x)$. ■

By Radon-Nikodym, there exist conditional densities, say $g(\ddot{x} | d, x)$ and $g(\ddot{x} | x)$ such that $dG(\ddot{x} | d, x) = g(\ddot{x} | d, x) d\nu(\ddot{x} | x)$ and $dG(\ddot{x} | x) = g(\ddot{x} | x) d\nu(\ddot{x} | x)$. Conditional dependence arises whenever $g(\ddot{x} | d, x)$ depends non-trivially on d .

Next, we impose differentiability and domination conditions.

Assumption A.5 For given j , $D_j g(\ddot{x} | d, x)$ is dominated on C_j by a function integrable with respect to $\nu(\cdot | x)$ at (d, x) . ■

We also impose the analog of A.3:

Assumption A.6 For given j and $k = 1, 2, \dots$, $(\mathbf{D}_j[(\tau_k \circ r)g])(d, x, \ddot{x})$ is dominated on C_j by a function integrable with respect to $\nu(\cdot | x)$ at (d, x) . ■

The differentiability of g implicit in A.5 and differentiability of $\tau_k \circ r$ with respect to d_j implicit in A.6 ensure existence of the product derivative $\mathbf{D}_j[(\tau_k \circ r)g]$.

Theorem 5.2 Suppose Assumptions A.1(a) and A.4 hold and let $s(d, x, \ddot{x}) \equiv 1 - g(\ddot{x} | x) / g(\ddot{x} | d, x) = 1 - g(d | x) / g(d | \ddot{x}, x)$ (i.a) Then for all $(d, x) \in \text{supp}(D, X)$ $\int s(d, x, \ddot{x}) g(\ddot{x} | d, x) d\nu(\ddot{x} | x) = 0$. (i.b) If A.5 also holds, then for the given (d, x)

$$\int \mathbf{D}_j \log g(\ddot{x} | d, x) g(\ddot{x} | d, x) d\nu(\ddot{x} | x) = 0.$$

(ii) Further, let A.1(c.i), A.6, and the remaining conditions of Theorem 4.1(i) hold. Then the functions $d_j \rightarrow \mu_k(d, x)$, $k = 1, 2, \dots$, are differentiable on C_j and

$$\begin{aligned} \mathbf{D}_j \mu_k(d, x) &= \int \mathbf{D}_j \tau_k(r(d, \tilde{x}, \ddot{x})) g(\ddot{x} | d, x) d\nu(\ddot{x} | x) \\ &+ \int \tau_k(r(d, \tilde{x}, \ddot{x})) \mathbf{D}_j \log g(\ddot{x} | d, x) g(\ddot{x} | d, x) d\nu(\ddot{x} | x), \end{aligned}$$

(iii) If, in addition for $k = 1, 2, \dots$ $E([\mathbf{D}_j \tau_k(r(D, \tilde{X}, \ddot{X}))]^2) < \infty$ and $E(s(D, X, \ddot{X})^2) < \infty$, then for $k = 1, 2, \dots$

$$\mathbf{D}_j \mu_k(d, x) = \xi_{k,j}(d, x) + \delta_{1,k,j}(d, x) + \delta_{2,k,j}(d, x),$$

where

$$\begin{aligned} \xi_{k,j}(d, x) &\equiv \int \mathbf{D}_j \tau_k(r(d, \tilde{x}, \ddot{x})) g(\ddot{x} | x) d\nu(\ddot{x} | x) \\ \delta_{1,k,j}(d, x) &\equiv \int \mathbf{D}_j \tau_k(r(d, \tilde{x}, \ddot{x})) s(d, x, \ddot{x}) g(\ddot{x} | d, x) d\nu(\ddot{x} | x) \\ \delta_{2,k,j}(d, x) &\equiv \int \tau_k(r(d, \tilde{x}, \ddot{x})) \mathbf{D}_j \log g(\ddot{x} | d, \ddot{x}) g(\ddot{x} | d, x) d\nu(\ddot{x} | x). \end{aligned}$$

(iv)(a) Letting $\sigma(d, x; \mathbf{D}_j(\tau_k \circ r)) \equiv [\text{var}(\mathbf{D}_j \tau_k(r(D, \tilde{X}, \ddot{X})) | (D, X) = (d, x))]^{1/2}$,

$$|\delta_{1,k,j}(d, x)| \leq \sigma(d, x; \mathbf{D}_j(\tau_k \circ r)) \sigma(d, x; s).$$

(b) If in addition $E([\mathbf{D}_j \log g(\ddot{X} | D, X)]^2) < \infty$ and $E(\tau_k(Y)^2) < \infty$, $k = 1, 2, \dots$, then

$$|\delta_{2,k,j}(d, x)| \leq \sigma(d, x; \tau_k) \sigma(d, x; \mathbf{D}_j \log g_d)$$

where $\sigma(d, x; \mathbb{D}_j \log g_d) \equiv [\int \{\mathbb{D}_j \log g(\ddot{x} \mid d, x)\}^2 g(\ddot{x} \mid d, x) d\nu(\ddot{x} \mid x)]^{1/2}$. ■

Thus, $\mathbb{D}_j \mu_k$ is a contaminated version of $\xi_{k,j}$, the covariate-conditioned average marginal τ_k moment effect. The causal discrepancy is $\delta_{k,j} \equiv \delta_{1,k,j} + \delta_{2,k,j}$. Conditional exogeneity is sufficient but not necessary for this to vanish. The causal discrepancy $\delta_{1,k,j}(d, x)$ vanishes if the discrepancy score $s(d, x, \ddot{x})$ vanishes for all \ddot{x} . The causal discrepancy $\delta_{2,k,j}(d, x)$ vanishes if the *marginal discrepancy score* $\mathbb{D}_j \log g(\ddot{x} \mid d, x)$ vanishes for all \ddot{x} . Result (iv) bounds the moment discrepancies. We thus have local identification and near identification results analogous to those of Theorem 5.1.

Results relating to $\mathbb{D}_j \mu_0$ to $\mathbb{D}_j \rho_0$ follow straightforwardly. For brevity, let τ_0 depend on a $\kappa \times 1$ vector, and write $\mu \equiv (\mu_1, \dots, \mu_\kappa)'$ and $\rho \equiv (\rho_1, \dots, \rho_\kappa)'$. The chain rule gives $\mathbb{D}_j \mu_0 = \nabla' \tau_0(\mu)(\mathbb{D}_j \mu)$, $\mathbb{D}_j \rho_0 = \nabla' \tau_0(\rho)(\mathbb{D}_j \rho)$, where $\nabla \tau_0$ is the $\kappa \times 1$ gradient vector of τ_0 with respect to its arguments, and $\mathbb{D}_j \mu$ and $\mathbb{D}_j \rho$ are $\kappa \times 1$ vectors (\mathbb{D}_j operates element by element). Adding and subtracting appropriately gives

$$\begin{aligned} \mathbb{D}_j \mu_0 - \mathbb{D}_j \rho_0 &= \nabla' \tau_0(\mu)[\mathbb{D}_j \mu - \mathbb{D}_j \rho] \\ &+ [\nabla' \tau_0(\mu) - \nabla' \tau_0(\rho)] (\mathbb{D}_j \mu) - [\nabla' \tau_0(\mu) - \nabla' \tau_0(\rho)][\mathbb{D}_j \mu - \mathbb{D}_j \rho] \\ &= \nabla' \tau_0(\mu) \delta_j + \nabla' \gamma_0(\mu, \gamma)(\mathbb{D}_j \mu) - \nabla' \gamma_0(\mu, \gamma) \delta_j, \\ &\equiv \delta_0(\mu, \gamma, \delta_j), \end{aligned}$$

where δ_j is the vector with elements $\delta_{k,j}$ and $\nabla' \gamma_0(\mu, \gamma) = \nabla' \tau_0(\mu) - \nabla' \tau_0(\rho)$ holds with $\gamma \equiv \mu - \rho$ and smoothness assumptions on τ_0 sufficient for the Fourier series approximation above to hold in a suitable Sobolev norm. The *marginal general moment effect discrepancy* $\delta_0(\mu, \gamma, \delta_j)$ can vanish for specific values of (d, x) under special circumstances. It vanishes for all (d, x) under conditional exogeneity.

