1

THE SCIENTIFIC MODEL OF CAUSALITY

*James J. Heckman**

Causality is a very intuitive notion that is difficult to make precise without lapsing into tautology. Two ingredients are central to any definition: (1) a set of possible outcomes (counterfactuals) generated by a function of a set of "factors" or "determinants" and (2) a manipulation where one (or more) of the "factors" or "determinants" is changed. An effect is realized as a change in the argument of a stable function that produces the same change in the outcome for a class of interventions that change the "factors" by the same amount. The outcomes are compared at different levels of the factors or generating variables. Holding all factors save one at a constant level, the change in the outcome associated with manipulation of the varied factor is called a causal effect of the manipulated factor. This definition, or some version of it, goes back to Mill (1848) and Marshall (1890). Haavelmo's (1943) made it more precise within the context of linear equations models. The phrase '*ceteris paribus*' (everything else held constant) is a mainstay of economic analysis

This research was supported by NSF 97-09-873, 00-99195, NSF SES-0241858, NIH R01-HD043411, and the American Bar Foundation. An earlier version of this paper was presented at the ISI meeting in Seoul, Korea, in August 2001. I am grateful to Jaap Abbring and Edward Vytlacil for very helpful discussions about the topics of this paper over the past five years. Yu Xie and especially T. N. Srinivasan made helpful comments on this version. Some of the material in this paper also appears in Heckman and Vytlacil (2006a,b).

*University of Chicago, University College London, and the American Bar Foundation

and captures the essential idea underlying causal models. This paper develops the scientific model of causality developed in economics and compares it to methods advocated in epidemiology, statistics, and in many of the social sciences outside of economics that have been influenced by statistics and epidemiology.

I make two main points that are firmly anchored in the econometric tradition. The first is that causality is a property of a model of hypotheticals. A fully articulated model of the phenomena being studied precisely defines hypothetical or counterfactual states.[1] A definition of causality drops out of a fully articulated model as an automatic by-product. A model is a set of possible counterfactual worlds constructed under some rules. The rules may be the laws of physics, the consequences of utility maximization, or the rules governing social interactions, to take only three of many possible examples. A model is in the mind. As a consequence, causality is in the mind.

In order to be precise, counterfactual statements must be made within a precisely stated model. Ambiguity in model specification implies ambiguity in the definition of counterfactuals and hence of the notion of causality. The more complete the model of counterfactuals, the more precise the definition of causality. The ambiguity and controversy surrounding discussions of causal models are consequences of analysts wanting something for nothing: a definition of causality without a clearly articulated model of the phenomenon being described (i.e., a model of counterfactuals). They want to describe a phenomenon as being modeled "causally" without producing a clear model of how the phenomenon being described is generated or what mechanisms select the counterfactuals that are observed in hypothetical or real samples. In the words of Holland (1986), they want to model the effects of causes without modeling the causes of effects. Science is all about constructing models of the causes of effects. This paper develops the scientific model of causality and shows its value in analyzing policy problems.

My second main point is that the existing literature on "causal inference" in statistics confuses three distinct tasks that need to be carefully distinguished:

[1]I will use the term *counterfactual* as defined in philosophy. A counterfactual need not be contrary to certain facts. It is just a hypothetical. The term *hypothetical* would be better and I will use the two concepts interchangeably.

- Definitions of counterfactuals.
- Identification of causal models from population distributions (infinite samples without any sampling variation). The hypothetical populations producing these distributions may be subject to selection bias, attrition, and the like. However, issues of sampling variability of empirical distributions are irrelevant for the analysis of this problem.
- Identification of causal models from actual data, where sampling variability is an issue. This analysis recognizes the difference between empirical distributions based on sampled data and population distributions generating the data.

Table 1 represents these three tasks.

The first task is a matter of science, logic, and imagination. It is also partly a matter of convention. A model of counterfactuals is more widely accepted the more widely accepted are its ingredients, which are

- the rules of the derivation of a model including whether or not the rules of logic and mathematics are followed;
- its agreement with other theories; and
- its agreement with the accepted interpretations of facts.

Models are not empirical statements or descriptions of actual worlds. They are descriptions of hypothetical worlds obtained by varying—hypothetically—the factors determining outcomes.

TABLE 1
Three Distinct Tasks Arising from Analysis of Causal Models

Task	Description	Requirements
1	Defining the Set of Hypotheticals or Counterfactuals	A Scientific Theory
2	Identifying Parameters (Causal or Otherwise) from Hypothetical Population Data	Mathematical Analysis of Point or Set Identification
3	Identifying Parameters from Real Data	Estimation and Testing Theory

The second task is one of inference in very large samples. Can we recover counterfactuals (or means or distributions of counterfactuals) from data that are free of sampling variation? This is the identification problem. It abstracts from any variability in estimates due to sampling variation. It is strictly an issue of finding unique mappings from population distributions, population moments or other population measures to causal parameters.

The third task is one of inference in practice. Can one recover a given model or the desired causal parameters from a given set of data? This entails issues of inference and testing in real world samples. This is the task most familiar to statisticians and empirical social scientists. This essay focuses on the first two tasks. Identification is discussed, but issues of sampling distributions of estimators, such as efficiency, are not.

Some of the controversy surrounding counterfactuals and causal models is partly a consequence of analysts being unclear about these three distinct tasks and often confusing solutions to each of them. Some analysts associate particular methods of estimation (e.g., matching or instrumental variable estimation) with causal inference and the definition of causal parameters. Such associations confuse the three distinct tasks of definition, identification, and estimation. Each method for estimating causal parameters makes some assumptions and forces certain constraints on the counterfactuals.

Many statisticians are uncomfortable with counterfactuals. Their discomfort arises in part from the need to specify models to interpret and identify counterfactuals. Most statisticians are not trained in science or social science and adopt as their credo that they "should stick to the facts." An extreme recent example of this discomfort is expressed by Dawid (2000), who denies the need for, or validity of, counterfactual analysis. Tukey (1986) rejects the provisional nature of causal knowledge—i.e., its dependence on *a priori* models to define the universe of counterfactuals and the mechanisms of selection and the dependence of estimators of causal parameters on *a priori*, untestable assumptions.[2] Cox (1992) appears to accept the provisional nature of causal knowledge (see also Cox and Wermuth 1996). Science is based on counterfactuals and theoretical models.

[2]The exchange between Heckman and Tukey in Wainer (1986) anticipates many of the issues raised in this paper.

Human knowledge is produced by constructing counterfactuals and theories. Blind empiricism unguided by a theoretical framework for interpreting facts leads nowhere.

Causal models which are widely used in epidemiology and statistics are incompletely specified because they do not delineate selection mechanisms for how hypothetical counterfactuals are realized or how hypothetical interventions are implemented even in hypothetical populations. They focus only on outcomes of treatment, leaving the model-selecting outcomes only implicitly specified. In addition, in this literature the construction of counterfactual outcomes is based on intuition and not on explicit formal models. Instead of modeling outcome selection mechanisms, a metaphor of "random selection" is adopted. This emphasis on randomization or its surrogates (like matching) rules out a variety of alternative channels of identification of counterfactuals from population or sample data. It has practical consequences because of the conflation of step one with steps two and three in Table 1. Since randomization is used to define the parameters of interest, this practice sometimes leads to the confusion that randomization is the only way—or at least the best way—to identify causal parameters from real data. In truth, this is not always so, as I show in this paper.

Another reason why epidemiological and statistical models are incomplete is that they do not specify the sources of randomness generating the unobservables in the models—i.e., they do not explain why observationally identical people make different choices and have different outcomes given the same choice. Modeling these unobservables greatly facilitates the choice of estimators to identify causal parameters. Statistical and epidemiological models are incomplete because they are recursive. They do not allow for simultaneous choices of outcomes of treatment that are at the heart of game theory and models of social interactions (e.g., see Tamer 2003; Brock and Durlauf 2001). They rule out the possibility that one outcome can cause another if all outcomes are chosen simultaneously. They are also incomplete because the ingredients of the "treatments" are not considered at a finer level. "Treatment" is usually a black box of many aggregate factors that are not isolated or related to underlying theory in a precise way. This makes it difficult to understand what factor or set of factors produces the "effect" of the intervention being analyzed. The treatment effects identified in the statistical literature cannot be used to forecast out-of-sample to new populations. They are incomplete because they do not

distinguish uncertainty from the point of view of the agent being analyzed from variability as analyzed by the observing social scientist.

Economists since the time of Haavelmo (1943, 1944) have recognized the need for precise models to construct counterfactuals and to answer "causal" questions and more general policy evaluation questions, including making out-of-sample forecasts. The econometric framework is explicit about how counterfactuals are generated and how interventions are assigned (the rules of assigning "treatment"). The sources of unobservables, in both treatment assignment equations and outcome equations, and the relationship between the unobservables are studied. Rather than leaving the rule governing selection of treatment implicit, the econometric approach explicitly models the relationship between the unobservables in outcome equations and selection equations to identify causal models from data and to clarify the nature of identifying assumptions. The theory of structural modeling in econometrics is based on these principles.

The goal of the econometric literature, like the goal of all science, is to model phenomena at a deeper level, to understand the causes producing the effects so that we can use empirical versions of the models to forecast the effects of interventions never previously experienced, to calculate a variety of policy counterfactuals, and to use scientific theory to guide the choices of estimators and the interpretation of the evidence. These activities require development of a more elaborate theory than is envisioned in the current literature on causal inference in epidemiology and statistics.

This essay is in five parts. Section 1 discusses policy evaluation questions as a backdrop against which to compare alternative approaches to causal inference. A notation is developed and both individual-level and population-level causal effects are defined. Population-level effects are defined both in terms of means and distributions. Uncertainty at the individual level is introduced to account for one source of randomness across persons in terms of outcomes and choices.

Section 2 is the heart of the paper. It defines causality using structural econometric models and analyzes both objective outcomes and subjective evaluations. It defines structural models and policy-invariant structural parameters. A definition of causality in models with simultaneously determined outcomes is presented. A distinction between conditioning and fixing variables is developed. The Neyman (1923)–Rubin (1978) model advocated in statistics is compared to the

scientific model. Marschak's maxim is defined. This maxim links the statistical treatment effect literature to the literature on structural models by showing that statistical treatment effects focus on answering one narrow question while the structural approach attempts to answer many questions. It is usually easier to answer one question well than to answer many questions at the same time but the narrowness of the question answered in the treatment effect literature limits the applicability of the answer obtained to address other questions.

Section 3 briefly discusses the identification problem at a general level (task 2 in Table 1). Section 4 applies the framework of the paper to the identification of four widely used estimators for causal inference and the implicit identifying assumptions that justify their application. This section is only intended as a comprehensive survey. Section 5 concludes.

1. POLICY EVALUATION QUESTIONS AND CRITERIA OF INTEREST

This paper discusses questions of causal inference in terms of policy evaluation and policy forecasting problems. Such a focus appears to limit the scope of the inquiry. In fact, it makes the discussion more precise by placing it in a concrete context. By focusing on policy questions, the discussion gains tangibility, something often lacking in the literature on causality. In social science, a major use of causal analysis is in determining "effects" of various policies. Causal analysis is almost always directed toward answering policy questions.

This section first presents three central policy evaluation questions. It then defines the notation used in this paper and the definition of individual-level causal effects or treatment effects. The policy evaluation problem is discussed in general terms. Population-level mean treatment parameters are then defined and distributional criteria are also presented. We discuss, in general terms, the type of data needed to construct the policy evaluation criteria.

1.1. *Three Policy Evaluation Problems*

Three broad classes of policy evaluation questions are of general interest. Policy evaluation question one is:

P1: *Evaluating the impact of historical interventions on outcomes including their impact in terms of welfare.*

By historical, I refer to interventions actually experienced. A variety of outcomes and welfare criteria might be used to form these evaluations. By impact, I mean constructing either individual-level or population-level counterfactuals and their valuations. By welfare, I mean the valuations of the outcomes obtained from the intervention by the agents being analyzed or some other party (e.g., the parents of the agent).

P1 is the problem of *internal validity*. It is the problem of identifying a given treatment parameter or a set of treatment parameters in a given environment (see Campbell and Stanley 1963). This is the policy question addressed in the epidemiological and statistical literature on causality. A drug trial for a particular patient population is the prototypical problem in that literature.

Most policy evaluation is designed with an eye toward the future and toward decisions about new policies and application of old policies to new environments. I distinguish a second task of policy analysis:

P2: *Forecasting the impacts (constructing counterfactual states) of interventions implemented in one environment in other environments, including their impacts in terms of welfare.*

Included in these interventions are policies described by generic characteristics (e.g., tax or benefit rates, etc.) that are applied to different groups of people or in different time periods from those studied in previous implementations of these policies. This is the problem of *external validity*: taking a treatment parameter or a set of parameters estimated in one environment to another environment. The "environment" includes the characteristics of individuals and of their social and economic setting.

Finally, the most ambitious problem is forecasting the effect of a new policy, never previously experienced:

P3: *Forecasting the impacts of interventions (constructing counterfactual states associated with interventions) never historically experienced to other environments, including their impacts in terms of welfare.*

This problem requires that one use past history to forecast the consequences of new policies. It is a fundamental problem in

knowledge.[3] I now present a framework within which one can address these problems in a systematic fashion. It is also a framework that can be used for causal inference.

1.2. *Notation and Definition of Individual-Level Treatment or Causal Effects*

To evaluate is to value and to compare values among possible outcomes. These are two distinct tasks that I distinguish in this essay. Define outcomes corresponding to state (policy, treatment) s for person ω as $Y(s, \omega)$, $\omega \in \Omega$. One can think of Ω as a universe of individuals each characterized by their own element ω. The ω encompass all features of individuals that affect Y outcomes. $Y(s, \omega)$ may be generated from a scientific or social science theory. $Y(s, \omega)$ may be vector valued. The components of $Y(s, \omega)$ may also be interdependent, as in the Cowles Commission simultaneous equations model developed by Haavelmo (1943, 1944) and discussed in Section 2. The components of $Y(s, \omega)$ may be discrete, continuous, or mixed discrete-continuous random variables.

I use "ω" as a shorthand descriptor of the state of a person. We (the analyst) may observe variables $X(\omega)$ that characterize the person as well. In addition, there may be model unobservables. I develop this distinction further in Section 2.

The $Y(s, \omega)$ are outcomes after treatment is chosen. In advance of treatment, agents may not know the $Y(s, \omega)$ but may make forecasts about them. These forecasts may influence their decisions to participate in a treatment or may influence the agents who make decisions about whether or not an individual participates in the treatment. Selection into the program based on actual or anticipated components of outcomes gives rise to the selection problem in the evaluation literature.

Let $\mathcal{S}$ be the set of possible treatments denoted by s. For simplicity of exposition, I assume that this set is the same for all ω.[4] For each choice of $s \in \mathcal{S}$ and for each person ω, we obtain a collection of possible outcomes given by $\{Y(s, \omega)\}_{s \in \mathcal{S}}$. The set $\mathcal{S}$ may be finite

[3]Knight (1921:313) succinctly summarizes the problem: "The existence of a problem in knowledge depends on the future being different from the past, while the possibility of a solution of the problem depends on the future being like the past."

[4]At the cost of a more cumbersome notation, this assumption can be modified so that $\mathcal{S}$ sets are ω-specific.

(e.g., J states with $\mathcal{S} = \{1, \ldots J\}$), countable, or may be defined on the continuum (e.g., $\mathcal{S} = [0, 1]$) so there are an uncountable number of states. For example, if $\mathcal{S} = \{0, 1\}$, there are two policies (or treatments), one of which may be a no-treatment state—for example, $Y(0, \omega)$ is the outcome for a person ω not getting a treatment like a drug, schooling, or access to a new technology, while $Y(1, \omega)$ corresponds to person ω getting the drug, schooling or access.

Each "state" (treatment, policy) may consist of a compound of subcomponent states. In this case, we can define s as a vector (e.g., $s = (s_1, s_2, \ldots, s_k)$) corresponding to the different components that comprise treatment. Thus a job training program typically consists of a package of treatments. We might be interested in the package or one (or more) of its components. Thus s_1 may be months of vocational education, s_2 quality of training and so forth. The outcomes may be time subscripted as well, with $Y_t(s, \omega)$ corresponding to outcomes of treatment measured at different times. The index set for t may be the integers, corresponding to discrete time, or an interval, corresponding to continuous time.[5] The $Y_t(s, \omega)$ are realized or *ex post* (after treatment) outcomes. When choosing treatment, these values may not be known. Gill and Robins (2001), Abbring and Van Den Berg (2003), Lechner (2004), Heckman and Vytlacil (2006a,b), and Heckman and Navarro (2006) develop models for dynamic counterfactuals.