5.2 Implicit Moment Effects

We subsume optimizer-based distributional aspect effects into the study of implicit moment effects, viewing optimizer-based distributional aspects as implicit moments defined by the first order conditions of the underlying optimization.

The implicit nonlinear definitions of μ_0 and ρ_0 present significant challenges to directly obtaining a tractable representation for the implicit moment discrepancy $\gamma_0 \equiv \mu_0 - \rho_0$, so for brevity we do not treat this here. Nevertheless, implicitly defined moments can often be well approximated by the explicit moments analyzed in Theorems 4.1 and 5.1.

An analysis of marginal effects analogous to Theorem 5.2 is more straightforward. To

succinctly state our next result, we now write

$$\int \tau_\mu g_d \, d\nu = \int \tau(r(d, \tilde{x}, \ddot{x}), \mu_0(d, x)) g(\ddot{x} \mid d, x) \, d\nu(\ddot{x} \mid x).$$

When the integral of Theorem 4.3(iii) exists, we write

$$\int \tau_\rho g \, d\nu = \int \tau(r(d, \tilde{x}, \ddot{x}), \rho_0(d, x; \tau)) g(\ddot{x} \mid x) \, d\nu(\ddot{x} \mid x).$$

When the referenced derivatives exist, we now write $\mathbf{D}_j \mu_0$ as the $\lambda \times 1$ vector containing the derivatives $\mathbf{D}_j \mu_{0,i}(d, x)$, $i = 1, \dots, \lambda$, and $\mathbf{D}_j \rho_0$ is now the $\lambda \times 1$ vector containing $\mathbf{D}_j \rho_{0,i}(d, x)$, $i = 1, \dots, \lambda$. For $\tau(r, m)$, let $\nabla_r \tau_\rho$ denote the $\lambda \times 1$ vector containing $(\partial/\partial r)\tau_i(r(d, \tilde{x}, \ddot{x}), \rho_0(d, x))$, $i = 1, \dots, \lambda$; let $\nabla_r \tau_\mu$ denote the $\lambda \times 1$ vector containing $(\partial/\partial r)\tau_i(r(d, \tilde{x}, \ddot{x}), \mu_0(d, x))$, $i = 1, \dots, \lambda$; let $\nabla'_m \tau_\mu$ denote the $\lambda \times \lambda$ matrix whose i th row has elements $(\partial/\partial m_j)\tau_i(r(d, \tilde{x}, \ddot{x}), \mu_0(d, x))$, $j = 1, \dots, \lambda$, $i = 1, \dots, \lambda$; and let $\nabla'_m \tau_\rho$ denote the $\lambda \times \lambda$ matrix whose i th row has elements $(\partial/\partial m_j)\tau_i(r(d, \tilde{x}, \ddot{x}), \rho_0(d, x))$, $j = 1, \dots, \lambda$, $i = 1, \dots, \lambda$. When the integrals and inverses exists, we define $Q_\mu \equiv - \int \nabla'_m \tau_\mu g_d \, d\nu$, $Q_\rho \equiv - \int \nabla'_m \tau_\rho g \, d\nu$.

Assumption A.7 (a) The elements of $\tau(r(d, \tilde{x}, \ddot{x}), \mu_0(d, x)) g(\ddot{x} \mid d, x)$ are dominated on C_j by a function integrable with respect to $\nu(\cdot \mid x)$ at (d, x) . (b) The elements of $\tau(r(d, \tilde{x}, \ddot{x}), \rho_0(d, x))$ are dominated on C_j by a function integrable with respect to $dG(\cdot \mid x)$ at (d, x) . ■

Theorem 5.3. (i) Suppose the conditions of Theorem 4.3(i) and A.1(c.i) hold, that $E(s(D, X, \ddot{X})^2) < \infty$, and that $E(\tau(Y, m)^2) < \infty$ for each $m \in M \subset \mathbb{R}^\lambda$. Then for each (d, x, m) in $\text{supp}(D, X) \times M$ the conditional expectation

$$\psi_\tau(d, x, m) = \int \tau(r(d, \tilde{x}, \ddot{x}), m) \, dG(\ddot{x} \mid x)$$

exists and is finite. (ii) Further, let τ, r , and $x \rightarrow G(\cdot \mid x)$ be such that for each (d, x, m) in $\text{supp}(D, X) \times M$, ψ_τ is differentiable on a neighborhood of (d, x, m) , the $\lambda \times \lambda$ matrix $\nabla_m \psi_\tau(d, x, m)$ is non-singular, and $\psi_\tau(d, x, m) = 0$. Then there exists a unique function ρ_0 such that for each $(d, x) \in \text{supp}(D, X)$, ρ_0 is differentiable at (d, x) and

$$\int \tau(r(d, \tilde{x}, \ddot{x}), \rho_0(d, x)) \, dG(\ddot{x} \mid x) = 0.$$

(iii) If A.4 and A.7 also hold, if τ is differentiable and for the given $(d_{(j)}, x)$, $(d_j, \ddot{x}) \rightarrow$

$D_j r(d, \tilde{x}, \tilde{x})$ exists on $C_j \times \text{supp}(\ddot{X} | \tilde{x})$, and if Q_ρ exists and is finite and non-singular, then

$$D_j \rho_0 = Q_\rho^{-1} \int \nabla_r \tau_\rho (D_j r) g \, d\nu.$$

(iv) If in addition A.4(a) and A.7(a) hold and if Q_μ exists and is finite and nonsingular, then $D_j \mu_0 = D_j \rho_0 + \delta_{0,j}$, where

$$\begin{aligned} \delta_{0,j} &\equiv \delta_{1,j} + \delta_{2,j} + \delta_{3,j} \\ \delta_{1,j} &\equiv Q_\rho^{-1} \int \nabla_r \tau_\rho (D_j r) s g_d \, dv \\ \delta_{2,j} &\equiv Q_\mu^{-1} \int \tau_\mu (D_j \log g_d) g_d \, dv \\ \delta_{3,j} &\equiv \int (Q_\mu^{-1} \nabla_r \tau_\mu - Q_\rho^{-1} \nabla_r \tau_\rho) D_j r g_d \, dv. \end{aligned}$$

(v)(a) Suppose $E[(\nabla_r \tau(Y, \rho_0(D, X)))' \nabla_r \tau(Y, \rho_0(D, X)) (D_j r(D, \tilde{X}, \ddot{X}))^2] < \infty$, define $\tilde{\sigma}(\cdot; \nabla_r \tau D_j r) \equiv [\int \nabla_r \tau'_\rho \nabla_r \tau_\rho (D_j r)^2 g_d \, dv]^{1/2}$, and let $\bar{\lambda}_\rho$ denote the largest eigenvalue of $(Q_\rho^{-1} Q_\rho^{-1})$. Then

$$\|\delta_{1,j}\| \equiv (\delta'_{1j} \delta_{1j})^{1/2} \leq \bar{\lambda}_\rho^{1/2} \tilde{\sigma}(\cdot; \nabla_r \tau D_j r) \sigma(\cdot; s).$$