Each policy regime $p \in \mathcal{P}$ consists of a collection of possible treatments $\mathcal{S}_p \subseteq \mathcal{S}$. Different policy regimes may include some of the same subsets of $\mathcal{S}$. Associated with each policy regime is an assignment mechanism $\tau \in \mathcal{T}_p$, where $\mathcal{T}_p$ is the set of possible mechanisms under policy p. (Some policy regimes may rule out some assignment mechanisms.) The assignment mechanism determines the allocation of persons $\omega \in \Omega$ to treatment. It implicitly sets the scale of the program. The mechanism could include randomization so that the assignment mechanism would assign probabilities $\pi_s^\tau \in [0, 1]$ to each treatment $s \in \mathcal{S}_p$. Let Π_p denote the set of families $(\pi_s)_{s \in \mathcal{S}_p}$, $\pi_s \in [0, 1]$, such that $\Sigma_{s \in \mathcal{S}_p} \pi_s = 1$. Then,

$$\Phi^p : \Omega \times \mathcal{T}_p \to \Pi_p,$$

[5]In principle, in addition to indexing $\mathcal{S}$ by ω ($\mathcal{S}_\omega$) so there are person-specific treatment possibility sets, we could index by t ($\mathcal{S}_{\omega,t}$), but we assume, for simplicity, a common $\mathcal{S}$ for all ω and t.

where $\Phi^p(\omega, \tau) \in \Pi_p$ is a family of probabilities which we note alternatively $\left(\pi_s^\tau(\omega)\right)_{s \in \mathcal{S}_p}$. This signifies that, under policy p with assignment mechanism τ, person ω receives treatment s_p with probability $\pi_{s_p}^\tau(\omega)$. For each person ω, the special case of deterministic assignment sets $\pi_{s_0}^\tau(\omega) = 1$ for exactly one treatment $s_0 \in \mathcal{S}_p$ and sets $\pi_{s_0}^\tau(\omega) = 0$ for all $s \in \mathcal{S}_p \backslash \{s_0\}$.

For deterministic policy assignment rules, a universal policy may consist of a single treatment ($\mathcal{S}_p$ may consist of a single element). Treatment can include direct receipt of some intervention (e.g., a drug, education) as well as the tax payment for financing the treatment. For some persons, the assigned treatment may only be the tax payment. In the special case where some get no treatment ($\omega \in \Omega_0$) and others get treatment ($\omega \in \Omega_1$), and there are two elements in $\mathcal{S}_p$ (e.g., $\mathcal{S}_p = \{0, 1\}$), we produce the classical binary treatment-control comparison.

Two assumptions are often invoked in the literature.[6] In our notation, they are:

$$Y(s, \omega, p, \tau) = Y(s, \omega, p', \tau) = Y(s, \omega, \tau) \text{ for } s \in \mathcal{S}_p \cap \mathcal{S}_{p'}, \quad \tau \in \mathcal{T}_p \cap \mathcal{T}_{p'}, \text{ for all } p, p' \in \mathcal{P} \text{ and } \omega \in \Omega. \tag{A-1}$$

This assumption says that outcomes for person ω under treatment s with assignment mechanism τ are the same in two different policy regimes which both include s as a possible treatment. It rules out social interactions and general equilibrium effects. A second assumption rules out any effect of the assignment mechanism on potential outcomes.

$$\text{Irrespective of assignment mechanism } \tau, \text{ for all policies } p \in \mathcal{P}, Y(s, \omega, \tau) = Y(s, \omega) \text{ for all } s \in \mathcal{S}_p \text{ and } \omega \in \Omega, \text{ so the outcome is not affected by the assignment.} \tag{A-2}$$

This assumption maintains that the outcome is the same no matter what the choice of assignment mechanism. (A-2) rules out, among other things, the phenomenon of randomization bias discussed in Heckman, LaLonde, and Smith (1999) where agent behavior is

[6]See, e.g., Holland (1986) or Rubin (1986).

affected by the act of participating in an experiment. Such effects are also called "Hawthorne" effects.

Heckman, LaLonde, and Smith (1999) discuss the evidence against both assumptions. In much of this essay, I maintain these strong assumptions mostly to simplify the discussion. But the reader should be aware of the strong limitations imposed by these assumptions. Recent work in economics tests and relaxes these assumptions (see Heckman and Vytlacil 2006a).

Under these assumptions, the *individual-level treatment effect* for person ω comparing outcomes from treatment s with outcomes from treatment s' is

$$Y(s,\omega) - Y(s',\omega), \quad s \neq s', \tag{1}$$

where two elements are selected $s, s' \in \mathcal{S}$.[7] This is also called an *individual-level causal effect*. This may be a random variable or a constant. Our framework accommodates both interpretations. Thus the same individual with the same choice set and characteristics may have the same outcome in a sequence of trials or it may be random across trials. We discuss intrinsic variability at the individual level in Section 2.[8]

Other comparisons might be made. Comparisons can be made in terms of utilities (personal, $V(Y(s, \omega), \omega)$, or in terms of planner preferences, V_G). Thus one can ask if $V(Y(s, \omega), \omega) > V(Y(s', \omega), \omega)$ or not (is the person better off as a result of treatment s compared to treatment s'?) Treatments s and s' may be bundles of components

[7]One could define the treatment effect more generally as

$$Y(s,\omega,p,\tau) - Y(s',\omega,p,\tau).$$

This makes clear that the policy treatment effect is defined under a particular policy regime and for a particular mechanism of selection within a policy regime. One could define treatment effects for policy regimes or regime selection mechanisms by varying the arguments p or τ respectively, holding the other arguments fixed.

[8]There is a disagreement in the literature on whether or not the individual-level treatment effects are constants or random at the individual level. I develop both cases in this paper.

as previously discussed. One could define the treatment effect as $\mathbf{1}[V(Y(s, \omega), \omega) > V(Y(s', \omega), \omega)]$ where $\mathbf{1}[\cdot] = 1$ if the argument in brackets is true and is zero otherwise. These definitions of treatment effects embody Marshall's notion of *ceteris paribus*. Holding ω fixed holds all features about the person fixed except the treatment assigned s.[9]

Social welfare theory constructs aggregates over Ω or subsets of Ω (Sen 1999). A comparison of two policies $\{s_p(\omega)\}_{\omega\in\Omega}$ and $\{s_{p'}(\omega)\}_{\omega\in\Omega}$, using the social welfare function $V_G(\{Y(s(\omega), \omega)\}_{\omega\in\Omega})$, can be expressed as

$$V_G(\{Y(s_p, \omega), \omega\}_{\omega\in\Omega}) - V_G(\{Y(s_{p'}, \omega), \omega\}_{\omega\in\Omega}).$$

We can use an indicator function to denote when this term is positive: $\mathbf{1}[V_G(\{Y(s_p(\omega), \omega)\}_{\omega\in\Omega}) > V_G(\{Y(s_{p'}(\omega), \omega)\}_{\omega\in\Omega})]$. A special case of this analysis is cost-benefit analysis in economics where willingness to pay measures $W(s(\omega), \omega)$ are associated with each person. The cost-benefit comparison of two policies is

$$\textbf{Cost Benefit : } \mathbf{CB}_{p,p'} = \int_\Omega W(Y(s_p(\omega), \omega))d\mu(\omega) - \int_\Omega W(Y(s_{p'}(\omega), \omega))d\mu(\omega),$$

[9]One might compare outcomes in different sets that are ordered. Thus, for a particular policy regime and assignment mechanism, if $Y(s, \omega)$ is scalar income and we compare outcomes for $s \in \mathcal{S}_A$ with outcomes for $s' \in \mathcal{S}_B$, where $\mathcal{S}_A \cap \mathcal{S}_B = \varnothing$, then one might compare $Y_{s_A} - Y_{s_B}$, where

$$s_A = \arg\max_{s\in\mathcal{S}_A}(Y(s, \omega)) \quad \text{and} \quad s_B = \arg\max_{s\in\mathcal{S}_B}(Y(s, \omega)).$$

This compares the best in one choice set with the best in the other. A particular case is the comparison of the best choice with the next best choice. To do so, define $s' = \arg\max_{s\in\mathcal{S}}(Y(s, \omega))$, $\mathcal{S}_B = \mathcal{S} \setminus \{s'\}$ and define the treatment effect as $Y_{s'} - Y_{s_B}$. This is the comparison of the highest outcome over $\mathcal{S}$ with the next best outcome. In principle, many different individual level comparisons might be constructed, and they may be computed using personal preferences, V_ω, using the preferences of the planner, V_G, or using the preferences of the planner over preferences of agents.

where p, p' are two different policies, p' may correspond to a benchmark of no policy, and $\mu(\omega)$ is the distribution of ω.[10] The distribution $\mu(\omega)$ is constructed over the individual characteristics ω (e.g., age, sex, race, income) The Benthamite criterion replaces $W(Y(s(\omega), \omega))$ with $V(Y(s(\omega), \omega))$ in the preceding expressions and integrates utilities across persons:

$$\textbf{Benthamite : } \mathbf{B}_{p,p'} = \int_{\Omega} V(Y(s_p(\omega), \omega)) d\mu(\omega) - \int_{\Omega} V(Y(s_{p'}(\omega), \omega)) d\mu(\omega).$$

I now discuss a fundamental problem that arises in constructing these and other criteria from data. This takes me to the problem of causal inference, the second task delineated in Table 1. Recall that I am talking about inference in a population, not in a sample, so no issues of sampling variability arise.

1.3. *The Evaluation Problem*

Operating purely in the domain of theory, I have assumed a world with a well-defined set of individuals $\omega \in \Omega$ and a universe of counterfactuals or hypotheticals defined for each person $Y(s, \omega)$, $s \in \mathcal{S}$. Different policies $p \in \mathcal{P}$ select treatment for persons. Each policy can in principle assign treatment to persons by different mechanisms $\tau \in \mathcal{T}$. In the absence of a theory, there are no well-defined rules for constructing counterfactual or hypothetical states or constructing the assignment to treatment rules Φ^p_τ.[11] Scientific theories provide algorithms for generating the universe of internally consistent, theory-consistent counterfactual states.

These hypothetical states are possible worlds. They are products of a purely mental activity. No empirical problem arises in constructing these theoretically possible worlds. Indeed, in forecasting new policies, or projecting the effects of old policies to new

[10]These willingness-to-pay measures are standard in the economics literature (e.g., see Boadway and Bruce 1984).

[11]Efforts like those of Lewis (1974) to define admissible counterfactual states without an articulated theory as "closest possible worlds" founder on the lack of any meaningful metric or topology to measure "closeness" among possible worlds. Statisticians often appeal to this theory, but it is not operational (e.g., see Gill and Robins 2001 for one such appeal).

environments, some of the $Y(s, \omega)$ may have never been observed for anyone. Different theories produce different outcomes $Y(s, \omega)$ and different $\Phi^p_\tau(\omega)$.

The evaluation problem, in contrast to the model construction problem, is an identification problem that arises in constructing the counterfactual states and treatment assignment rules produced by abstract models from population data. This is the second task presented in Table 1.

This problem is not precisely stated until the data available to the analyst are precisely defined. Different subfields in science and social science assume access to different types of data. They also make different assumptions about the underlying models generating the counterfactuals and mechanisms for selecting which counterfactuals are actually observed.

At any point in time, we can observe person ω in one state but not in any of the other states. The states are mutually exclusive. Thus we do not observe $Y(s', \omega)$ for person ω if we observe $Y(s, \omega)$, $s \neq s'$. Let $D(s, \omega) = 1$ if we observe person ω in state s. Then $D(s', \omega) = 0$ for $s \neq s'$. $D(s, \omega)$ is generated by $\Phi^p_\tau(\omega) : D(s, \omega) = 1$ if $\Phi^p(\omega) = s$.

We observe $Y(s, \omega)$, if $D(s, \omega) = 1$ but we do not observe $Y(s', \omega)$, $s \neq s'$. We can define observed $Y(\omega)$ as

$$Y(\omega) = \sum_{s \in \mathcal{S}} D(s, \omega) Y(s, \omega).^{12} \tag{2}$$

Without further assumptions, constructing an empirical counterpart to equation (1) is impossible from the data on $(Y(\omega), D(\omega))$, $\omega \in \Omega$. This formulation of the evaluation problem is known as Quandt's switching regression model (Quandt 1958, 1974) and is attributed in statistics to Neyman (1923), Cox (1958), and Rubin (1978). A revision of it is formulated in a linear equations context for a continuum of treatments by Haavelmo (1943). The Roy model (Roy 1951) in economics is another version of it with two possible treatment outcomes ($\mathcal{S} = \{0,1\}$) and a scalar outcome measure and a particular selection mechanism $\tau \in \mathcal{T}$ which is that $D(1, \omega) = \mathbf{1}\,(Y(1, \omega) > Y(0, \omega))$ where "$\mathbf{1}[\cdot]$" is an indicator function which equals 1 when the event inside the

[12]In the general case, $Y(\omega) = \int_{\mathcal{S}} D(s, \omega) Y(s, \omega) ds$ where $D\ (s, \omega)$ is a Dirac function.

parentheses is true and is zero otherwise.[13] The mechanism of selection depends on the potential outcomes. Agents choose the sector with the highest outcome, so the actual selection mechanism is not a randomization.

Social experiments attempt to create assignment rules so that $D(s, \omega)$ is random with respect to $\{Y(s, \omega)\}_{s \in \mathcal{S}}$ for each ω (i.e., so that receipt of treatment is independent of the outcome of treatment). When agents self-select into treatment, rather than being randomly assigned, in general the $D(s, \omega)$ are not independent of $\{Y(s, \omega)\}_{s \in \mathcal{S}}$. This arises in the Roy model example. This selection rule creates the potential for *self-selection bias in inference*. We discuss this problem at length in Section 4.

The problem of self-selection is an essential aspect of the evaluation problem when data are generated by choices of agents. The agents making choices may be different from the agents receiving treatment (e.g., parents making choices for children). Such choices can include compliance with the protocols of a social experiment as well as ordinary choices about outcomes that people make in everyday life. Observe that in the Roy model, the choice of treatment (including the decision not to attrite from a program) is informative on the relative valuation of the $Y(s, \omega)$. This point is more general and receives considerable emphasis in the econometric literature but none in the statistical or epidemiological literature. Choices of treatment provide information on subjective relative evaluations of treatment by the decision maker and provides analysts with information on agent valuations of outcomes that are of independent interest.

A central problem considered in the literature on causal inference is the absence of information on outcomes for person ω other than the outcome that is observed. Even a perfectly implemented social experiment does not solve this problem (Heckman 1992) and, even under ideal conditions, randomization identifies only one component of $\{Y(s, \omega)\}_{s \in \mathcal{S}}$. In addition, even with ideal data and infinite samples some of the $s \in \mathcal{S}$ may not be observed if one is seeking to evaluate policies that produce new outcome states.

There are two main avenues of escape from this problem. The first, featured in explicitly formulated econometric models, often called "structural econometric analysis," is to model $Y(s, \omega)$ explicitly in terms of its determinants as specified by theory. This entails

[13] In terms of the assignment mechanism, $\Phi^p(\omega, \tau) = 1$ for ω such that $Y(1, \omega) > Y(0, \omega)$.

describing ω and carefully distinguishing what agents know and what the analyst knows. This approach also models $D(s, \omega)$—or $\Phi^p(\omega)$—and the dependence between $Y(s, \omega)$ and $D(s, \omega)$ produced from variables common to $Y(s, \omega)$ and $D(s, \omega)$. The Roy model, previously discussed, explicitly models this dependence.[14] Like all scientific models, this approach seeks to understand the factors underlying outcome, choice of outcome equations, and their relationship. Empirical models explicitly based on economic or social theory pursue this avenue of investigation. Some statisticians call this the "scientific approach" and are surprisingly hostile to it (Holland 1986).[15]

A second avenue of escape, and the one pursued in the recent epidemiological and statistical treatment effect literature, defines the problem away from estimating $Y(s, \omega)$ to be one of estimating some population version of equation (1), most often a mean, without modeling those factors giving rise to the outcome or the relationship between the outcomes and the mechanism selecting outcomes. Agent valuations of outcomes are ignored. The treatment effect literature focuses almost exclusively on policy problem P1 for the subset of outcomes that is observed. It ignores the problems of forecasting a policy in a new environment (problem P2) or a policy never previously experienced (problem P3). Forecasting the effects of new policies is a central task of science and public policy analysis that the treatment effect literature ignores.[16]

1.4. *Population-Level Treatment Parameters*

Constructing equation (1) or any of the other individual-level parameters defined in Section 1.2 for a given person is a difficult task because we rarely observe the same person ω in distinct s states. In addition, some of the states in $\mathcal{S}$ may not be experienced by anyone. The conventional approach in the treatment effect literature is to reformulate the parameter of interest to be some summary measure of the population distribution of treatment

[14]See Heckman and Honoré (1990) for a discussion of this model.

[15]I include in this approach methods based on panel data or more generally the method of paired comparisons, as applications of the scientific approaches. Under special conditions discussed in Heckman and Smith (1998), we can observe the same person in states s and s' in different time periods and can construct (1) for all ω.

[16]See Heckman and Vytlacil (2005) for one synthesis of the treatment effect and the structural literatures.

effects, most often a mean, or sometimes the distribution itself, rather than attempting to identify individual treatment effects. This approach focuses on presenting some summary measure of outcomes, not analyzing determinants of outcomes.[17] This approach also confines attention to the subsets of $\mathcal{S}$ that are observed states. Thus the objects of interest are redefined to be the distributions of $(Y(j, \omega) - Y(k, \omega))$ over ω, conditional on known components of ω, or certain means (or quantiles) of the distribution of $(Y(j, \omega) - Y(k, \omega))$ over ω, conditional on known components of ω (Heckman, Smith, and Clements 1997) or of $Y(j, \omega)$ and $Y(k, \omega)$ separately (Abadie, Angrist, and Imbens 2002). The standard assumptions in the treatment effect literature are that all states in $\mathcal{S}$ are observed, and that assumptions (A-1) and (A-2) hold (see Holland 1986; Rubin 1986).

The conventional parameter of interest, and the focus of many investigations in economics and statistics, is the average treatment effect (ATE). For program (treatment) j compared to program (treatment) k, this parameter is

$$ATE(j,k) = E_\omega(Y(j,\omega) - Y(k,\omega)), \tag{3a}$$

where "E_ω" means that we take expectations with respect to distribution of the factors generating outcomes and choices that characterize ω. Conditioning on covariates X, which are observed components associated with ω (and hence working with conditional distributions), this parameter is

$$ATE(j,k \mid x) = E_\omega(Y(j,\omega) - Y(k,\omega) \mid X = x). \tag{3b}$$

This is the effect of assigning a person to a treatment—taking someone from the overall population (3a) or a subpopulation conditional on X (3b) and determining the mean gain of the move from base state k, averaging over the factors that determine Y but are not captured by X. This parameter is also the effect of moving the society from a universal policy (characterized by policy k) and moving to a universal policy of j (e.g., from no social security to full population coverage). Such a policy would likely induce social interaction and general equilibrium effects that are

[17]The effects of causes and not the causes of effects, in the language of Holland (1986).

assumed away by (A-1) in the treatment effect literature and which, if present, fundamentally alter the interpretation placed on this parameter.

A second conventional parameter in this literature is the average effect of treatment on the treated. Letting $D(j, \omega) = 1$ denote receipt of treatment j, the conventional parameter is

$$TT(j,k) = E_\omega(Y(j,\omega) - Y(k,\omega) \mid D(j,\omega) = 1). \tag{4a}$$

For a population conditional on $X = x$, it is

$$TT(j,k \mid x) = E_\omega(Y(j,\omega) - Y(k,\omega) \mid D(j,\omega) = 1, X = x). \tag{4b}$$

These are, respectively, the mean impact of moving persons from k to j for those people who get treatment, unconditional and conditional on $X = x$.

A parallel pair of parameters for nonparticipants is treatment on the untreated, where $D(j, \omega) = 0$ denotes no treatment at level j:

$$TUT(j,k) = E_\omega(Y(j,\omega) - Y(k,\omega) \mid D(j,\omega) = 0) \tag{5a}$$

$$TUT(j,k \mid x) = E_\omega(Y(j,\omega) - Y(k,\omega) \mid D(j,\omega) = 0,\ X = x). \tag{5b}$$

These parameters answer (conditionally and unconditionally) the question of how extension of a program to nonparticipants as a group would affect their outcomes.[18]

The population treatment parameters just discussed are average effects: how the average in one treatment group compares with the average for another. The distinction between the marginal and average return has wide applicability in many areas of social science. The average student going to college may have higher earnings than the marginal student who is indifferent between going to school or not. It is often of interest to evaluate the impact of marginal extensions (or contractions) of a program. Incremental cost-benefit analysis is conducted in terms of marginal gains and benefits. The *effect of treatment for people at the margin of indifference* (*EOTM*) between

[18]Analogous to the pairwise comparisons, we can define setwise comparisons as is done in footnote 9.

j and k, given that these are the best two choices available is, with respect to personal preferences, and with respect to choice-specific costs $P(j, \omega)$,

$$EOTM_{\omega}^{V}(Y(j,\omega) - Y(k,\omega))$$
$$= E_{\omega}\left(Y(j,\omega) - Y(k,\omega) \,\middle|\, \begin{array}{l} V(Y(j,\omega),P(j,\omega),\omega) = V(Y(k,\omega),P(k,\omega),\omega); \\ V(Y(j,\omega),P(j,\omega),\omega) \\ V(Y(k,\omega),P(k,\omega),\omega) \end{array} \right\} \geq V(Y(l,\omega),P(l,\omega),\omega), \quad l \neq j,k \Bigg) . \tag{6}$$

This is the mean gain to people indifferent between j and k, given that these are the best two options available. In a parallel fashion, we can define $EOTM_{\omega}^{V_G}(Y(j) - Y(k))$ using the preferences of another person (e.g., the parent of a child or a paternalistic bureaucrat).[19]

A generalization of this parameter called the *marginal treatment effect*—developed in Heckman and Vytlacil (1999, 2000, 2005, 2006b), Heckman (2001), and estimated in Carneiro, Heckman, and Vytlacil (2005)—plays a central role in organizing and interpreting a wide variety of evaluation estimators. Many other mean treatment parameters can be defined depending on the choice of the conditioning set. Analogous definitions can be given for median and other quantile versions of these parameters (see Heckman, Smith, and Clements 1997; Abadie, Angrist, and Imbens 2002). Although means are conventional, distributions of treatment parameters are also of considerable interest, and we consider them in the next section.

Mean treatment effects play a special role in the statistical approach to causality. They are the centerpiece of the Rubin (1986)–Holland (1986) model and in many other studies in statistics and epidemiology. Social experiments with full compliance and no disruption can identify these means because of a special mathematical property of means. If we can identify the mean of $Y(j, \omega)$ and the mean of $Y(k, \omega)$ from an experiment where j is the treatment and k is the baseline, we can form the average treatment effect for j compared

[19]An analogous parameter can be defined for mean setwise comparisons as in footnote 9.

with k (3a). These can be formed over two different groups of people classified by their X values. By a similar argument, we can form the treatment on the treated parameter (TT) (4a) or (TUT) (5a) by randomizing over particular subsets of the population ($D = 1$ or $D = 0$, respectively) assuming full compliance and no randomization (disruption) bias. Disruption bias arises when the experiment itself affects outcomes $(Y(s, \omega))_{\omega \in \Omega}$ and (A-2) is violated.[20]

The case for randomization is weaker if the analyst is interested in other summary measures of the distribution, or the distribution itself. Experiments do not solve the problem that we cannot form $Y(s, \omega) - Y(s', \omega)$ for any person. Randomization is not an effective procedure for identifying median gains, or the distribution of gains, under general conditions. The elevation of population means to be the central population-level "causal" parameters promotes randomization as an ideal estimation method. By focusing exclusively on mean outcomes, the statistical literature converts a metaphor for outcome selection—randomization—into an ideal.

1.5. *Criteria of Interest Besides the Mean: Distributions of Counterfactuals*

Although means are traditional, the answer to many interesting policy evaluation questions requires knowledge of features of the distribution of program gains other than some mean. It is also of interest to know the following for scalar outcomes

a. The proportion of people taking the program j who benefit from it relative to some alternative k, $\Pr_\omega(Y(j, \omega) > Y(k, \omega) | D(j, \omega) = 1)$;
b. The proportion of the total population that benefits from the program k compared with program j, $\Pr_\omega(Y(j, \omega) > Y(k, \omega))$, sometimes called the *voting criterion*;
c. Selected quantiles of the impact distribution;[21]
d. The distribution of gains at selected base state values, (the distribution of $Y(j, \omega) - Y(k, \omega)$ given $Y(k, \omega) = y(k)$).

[20]Such disruptions leading to changed outcomes are also called Hawthorne effects; see Heckman (1992) and Heckman, LaLonde, and Smith (1999).

[21]$\inf \{\delta : F_\Delta(\delta) \geq q\}$ where q is a quantile of the distribution and F_Δ is the distribution function of $\Delta = Y(j, \omega) - Y(k, \omega)$.

Each of these measures can be defined conditional on observed characteristics X. Measure (a) is of interest in determining how widely program gains are distributed among participants. Voters in an electorate in a democratic society are unlikely to assign the same weight to two programs with the same mean outcome, one of which produced large favorable outcomes for only a few persons while the other distributed smaller gains more broadly. This issue is especially relevant if program benefits are not transferrable or if restrictions on feasible social redistributions prevent distributional objectives from being attained.

Measure (b) is the proportion of the entire population that benefits from a program. In a study of the political economy of interest groups, it is useful to know which groups benefit from a program and how widely distributed the program benefits are. Measure (c) reveals the gains at different percentiles of the impact distribution. Criterion (d) focuses on the distribution of impacts for subgroups of participants with particular outcomes in the nonparticipation state. Concerns about the impact of policies on the disadvantaged emphasize such criteria (Rawls 1971). All of these measures require knowledge of features of the joint distribution of outcomes for participants for their construction, not just the mean. Identifying distributions is a more demanding task than identifying means.

Distributions of counterfactuals are also required in computing the option values conferred by social programs.[22] Heckman and Smith (1998), Aakvik, Heckman, and Vytlacil (1999, 2005), Carneiro, Hansen, and Heckman (2001, 2003), and Cunha, Heckman, and Navarro (2005a) develop methods for identifying distributions of counterfactuals.

1.6. *Accounting for Private and Social Uncertainty*

Persons do not know the outcomes associated with possible states not yet experienced. If some potential outcomes are not known at the time treatment decisions are made, the best that agents can do is to forecast them with some rule. Even if, *ex post*, agents know their outcome in a benchmark state, they may not know it *ex ante*, and they may always

[22]Heckman, Smith, and Clements (1997) present estimates of the option values of social programs.

be uncertain about what they would have experienced in alternative states. This creates a further distinction between *ex ante* and *ex post* evaluations of both subjective and objective outcomes. This distinction is missing from the statistical treatment effect literature.

In the literature on social choice, one form of decision-making under uncertainty plays a central role. The *Veil of Ignorance* of Vickrey (1945, 1960) and Harsanyi (1955, 1975) postulates that individuals are completely uncertain about their position in the distribution of outcomes under each policy considered, or should act as if they are completely uncertain, and they should use expected utility criteria (Vickrey-Harsanyi) or a maximin strategy (Rawls 1971) to evaluate their welfare under alternative policies. Central to this viewpoint is the anonymity postulate that claims the irrelevance of any particular person's outcome to the overall evaluation of social welfare. This form of ignorance is sometimes justified as an ethically correct position that captures how an objectively detached observer should evaluate alternative policies even if actual participants in the political process use other criteria. An approach based on the Veil of Ignorance is widely used in applied work in evaluating different income distributions (see Foster and Sen 1998). It only requires information about the marginal distributions of outcomes produced under different policies. If the outcome is income, policy j is preferred to policy k if the income distribution under j stochastically dominates the income under k.[23]

An alternative criterion is required if it is desired to model social choices where persons act in their own self-interest, or in the interest of certain other groups (e.g., the poor, the less able) and have at least partial knowledge about how they (or the groups they are interested in) will fare under different policies. The outcomes in different regimes may be dependent so that persons who benefit under one policy may also benefit under another (see Carneiro, Hansen, and Heckman 2001, 2003).

Because agents typically do not possess perfect information, the simple voting criterion assuming perfect foresight discussed in Section 1.5 may not accurately predict choices and requires

[23]See Foster and Sen (1998) for a definition of stochastic dominance. It compares one distribution with another and determines which, if either, has more mass at favorable outcomes.

modification. Let $\mathcal{I}_\omega$ denote the information set available to agent ω. The agent evaluates policy j against k using that information. Under an expected utility criterion, person ω prefers policy j over k if

$$E_\omega(V(Y(j,\omega),\omega) \mid \mathcal{I}_\omega) > E_\omega(V(Y(k,\omega),\omega) \mid \mathcal{I}_\omega).$$

The proportion of people who prefer j is

$$PB(j \mid j,k) = \int \mathbf{1}\begin{bmatrix} E_\omega(V(Y(j,\omega),\omega)|\mathcal{I}_\omega) > \\ E_\omega(V(Y(k,\omega),\omega) \mid \mathcal{I}_\omega) \end{bmatrix} d\mu(\omega), \tag{7}$$

where $\mu(\omega)$ is the distribution of ω in the population.[24] The voting criterion previously discussed in Section 1.5 is the special case where $I_\omega = (Y(j, \omega), Y(k, \omega))$, so there is no uncertainty about $Y(j, \omega)$ and $Y(k, \omega)$. In the more general case, the expectation is computed against the distribution of $(E_\omega(V(Y(j, \omega), \omega|\mathcal{I}_\omega)), E_\omega(V(Y(k, \omega), \omega)|\mathcal{I}_\omega))$.[25]

Accounting for uncertainty in the analysis makes it essential to distinguish between *ex ante* and *ex post* evaluations. *Ex post*, part of the uncertainty about policy outcomes is resolved although individuals do not, in general, have full information about what their potential outcomes would have been in policy regimes they have not experienced and may have only incomplete information about the policy they have experienced (e.g., the policy may have long run consequences extending after the point of evaluation). It is useful to index the information set $\mathcal{I}_\omega$ by t, $\mathcal{I}_{\omega,t}$, to recognize that information about the outcomes of policies may accrue over time. *Ex ante* and *ex post* assessments of a voluntary program need not agree. *Ex post* assessments of a program through surveys administered to persons who have completed it (see Katz, Gutek, Kahn, and Barton 1975) may disagree with *ex ante* assessments of the program. Both may reflect honest valuations of the program but they are reported when agents have different information about it or have their preferences

[24]Persons would not necessarily vote "honestly," although in a binary choice setting they do and there is no scope for strategic manipulation of votes (see Moulin 1983). *PB* is simply a measure of relative satisfaction and need not describe a voting outcome when other factors come into play.

[25]See Cunha, Heckman, and Navarro (2005b) for computations regarding both types of joint distributions.

altered by participating in the program. Before participating in a program, persons may be uncertain about the consequences of participation. A person who has completed program j may know $Y(j, \omega)$ but can only guess at the alternative outcome $Y(k, \omega)$ which they have not experienced. In this case, *ex post* "satisfaction" with j relative to k for agent ω is synonymous with the inequality

$$V(Y(j,\omega),\omega) > E_\omega(V(Y(k,\omega),\omega) \mid \mathcal{I}_\omega), \tag{8}$$

where the information is post-treatment. Survey questionnaires about "client" satisfaction with a program may capture subjective elements of program experience not captured by "objective" measures of outcomes that usually exclude psychic costs and benefits. (Heckman, Smith, and Clements 1997 and Heckman and Smith 1998 present evidence on this question.) Carneiro, Hansen, and Heckman (2001, 2003), Cunha, Heckman, and Navarro (2005a,b), and Heckman and Navarro (2004, 2006) develop econometric methods for distinguishing *ex ante* from *ex post* evaluations of programs.

1.7. *Information Needed to Construct Various Criteria*

Four ingredients are required to implement the criteria discussed in this section: (1) private preferences, including preferences over outcomes by the decision maker; (2) social preferences, as exemplified by social welfare function $V_G(\{Y(s_p(\omega), \omega)\}_{\omega\in\Omega})$; (3) distributions of outcomes in alternative states, and for some criteria, such as the voting criterion, *joint* distributions of outcomes *across* policy states; and (4) *ex ante* and *ex post* information about outcomes. Cost-benefit analysis requires only information about means of measured outcomes and for that reason is easier to implement. The treatment effect literature in epidemiology and statistics largely focuses on means. Recent work in econometrics analyzes distributions of treatment effects (see Heckman, Smith, and Clements 1997; Carneiro, Hansen, and Heckman 2001, 2003; Cunha, Heckman, and Navarro 2005a). The rich set of questions addressed in this section contrasts sharply with the focus on mean outcome parameters in the epidemiology and statistics literatures, which ignore private and social preferences and ignore distributions of outcomes. Carneiro, Hansen, and Heckman (2001, 2003), Cunha, Heckman, and Navarro (2005a,b), and

Heckman and Navarro (2006) present methods for extracting private information on evaluations and their evolution over time. I now exposit more formally the econometric approach to formulating causal models.