(b) Suppose $E[\tau(Y, \mu_0(D, X))' \tau(Y, \mu_0(D, X))] < \infty$, define $\tilde{\sigma}(\cdot; \tau_\mu) \equiv [\int \tau'_\mu \tau_\mu g_d \, dv]^{1/2}$, and let $\bar{\lambda}_\mu$ denote the largest eigenvalue of $(Q_\mu^{-1} Q_\mu^{-1})$. Then

$$\|\delta_{2,j}\| \leq \bar{\lambda}_\mu^{1/2} \tilde{\sigma}(\cdot; \tau_\mu) \sigma(\cdot; D_j \log g_d).$$

(c) Suppose $E[\nabla_r \tau(Y, \mu_0(D, X))' \nabla_r \tau(Y, \mu_0(D, X))] < \infty$ and $E[\nabla_r \tau(Y, \rho_0(D, X))' \nabla_r \tau(Y, \rho_0(D, X))] < \infty$, and define $\tilde{\sigma}(\cdot; Q_\mu^{-1} \nabla_r \tau_\mu - Q_\rho^{-1} \nabla_r \tau_\rho) \equiv [\int (Q_\mu^{-1} \nabla_r \tau_\mu - Q_\rho^{-1} \nabla_r \tau_\rho)' (Q_\mu^{-1} \nabla_r \tau_\mu - Q_\rho^{-1} \nabla_r \tau_\rho) g_d \, dv]^{1/2}$ and $\tilde{\sigma}(\cdot; D_j r) \equiv [\int (D_j r)^2 g_d \, dv]^{1/2}$. Then

$$\|\delta_{3j}\| \leq \tilde{\sigma}(\cdot; Q_\mu^{-1} \nabla_r \tau_\mu - Q_\rho^{-1} \nabla_r \tau_\rho) \tilde{\sigma}(\cdot; D_j r). \quad \blacksquare$$

The functions in (iii) - (v) above are implicitly evaluated at (d, x) as specified in A.7.

To interpret the *marginal implicit moment effect discrepancy* $\delta_{0,j}$, note that each of its components vanishes under conditional exogeneity, as then $s = 0$, $D_j \log g_d = 0$, and $Q_\mu^{-1} \nabla_r \tau_\mu - Q_\rho^{-1} \nabla_r \tau_\rho = 0$ (Theorem 4.3). If conditional exogeneity fails but the marginal effect $D_j r(d, \tilde{x}, \tilde{x})$ is zero for all \tilde{x} in $\text{supp}(\ddot{X} | d, x)$, the true marginal implicit moment

effect vanishes ($\mathbb{D}_j \rho_0(d, x) = 0$), but the apparent effect becomes

$$\delta_{0,j} \equiv Q_\mu^{-1} \int \tau_\mu (\mathbb{D}_j \log g_d) g_d \, d\nu.$$

The important special case $\tau(r, m) = r - m$ offers further insight. Here $Q_\mu = Q_\rho = 1$ and $\nabla_r \tau_\mu = \nabla_r \tau_\rho = 1$. In this case, we have (cf. Theorem 5.2(iii))

$$\delta_{0,j} \equiv \int (\mathbb{D}_j r) s g_d \, d\nu + \int \tau_\mu (\mathbb{D}_j \log g_d) g_d \, d\nu.$$

Theorem 5.3 affords considerable opportunity to explore special cases of interest (for example, when $\lambda = 1$, consider the anti-symmetric case in which $\tau(r, m) = -\tau(m, r)$, which applies to the conditional median). For brevity, we leave this aside here.

The first two components of $\delta_{0,j}$ are clearly conditional covariances, given that s and $\mathbb{D}_j \log g_d$ have conditional mean zero. The third term is generally not a covariance because there is no need for $\mathbb{D}_j r$ or $Q_\mu^{-1} \nabla_r \tau_\mu - Q_\rho^{-1} \nabla_r \tau_\rho$ to have conditional mean zero. Result (v) provides a near identification and continuity result, generalizing that of Theorem 5.2(iv).

6 Constructing Covariates

Recall that Proposition 3.2 ensures conditional exogeneity ($\ddot{X} \perp D \mid X$) provided $D \stackrel{c}{=} c(X, U)$, where $\ddot{X} \perp U \mid X$. Thus, when the researcher controls D , conditional exogeneity can be ensured using conditional randomization. The covariates are known to the researcher, as they are part of the experimental design. The control necessary for this is common for experimental or clinical researchers, and use of experimental methods to study economic behavior is a rapidly expanding field (see, e.g. Lucking-Reiley, 1999; List and Lucking-Reiley, 2000, 2002; Fershtman and Greezy, 2001; Harrison, Lau and Williams, 2002; Karlan, 2003; Eckel, Johnson, and Montmarquette, 2003; List, 2004; and Bettinger and Slonin, 2004). Our results are thus directly relevant there.

Outside the laboratory, economists can sometimes exploit randomization (e.g., Angrist, 1998), but in the absence of experimental control, economists typically cannot rely on randomization alone to identify effects of interest. We now study how Proposition 3.2 can be used to guide the selection of covariates in the absence of experimental control.

When D is a response generated beyond researcher control, we have

$$D \stackrel{c}{=} c^*(\tilde{X}^*, \ddot{X}^*),$$

for some unknown measurable function c^* and “ D -relevant” explanatory variables $(\tilde{X}^*, \ddot{X}^*)$, say, where \tilde{X}^* is observable and \ddot{X}^* is not. The D -relevant explanatory variables may

include elements of (Z, A) as well as other variables. These other variables cannot be caused by \mathcal{Y} , as this would violate the requirement that $\mathcal{Y} \leftarrow_S \mathcal{D}$; nor can these be caused by \mathcal{D} , as this would violate A.1(c.i).

Examining this expression for D , we see that Proposition 3.2 applies with $c = c^*$, $X = \tilde{X}^*$, and $U = \ddot{X}^*$, in the special case where $\ddot{X} \perp \ddot{X}^* \mid \tilde{X}^*$. Here, the covariates are just $X = \tilde{X}^*$. Generally, however, we might expect \ddot{X}^* to contain useful predictive information for \ddot{X} , even given \tilde{X}^* , as \ddot{X}^* and \ddot{X} could contain common elements. We proceed, therefore, by augmenting \tilde{X}^* with additional observables \tilde{X}^+ , say, sufficiently predictive for \ddot{X} so that given $X = (\tilde{X}^*, \tilde{X}^+)$, \ddot{X}^* no longer contains useful predictive information for \ddot{X} : that is, $\ddot{X} \perp \ddot{X}^* \mid X$. Further, the presence of “ D -irrelevant” variables \tilde{X}^+ does not adversely impact the representation of D . Formally,

$$D \stackrel{c}{=} c(\tilde{X}^*, \tilde{X}^+, \ddot{X}^*) = c^*(\tilde{X}^*, \ddot{X}^*),$$

where the function c acts to “ignore” the D -irrelevant variables. The additional variables \tilde{X}^+ thus permit us to satisfy the conditions of Proposition 3.2 with $U = \ddot{X}^*$.

An obvious source of such variables is \ddot{X} , as \ddot{X} and \tilde{X} are explicitly permitted to be dependent under A.1(i). What other predictors for \ddot{X} are there? To proceed, we separately consider predictors for attributes and for causes. For \ddot{A} , useful proxies are observable attributes other than \tilde{A} , say \tilde{B} , correlated with \ddot{A} . For example, the attribute of height could be used as a proxy for the attribute of gender and vice-versa.

In considering proxies for \ddot{Z} , it is useful to apply Reichenbach’s (1956) *principle of common causality*, as in White (2006). This holds that dependence between two causal variables arises either because one causes the other or because the two share a common cause. As explained in White (2006), it follows that the only relevant channel for the dependence sought here is for the observable proxy to be a response to \ddot{Z} . Specifically, if \ddot{Z} were instead a response to a potential proxy, then a suitable substitution in r yields a modified response function containing the potential proxy as a structurally relevant observable cause. Although this may be helpful, this produces an observable cause, not a proxy for an unobservable cause. Or if both \ddot{Z} and the observable proxy respond to a common unobservable cause, then substituting the response function for \ddot{Z} into r yields a proxy that is again a response to an unobservable cause, the underlying common cause. Measurement error-laden versions of \ddot{Z} are useful proxies of this sort.

Similarly $\ddot{X} \perp \ddot{X}^* \mid X$ implies the symmetrical requirement that \tilde{X}^+ is sufficiently predictive for \ddot{X}^* that given $X = (\tilde{X}^*, \tilde{X}^+)$, \ddot{X} no longer usefully predicts $\ddot{X}^* \equiv (\ddot{Z}^*, \ddot{A}^*)$. Parallel to the discussion above, one may seek observable attributes \tilde{B} correlated with

the unobservable D -relevant attributes, \ddot{A}^* . Similarly, one may seek predictive proxies responding to the unobservable D -relevant causes \ddot{Z}^* .

This discussion justifies Assumption A.1(c.ii): $\mathcal{W} \leftarrow_S \mathcal{Z}$. Thus, the referenced causal variables obey the recursive structure $\mathcal{W} \leftarrow_S \mathcal{Z}$, $\mathcal{D} \leftarrow_S (\mathcal{W}, \mathcal{Z})$, and $\mathcal{Y} \leftarrow_S (\mathcal{D}, \mathcal{W}, \mathcal{Z})$. This does not completely characterize the recursive structure of the settable system, as not every cause of \mathcal{W} is referenced by \mathcal{Z} , and not every cause of \mathcal{D} is referenced by $(\mathcal{W}, \mathcal{Z})$. Assumption A.1(c.ii) is not necessary for our results; we include it in A.1 to emphasize the natural recursive structure that arises when using predictive proxies.

Summarizing, the covariates should include: (1) observable structurally relevant and D -relevant explanatory variables \tilde{X} and \tilde{X}^* ; (2) observable attributes \tilde{B} correlated with unobservable structurally and D -relevant attributes \ddot{A} and \ddot{A}^* ; and (3) observable responses to unobservable structurally relevant and D -relevant causes \ddot{Z} and \ddot{Z}^* . In A.1, the covariates are $X = (W, \tilde{X}, \tilde{B})$. Thus, proxy settings W can be generated as

$$W \stackrel{c}{=} w(\tilde{X}^*, \ddot{X}^*, \ddot{X}, \tilde{X}, \tilde{B}, \ddot{B}, V),$$

where w is a measurable vector-valued function. New variables here are \ddot{B} , a vector of W -relevant unobservable attributes, and V , a vector of unobservable random variables (settings) such that $V \perp D \mid \tilde{X}^*, \ddot{X}^*, \ddot{X}, \tilde{X}, \tilde{B}, \ddot{B}$ (e.g., measurement errors).

It is important to emphasize that economic theory and application domain knowledge play the key roles in identifying legitimate and illegitimate elements of the covariates. Economic theory identifies structurally relevant explanatory variables, D -relevant explanatory variables, and the responses yielding W . Domain knowledge identifies which variables are observable or unobservable, and which observable attributes may be correlated with unobservable structurally or D -relevant attributes. Economic theory also identifies variables caused by \mathcal{Y} or \mathcal{D} , not legitimate for constructing predictive proxies.

Although this provides specific guidance not elsewhere available for choosing covariates, constructing (W, \tilde{B}) as just described cannot be proven to ensure $\ddot{X} \perp \ddot{X}^* \mid X$, so conditional exogeneity cannot be affirmatively verified. Nevertheless, it is falsifiable: one can test conditional exogeneity (see White and Chalak, 2006).

7 Unconfoundedness and Conditional Exogeneity

Unconfoundedness is a key notion in the treatment effects literature, introduced by Rosenbaum and Rubin (1983) for a binary treatment, D , as “ignorable treatment assignment.” It requires

$$(Y(0), Y(1)) \perp D \mid X,$$

where $Y(0)$ is response without treatment ($D = 0$) and $Y(1)$ is response with treatment ($D = 1$). As Imbens (2004, p.7) notes, this is also called the “conditional independence condition” by Lechner (1999, 2001), and “selection on observables” (Heckman and Robb, 1985, following work of Barnow, Cain, and Goldberger, 1980).

In the treatment effects literature, the response function generating $Y(0)$ and $Y(1)$ is typically not specified. In contrast, the settable systems framework explicitly imposes theory-based structure to represent the response. With a single binary treatment, we write $Y(0) = r(0, Z, A)$, $Y(1) = r(1, Z, A)$. As in White (2006, proposition 3.2), Assumptions A.1 and A.2 then imply $(Y(0), Y(1)) \perp D \mid X$.

Hirano and Imbens (2004) extend unconfoundedness to a continuous treatment D taking values in $\mathbf{D} = \text{supp}(D)$. Their “weak unconfoundedness” condition is

$$Y(d) \perp D \mid X \text{ for all } d \text{ in } \mathbf{D},$$

where $Y(d)$ represents the “potential” or “counterfactual” response to treatment d . As is standard in the treatment effects literature, Hirano and Imbens (2004) do not specify the response function generating $Y(d)$. Nevertheless, by taking $Y(d) = r(d, Z, A)$, one can show that A.1 and conditional exogeneity imply weak unconfoundedness.

By not specifying a response function, one operates in a general context, as there is no need to commit to any theory about how the response is determined, apart from its potential dependence on the causes of interest. This appealing generality is costless in experimental contexts, where the researcher can assign treatments by conditional randomization with known covariates. In non-experimental contexts, however, the price paid for this agnosticism is that one has correspondingly little guidance as to the construction of proper covariates. Moreover, in the absence of randomization, the availability of covariates is generally necessary for the identification of causal effects. Even if one desired to proceed atheoretically, this would fail in practice, as economic theory must inevitably play some role in the selection of covariates.

In contrast, by allowing economic theory to play its full and natural role in specifying the variables potentially determining the response and causes of interest one gains broad insight into the selection of legitimate and illegitimate covariates, as Section 6 shows. Settable systems provide a framework in which economic theory can operate compatibly with the higher-level atheoretic structures of the treatment effects framework. Given such theory-based structure (Assumption A.1), the higher-level unconfoundedness condition can play its key role, as ensured by conditional exogeneity.

The machine learning DAG approach also refers to “conditional independence” when

identifying causal effects (e.g., Pearl, 2000). Nevertheless, we prefer to refer to A.2 as conditional exogeneity rather than as simply a conditional independence condition, as conditional exogeneity is a natural extension of terminology entrenched in econometrics, aptly referencing causes of interest, unobserved explanatory variables, and covariates. Referring simply to conditional independence does not afford sufficient specificity.

8 Summary and Conclusion

Settable systems unify three complementary approaches to defining and identifying causal effects: the classical structural equations approach of the Cowles Commission; methods developed in the labor economics and treatment effects literatures; and the structural Directed Acyclic Graph (DAG) approach of the machine learning literature. Settable systems not only nest each of these, but also afford significant improvements to each. Among other things, the settable system approach permits identification of causal effects without exogenous instruments, generalizing the classical structural equations approach; it relaxes the stable unit treatment value assumption of the treatment effect approach; and it accommodates mutual causality and attributes, generalizing the DAG approach.