2. COUNTERFACTUALS, CAUSALITY, AND STRUCTURAL ECONOMETRIC MODELS

This section formally defines structural models as devices for generating counterfactuals. I consider both outcome and treatment choice equations. The scientific model of econometrics is compared with the Neyman (1923)–Rubin (1978) model of causality that dominates discussions in epidemiology, in statistics, and in certain social sciences outside of economics. The structural equations approach and treatment effects approach are compared and evaluated.

2.1. *Generating Counterfactuals*

The treatment effect and structural approaches differ in the detail with which they specify counterfactual outcomes, $Y(s, \omega)$. The scientific approach embodied in the structural economics literature models the counterfactuals more explicitly than is common in the statistical treatment effect literature. This facilitates the application of theory to provide interpretation of counterfactuals and comparison of counterfactuals across empirical studies using basic parameters of social theory. These models also suggest strategies for identifying parameters (task 2 in Table 1). Models for counterfactuals are the basis for extending historically experienced policies to new environments and for forecasting the effects of new policies never previously experienced. These are policy questions P2 and P3 stated in Section 1.

Models for counterfactuals are in the mind. They are internally consistent frameworks derived from theory. Verification and identification of these models from data are separate tasks from the purely theoretical act of constructing internally consistent models. No issue of sampling, inference, or selection bias is entailed in constructing theoretical models for counterfactuals.

The traditional model of econometrics is the "all causes" model.[26] It writes outcomes as a deterministic function of inputs:

$$y(s) = g_s(x, u_s), \tag{9}$$

where x and u_s are fixed variables specified by the relevant economic theory for person ω.[27] All outcomes are explained in a functional sense by the arguments of g_s in equation (9). If we model the *ex post* realizations of outcomes, it is entirely reasonable to invoke an all causes model because *ex post* all uncertainty has been resolved. Equation (9) is a "production function" relating inputs (factors) to outputs (outcomes). The notation x and u_s anticipates the econometric problem that some arguments of functional relationship (9) are observed while other arguments may be unobserved by the analyst. In the analysis of this section, their roles are symmetric.

My notation allows for different unobservables from a common list u to appear in different outcomes.[28] g_s maps (x, u_s) into y. The domain of definition $\mathcal{D}$ of g_s may differ from the empirical support. Thus we can think of (9) as mapping logically possible inputs into logically possible *ex post* outcomes, but in a real sample we may observe only a subset of the domain of definition.

A "deep structural" version of (9) models the variation across the g_s in terms of s as a function of generating characteristics c_s that capture what "s" is:[29]

$$y(s) = g(c_s, x, u_s). \tag{10}$$

The components c_s provide the basis for generating the counterfactuals across treatments from a base set of characteristics. This approach models different treatments as consisting of different bundles of characteristics. g maps c, s, u_s into $y(s)$, where the domain of definition $\mathcal{D}$ of g may differ from its empirical support. Different treatments s are characterized by different bundles of the same characteristics that generate all outcomes. This framework provides the

[26]This term is discussed in Dawid (2000).

[27]Denote $\mathcal{D}$ as the domain of $g_s : \mathcal{D} \rightarrow \mathcal{R}^y$ where $\mathcal{R}^y$ is the range of y.

[28]An alternative notation would use a common u and let g_s select out s-specific components.

[29]Now the domain of g, $\mathcal{D}$, is defined for c_s, x, u_s and $g : \mathcal{D} \rightarrow \mathcal{R}^y$.

basis for solving policy problem P3 since new policies (treatments) are generated as different packages of common characteristics, and all policies are put on a common basis. If a new policy is characterized by known transformations of (c, x, u_s) that lie in the known empirical support of g, policy forecasting problem P3 can be solved.[30] This point is discussed further in the Appendix.

Part of the *a priori* specification of a causal model is the choice of the arguments of the functions g_s and g. Analysts may disagree about appropriate arguments to include based on alternative theoretical frameworks. One benefit of the statistical approach that focuses on problem P1 is that it works solely with the outcomes rather than the inputs. However, it is silent on how to solve problems P2 and P3 and provides no basis for interpreting the population-level treatment effects.

Consider alternative models of schooling outcomes of pupils where s indexes the schooling type (e.g., regular public, charter public, private secular, and private parochial). The c_s are the observed characteristics of schools of type s. The x are the observed characteristics of the pupil. The u_s are the unobserved characteristics of both the schools and the pupil. If we can characterize a proposed new type of school as a new package of different levels of the same ingredients x, c_s, and u_s and we can identify (10) over the domain defined by the new package, we can solve problem P3. If the same schooling input (same c_s) is applied to different students (those with different x) and we can identify (9) or (10) over the new domain of definition, we solve problem P2. By digging deeper into the "causes of the effects" we can do more than just compare the effects of treatments in place with each other. In addition, as we shall see, modeling the u_s and its relationship with the corresponding unobservables in the treatment choice equation is informative on appropriate identification strategies.

Equations (9) and (10) describing *ex post* outcomes are sometimes called Marshallian causal functions (see Heckman 2000). Assuming that the components of (x, u_s) or (c_s, x, u_s) can be independently varied or are variation-free,[31] a feature that may or may not be

[30]See Heckman and Vytlacil (2005, 2006a).

[31]The requirement is that if $(\mathcal{X},\mathcal{U})$ or $(\mathcal{C},\mathcal{X},\mathcal{U})$ are the domains of (9) and (10), $(\mathcal{X},\mathcal{U}) = (\mathcal{X}_1 \times \cdots \times \mathcal{X}_N \times \mathcal{U}_1 \times \cdots \times \mathcal{U}_M)$ or $(\mathcal{C},\mathcal{X},\mathcal{U}) = (\mathcal{C}_1 \times \cdots \times \mathcal{C}_K \times \mathcal{X}_1 \times \cdots \times \mathcal{X}_N \times \mathcal{U}_1 \times \cdots \times \mathcal{U}_M)$, where we assume K components in $\mathcal{C}$, N components in $\mathcal{X}$, and M components in $\mathcal{U}$. This means that we can vary one variable without necessarily varying another.

produced by the relevant theory, we may vary each argument of these functions to obtain a causal effect of that argument on the outcome. These thought experiments are for hypotheticals.

Changing one coordinate while fixing the others produces a Marshallian *ceteris paribus* causal effect of a change in that coordinate on the variable. Varying c_s sets different treatment levels. Variations in x,u_s among persons explains why people facing the same characteristics c_s respond differently to the same treatment s. Variations in u_s not observed by the analyst explain why people with the same x values respond differently.

The *ceteris paribus* variation used to define causal effects need not be for a single variable of the function. A treatment generally consists of a package of characteristics and if we vary the package from c_s to $c_{s'}$, we get different treatment effects.

I use lowercase notation produced from the theory to denote fixed values. I use uppercase notation to denote random variables. In defining equations (9) and (10), I have explicitly worked with fixed variables that are manipulated in a hypothetical way as in algebra or elementary physics. In a purely deterministic world, agents would act on these nonstochastic variables. Even if the world is uncertain, *ex post*, after the realization of uncertainty, the outcomes of uncertain inputs are deterministic. Some components of u_s may be random shocks realized after decisions about treatment are made.

Thus if uncertainty is a feature of the environment, equations (9) and (10) can be interpreted as *ex post* realizations of the counterfactual as uncertainty is resolved. *Ex ante* versions of these relationships may be different. From the point of view of agent ω with information set $\mathcal{I}_\omega$, the *ex ante* expected value of $Y(s, \omega)$ is,[32]

$$E(Y(s,\omega) \mid \mathcal{I}_\omega) = E(g(C_s(\omega), X(\omega), U(s,\omega)) \mid \mathcal{I}_\omega), \tag{11}$$

where C_s, X, U_s are random variables generated from a distribution that depends on the agent's information set, indexed by $\mathcal{I}_\omega$. This distribution may differ from the distribution produced by "reality"

[32]The expectation might be computed using the information sets of the relevant decision maker (e.g., the parents in the case of the outcomes of the child) who might not be the agent whose outcomes are measured. These random variables are drawn from agent ω's subjective distribution.

or nature if agent expectations are different from objective reality.[33] In the presence of intrinsic uncertainty, the relevant decision maker acts on equation (11), but the *ex post* counterfactual is

$$Y(s,\omega) = E(Y(s,\omega) \mid \mathcal{I}_\omega) + \nu(s,\omega), \tag{12}$$

where $\nu(s, \omega)$ satisfies $E(\nu(s, \omega)|\mathcal{I}_\omega) = 0$. In this interpretation, the information set of agent ω before realizations occur, $\mathcal{I}_\omega$, is part of the model specification. This discussion clarifies the distinction between deterministic (*ex post*) outcomes and intrinsically random (*ex ante*) outcomes discussed in Section 1.

This statement of the basic deterministic model reconciles the all causes model (9) and (10) with a model of intrinsic uncertainty favored by some statisticians (see Dawid 2000 and the following discussion). *Ex ante*, there is uncertainty at the agent (ω) level but *ex post* there is not. Realization $\nu(s, \omega)$ is an ingredient of the *ex post* all causes model but not the subjective *ex ante* all causes model. The probability law used by the agent to compute the expectation of $C_s(\omega)$, $X(\omega)$, $U_s(\omega)$ may differ from the objective distribution, i.e., the distribution that generates the observed data. In the *ex ante* all causes model, manipulations of $\mathcal{I}_\omega$ define the *ex ante* Marshallian causal parameters.

Thus from the point of view of the agent we can vary elements in $\mathcal{I}_\omega$ to produce Marshallian *ex ante* causal response functions. The *ex ante* treatment effect from the point of view of the agent for treatment s and s' is

$$E(Y(s,\omega)|\mathcal{I}_\omega) - E(Y(s',\omega) \mid \mathcal{I}_\omega). \tag{13}$$

However, agents may not act on these *ex ante* effects if they have decision criteria (utility functions) that are not linear in $Y(s, \omega)$, $s = 1, \ldots, \bar{S}$. I discuss *ex ante* valuations of outcomes in the next section.

The value of the scientific (or explicitly structural) approach to the construction of counterfactuals is that it explicitly models the unobservables and the sources of variability among observationally

[33]Thus agents do not necessarily use rational expectations, so the distribution used by the agent to make decisions need not equal the distribution generating the data.

identical people. Since it is the unobservables that give rise to selection bias and problems of inference that are central to empirically rigorous causal analysis, analysts using the scientific approach can draw on scientific theory and in particular choice theory to design and justify methods to control for selection bias. This avenue is not available to adherents of the statistical approach. Statistical approaches that are not explicit about the sources of the unobservables make strong implicit assumptions which, when carefully exposited, are often unattractive. We exposit some of these assumptions in Section 5.

The models for counterfactuals—equations (9)–(13)—are derived from theory. The arguments of these functions are varied by hypothetical manipulations to produce outcomes. These are thought experiments. When analysts attempt to construct counterfactuals empirically, they must carefully distinguish between these theoretical relationships and the empirical relationships determined by the available evidence.

The data used to determine these functions may be limited in their support. (The support is the region of the domain of definition where we have data on the function.)[34] In this case we cannot fully identify the theoretical relationships. In addition, in the support, the components of X, U_s and $\mathcal{I}_\omega$ may not be variation-free even if they are in the hypothetical domain of definition of the function. A good example is the problem of multicollinearity. If the X in a sample are linearly dependent, it is not possible to identify the Marshallian causal function with respect to variations in x over the available support even if we can imagine hypothetically varying the components of x over the domains of definition of the functions (9) or (10).

Thus in the available data (i.e., over the empirical support), one of the X (gender) may be perfectly predictable by the other X. With limited empirical supports that do not match the domain of definition of the outcome equations, one may not be able to identify the Marshallian causal effect of gender even though one can define it in some hypothetical model. In empirical samples, gender may be predictable in a statistical sense by other empirical factors. Holland's 1986 claim that the causal effects of race or gender are meaningless conflates an empirical problem (task 2 in Table 1) with a problem of theory (task 1 in Table 1). The scientific

[34]Thus if $\mathcal{D}_x$ is the domain of x, the support of x is the region $Supp\ (x) \subset \mathcal{D}_x$ such that the data density $f\ (x)$ satisfies the condition $f\ (x) > 0$ for $x \in Supp\ (x)$.

approach sharply distinguishes these two issues. One can in theory define the effect even if one cannot identify it from population or sample data.

I next turn to an important distinction between fixing and conditioning on factors that gets to the heart of the distinction between causal models and correlational relationships. This point is independent of any problem with the supports of the samples compared to the domains of definition of the functions.

2.2. *Fixing Versus Conditioning*

The distinction between *fixing* and *conditioning* on inputs is central to distinguishing true causal effects from spurious causal effects. In an important paper, Haavelmo (1943) made this distinction in linear equations models. It is the basis for Pearl's (2000) book on causality that generalizes Haavelmo's analysis to nonlinear settings. Pearl defines an operator "do" to represent the mental act of fixing a variable to distinguish it from the action of conditioning which is a statistical operation. If the conditioning set is sufficiently rich, fixing and conditioning are the same in an *ex post* all causes model.[35] Pearl suggests a particular physical mechanism for fixing variables and operationalizing causality, but it is not central to his or any other definition of causality. Pearl's analysis conflates the three tasks of Table 1.

An example of fixing versus conditioning is most easily illustrated in a linear regression model of the type analyzed by Haavelmo (1943). Let $y = x\beta + u$. Although both y and u are scalars, x may be a vector. The linear equation maps (x, u) into y: $(x, u) \mapsto y$. Suppose that the support of random variable (X, U) in the data is the same as the domain of (x, u) that are fixed in the hypothetical thought experiment and that the (x, u) are variation-free (i.e., they can be independently varied coordinate by coordinate). Thus we abstract from the problem of limited support that is discussed in the preceding section. We may write (dropping the "ω" notation for random variables)

$$Y = X\beta + U.$$

[35]Florens and Heckman (2003) carefully distinguish conditioning from fixing, and generalize Pearl's analysis to both static and dynamic settings.

Here "nature" or the "real world" picks (X, U) to determine Y. X is observed by the analyst and U is not observed, and (X, U) are random variables. This is an all causes model in which $(X, U) \mapsto Y$. The variation generated by the hypothetical model varies one coordinate of (X, U), fixing all other coordinates to produce the effect of the variation on the outcome Y. Nature (as opposed to the model) may not permit such variation.

Formally, we can write this model formulated at the population level as a conditional expectation,

$$E(Y|X, U) = X\beta + U.$$

Since we condition on both X and U, there is no further source of variation in Y. This is a deterministic model that coincides with the all causes model. Thus on the support, which is also assumed to be the domain of definition of the function, this model is the same model as the deterministic, hypothetical model, $y = x\beta + u$. Fixing X at different values corresponds to doing different thought experiments with the X. Fixing and conditioning are the same in this case.

If, however, we only condition on X in the sample, we obtain

$$E(Y|X) = X\beta + E(U|X).^{36} \tag{14}$$

This relationship does not generate U-constant (Y, X) relationships. It generates only an X-constant relationship. Unless we condition on all of the "causes" (the right hand side variables), the empirical relationship (14) does not identify causal effects of X on Y. The variation in X also moves the conditional mean of U unless U is independent of X.

This analysis readily generalizes to a general nonlinear model $y = g(c, x, u)$. A model specified in terms of random variables C, X, U with the same support as c, x, u has as its conditional expectation $g(C, X, U)$ under general conditions. Conditioning only on C, X does not in principle identify $g(c, x, u)$ or any of its derivatives (if they exist) or differences of outcomes defined in terms of c and x.

[36] I assume that the mean of U is finite.

Conditioning and fixing on the arguments of g or g_s are the same in an "all causes" model if all causes are accounted for. Otherwise, they are not the same. This analysis can be generalized to account for the temporal resolution of uncertainty if we include $\nu(s, \omega)$ as an argument in the *ex post* causal model. The outcomes can include both objective outcomes $Y(s, \omega)$ and subjective outcomes $V(Y(s, \omega), \omega)$.

Statisticians and epidemiologists have great difficulty with the distinction between fixing and conditioning because they typically define the models they analyze in terms of some type of conditioning. However, thought experiments in models of hypotheticals that vary factors are distinct from variations in conditioning variables that conflate the effects of variation in X, holding U fixed, with the effects of X in predicting the unobserved factors (the U) in the outcome equations.

2.3. *Modeling the Choice of Treatment*

Parallel to the models for outcomes are models for the choice of treatment. Consider *ex ante* personal valuations of outcomes based on expectations of gains from receiving treatment s:

$$E[V(Y(s,\omega), P(s,\omega), C_s(\omega), \omega)|\mathcal{I}_\omega], s \in S,$$

where $P(s, \omega)$ is the price or cost the agent must pay for participation in treatment s. We write $P(s, \omega) = K(Z(s, \omega), \eta(s, \omega))$. I allow utility V to be defined over the characteristics that generate the treatment outcome (e.g., quality of teachers in a schooling choice model) as well as other attributes of the consumer. In parallel with the g_s function generating the $Y(s, \omega)$, we write

$$V(Y(s,\omega), P(s,\omega), C_s(\omega), \omega) = f(Y(s,\omega), Z(s,\omega), C_s(\omega), \eta(s,\omega), \omega).$$

Parallel to the analysis of outcomes, we may keep $C_s(\omega)$ implicit and use f_s functions instead of f.

My analysis includes both measured and unmeasured attributes. The agent computes expectations against his/her subjective distribution of information. I allow for imperfect information by postulating an ω-specific information set. If agents know all

components of future outcomes, the uppercase letters become lowercase variables that are known constants. The $\mathcal{I}_\omega$ are the causal factors for ω. In a utility-maximizing framework, choice $\hat{s}$ is made if $\hat{s}$ is maximal in the set of valuations of potential outcomes:

$$\{E[V(Y(s,\omega), P(s,\omega), C_s(\omega), \omega)|\mathcal{I}_\omega] : s \in S\}.$$

In this interpretation, the information set plays a key role in specifying agent preferences. Actual realizations may not be known at the time decisions are made. Accounting for uncertainty and subjective valuations of outcomes (e.g., pain and suffering for a medical treatment) is a major contribution of the scientific approach. The factors that lead an agent to participate in treatment s may be dependent on the factors affecting outcomes. Modeling this dependence is a major source of information used in the scientific approach to constructing counterfactuals from real data, as I demonstrate in Section 4. A parallel analysis can be made if the decision maker is not the same as the agent whose objective outcomes are being evaluated.

2.4. *The Scientific Model Versus the Neyman–Rubin Model*

Many statisticians and social scientists invoke a model of counterfactuals and causality attributed to Donald Rubin by Paul Holland (1986) but which actually dates back to Neyman (1923).[37] Neyman and Rubin postulate counterfactuals $\{Y(s, \omega)\}_{s\in\mathcal{S}}$ without modeling the factors determining the $Y(s, \omega)$ as I have done in equations (9)–(12), using the scientific, structural approach. Rubin and Neyman offer no model of the choice of which outcome is selected. Thus there no "lowercase," all causes models explicitly specified in this approach, nor is there any discussion of the science or theory producing the outcomes studied.

In my notation, Rubin assumes (A-1) and (A-2) as presented in Section 1.[38] Recall that (A-1) assumes no general equilibrium effects or social interactions among agents. Thus the outcome for the person is the

[37]The framework attributed to Rubin was developed in statistics by Neyman (1923), Cox (1958), and others. Parallel frameworks were independently developed in psychometrics (Thurstone 1930) and economics (Haavelmo 1943; Roy 1951; Quandt 1958, 1972).

[38]Rubin (1986) calls these two assumptions "SUTVA" for Stable Unit Treatment Value Assumption.

same whether one person receives treatment or many receive treatment. (A-2) says that however ω receives s, the same outcome arises. (A-2) also rules out randomization bias where the act of randomization affects the potential outcomes.[39]

More formally, the Rubin model assumes the following:

> **R-1** $\{Y(s, \omega)\}_{s \in \mathcal{S}}$, a set of counterfactuals defined for *ex post* outcomes (no valuations of outcomes or specification of treatment selection rules).
> **R-2** (A-1) (No social interactions).
> **R-3** (A-2) (Invariance of counterfactual to assignment mechanism of treatment).
> **R-4** P1 is the only problem of interest.
> **R-5** Mean causal effects are the only objects of interest.
> **R-6** There is no simultaneity in causal effects, i.e., outcomes cannot cause each other reciprocally (see Holland 1988).

The scientific model (1) decomposes the $Y(s, \omega)$, $s \in \mathcal{S}$ into its determinants; (2) considers valuation of outcomes as an essential ingredient of any study of causal inference; (3) models the choice of treatment and uses choice data to infer subjective valuations of treatment; (4) uses the relationship between outcomes and treatment choice equations to motivate, justify, and interpret alternative identifying strategies; (5) explicitly accounts for the arrival of information through *ex ante* and *ex post* analyses; (6) considers distributional causal parameters as well as mean effects; (7) addresses problems P1–P3; (8) allows for nonrecursive (simultaneous) causal models. I develop nonrecursive models in the next section.

In the Neyman–Rubin model, the sources of variability generating $Y(s, \omega)$ as a random variable are not specified. The "causal effect" of s compared to s' is defined as the treatment effect in equation (1). Holland (1986, 1988) argues that it is an advantage of the Rubin model that it is not explicit about the sources of variability among observationally identical people, or about the factors that

[39]See Heckman (1992) or Heckman, LaLonde, and Smith (1999) for discussions and evidence on this question.

generate $Y(s, \omega)$. Holland and Rubin focus on mean treatment effects as the interesting causal parameters.

The scientific (econometric) approach to causal inference supplements the model of counterfactuals with models of the choice of counterfactuals $\{D(s, \omega)\}_{s \in \mathcal{S}}$ generated by the maps $\Phi^p_\tau(\omega)$ and the relationship between choice equations and the counterfactuals. The $D(s, \omega)$ are assumed to be generated by the collection of random variables $(C_s(\omega), Z(s, \omega), \eta(s, \omega), Y(s, \omega)|\mathcal{I}_\omega)$, $s \in \mathcal{S}$, where $C_s(\omega)$ is the characteristic of the treatment s for person ω, $Z(s, \omega)$ are observed determinants of costs, the $\eta(s, \omega)$ are unobserved (by the analyst) cost (or preference) factors and $Y(s, \omega)$ are the outcomes, and the "|" denotes that these variables are defined conditional on $\mathcal{I}_\omega$ (the agent's information set).[40] Along with the *ex ante* valuations that generate $D(s, \omega)$ are the *ex post* valuations discussed in Section 1.6.

Random utility models generating $D(s, \omega)$ go back to Thurstone (1930) and McFadden (1974, 1981).[41] The full set of counterfactual outcomes for each agent is assumed to be unobserved by the analyst. It is the dependence of unmeasured determinants of treatment choices with unmeasured determinants of potential outcomes that gives rise to selection bias in empirically constructing counterfactuals and treatment effects, even after conditioning on the observables. Knowledge of the relationship between choices and counterfactuals suggests appropriate methods for solving selection problems. By analyzing the relationship of the unobservables in the outcome equation, and the unobservables in the treatment choice equation, the analyst can use *a priori* theory to devise appropriate estimators to identify causal effects.

The scientific approach is more general than the Neyman–Rubin model because it emphasizes the welfare of the agents being studied (through V_G or $V(Y(s, \omega), \omega)$)—the "subjective evaluations"—as well as the objective evaluations. The econometric approach also

[40] If other agents make the treatment assignment decisions, then the determinants of $D(s, \omega)$ are modified according to what is in their information set.

[41] Corresponding to these random variables are the deterministic all causes counterparts $d(s)$, c_s, $z(s)$, $\eta(s)$, $\{y(s)\}$, i, where the $(\{z(s)\}_{s \in \mathcal{S}}, \{c_s\}_{s \in \mathcal{S}}, \{\eta(s)\}_{s \in \mathcal{S}}, \{y(s)\}_{s \in \mathcal{S}}, i)$ generate the $d(s) = 1$ if $(\{z(s)\}_{s \in \mathcal{S}}, \{c_s\}_{s \in \mathcal{S}}, \{\eta(s)\}_{s \in \mathcal{S}}, \{y(s)\}_{s \in \mathcal{S}}) \in \Psi$, a subset of the domain of the generators of $d(s)$. Again the domain of definition of $d(s)$ is not necessarily the support of $z(s, \omega), c_s(\omega), \eta(s, \omega), \{Y(s, \omega)\}_{s \in \mathcal{S}}$ and $\mathcal{I}_\omega$.

distinguishes *ex ante* from *ex post* subjective evaluations, so it can measure both agent satisfaction and regret.[42]

In addition, modelling $Y(s, \omega)$ in terms of characteristics of treatment, and of the treated, facilitates comparisons of counterfactuals and derived causal effects across studies where the composition of programs and treatment group members may vary. It also facilitates the construction of counterfactuals on new populations and the construction of counterfactuals for new policies. The Neyman–Rubin framework focuses exclusively on population-level mean "causal effects" or treatment effects for policies actually experienced and provides no framework for extrapolation of findings to new environments or for forecasting new policies (problems P2 and P3). Its focus on population mean treatment effects elevates randomization and matching to the status of preferred estimators. Such methods cannot identify distributions of treatment effects or general quantiles of treatment effects.

Another feature of the Neyman–Rubin model is that it is recursive. It cannot model causal effects of outcomes that occur simultaneously. I now present a model of simultaneous causality.

2.5. *Nonrecursive (Simultaneous) Models of Causality*

A system of linear simultaneous equations captures interdependence among outcomes Y. For simplicity, I focus on *ex post* outcomes so I ignore the revelation of information over time. To focus on the main ideas of this section, I assume that the domain of definition of the model is the same as the support of the population data. Thus the model for values of uppercase variables has the same support as the domain of definition for the model in terms of lowercase variables.[43] The model developed in this section is rich enough to model interactions among agents.[44] I write this model in terms of parameters (Γ, B), observables (Y, X), and unobservables U as

$$\Gamma Y + BX = U, \qquad E(U) = 0, \tag{15}$$

[42]See Cunha, Heckman, and Navarro (2005a,b) for estimates of subjective evaluations and regret in schooling choices.

[43]This approach merges tasks 1 and 2 in Table 1. I do this here because the familiarity of the simultaneous equations model as a statistical model makes the all causes *ex post* version confusing to many readers familiar with this model.

[44]For simplicity, I work with the linear model in the text, developing the nonlinear case in footnotes.

where Y is now a vector of endogenous and interdependent variables, X is exogenous ($E(U|X) = 0$), and Γ is a full rank matrix. A better nomenclature, suggested by Leamer (1985), is that the Y are internal variables determined by the model and the X are external variables specified outside the model.[45] This definition distinguishes two issues: (1) defining variables (Y) that are determined from inputs outside the model (the X) and (2) determining the relationship between observables and unobservables.[46] When the model is of full rank (Γ^{-1} exists), it is said to be "complete." A complete model produces a unique Y from a given (X, U). A complete model is said to be in reduced form when equation (15) is multiplied by Γ^{-1}. The reduced form is $Y = \Pi X + R$ where $\Pi = -\Gamma^{-1}B$ and $R = \Gamma^{-1}U$.[47] This is a linear-in-parameters "all causes" model for vector Y, where the causes are X and R. The "structure" is (Γ, B), Σ_U, where Σ_U is the variance-covariance matrix of U. The reduced form slope coefficients are Π, and Σ_R is the variance-covariance matrix of R.[48] In the population generating (15), least squares recovers Π provided Σ_X, the variance of X, is nonsingular (no multicollinearity). In this linear-in-parameters equation setting, the full rank condition for Σ_X is a variation-free condition on the external variables. The reduced form solves out for the dependence among the Y. The linear-in-parameters model is traditional. Nonlinear versions are available (Fisher 1966; Matzkin 2004).[49] For simplicity, I stick to the linear version, developing the nonlinear version in footnotes.

The structural form (15) is an all causes model that relates in a deterministic way outcomes (internal variables) to other outcomes (internal variables) and external variables (the X and U). Without some restrictions, certain *ceteris paribus* manipulations associated

[45]This formulation is static. In a dynamic framework, Y_t would be the internal variables and the lagged Y, Y_{t-k}, $k > 0$, would be external to period t and be included in the X_t. Thus we could work with lagged dependent variables. The system would be $\Gamma Y_t + BX_t = U_t$, $E(U_t) = 0$.

[46]In a time-series model, the internal variables are Y_t determined in period t.

[47]In this section only, Π refers to the reduced form coefficient matrix and not the set of policies Π_p, as in earlier sections.

[48]The original formulations of this model assumed normality so that only means and variances were needed to describe the joint distributions of (Y, X).

[49]The underlying all causes model writes $\Gamma y + Bx = u$, $y = \Pi x + r$ and $\Pi = -\Gamma^{-1}\ B$, $r = \Gamma^{-1}u$. Recall that I assume that the domain of the all causes model is the same as the support of (x, u). Thus there is a close correspondence between these two models.

with the effect of some components of Y on other components of Y are not possible within the model. I now demonstrate this point.

For specificity, consider a two-person model of social interactions. Y_1 is the outcome for person 1; Y_2 is the outcome for person 2. This could be a model of interdependent consumption where the consumption of person 1 depends on the consumption of person 2 and other person-1-specific variables (and possibly other person-2-specific variables). It could also be a model of test scores. We can imagine populations of data generated from sampling the same two-person interaction over time or sampling different two-person couplings at a point in time.

Assuming that the preferences are interdependent, we may write

$$Y_1 = \alpha_1 + \gamma_{12} Y_2 + \beta_{11} X_1 + \beta_{12} X_2 + U_1 \tag{16a}$$

$$Y_2 = \alpha_2 + \gamma_{21} Y_1 + \beta_{21} X_1 + \beta_{22} X_2 + U_2. \tag{16b}$$

This model is sufficiently flexible to capture the notion that the consumption of person 1 (Y_1) depends on the consumption of person 2 (if $\gamma_{12} \neq 0$), as well as person 1's value of X (if $\beta_{11} \neq 0$), X_1 (assumed to be observed), person 2's value of X, X_2 (if $\beta_{12} = 0$), and unobservable factors that affect person 1 (U_1). The determinants of person 2's consumption are defined symmetrically. I allow U_1 and U_2 to be freely correlated. I assume that U_1 and U_2 are mean independent of (X_1, X_2) so

$$E(U_1 | X_1, X_2) = 0 \tag{17a}$$

and

$$E(U_2 | X_1, X_2) = 0. \tag{17b}$$

Completeness guarantees that (16a) and (16b) have a determinate solution for (Y_1, Y_2).

Applying Haavelmo's argument to (16a) and (16b), the causal effect of Y_2 on Y_1 is γ_{12}. This is the effect on Y_1 of fixing Y_2 at different values, holding constant the other variables in the equation. Symmetrically, the causal effect of Y_1 on Y_2 is γ_{21}. Conditioning,—that is, using least squares—which is the method of matching, in general fails to identify these causal effects because U_1 and U_2 are correlated with Y_1 and Y_2. This is a traditional argument. It is based on the correlation between Y_2 and U_1. But even if $U_1 = 0$ and $U_2 = 0$, so that there are no

unobservables, matching or least squares breaks down because Y_2 is perfectly predictable by X_1 and X_2. We cannot simultaneously vary Y_2, X_1, and X_2. This is the essence of the problem of defining a causal effect. To see why, we derive the reduced form of this model.

Assuming completeness, the reduced form outcomes of the model after social interactions are solved out can be written as

$$Y_1 = \pi_{10} + \pi_{11}X_1 + \pi_{12}X_2 + R_1 \tag{18a}$$

$$Y_2 = \pi_{20} + \pi_{21}X_1 + \pi_{22}X_2 + R_2. \tag{18b}$$

Least squares (matching) can identify the *ceteris paribus* effects of X_1 and X_2 on Y_1 and Y_2 because $E(R_1|X_1, X_2) = 0$ and $E(R_2|X_1, X_2) = 0$. Simple algebra informs us that

$$\begin{aligned} \pi_{11} &= \frac{\beta_{11} + \gamma_{21}\beta_{21}}{1 - \gamma_{12}\gamma_{21}} \quad \pi_{12} = \frac{\beta_{12} + \beta_{22}\gamma_{12}}{1 - \gamma_{12}\gamma_{21}} \\ \pi_{21} &= \frac{\gamma_{21}\beta_{11} + \beta_{21}}{1 - \gamma_{12}\gamma_{21}} \quad \pi_{22} = \frac{\gamma_{12}\beta_{12} + \beta_{22}}{1 - \gamma_{12}\gamma_{21}} \end{aligned}. \tag{19}$$

and

$$\begin{aligned} R_1 &= \frac{U_1 + \gamma_{21}U_2}{1 - \gamma_{12}\gamma_{21}} \\ R_2 &= \frac{\gamma_{12}U_1 + U_2}{1 - \gamma_{12}\gamma_{21}} \end{aligned}.$$

Observe that because R_2 depends on both U_1 and U_2 in the general case, Y_2 is correlated with U_1 (through the direct channel of U_1 and through the correlation between U_1 and U_2). Without any further information on the variances of (U_1, U_2) and their relationship to the causal parameters, we cannot isolate the causal effects γ_{12} and γ_{21} from the reduced form regression coefficients. This is so because holding X_1, X_2, U_1, and U_2 fixed in (16a) or (16b), it is not *in principle* possible to vary Y_2 or Y_1, respectively, because they are exact functions of X_1, X_2, U_1, and U_2.