This work thus complements and creates the opportunity for further extensions of the work of Heckman, Ichimura, and Todd (1997, 1998), Angrist (1998), Hahn (1998), Hirano and Imbens (2001), Hirano, Imbens, and Ridder (2003), Imbens and Newey (2003), Matzkin (2003, 2004, 2005), Heckman, Urzua, and Vytlačil (2005), and Heckman and Vytlačil (2005), among others.

The settable system framework consists of a stochastic structure and a causal structure that rests on this stochastic foundation. It permits precise definitions of cause and effect and of interventions and counterfactuals, and it identifies the objects of experimental control. It delivers straightforward conditions for structural identification of a broad range of causal effects, based on the equality of counterfactual objects with standard stochastic counterparts. The settable system framework yields extensive guidance, not previously available, as to the construction of suitable and unsuitable covariates, making clear the crucial roles played by economic theory and domain knowledge in this construction.

Given structural identification, identified effects can be informatively estimated by standard methods. We pursue this in White and Chalak (2006), where we also study tests for conditional exogeneity.

Our focus here is on recursive settable systems. Analysis of other partitioned settable systems should yield identification results for effects associated with mutually causal relationships, direct or indirect. Versions of instrumental variables methods may yield effective estimators of such effects.

A related direction for further research is to study identification and estimation in the

absence of the requirement that the causes of interest do not cause the ancillary causes (Assumption A.1(c.i)). This raises the possibility of decomposing total effects into direct and indirect components. Chalak and White (2006a) constitutes a beginning for this inquiry, revealing further extensions of the concept of instrumental variables. Also of interest is to construct statistical tests for validity of A.1(c.i).

Finally, it is our hope that empirical application of the settable system framework and the reinterpretation of standard methods that it affords will enable clearer and more robust understanding of causal effects in economic science.

9 Mathematical Appendix

Proof of Proposition 3.1 (i) Given A.1(a, b) and $E(Y) < \infty$, $E(Y | D, X)$ exists and is finite by Billingsley (1979, p.395). (ii) Apply theorem 34.5 of Billingsley (1979). ■

Proof of Proposition 3.2 X and \ddot{X} exist by Assumption A.1(a). Given $\ddot{X} \perp U | X$ (A.2'), it follows from Dawid (1979, lemma 4.1) that $(X, \ddot{X}) \perp (X, U) | X$. Applying Dawid (1979, lemma 4.2(i)) twice gives $\ddot{X} \perp c(X, U) | X$. ■

Proof of Theorem 3.3 (i) Given A.1(a, b) and $E(Y) < \infty$, Proposition 3.1 gives $\mu(d, \tilde{x}) = \int r(d, \tilde{x}, \ddot{x}) dG(\ddot{x} | d, x)$. Assumptions A.1(c.i) and A.2 then imply $\int r(d, \tilde{x}, \ddot{x}) dG(\ddot{x} | d, x) = \int r(d, \tilde{x}, \ddot{x}) dG(\ddot{x} | x) = \rho(d, x)$, ensuring both existence of $\rho(d, x)$ and $\rho(d, x) = \mu(d, x)$. (ii) Assumption A.3 permits application of Bartle (1966, corollary 5.9) establishing the differentiability of ρ and μ (by (i)) and ensuring the validity of an interchange of derivative and integral, giving

$$\begin{aligned} D_j \mu(d, x) &= D_j \rho(d, x) \\ &= \int D_j r(d, \tilde{x}, \ddot{x}) dG(\ddot{x} | x) \quad (\equiv \xi_j(d, x)) \\ &= \int D_j r(d, \tilde{x}, \ddot{x}) dG(\ddot{x} | d, x). \end{aligned}$$

The first equality holds by (i). In the second, A.3 ensures the interchange of derivative and integral, and A.1(c.i) ensures the absence of terms involving $(\partial z / \partial d_j)$. The third equality follows by A.2. ■

Proof of Theorem 4.1. (i) For given $k = 1, 2, \dots$, the proof is identical to that of Proposition 3.1(i), *mutatis mutandis* (replacing r with $\tau_k \circ r$). The measurability of μ_0 follows by measurability of compositions of measurable functions. (ii) For given $k = 1, 2, \dots$, the proof is identical to that of Theorem 3.3(i), *mutatis mutandis*. Measurability follows for ρ_0 just as for μ_0 . That $\rho_0 = \mu_0$ follows immediately from $\rho_k = \mu_k$, $k = 1, 2, \dots$. ■

Proof of Theorem 4.2. (i) Given $E(\tau(Y, m)) < \infty$, the existence and finiteness of $\varphi_{\tau, d}(d, x, m)$ follow from Billingsley (1979, p.395). (ii) For each (d, x) in $\text{supp}(D, X)$, the existence of the non-empty, compact-valued, upper hemi-continuous correspondence $\mu_0(d, x)$ follows from the Theorem of the Maximum (Berge, 1963) under the stated conditions. (iii) If A.1(c.i) and A.2 also hold, then for each $(d, x, m) \in (D, X) \times \mathbb{R}^\lambda$ $\varphi_{\tau, d}(d, x, m) = \varphi_\tau(d, x, m)$. Setting $\rho_0(d, x) = \mu_0(d, x)$ completes the proof. ■

Proof of Theorem 4.3. (i) Given $E(\tau(Y, m)) < \infty$, the existence and finiteness of $\psi_{\tau, d}(d, x, m)$ follow from Billingsley (1979, p. 395). (ii) The existence, uniqueness, and differentiability of μ_0 follow immediately under the given conditions from the implicit function theorem (e.g., Chiang, 1984, pp. 210-211). (iii) If A.1(c.i) and A.2 also hold, then for each $(d, x) \in \text{supp}(D, X)$

$$\int \tau(r(d, \tilde{x}, \ddot{x}), \mu_0(d, x)) \, dG(\ddot{x} \mid d, x) = \int \tau(r(d, \tilde{x}, \ddot{x}), \mu_0(d, x)) \, dG(\ddot{x} \mid x).$$

Setting $\rho_0(d, x) = \mu_0(d, x)$ completes the proof. ■

Proof of Theorem 5.1. Given A.1(a), the densities $dG(\ddot{x} \mid d, x)$ and $dG(\ddot{x} \mid x)$ exist and can be written $dG(\ddot{x} \mid d, x) = dG(d, x, \ddot{x}) / dG(d, x)$ and $dG(\ddot{x} \mid x) = dG(x, \ddot{x}) / dG(x)$. Consequently, $dG(\ddot{x} \mid x) / dG(\ddot{x} \mid d, x) = [dG(x, \ddot{x}) / dG(x)] / [dG(d, x, \ddot{x}) / dG(d, x)] = [dG(d, x) / dG(x)] / [dG(d, x, \ddot{x}) / dG(x, \ddot{x})] = dG(d \mid x) / dG(d \mid \ddot{x}, x)$. (i) We have $\int s(d, x, \ddot{x}) \, dG(\ddot{x} \mid d, x) = \int \{[dG(\ddot{x} \mid d, x) - dG(\ddot{x} \mid x)] / dG(\ddot{x} \mid d, x)\} \, dG(\ddot{x} \mid d, x) = \int [dG(\ddot{x} \mid d, x) - dG(\ddot{x} \mid x)] = 0$, given that $dG(\ddot{x} \mid d, x)$ and $dG(\ddot{x} \mid x)$ are each conditional densities. (ii) Given A.1(a, b) and the conditions on τ_k , Theorem 4.1(i) gives $\mu_k(d, x) = \int \tau_k(r(d, \tilde{x}, \ddot{x})) \, dG(\ddot{x} \mid d, x)$. Adding and subtracting appropriately, with A.1(c.i) we have

$$\begin{aligned} \rho_k(d, x) &\equiv \int \tau_k(r(d, \tilde{x}, \ddot{x})) \, dG(\ddot{x} \mid x) \\ &= \int \tau_k(r(d, \tilde{x}, \ddot{x})) \, dG(\ddot{x} \mid d, x) + \int \tau_k(r(d, \tilde{x}, \ddot{x})) [dG(\ddot{x} \mid x) - dG(\ddot{x} \mid d, x)] \\ &= \mu_k(d, x) + \int \tau_k(r(d, \tilde{x}, \ddot{x})) \{[dG(\ddot{x} \mid x) - dG(\ddot{x} \mid d, x)] / dG(\ddot{x} \mid d, x)\} \, dG(\ddot{x} \mid d, x) \\ &= \mu_k(d, x) - \int \tau_k(r(d, \tilde{x}, \ddot{x})) \, s(d, x, \ddot{x}) \, dG(\ddot{x} \mid d, x) \\ &= \mu_k(d, x) - \gamma_k(d, x). \end{aligned}$$

The existence of $\mu_k(d, x)$ follows given $E(\tau_k(Y)) < \infty$ and the existence of $\gamma_k(d, x)$ follows from the imposed second moment conditions and the Cauchy-Schwarz inequality. It follows that $\rho_k(d, x)$ exists and $\mu_k(d, x) = \rho_k(d, x) + \gamma_k(d, x)$. (iii) The result follows

immediately from the Cauchy-Schwartz inequality, applied to (ii) and using (i). ■

Proof of Theorem 5.2. (i) Assumptions A.1(a) and A.4(a) ensure that for each (d, x) in $\text{supp}(D, X)$, $g(\ddot{x} | d, x) = dG(\ddot{x} | d, x) / d\nu(\ddot{x} | x)$ is a density by the Radon-Nikodym theorem (e.g., Bartle, 1966, theorem 8.9), so $\int g(\ddot{x} | d, x) d\nu(\ddot{x} | x) = 1$. Assumption A.5 ensures that the left hand expression above is differentiable with respect to d_j by Bartle (1966, corollary 5.9). Differentiating both sides of this equality with respect to d_j gives

$$D_j \int g(\ddot{x} | d, x) d\nu(\ddot{x} | x) = 0.$$

Assumption B.4 further justifies interchanging the derivative and integral on the left by Bartle (1966, corollary 5.9), so that

$$\int D_j g(\ddot{x} | d, x) d\nu(\ddot{x} | x) = 0.$$

Substituting $D_j g(\ddot{x} | d, x) = D_j \log g(\ddot{x} | d, x) g(\ddot{x} | d, x)$ delivers the desired result. (ii) Given A.1(a, b) and the conditions on τ_k , Theorem 4.1(i) gives $\mu_k(d, x) = \int \tau_k(r(d, \tilde{x}, \ddot{x})) dG(\ddot{x} | d, x)$. Substituting $dG(\ddot{x} | d, x) = g(\ddot{x} | d, x) d\nu(\ddot{x} | x)$ gives

$$\mu_k(d, x) = \int \tau_k(r(d, \tilde{x}, \ddot{x})) g(\ddot{x} | d, x) d\nu(\ddot{x} | x).$$

Assumption A.6 permits application of Bartle (1966, corollary 5.9) establishing the differentiability of μ_k and ensuring the validity of an interchange of derivative and integral, giving

$$\begin{aligned} D_j \mu_k(d, x) &= \int D_j ((\tau_k \circ r)g)(d, x, \ddot{x}) d\nu(\ddot{x} | x) \\ &= \int D_j (\tau_k \circ r)(d, \tilde{x}, \ddot{x}) g(\ddot{x} | d, x) d\nu(\ddot{x} | x) \\ &\quad + \int \tau_k(r(d, \tilde{x}, \ddot{x})) D_j g(\ddot{x} | d, x) d\nu(\ddot{x} | x), \end{aligned}$$

where A.1(c.i) ensures the absence of terms involving $(\partial x / \partial d_j)$ in the second equality. Substituting $D_j g(\ddot{x} | d, x) = D_j \log g(\ddot{x} | d, x) g(\ddot{x} | d, x)$ delivers the result. (iii) The moment conditions ensure $\left| E(D_j \tau_k(r(D, \tilde{X}, \ddot{X}))s(D, X, \ddot{X})) \right| < \infty$ by Cauchy-Schwarz, ensuring the existence of $\delta_{1,k,j}(d, x)$ and thus of

$$\xi_{k,j}(d, x) = \int D_j \tau_k(r(d, \tilde{x}, \ddot{x})) g(\ddot{x} | d, x) d\nu(\ddot{x} | x) - \delta_{1,k,j}(d, x).$$

The result now follows from (ii). (iv) The result follows immediately from the Cauchy-Schwarz inequality, applied to (iii) and using (i). ■

Proof of Theorem 5.3 (i) We write

$$\begin{aligned}\psi_\tau(d, x, m) &\equiv \int \tau(r(d, \tilde{x}, \ddot{x}), m) \, dG(\ddot{x} \mid x) \\ &= \int \tau(r(d, \tilde{x}, \ddot{x}), m) [dG(\ddot{x} \mid x) / dG(\ddot{x} \mid d, x)] \, dG(\ddot{x} \mid d, x).\end{aligned}$$

The imposed second moment conditions ensure the existence and finiteness of this integral by Cauchy-Schwarz. (ii) The existence, uniqueness, and differentiability of ρ_0 follow immediately under the given conditions from the implicit function theorem (e.g., Chiang, 1984, pp. 210-211). (iii) Given the assumed differentiability of ψ_τ and Assumption A.7(b), we have

$$\mathbf{D}_j \psi_\tau(d, x, \rho_0(d, x)) = \int \mathbf{D}_j \tau(r(d, \tilde{x}, \ddot{x}), \rho_0(d, x,)) \, dG(\ddot{x} \mid x) = 0,$$

where the interchange of integral and derivative is justified by Bartle (1966, corollary 5.9), and the equality holds because $\psi_\tau(d, x, \rho_0(d, x)) = 0$ for all (d, x) in $\text{supp}(D, X)$. Using the assumed differentiability of τ and r , the differentiability of ρ_0 ensured by (ii), and the chain rule gives

$$\int [\nabla_r \tau_\rho \, \mathbf{D}_j r + \nabla'_m \tau_\rho \, \mathbf{D}_j \rho_0] \, dG(\ddot{x} \mid x) = 0,$$

where we exploit the notation introduced preceding Theorem 5.3 in the text. Solving for $\mathbf{D}_j \rho_0$ given the assumed existence of $Q_\rho^{-1} = -[\int \nabla'_m \tau_\rho \, dG(\ddot{x} \mid x)]^{-1}$ yields

$$\mathbf{D}_j \rho_0 = Q_\rho^{-1} \int \nabla_r \tau_\rho \, \mathbf{D}_j r \, dG(\ddot{x} \mid x) = Q_\rho^{-1} \int \nabla_r \tau_\rho \, \mathbf{D}_j r \, g \, d\nu,$$

where the second equality holds given A.4(b). (iv) A similar argument invoking A.4(a) and A.7(a) instead of A.4(b) and A.7(b) gives

$$\mathbf{D}_j \mu_0 = Q_\mu^{-1} \int [\nabla_r \tau_\mu \, \mathbf{D}_j r + \tau_\mu (\mathbf{D}_j \log g_d)] \, g_d \, d\nu,$$

given the assumed existence of $Q_\mu^{-1} = -[\int \nabla'_m \tau_\mu \, g_d \, d\nu]^{-1}$. It follows that

$$\mathbf{D}_j \mu_0 - \mathbf{D}_j \rho_0 = \int [Q_\mu^{-1} \nabla_r \tau_\mu - Q_\rho^{-1} \nabla_r \tau_\rho \, g/g_d] \, \mathbf{D}_j r \, g_d \, d\nu + Q_\mu^{-1} \int \tau_\mu (\mathbf{D}_j \log g_d) \, g_d \, d\nu.$$