This exact dependence holds true even if $U_1 = 0$ and $U_2 = 0$ so that there are no unobservables.[50] In this case, which is thought to be the most favorable to the application of least squares or matching to (16a) and (16b), it is evident from (18a) and (18b) that when $R_1 = 0$ and

[50]See Fisher (1966).

$R_2 = 0$, Y_1 and Y_2 are exact functions of X_1 and X_2. There is no mechanism yet specified within the model to independently vary the right-hand sides of equations (16a) and (16b).[51] The X effects on Y_1 and Y_2, identified through the reduced forms, combine the direct effects (through β_{ij}) and the indirect effects (as they operate through Y_1 and Y_2, respectively).

If we assume exclusions ($\beta_{12} = 0$) or ($\beta_{21} = 0$) or both, we can identify the *ceteris paribus* causal effects of Y_2 on Y_1 and of Y_1 on Y_2 respectively. Thus if $\beta_{12} = 0$ from the reduced form,

$$\frac{\pi_{12}}{\pi_{22}} = \gamma_{12}.$$

If $\beta_{21} = 0$, we obtain

$$\frac{\pi_{21}}{\pi_{11}} = \gamma_{21}.$$

These exclusions say that the social interactions only operate through the Y'*s*. Person 1's consumption depends only on person 2's consumption and not on his or her X_2 or directly through his or her U_2. Person 2 is modeled symmetrically versus person 1. Observe that I have *not* ruled out correlation between U_1 and U_2. When the procedure for identifying causal effects is applied to samples, it is called indirect least squares. The method traces back to Haavelmo (1943, 1944).[52]

The intuition for these results is that if $\beta_{12} = 0$, we can vary Y_2 in equation (16a) by varying the X_2. Since X_2 does not appear in the

[51]Some readers of an earlier draft of this paper suggested that the mere fact that we can write (16a) and (16b) means that we "can imagine" independent variation. By the same token, we can imagine a model

$$Y = \varphi_0 + \varphi_1 X_1 + \varphi_2 X_2,$$

but if part of the model is (*) $X_1 = X_2$, the rules of the model constrain $X_1 = X_2$. No causal effect of X_1 holding X_2 constant is possible. If we break restriction (*) and permit independent variation in X_1 and X_2, we can define the causal effect of X_1 holding X_2 constant.

[52]The analysis for social interactions in this section is of independent interest. It can be generalized to the analysis of N person interactions if the outcomes are continuous variables. For binary outcomes variables, the same analysis goes through for the special case analyzed by Heckman and MaCurdy (1985). However, in the general case, for discrete outcomes generated by latent variables it is necessary to modify the system to obtain a coherent probability model; see Heckman (1978).

equation, under exclusion, we can keep U_1, X, fixed and vary Y_2 using X_2 in (18b) if $\beta_{22} \neq 0$.[53] Symmetrically, by excluding X_1 from (16b), we can vary Y_1, holding X_2 and U_2 constant. These results are more clearly seen when $U_1 = 0$ and $U_2 = 0$.

Observe that in the model under consideration, where the domain of definition and the supports of the variables coincide, the causal effects of simultaneous interactions are defined if the parameters are identified in the traditional Cowles definition of identification (e.g., see Ruud 2000 for a modern discussion of these conditions). A hypothetical thought experiment justifies these exclusions. If agents do not know or act on the other agents X, these exclusions are plausible.

An implicit assumption in using (16a) and (16b) for causal analysis is invariance of the parameters (Γ, β, Σ_U) to manipulations of the external variables. This invariance embodies the key idea in assumption (A-2). Invariance of the coefficients of equations to classes of manipulation of the variables is an essential part of the definition of structural models that I develop more formally in the next section.

This definition of causal effects in an interdependent system generalizes the recursive definitions of causality featured in the statistical treatment effect literature (Holland 1988; Pearl 2000). The key to this definition is manipulation of external inputs and exclusion, not randomization or matching. Indeed matching or, equivalently, *OLS*, using the right-hand side variables of (16a) and (16b), does not identify causal effects as Haavelmo (1943) established long ago. We can use the population simultaneous equations model to define the class of admissible variations and address problems of definitions (task 1 in Table 1). If for a given model, the parameters of (16a) or (16b) shift when external variables are manipulated, or if external variables cannot be independently manipulated, causal effects of one internal variable on another cannot be defined *within that model*. If people were randomly assigned to pair with their neighbors, and the parameters of (16a) were not affected by the randomization, then Y_2 would be exogenous in equation (16b) and we could identify causal

[53]Notice that we could also use U_2 as a source of variation in (18b) to shift Y_2. The roles of U_2 and X_2 are symmetric. However, if U_1 and U_2 are correlated, shifting U_2 shifts U_1 unless we control for it. The component of U_2 uncorrelated with U_1 plays the role of X_2.

effects by least squares. At issue is whether such a randomization would recover γ_{12}. It might fundamentally alter agent 1's response to Y_2 if that person is randomly assigned as opposed to being selected by the agent. Judging the suitability of an invariance assumption entails a thought experiment—a purely mental act.

Controlled variation in external forcing variables is the key to defining causal effects in nonrecursive models. It is of some interest to readers of Pearl (2000) to compare my use of the standard simultaneous equations model of econometrics in defining causal parameters to his. In the context of equations (16a) and (16b), Pearl defines a causal effect by "shutting one equation down" or performing "surgery" in his colorful language.

He implicitly assumes that "surgery," or shutting down an equation in a system of simultaneous equations, uniquely fixes one outcome or internal variable (the consumption of the other person in my example). In general, it does not. Putting a constraint on one equation places a restriction on the entire set of internal variables. In general, no single equation in a system of simultaneous equations uniquely determines any single outcome variable. Shutting down one equation might also affect the parameters of the other equations in the system and violate the requirements of parameter stability.

A clearer manipulation is to assume that it is possible to fix Y_2 by setting $\gamma_{12} = 0$. Assume that U_1 and U_2 are uncorrelated.[54] This makes the model recursive. It assumes that person 1 is unaffected by the consumption of person 2. Under these assumptions, we can regress Y_1 on Y_2, X_1, and X_2 in the population and recover all of the causal parameters of (16a). Variation in U_2 breaks the perfect collinearity among Y_2, X_1, and X_2. It is far from obvious, however, that one can freely set parameters without affecting the rest of the parameters of the model.

Shutting down an equation or fiddling with the parameters in Γ is not required to *define* causality in an interdependent, nonrecursive system or to identify causal parameters. The more basic idea is *exclusion* of different external variables from different equations which, when manipulated, allow the analyst to construct the desired causal quantities.

[54]Alternatively, we can assume that it is possible to measure U_1 and control for it.

One can move from the problem of definition (task 1 in Table 1) to identification (task 2) by using population analog estimation methods—in this case the method of indirect least squares.[55] There are many ways other than through exclusions of variables to identify this and more general systems. Fisher (1966) presents a general analysis of identification in both linear and nonlinear simultaneous equations systems. Matzkin (2004) is a recent substantial extension of this literature.

In the context of the basic nonrecursive model, there are many possible causal variations, richer than what can be obtained from the reduced form. Using the reduced form ($Y = X\Pi + R$), we can define causal effects as *ceteris paribus* effects of variables in X or R on Y. This definition solves out for all of the intermediate effects of the internal variables on each other. Using the structure in equation (15), we can define the effect of one internal variable on another holding constant the remaining internal variables and (X, U). It has just been established that such causal effects may not be defined within the rules specified for a particular structural model. Exclusions and other restrictions discussed in Fisher (1966) make definitions of causal effects possible under certain conditions.

One can, in general, solve out from the general system of equations for subsets of the Y (e.g., Y^* where $Y = (Y^*, Y^{**})$) using the reduced form of the model and use *quasi-structural* models to define a variety of causal effects that solve out for some but not all of the possible causal effects of Y on each other. These quasi-structural models may be written as

$$\Gamma^{**} Y^{**} = \Pi^{**} X + U^{**}.$$

This expression is obtained by using the reduced form for component Y^*: $Y^* = \Pi^* X + R^*$ and substituting for Y^* in (15). U^{**} is the error term associated with this representation. There are many possible quasi-structural models. Causal effects of internal variables may or may not be defined within them, depending on the assumed *a priori* information.

The causal effect of one component of Y^{**} on another does not fix Y^* but allows the Y^* components to adjust as the components of Y^{**} and the X are varied. Thus the Y^* are not being held fixed when

[55]Two-stage least squares would work as well.

X and/or components of the Y^{**} are varied. Viewed in this way, the reduced form and the whole class of quasi-structural models do not define any *ceteris paribus* causal effect relative to all of the variables (internal and external) in the system since they do not fix the levels of the other Y or Y^{*} in the case of the quasi-structural models. Nonetheless, the reduced form may provide a good guide to forecasting the effects of certain interventions that affect the external variables. The quasi-structural models may also provide a useful guide for predicting certain interventions, where Y^{**} are fixed by policy. The reduced form defines a net causal effect of variations in X as they affect the internal variables. There are many quasi-structural models and corresponding thought experiments.

This discussion demonstrates another reason why causal knowledge is provisional. Different analysts may choose different subsystems of equations derived from equation (15) to work with and define different causal effects within the different possible subsystems. Some of these causal effects may not be identified, while others may be. Systems smaller or larger than (15) can be imagined. The role of *a priori* theory is to limit the class of models and the resulting class of counterfactuals and to define which ones are interesting.

I now present a basic definition of structure in terms of invariance of equations to classes of interventions. Invariance is a central idea in causal analysis and in policy analysis.

2.6. *Structure as Invariance*

A basic definition of a system of structural relationships is that it is a system of equations invariant to a class of modifications or interventions. In the context of policy analysis, this means a class of policy modifications. This is the definition that was proposed by Hurwicz (1962). It is implicit in Marschak (1953) and it is explicitly utilized by Sims (1977), Lucas and Sargent (1981), and Leamer (1985), among others. This definition requires a precise definition of a policy, a class of policy modifications, and specification of a mechanism through which policy operates.

The mechanisms generating counterfactuals and the choices of counterfactuals have already been characterized in Sections 2.1 and 2.3. Policies can act on preferences and the arguments of preferences (and hence choices), on outcomes $Y(s, \omega)$ and the determinants

affecting outcomes or on the information facing agents. Recall that g_s, $s \in \mathcal{S}$, generates outcomes while f_s, $s \in \mathcal{S}$, generates evaluations. Specifically,

1. Policies can shift the distributions of the determinants of outcomes and choices (C, Z, X, U, η), where $C = \{C_s(\omega)\}_{s\in\mathcal{S}}$, $Z = \{Z(s, \omega)\}_{s\in\mathcal{S}}$, $\eta = \{\eta(s, \omega)\}_{s\in\mathcal{S}}$ and $U = \{U_s(\omega)\}_{s\in\mathcal{S}}$ in the population. This may entail defining the g_s and f_s over new domains. Let $Q = (C, Z, X, U, \eta)$. Policies shifting the distributions of these variables are characterized by maps $T_Q : Q \longmapsto Q'$.
2. Policies may select new f, g or $\{f_s, g_s\}_{s\in\mathcal{S}}$ functions.[56] In particular, new arguments (e.g., amenities or characteristics of programs) may be introduced as a result of policy actions creating new attributes. Policies shifting functions map f, g or $\{f_s, g_s\}_{s\in\mathcal{S}}$ into new functions $T_f : f_s \longmapsto f'_s$; $T_g : g_s \longmapsto g'_s$. This may entail changes in functional forms with a stable set of arguments as well as changes in arguments of functions.
3. Policies may affect individual information sets $(\mathcal{I}_\omega)_{\omega\in\Omega}$. $T_{\mathcal{I}\omega} : \mathcal{I}_\omega \longmapsto \mathcal{I}'_\omega$.

Clearly, any particular policy may incorporate elements of all three types of policy shifts.

Parameters of a model or parameters derived from a model are said to be policy invariant if they are not changed (are invariant) when policies are implemented. This notion is partially embodied in assumption (A-2), which is defined solely in terms of *ex post* outcomes. More generally, policy invariance for f,g or $\{f_s, g_s\}_{s\in\mathcal{S}}$ requires the following:

> **(A-3)** The functions f,g or $\{f_s, g_s\}_{s\in\mathcal{S}}$ are the same for all values of the arguments in their domain of definition no matter how their arguments are determined.

This definition can be made separately for f, g, f_s, g_s or any function derived from them. It requires that when we change an argument of a function it does not matter how we change it.

[56] By f_s, we mean s-specific valuation functions.

In the simultaneous equations model analyzed in the last section, invariance requires stability of Γ, B, and Σ_U to interventions. Such models can be used to accurately forecast the effects of policies that can be cast as variations in the inputs to the model. Policy-invariant parameters are not necessarily causal parameters, as we noted in our analysis of reduced forms in the preceding section. Thus, in the simultaneous equations model, depending on the *a priori* information available, no causal effect of one internal variable on another may be defined but if Π is invariant to modifications in X, the reduced form is policy invariant for those modifications. The class of policy-invariant parameters is thus distinct from the class of causal parameters, but invariance is an essential attribute of a causal model. For counterfactuals $Y(s, \omega)$, if assumption (A-3) is not postulated, all of the treatment effects defined in Section 1 would be affected by policy shifts. Rubin's assumption (A-2) makes $Y(s, \omega)$ invariant to policies that change f but not policies that change g or the support of Q. Within the treatment effects framework, a policy that adds a new treatment to $\mathcal{S}$ is not policy invariant for treatment parameters comparing the new treatment to any other treatment unless the analyst can model all policies in terms of a generating set of common characteristics specified at different levels. The lack of policy invariance makes it difficult to forecast the effects of new policies using treatment effect models within the framework of the Appendix.

"Deep structural" parameters generating the f and g are invariant to policy modifications that affect technology, constraints, and information sets except when the policies extend the historical supports. Invariance can only be defined relative to a class of modifications and a postulated set of preferences, technology, constraints, and information sets. Thus causal parameters can be precisely identified only within a class of modifications.

2.7. *Marschak's Maxim and the Relationship Between Structural Literature and Statistical Treatment Effect Literature*

The absence of explicit models is a prominent feature of the statistical treatment effect literature. Scientifically well-posed models make explicit the assumptions used by analysts regarding preferences, technology, the information available to agents, the constraints under which they operate, and the rules of interaction among agents in

market and social settings and the sources of variability among persons. These explicit features make these models, like all scientific models, useful vehicles: (1) for interpreting empirical evidence using theory; (2) for collating and synthesizing evidence using theory; (3) for measuring the welfare effects of policies; and (4) for forecasting the welfare and direct effects of previously implemented policies in new environments and the effects of new policies.

These features are absent from the modern treatment effect literature. At the same time, this literature makes fewer statistical assumptions in terms of exogeneity, functional form, exclusion, and distributional assumptions than the standard structural estimation literature in econometrics. These are the attractive features of this approach.

In reconciling these two literatures, I reach back to a neglected but important paper by Jacob Marschak. Marschak (1953) noted that for many specific questions of policy analysis, it is unnecessary to identify full structural models, where by structural I mean parameters invariant to classes of policy modifications as defined in the last section. All that is required are combinations of subsets of the structural parameters, corresponding to the parameters required to forecast particular policy modifications, which are much easier to identify (i.e., require fewer and weaker assumptions). Thus in the simultaneous equations system examples, policies that only affect X may be forecast using reduced forms, not knowing the full structure, provided that the reduced forms are invariant to the modifications.[57] Forecasting other policies may require only partial knowledge of the system. I call this principle *Marschak's maxim* in honor of this insight. I interpret the modern statistical treatment effect literature as implicitly implementing Marschak's maxim where the policies analyzed are the treatments and the goal of policy analysis is restricted to evaluating policies in place (task 1; P1) and not in forecasting the effects of new policies or the effects of old policies on new environments.

Population mean treatment parameters are often identified under weaker conditions than are traditionally assumed in econometric structural analysis. Thus to identify the average

[57]Thus we require that the reduced form Π does not change when we change the X.

treatment effect for s and s' we require only $E(Y(s, \omega) \mid X = x) - E(Y(s', \omega) \mid X = s)$. We do not have to know the full functional form of the generating g_s functions nor does X have to be exogenous. The treatment effects may, or may not, be causal parameters depending on what else is assumed about the model.

Considerable progress has been made in relaxing the parametric structure assumed in the early structural models in econometrics (see Matzkin 2006). As the treatment effect literature is extended to address the more general set of policy forecasting problems entertained in the structural literature, the distinction between the two literatures will vanish although it is currently very sharp. Heckman and Vytlacil (2005, 2006a,b) and Heckman (2006) are attempts to bridge this gulf.