The expression for $\delta_{0,j}$ holds by adding and subtracting terms appropriately. (v)(a)

$\delta'_{1j}\delta_{1j} = [\int \nabla'_r \tau_\rho(\mathbf{D}_j r) s g_d dv] Q_\rho^{-1} Q_\rho^{-1} [\int \nabla_r \tau_\rho(\mathbf{D}_j r) s g_d dv] \leq \bar{\lambda}_\rho [\int \nabla_r \tau_\rho(\mathbf{D}_j r) s g_d dv]' [\int \nabla_r \tau_\rho(\mathbf{D}_j r) s g_d dv]$ by the Rayleigh inequality. The result then follows by Cauchy-Schwarz. (b) Analogous to (a). (c) Analogous to (a). ■

References

Abadie, A., J. Angrist, and G. Imbens (2002), “Instrumental Variables Estimates of the Effects of Subsidized Training on the Quantiles of Trainee Earnings,” *Econometrica*, 70, 91-117.

Abadie, A. and G. Imbens (2002), “Simple and Bias-Corrected Matching Estimators for Average Treatment Effects,” NBER Technical Working Paper No. 283.

Ai, C. and Chen, X. (2003), “Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions,” *Econometrica*, 71, 1795-1843.

Angrist, J. (1998) “Using Social Security Data on Military Applicants to Estimate the Effect of Voluntary Military Service on Earnings,” *Econometrica*, 66, 249-288.

Angrist, J. and A. Krueger (1999), “Empirical Strategies in Labor Economics,” in: O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*, Vol 3A. Amsterdam: Elsevier, pp. 1277 - 1368.

Bajari, P., H. Hong, and J. Ryan (2004), “Identification and Estimation of Discrete Games of Complete Information,” University of Michigan Department of Economics Working Paper.

Bang-Jensen, J. and G. Gutin (2001). *Digraphs: Theory, Algorithms, and Applications*. London: Springer Verlag.

Barnow, B., G. Cain, and A. Goldberger (1980), “Issues in the Analysis of Selectivity Bias,” in E. Stromsdorfer and G. Farkas (eds.), *Evaluation Studies*, vol 5. San Francisco: Sage, pp. 43-59.

Bartle, R. (1966), *Elements of Integration*. New York: Wiley.

Berge, C. (1963), *Espaces Topologiques*. Paris: Dunod (translation by E.M. Patterson, *Topological Spaces*. Edinburgh: Oliver and Boyd).

Bettinger, E. and R. Slonin (2004), “Using Experimental Economics to Measure the Effect of a Natural Educational Experiment on Altruism,” NBER Working Paper.

Billingsley, P. (1979). *Probability Theory and Measure*. New York: Wiley.

Blundell, R. and J. Powell (2003), “Endogeneity in Nonparametric and Semiparametric Regression Models,” in M. Dewatripoint, L. Hansen, and S. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, vol II. New York: Cambridge University Press, pp. 312-357.

Cartwright, N. (1989). *Nature’s Capacities and Their Measurement*. Oxford: Oxford University Press.

Chalakov, K. and H. White (2006a), "An Extended Class of Instrumental Variables for the Estimation of Causal Effects," UCSD Department of Economics Discussion Paper.

Chalakov, K. and H. White (2006b), "Independence and Conditional Independence in Causal Systems," UCSD Department of Economics Discussion Paper.

Chen, X. (2005), "Large Sample Sieve Estimation of Semi-Nonparametric Models," New York University C.V. Starr Center Working Paper.

Chiang, A. (1984). *Fundamental Methods of Mathematical Economics*. New York: McGraw-Hill.

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

Dawid, A.P. (2000), "Causal Inference without Counterfactuals," *Journal of the American Statistical Association*, 95, 407-448.

Dawid, A.P. (2002), "Influence Diagrams for Causal Modeling and Inference," *International Statistical Review*, 70, 161-189.

Eckel, C., C. Johnson, and C. Montmarquette (2003), "Human Capital Investment by the Poor: Calibrating Policy with Laboratory Experiments," VPI Department of Economics Working Paper.

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

Fershtman, C. and U. Gneezy (2001), "Discrimination in a Segmented Society: An Experimental Approach," *The Quarterly Journal of Economics*, 116, 351-377.

Firpo, S., N. Fortin, and T. Lemieux (2005), "Decomposing Wage Distributions: Estimation and Inference," UBC Department of Economics Working Paper.

Fisher, F. (1961), "On the Cost of Approximate Specification in Simultaneous Equation Estimation," *Econometrica*, 29, 139-170.

Fisher, F. (1966). *The Identification Problem in Econometrics*. New York: McGraw-Hill.

Fisher, F. (1970), "A Correspondence Principle for Simultaneous Equations Models," *Econometrica*, 38, 73-92.

Fisher, R.A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Gibbons, R. (1992). *Game Theory for Applied Economists*. Princeton: Princeton University Press.

Gourieroux, C., A. Monfort, and A. Trognon (1984), "Pseudo Maximum Likelihood Methods: Theory," *Econometrica*, 52, 681-700.

Grenander, U. (1981). *Abstract Inference*. New York: Wiley.

Hahn J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric

Estimation of Average Treatment Effect,” *Econometrica*, 66, 315-331.

Hansen, L.P. (1982), “Large Sample Properties of Generalized Method of Moments Estimators,” *Econometrica*, 50, 1029-1054.

Harrison, G., M. Lau, and M. Williams (2002), “Estimating Individual Discount Rates for Denmark: A Field Experiment,” *American Economic Review*, 92, 1606-1617.

Heckman, J. (2005), “Econometric Causality,” University of Chicago Department of Economics, manuscript.

Heckman, J. and J. Hotz (1989), “Alternative Methods for Evaluating the Impact of Training Programs,” (with discussion), *Journal of the American Statistical Association*, 84, 862-874.

Heckman, J., H. Ichimura, and P. Todd (1997), “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program,” *Review of Economic Studies*, 64, 605-654.

Heckman, J., H. Ichimura, and P. Todd (1998), “Matching as an Econometric Evaluation Estimator,” *Review of Economic Studies*, 65, 261-294.

Heckman, J. and R. Robb (1985), “Alternative Methods for Evaluating the Impact of Interventions,” in J. Heckman and B. Singer (eds.), *Longitudinal Analysis of Labor Market Data*. Cambridge: Cambridge University Press, pp. 156-245.

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

Heckman, J., S. Urzua, and E. Vytlačil (2005), “Understanding Instrumental Variables in Models with Essential Heterogeneity,” University of Chicago Department of Economics, manuscript.

Heckman J. and E. Vytlačil (2005), “Structural Equations, Treatment Effects, and Econometric Policy Evaluation,” *Econometrica*, 73, 669-738.

Hirano, K. and G. Imbens (2001), “Estimation of Causal Effects using Propensity Score Weighting: An Application to Right Heart Catheterization,” *Health Services and Outcomes Research*, 2, 259-278.

Hirano, K. and G. Imbens (2004), “The Propensity Score with Continuous Treatments,” in A. Gelman and X.-L. Meng (eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. New York: Wiley, pp. 73-84.

Hirano, K., G. Imbens, and G. Ridder (2003), “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,” *Econometrica*, 71, 1161-1189.

Holland, P.W. (1986), “Statistics and Causal Inference” (with Discussion), *Journal of the American Statistical Association*, 81, 945-970.

Hoover, K. (2001). *Causality in Macroeconomics*. Cambridge: Cambridge University Press.

Hoover, K. (2004), "Lost Causes," *Journal of the History of Economic Thought*, 26, 149-164.

Imbens, G. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86, 4-29.

Imbens, G. and W. Newey (2003), "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," MIT Department of Economics Working Paper.

Karlan, D. (2005), "Using Experimental Economics to Measure Social Capital and Predict Finance Decisions," *American Economic Review*, 95, 1688-1699.

Lauritzen, S. and T. Richardson (2002), "Chain Graph Models and Their Causal Interpretations," *Journal of the Royal Statistical Society, Series B*, 64, 321-361 (with discussion).

Lechner, M. (1999), "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification," *Journal of Business and Economic Statistics*, 17, 74-90.

Lechner, M. (2001), "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption," in M. Lechner and F. Pfeiffer (eds.), *Econometric Evaluations of Active Labor Market Policies in Europe*. Heidelberg: Physica-Springer, pp. 43-58.

Lehmann, E. (1974). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day.

List, J. (2004), "The Nature and Extent of Discrimination in the Market Place: Evidence from the Field," *Quarterly Journal of Economics*, 119, 49-89.

List, J. and D. Lucking-Reiley (2000), "Demand Reduction in Multi-Unit Auctions: Evidence from a Sportscard Field Experiment," *American Economic Review*, 90, 961-972.

List, J. and D. Lucking-Reiley (2002), "The Effects of Seed Money and Refunds on Charitable Giving: Experimental Evidence from a University Capital Campaign," *Journal of Political Economy*, 90, 215-233.

Lucking-Reiley, D. (1999), "Using Field Experiments to Test Equivalence Between Auction Formats: Magic on the Internet," *American Economic Review*, 89, 1063-1080.

Matzkin, R. (2003), "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339-1375.

Matzkin, R. (2004), "Unobservable Instruments," Northwestern University Department of Economics Working Paper.

Matzkin, R. (2005), "Identification of Nonparametric Simultaneous Equations," Northwestern University Department of Economics Working Paper.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufman.

Pearl, J. (1993a), "Aspects of Graphical Methods Connected with Causality," in *Proceedings of the 49th Session of the International Statistical Institute*, pp. 391-401.

Pearl, J. (1993b), "Comment: Graphical Models, Causality, and Intervention," *Statistical Science*, 8, 266-269.

Pearl, J. (1995), "Causal Diagrams for Experimental Research" (with Discussion), *Biometrika*, 82, 669-710.

Pearl, J. (1998), "Graphs, Causality, and Structural Equation Models," *Sociological Methods and Research*, 27, 226-284.

Pearl, J. (2000). *Causality*. New York: Cambridge University Press.

Powell, J., J. Stock, and T. Stoker (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403-1430.

Ragusa, G. (2005), "Alternatives to GMM: Properties of Minimum Divergence Estimators," UCSD Department of Economics Discussion Paper.

Reichenbach, H. (1956). *The Direction of Time*. Berkeley: University of California Press.

Robins, J. (1989), "The Control of Confounding by Intermediate Variables," *Statistics in Medicine*, 8, 679-701.

Robins, J., S. Greenland, and F.-C. Hu (1999), "Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome," *Journal of the American Statistical Association*, 94, 687-712.

Robins, J., M. Hernan, and B. Brumback (2000), "Marginal Structural Models and Causal Inference in Epidemiology," *Epidemiology*, 11, 550-560.

Robins, J., M. Hernan, and U. Siebert (2004), "Effects of Multiple Interventions," in M. Ezzati, A. Lopez, A. Rodgers, and C. Murray (eds.), *Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors*, Vol 1. Geneva: World Health Organization, pp. 2191-2230.

Rosenbaum, P. (1984), "The Consequences of Adjustment for a Concomitant Variable that has been Affected by Treatment," *Journal of the Royal Statistical Society, Series A*, 147, 656-666.

Rosenbaum, P. (1987), "The Role of a Second Control Group in an Observational Study," (with discussion), *Statistical Science*, 2:3, 292-316.

Rosenbaum, P. and D. Rubin (1983), "The Central Role of the Propensity Score in

Observational Studies for Causal Effects,” *Biometrika*, 70, 41-55.

Rubin, D. (1974), “Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies,” *Journal of Educational Psychology*, 66, 688-701.

Rubin, D. (1980), “Comment on ‘Randomization Analysis of Experimental Data: The Fisher Randomization Test,’ by D. Basu,” *Journal of the American Statistical Association*, 75, 591-593.

Rubin D. (1986), “Which Ifs have Causal Answers?” (Comment on “Statistics and Causal Inference,” by P. Holland), *Journal of the American Statistical Association*, 81, 961-962.

Shennach, S. (2004), “Exponentially Tilted Empirical Likelihood,” University of Chicago Working Paper.

Shipley, W. (2000). *Cause and Correlation in Biology: A User’s Guide to Path Analysis, Structural Equations, and Causal Inference*. Cambridge: Cambridge University Press.

Simon, H. (1953), “Causal Ordering and Identifiability,” in W.C. Hood and T.C. Koopmans (eds.), *Studies in Econometric Methods*. New York: Wiley, pp. 49-74.

Spirtes, P., C. Glymour and R. Scheines (1993). *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

Stoker, T. (1986), “Consistent Estimation of Scaled Coefficients,” *Econometrica*, 54, 1461-1481.

Strotz, R. and H. Wold (1960), “Recursive vs. Nonrecursive Systems: An Attempt at Synthesis,” *Econometrica*, 28, 417-427.

Verma, T. and J. Pearl (1991), “Equivalence and Synthesis of Causal Models,” in P.P. Bonissone, M. Henrion, L.N. Kanal, and J.F. Lemmer (eds.), *Uncertainty in Artificial Intelligence, vol. 6*. Amsterdam: Elsevier, pp. 255-268.

White, H. (1980), “Using Least Squares to Approximate Unknown Regression Functions,” *International Economic Review*, 21, 149-170.

White, H. (1994). *Estimation, Inference and Specification Analysis*. New York: Cambridge University Press.

White, H. (2006), “Time Series Estimation of the Effects of Natural Experiments,” *Journal of Econometrics* (in press).

White, H. and K. Chalak (2006), “Parametric and Nonparametric Estimation of Covariate-Conditioned Average Causal Effects,” UCSD Department of Economics Discussion Paper.

Wold, H. (1954), “Causality and Econometrics,” *Econometrica*, 162-177.

Wold, H. (1956), “Causal Inference from Observational Data: A Review of Ends and Means,” *Journal of the Royal Statistical Society, Series A*, 119, 28-50.

Wooldridge, J. (2002). *Econometric Analysis of Cross-Section and Panel Data*. Cambridge MA: MIT Press.

Wooldridge, J. (2005), "Violating Ignorability of Treatment by Controlling for Too Many Factors," *Econometric Theory*, 21, 1026-1029.