Up to this point in the essay, everything that has been discussed precisely is purely conceptual, although I have alluded to empirical problems and problems of identification going from data of various forms to conceptual models. Models are conceptual and so are the treatment effects derived from them. The act of defining a model is distinct from identifying it or estimating it although statisticians often conflate these distinct issues. I now discuss the identification problem, which must be solved if causal models are to be empirically operational.

3. IDENTIFICATION PROBLEMS: DETERMINING MODELS FROM DATA

Unobserved counterfactuals are the source of the problems considered in this paper. For a person in state s, we observe $Y(s, \omega)$ but not $Y(s', \omega)$, $s' \neq s$. A central problem in the literature on causal inference is how to identify counterfactuals and the derived treatment parameters. Unobservables, including missing data, are at the heart of the identification problem considered here.

Estimators differ in the amount of knowledge they assume that the analyst has relative to what the agents being studied have when making their program enrollment decisions (or their decisions are made for them as a parent for a child). This is strictly a matter of the quality of the available data. Unless the analyst has access to all of the relevant information that produces the dependence between

outcomes and treatment rules (i.e., that produces selection bias), he or she must devise methods to control for the unobserved components of relevant information. Heckman and Vytlacil (2006b) and Heckman and Navarro (2004) define relevant information precisely. Relevant information is the information which, if available to the analyst and conditioned on, would eliminate selection bias. Intuitively, there may be a lot of information known to the agent but not known to the observing analyst that is irrelevant in creating the dependence between outcomes and choices. It is the information that gives rise to the dependence between outcomes and treatment choices that matters for eliminating selection bias.

A priori one might think that the analyst knows a lot less than the agent whose behavior is being analyzed. At issue is whether the analyst knows less *relevant* information, which is not so obvious, if only because the analyst can observe the outcomes of decisions in a way that agents making decisions cannot. This access to *ex post* information can sometimes give the analyst a leg up on the information available to the agent.

The policy forecasting problems P2 and P3 raise the additional issue that the support over which treatment parameters and counterfactuals are identified may not correspond to the support to which the analyst seeks to apply them. Common to all scientific models, there is the additional issue of how to select (X, Z), the conditioning variables, and how to deal with them if they are endogenous. Finally, there is the problem of lack of knowledge of functional forms of the models. Different econometric methods solve these problems in different ways. I now present a precise discussion of identification.

3.1. *The Identification Problem*

The identification problem asks whether theoretical constructs have any empirical content in a hypothetical population or in real samples. This formulation considers tasks 2 and 3 in Table 1 together, although some analysts like to separate these issues, focusing solely on task 2. The identification problem considers what particular models within a broader class of models are consistent with a given set of data or facts. Specifically, we can consider a model space M. This is the set of admissible models that are produced by some theory for generating counterfactuals. Elements $m \in M$ are admissible theoretical models.

We may be interested in only some features of a model. For example, we may have a rich model of counterfactuals $\{Y(s, \omega)\}_{s \in \mathcal{S}}$, but we may be interested in only the average treatment effect $E_\omega[Y(s, \omega) - Y(s', \omega)]$. Let the objects of interest be $t \in T$, where "t" stands for the target—the goal of the analysis. The target space T may be the whole model space M or something derived from it.

Define map g: $M \to T$. This maps an element $m \in M$ into an element $t \in T$. In the example in the preceding paragraph, T is the space of all average treatment effects produced by the models of counterfactuals. I assume that g is into.[58] Associated with each model is an element t derived from the model, which could be the entire model itself. Many models may map into the same t so the inverse map (g^{-1}), mapping T to M, may not be well-defined. Thus many different models may produce the same average treatment effect.

Let the class of possible information or data be I. Define a map h: $M \to I$. For an element $i \in I$, which is a given set of data, there may be one or more models m consistent with i. If i can be mapped only into a single m, the model is exactly identified.[59] If there are multiple m's, consistent with i, these models are not identified. Thus, in Figure 1, many models (elements of M) may be consistent with the same data (single element of I).

Let $M_h(i)$ be the set of models consistent with i. $M_h(i) = h^{-1}(\{i\}) = \{m \in M : h(m) = i\}$. The data i reject the other models $M \backslash M_h(i)$, but are consistent with all models in $M_h(i)$. If $M_h(i)$ contains more than one element, the data produce set-valued instead of point-valued identification. If $M_h(i) = \emptyset$, the empty set, no

[58]By this, we mean that for every $t \in T$, there is an element $m \in M$ such that g sends m to t, i.e., the image of g is the entire set T. Of course, g may send many elements of M to a single element of T.

[59]Associated with each data set i is a collection of random variables $Q(i)$, which may be a vector. Let $F_Q(q|m)$ be the distribution of q under model m. To establish identification on nonnegligible sets, one needs that, for some true model m^*,

$$\Pr(|F_Q(q|m^*) - F_Q(q|m)| > \varepsilon) > 0$$

for some $\varepsilon > 0$ for all $m \neq m^*$. This guarantees that there are observable differences between the data generating process for Q given m and for Q given m^*. We can also define this for $F_Q(q|t^*)$ and $F_Q(q|t)$.

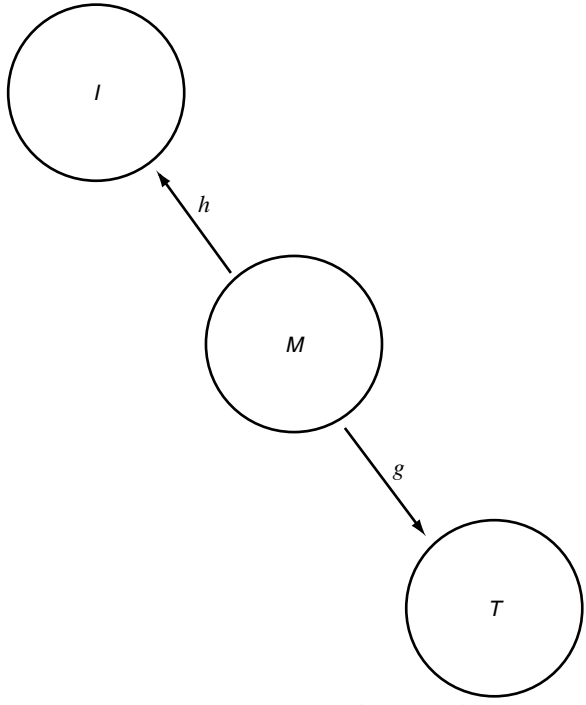


FIGURE 1. Are elements in T uniquely determined from elements in I? Sometimes $T = M$. Usually T consists of elements derived from M.

model is consistent with the data. By placing restrictions on models, we can sometimes reduce the number of elements in $M_h(i)$ if it has multiple members. Let $R \subset M$ be a set of restricted models. It is sometimes possible by imposing restrictions to reduce the number of models consistent with the data. Recall that in the two-person model of social interactions, if $\beta_{12} = 0$ and $\beta_{21} = 0$ we could uniquely identify the remaining parameters under the other conditions maintained in Section 2.5. Thus $R \cap M_h(i)$ may contain only a single element. Another way to solve this identification problem is to pick another data source $i' \in I$, which may produce more restrictions on the class of admissible models. More information provides more hoops for the model to jump through.

Going after a more limited class of objects such as features of a model ($t \in T$) rather than the full model ($m \in M$) is another way to secure unique identification. Let $M_g(t) = g^{-1}(\{t\}) = \{m \in M: g(m) = t\}$.

Necessary and sufficient conditions for the existence of a unique map $f: I \to T$ with the property $f \circ h = g$ are (a) h must map M *onto* I and (b) for all $i \in I$, there exists $t \in T$ such that $M_h(i) \subseteq M_g(t)$. Condition (b) means that even though one element $i \in I$ may be consistent with many elements in M, so that $M_h(i)$ consists of more than one element, it may be that all elements in $M_h(i)$ are mapped by g into a single element of T. The map f is onto since $g = f \circ h$ and g is onto by assumption. In order for the map f to be one-to-one, it is necessary and sufficient to have equality of $M_h(i)$ and $M_g(t)$ instead of simply inclusion.

If we follow Marschak's maxim and focus on a smaller target space T, it is possible that g maps the admissible models into a smaller space. Thus the map f described above may produce a single element even if there are multiple models m consistent with the data source i. This would arise, for example, if for a given set of data i, we could only estimate the mean μ_1 of Y_1 up to a constant c and the mean μ_2 of Y_2 up to the same constant c. But we could uniquely identify the element $\mu_1 - \mu_2 \in T$.[60] In general, identifying elements of T is easier than identifying elements of M. Thus, in Figure 1, even though many models (elements of M) may be consistent with the same $i \in I$, only one element of T may be consistent with that i. I now turn to empirical causal inference and illustrate the provisional nature of causal inference.

4. THE PROVISIONAL NATURE OF CAUSAL INFERENCE[61]

This section develops the implicit assumptions underlying four widely used methods of causal inference applied to data: (1) matching, (2) control functions, (3) instrumental variable methods, and (4) the method of directed acyclic graphs promoted by Pearl (2000) (or the *g*-computation method of Robins 1989). It is not intented as an

[60]Most modern analyses of identification assume that sample sizes are infinite, so that enlarging the sample size is not informative. However, in any applied problem this distinction is not helpful. Having a small sample (e.g. fewer observations than regressors) can produce an identification problem. This definition combines task 3 and task 2 if we allow for samples to be finite.

[61]Portions of this section are based on Heckman and Navarro (2004).

exhaustive survey of the literature. I demonstrate the value of the scientific approach to causality by showing how explicit analysis of the choice of treatment (or the specification of the selection equations) and the outcomes, including the relationship between the unobservables in the outcome and selection equations clarifies the implicit assumptions being made in each method. This enables the analyst to use behavioral theory aided by statistics to choose estimators and interpret their output. This discussion also clarifies that each method of inference makes implicit identifying assumptions in going from samples to make inferences about models. There is no assumption-free method of causal inference.[62]

I do not discuss randomization systematically except to note that randomization does not in general identify distributions of treatment effects (Heckman 1992; Heckman and Smith 1998; Heckman, Smith, and Clements 1997; Heckman and Vytlacil 2006b). Matching implicitly assumes a randomization by nature in the unobservables producing the choice treatment equation relative to the outcome equation, so my analysis of matching implicitly deals with randomization.

I focus primarily on identification of mean treatment effects in this paper. Discussions of identification of distributions of treatment effects are presented in Aakvik, Heckman, and Vytlacil (1999, 2005), Carneiro, Hansen, and Heckman (2001, 2003), and Heckman and Navarro (2006). I start by presenting a prototypical econometric selection model.

4.1. *A Prototypical Model of Treatment Choice and Outcomes*

To focus the discussion, and to interpret the implicit assumptions underlying the different estimators presented in this paper, I present a benchmark model of treatment choice and treatment outcomes. For simplicity I consider two potential outcomes (Y_0, Y_1). I drop the individual (ω) subscripts to avoid notational clutter. $D = 1$ if Y_1 is selected; $D = 0$ if Y_0 is selected. Agents pick the realized outcome based on their evaluation of the outcomes, given their information. The agent picking the treatment might be different from the person experiencing the outcome

[62]This is true for experiments as well. See Heckman (1992).

(e.g., the agent could be a parent choosing outcomes for the child). Let V be the agent's valuation of treatment. I write

$$V = \mu_V(W,\ U_V) \quad D = \mathbf{1}(V > 0), \tag{20}$$

where the W are factors (observed by the analyst) determining choices, U_V are the unobserved (by the analyst) factors determining choice. Valuation function (20) is a centerpiece of the scientific model of causality but is not specified in the statistical approach.

Potential outcomes are written in terms of observed variables (X) and unobserved (by the analyst) outcome-specific variables

$$Y_1 = \mu_1(X,\ U_1) \tag{21a}$$

$$Y_0 = \mu_0(X, U_0). \tag{21b}$$

I assume throughout that U_0, U_1, and U_V are continuous random variables and that all means are finite.[63] The individual level treatment effect is thus

$$\Delta = Y_1 - Y_0.$$

More familiar forms of (20), (21a), and (21b) are additively separable expressions,

$$V = \mu_V(W) + U_V \quad E(U_V) = 0, \tag{22a}$$

$$Y_1 = \mu_1(X) + U_1 \quad E(U_1) = 0, \tag{22b}$$

$$Y_0 = \mu_0(X) + U_0 \quad E(U_0) = 0. \tag{22c}$$

Additive separability is not strictly required in modern econometric models (e.g., see Matzkin 2003). However, I use the additively separable representation throughout most of this section because of its familiarity, noting when it is a convenience and when it is an essential part of a method.

The distinction between X and Z is crucial to the validity of many econometric procedures. In matching as conventionally

[63]Strictly speaking, absolutely continuous with respect to the Lebesgue measure.

formulated there is no distinction between X and Z. The roles of X and Z in alternative estimators are explored in this section.

A simple example will serve to fix ideas. It will enable me to synthesize the main results of the first three sections of this paper and lay the ground for this section.

Suppose that we use linear-in-parameters expressions. We write the potential outcomes for the population as

$$Y_1 = X\beta_1(C_1) + U_1 \tag{23a}$$

$$Y_0 = X\beta_0(C_0) + U_0, \tag{23b}$$

where we let X be the characteristics of persons and we let the β depend on C_1 and C_0, the characteristics of the programs. These are linear-in-parameters versions of equation (10) for $s = 0,1$. The U_1 and U_0 are the unobservables arising from omitted X, C_1, and C_0 components. Included among the X is "1" so that the characteristics of the programs are allowed to enter directly and in interaction with the X. By modeling how β_1 and β_0 depend on C_1 and C_0, we can answer policy question P3 for new programs that offer new packages of C, assuming we can account for the effects C_i on generating U_1 and U_0.

A version of the model most favorable to solving problems P2 and P3 writes

$$\beta_1(C_1) = \Lambda C_1'$$
$$\beta_0(C_0) = \Lambda C_0',$$

where C_1 and C_0 are $1 \times J$ vectors of characteristics of programs, and C_1' and C_0' are their transposes. Assuming that X is a $1 \times K$ vector of person-specific characteristics, Λ is a $K \times J$ matrix. This specification enables us to represent all of the coefficients of the outcome equations in terms of a base set of generator characteristics.

For each fixed set of characteristics of a program, we can model how outcomes are expected to differ when we change the characteristics of the people participating in them (the X). This is an ingredient for solving problem P3.

Equations (23a) and (23b) are in *ex post* all causes form. For information set $\mathcal{I}$, we can write the *ex ante* version as $E(Y_1|\mathcal{I})$ and $E(Y_0|\mathcal{I})$ (see equation 11). The decision-making agent may be uncertain about the X, the β_i, the C_i, and the U_i. The *ex ante* version reflects this uncertainty. Cunha, Heckman, and Navarro (2005a,b)

provide examples of *ex ante* outcome models. *Ex ante* Marshallian causal functions are defined in terms of variations in $\mathcal{I}$. *Ex post* and *ex ante* outcomes are connected by shock $\nu(s, \omega)$, as in equation (12).

The choice equation may depend on expected rewards and costs, as in Section 2.3. Let

$$V = E(Y_1 - Y_0 - (P_1 - P_0)|\mathcal{I}), \tag{24}$$

where P_i is the price of participating in i and $P_i = Z\varphi_i + \eta_i$. In the special case of perfect foresight, $\mathcal{I} = (U_1, U_0, C_1, C_0, X, Z, \Lambda, \varphi_1, \varphi_2)$.

To focus on some main ideas, suppose that we work with β_1 and β_0, leaving the C_i implicit. Substituting for the P_i in equation (24) and for the outcomes (23a) and (23b), we obtain after some algebra

$$V = E[X(\beta_1 - \beta_0) - Z(\varphi_1 - \varphi_0) + (U_1 + U_0) - (\eta_1 - \eta_0) \mid \mathcal{I}],$$

where $\mathcal{I}$ is the information set at the time the agent is making the participation decision. Let $W = (X, Z)$, $U_W = (U_1 - U_0) - (\eta_1 - \eta_0)$, and $\gamma = (\beta_1 - \beta_0, -(\varphi_1 - \varphi_0))$. We can then represent the choice equation as

$$V = E[W\gamma + U_W|\mathcal{I}],$$

where

$$D = \mathbf{1}(V > 0).$$

Let U_V be the random variable of U_W conditional on $\mathcal{I}$. For simplicity, we assume that agents know $W = (X, Z)$ but not all of the components of U_W when they make their treatment selection decisions. We also assume that the analyst knows $W = (X, Z)$.

The selection problem arises when D is correlated with (Y_0, Y_1). This can happen if the observables or the unobservables in (Y_0, Y_1) are correlated with or dependent on D. Thus there may be common observed or unobserved factors connecting V and (Y_0, Y_1).

If D is not independent of (Y_0, Y_1), the observed (Y_0, Y_1) are not randomly selected from the population distribution of (Y_0, Y_1). In the Roy model, discussed in Section 1, $\varphi_1 = \varphi_0 = 0$, $\eta_1 = \eta_0 = 0$, and selection is based on Y_1 and Y_0 ($D = \mathbf{1}(Y_1 > Y_0)$). Thus we observe Y_1 if $Y_1 > Y_0$ and we observe Y_0 if $Y_0 \geq Y_1$.

If conditioning on W makes (Y_0, Y_1) independent of D, selection on observables is said to characterize the selection process.[64] This is the motivation for the method of matching. If conditional on W, (Y_0, Y_1) are not independent of D, then we have selection on unobservables and alternative methods must be used.

For the Roy model, Heckman and Honoré (1990) show that it is possible to identify the distribution of treatment outcomes $(Y_1 - Y_0)$ under the conditions they specify. Randomization can identify only the marginal distributions of Y_0 and of Y_1, not the joint distribution of $(Y_1 - Y_0)$ or the quantiles of $(Y_1 - Y_0)$. Thus, under its assumptions, the Roy model is more powerful than randomization in producing the distributional counterfactuals.[65]

The role of the choice equation is to motivate and justify the choice of an evaluation estimator. This is a central feature of the econometric approach that is missing from the statistical and epidemiological literature on treatment effects. Heckman and Smith (1998), Aakvik, Heckman, and Vytlacil (2005), Carneiro, Hansen, and Heckman (2003), and Cunha, Heckman, and Navarro (2005a,b) extend these results to estimate distributions of treatment effects.

4.2. *Parameters of Interest*

There are many different treatment parameters that can be derived from this model if $U_1 \neq U_0$ and agents know or partially anticipate U_0, U_1 in making their decisions (Heckman and Robb 1985; Heckman 1992; Heckman, Smith, and Clements 1997: Heckman 2001; Heckman and Vytlacil 2000; Cunha, Heckman, and Navarro 2005a,b). For specificity, I focus on certain means because they are traditional. As noted in Section 2 and in Heckman and Vytlacil (2000, 2005) and Heckman (2001), the traditional means often do not answer interesting social and economic questions.

[64]See Heckman and Robb (1985).

[65]The same analysis applies to matching, which cannot identify the distributions of $(Y_1 - Y_0)$ or derived quantiles.

The traditional means conditional on covariates are as follows:

$$\begin{aligned}\text{Average Treatment Effect } (ATE) &: E(Y_1 - Y_0|X)\\ \text{Treatment on the Treated } (TT) &: E(Y_1 - Y_0|X, D = 1)\\ \text{Marginal Treatment Effect } (MTE) &: E(Y_1 - Y_0|X, Z, V = 0).\end{aligned}$$

The MTE is the marginal treatment effect introduced into the evaluation literature by Björklund and Moffitt (1987). It is the average gain to persons who are indifferent to participating in sector 1 or sector 0 given X, Z. These are persons at the margin, defined by (W) so Z plays a role in the definition of the parameter by fixing $\mu_V(W)$ in equation (20) or equation (22a) and hence fixing U_V. It is a version of $EOTM$ as defined in Section 1. An alternative definition in this setup is $MTE = E(Y_1 - Y_0|X,\ U_V)$. Heckman and Vytlacil (1999, 2005, 2006b) show how the MTE can be used to construct all mean treatment parameters, including the policy relevant treatment parameters, under the conditions specified in their papers. These parameters can be defined for the population as a whole not conditioning on X or Z.[66]

4.3. *The Selection Problem Stated in Terms of Means*

Let $Y = DY_1 + (1 - D)Y_0$. Samples generated by choices have the following means which are assumed to be known:

$$E(Y|X, Z, D = 1) = E(Y_1|X, Z, D = 1)$$

and

$$E(Y|X, Z, D = 0) = E(Y_0|X, Z, D = 0)$$

for outcomes Y_1 for participants and the outcomes Y_0 for nonparticipants, respectively. In addition, choices are observed so that in large samples $\Pr(D = 1|X, Z)$ is known—that is, the probability of choosing treatment is known. From the sample data, we can also construct

$$E(Y_1|X, D = 1) \quad \text{and} \quad E(Y_0|X, D = 0).$$

[66]The average marginal treatment effect is

$$E(Y_1 - Y_0|V = 0) = \int E(Y_1 - Y_0|X, Z, V = 0)f(X, Z|V = 0)dXdZ.$$

The conditional biases from using the difference of these means to construct the three parameters studied in this paper are

$$\begin{aligned}\text{Bias } TT &= [E(Y|X, D=1) - E(Y|X, D=0)] - E(Y_1 - Y_0|X, D=1)\\ &= [E(Y_0|X, D=1) - E(Y_0|X, D=0)].\end{aligned}$$

In the case of additive separability

$$\text{Bias } TT = E(U_1|X,\ D=1) - E(U_0|X,\ D=0).$$

For ATE,

$$\text{Bias } ATE = E(Y|X,\ D=1) - E(Y|X,\ D=0) - [E(Y_1 - Y_0|X)].$$

In the case of additive separability

$$\text{Bias } ATE = [E(U_1|X, D=1) - E(U_1|X)] - [E(U_0|X, D=0) - E(U_0|X)].$$

For MTE,

$$\begin{aligned}\text{Bias } MTE &= E(Y|X, Z, D=1) - E(Y|X, Z, D=0)\\ &\quad - E(Y_1 - Y_0|X, Z, V=0)\\ &= [E(U_1|X, Z, D=1) - E(U_1|X, Z, V=0)]\\ &\quad - [E(U_0|X, Z, D=0) - E(U_0|X, Z, V=0)],\end{aligned}$$

for the case of additive separability in outcomes. The MTE is defined for a subset of persons indifferent between the two sectors and so is defined for X and Z. The bias is the difference between average U_1 for participants and marginal U_1 minus the difference between average U_0 for nonparticipants and marginal U_0. Each of these terms is a bias that can be called a selection bias. These biases can be defined conditional on X (or X and Z or X, Z, and V in case of the MTE) or unconditionally.

4.4. *How Different Methods Eliminate the Bias*

In this section I consider the identification conditions that underlie matching, control functions, and instrumental variable methods to

identify the three parameters using the data on mean outcomes. I also briefly discuss the method of directed acyclic graphs or the *g*-computation method for one causal parameter. I discuss sources of unobservables, implicit assumptions about how unobservables are eliminated as sources of selection problems, and the assumed relationship between outcomes and choice equations. I start with the method of matching.

4.4.1. *Matching*

The method of matching as conventionally formulated makes no distinction between X and Z. Define the conditioning set as $W = (X, Z)$. The strong form of matching advocated by Rosenbaum and Rubin (1983) and in numerous predecessor papers, assumes that

$$(Y_1, Y_0) \perp\!\!\!\perp D|W \tag{M-1}$$

and

$$0 < \Pr(D = 1|W) = P(W) < 1, \tag{M-2}$$

where "$\perp\!\!\!\perp$" denotes independence given the conditioning variables after "|". $P(W)$ is the probability of selection into treatment and is sometimes called the propensity score. Condition (M-2) implies that the mean treatment parameters can be defined for all values of W (i.e., for each W, in very large samples, there are observations for which we observe a Y_0 and other observations for which we observe a Y_1). Rosenbaum and Rubin (1983) show that under (M-1) and (M-2)

$$(Y_1, Y_0) \perp\!\!\!\perp D|P(W). \tag{M-3}$$

This reduces the dimensionality of the matching problem. They assume that P is known. When it is not known, it is necessary to estimate it. Nonparametric estimation of $P(W)$ restores the dimensionality problem but shifts it to the estimation of $P(W)$.[67] Under these

[67]Rosenbaum (1987) or Rubin and Thomas (1992) consider the distribution of the matching estimator when P is estimated under special assumptions about the distribution of the matching variables. Papers that account for estimated P under general conditions include Heckman, Ichimura, and Todd (1997, 1998) and Hahn (1998).

assumptions, conditioning on P eliminates all three biases defined in Section 4.3 for parameters defined conditional on P because

$$E(Y_1|D=0,P(W)) = E(Y_1|D=1,P(W)) = E(Y_1|P(W))$$
$$E(Y_0|D=1,P(W)) = E(Y_0|D=0,P(W)) = E(Y_0|P(W)).$$

Thus for TT one can identify counterfactual mean $E(Y_0|D=1,\ P(W))$ from $E(Y_0|D=0,\ P(W))$. In fact, one only needs the weaker condition $Y_0 \perp\!\!\!\perp D|P(W)$ to remove the bias[68] because $E(Y_1|D=1,\ P(W))$ is known, and only $E(Y_0|D=1,P(W))$ is unknown. From the observed conditional means one can form ATE. Since the conditioning is on $P(W)$, the parameter is defined conditional on it and not X or (X, Z). Integrating out $P(W)$ produces unconditional ATE. Integrating out $P(W)$ given $D=1$ produces unconditional TT.[69]

Observe that since $ATE = TT$ for all X, Z under (M-1) and (M-2), the effect for the average person participating in a program is the same as the effect for the marginal person, conditional on W, and there is no bias in estimating MTE.[70] The strong implicit assumption that the marginal participant in a program gets the same return as the average participant in the program, conditional on W, is an unattractive implication of these assumptions (see Heckman 2001 and Heckman and Vytlacil 2005, 2006a,b). The method assumes that all of the dependence between U_V and $(U_1,\ U_0)$ is eliminated by conditioning on W,

$$U_V \perp\!\!\!\perp (U_1, U_0)|W.$$

This motivates the term "selection on observables" introduced in Heckman and Robb (1985, 1986).

Assumption (M-2) has the unattractive feature that if the analyst has too much information about the decision of who takes treatment so that $P(W)=1$ or 0, the method breaks down because people cannot be compared at a common W. The method of matching

[68]See Heckman, Ichimura, and Todd (1997) and Abadie (2003).

[69]To estimate the parameters conditional on W, one cannot use $P(W)$ but must use the full W vector.

[70]As demonstrated in Carneiro (2002), one can still distinguish marginal and average effects in terms of observables.

assumes that, given W, some unspecified randomization device allocates people to treatment. The fact that the cases $P(W) = 1$ and $P(W) = 0$ must be eliminated suggests that methods for choosing which variables enter W based on the fit of the model to data on choices (D) are potentially problematic; see Heckman and Navarro (2004) and Heckman and Vytlacil (2005) for further discussion of this point.

What justifies (M-1) or (M-3)? Absent an explicit theoretical model of treatment assignment and an explicit model of the sources of randomness, analysts are unable to justify the assumption except by appeal to convenience. Because there are no exclusion restrictions in the observables, the only possible source of variation in D given W are the unobservable elements generating D. These elements are assumed to act like an ideal randomization that assigns person to treatment but is independent of (U_1, U_2), the unobservables generating (Y_0, Y_1), given W.

If agents partially anticipate the benefits of treatment and make enrollment decisions based on these anticipations, (M-1) or (M-3) is false. In the extreme case of the Roy model, where $D = \mathbf{1}(Y_1 > Y_0)$, (M-1) or (M-3) is certainly false. Even if agents are only imperfectly prescient but can partially forecast (Y_1, Y_0) and use that information in deciding whether or not to participate, (M-1) or (M-3) is false.

Without a model of interventions justifying these assumptions, and without a model of the sources of unobservables, (M-1) or (M-3) cannot be justified. The model cannot be tested without richer sources of data.[71] Judgments about whether agents are as ignorant about potential outcomes given W, as is assumed in (M-1) or (M-3), can only be settled by the theory unless it is possible to randomize persons into treatment, and randomization does not change the outcome—that is, under assumption (A-2). The matching model makes strong implicit assumptions about the unobservables.

In the recent literature, the claim is sometimes made that matching is "for free" (e.g., see Gill and Robins 2001). The idea underlying this claim is that since $E(Y_0|D = 1, W)$ is not observed, we might as well set it to $E(Y_0|D = 0, W)$, an implication of (M-1). This argument

[71]See Heckman, Ichimura, Smith, and Todd (1998) for a test of matching assumptions using data from randomized trials.

is correct so far as data description goes. Matching imposes just-identifying restrictions and in this sense—at a purely empirical level—is as good as any other just-identifying assumption in describing the data.

However, the implied behavioral restrictions are not "for free." Imposing that—conditional on X and Z or conditional on $P(W)$ the marginal person entering a program is the same as the average person—is a strong and restrictive implication of the conditional independence assumptions and is not a "for free" assumption in terms of its behavioral content.[72] In the context of estimating the economic returns to schooling, it implies that, conditional on W, the economic return to schooling for persons who are just at the margin of going to school are the same as the return for persons with strong preferences for schooling.

Introducing a distinction between X and Z allows the analyst to overcome the problem arising from perfect prediction of treatment assignment for some values of (X, Z) if there are some variables Z not in X. If P is a nontrivial function of Z (so $P(X, Z)$ varies with Z for all X) and Z can be varied independently of X for all points of support of X,[73] and if outcomes are defined solely in terms of X, the problem of perfect classification can be solved. Treatment parameters can be defined for all support values of X since for any value (X, Z) that perfectly classifies D, there is another value (X, Z'), $Z' \neq Z$, that does not (see Heckman, Ichimura, and Todd 1997).

Offsetting the disadvantages of matching, the method of matching with a known conditioning set that satisfies (M-1) does not require separability of outcome or choice equations into observable and unobservable components, exogeneity of conditioning variables, exclusion restrictions, or adoption of specific functional forms of outcome equations. Such assumptions are commonly used in conventional selection (control function) methods and conventional applications of *IV* although recent work in semiparametric estimation

[72]As noted by Heckman, Ichimura, Smith, and Todd (1998), if one seeks to identify $E(Y_1 - Y_0|D = 1, W)$ one only needs to impose a weaker condition $[E(Y_0|D = 1, W) = E(Y_0|D = 0, W)]$ or $Y_0 \perp\!\!\!\perp D|W$ rather than (M-1). This imposes the assumption of no selection on levels of Y_0 (given W) and not the assumption of no selection on levels of Y_1 or on $Y_1 - Y_0$, as (M-1) does. Marginal can be different from average in this case.

[73]A precise sufficient condition is that $Supp(Z|X) = Supp(Z)$. We can get by with a weaker condition that in any neighborhood of X, there is a Z^* such that $0 < \Pr(D = 1|X, Z^*) < 1$, and that Z^* is in the support of $Z|X$.

relaxes many of these assumptions, as I note below (see also Heckman and Vytlacil 2005, 2006b). Moreover, the method of matching does not strictly require (M-1). One can get by with weaker mean independence assumptions,

$$\begin{aligned} E(Y_1|W, D = 1) &= E(Y_1|W), \\ E(Y_0|W, D = 0) &= E(Y_0|W), \end{aligned} \tag{M-1$'$}$$

in the place of the stronger (M-1) conditions. However, if (M-1$'$) is invoked, the assumption that one can replace W by $P(W)$ does not follow from the analysis of Rosenbaum and Rubin, and is an additional new assumption.

4.4.2. *Control Functions*

The principle motivating the conventional method of control functions is different. (See Heckman 1976, 1978, 1980 and Heckman and Robb 1985, 1986, where this principle was first developed.) Like matching, it works with conditional expectations of (Y_1, Y_0) given (X, Z and D). Conventional applications of the control function method assume additive separability that is not required in matching. Strictly speaking, additive separability in the outcome equation is not required in the application of control functions either.[74] What is required is a model relating the outcome unobservables to the observables, including the choice of treatment. The method of matching assumes that, conditional on the observables (X, Z), the unobservables are independent of D.[75] For the additively separable case, control functions based on the principle of modeling the conditional expectations of Y_1 and Y_0 given X, Z, and D can be written as

$$\begin{aligned} E(Y_1|X, Z, D = 1) &= \mu_1(X) + E(U_1|X, Z, D = 1) \\ E(Y_0|X, Z, D = 0) &= \mu_0(X) + E(U_0|X, Z, D = 0). \end{aligned}$$

[74]Examples of nonseparable selection models are found in Cameron and Heckman (1998).

[75]Or mean independent in the case of mean parameters.

In the method of control functions if one can model $E(U_1|X,\ Z,\ D=1)$ and $E(U_0|X,\ Z,\ D=0)$ and these functions can be independently varied against $\mu_1\ (X)$ and $\mu_0\ (X)$ respectively, one can identify $\mu_1\ (X)$ and $\mu_0\ (X)$ up to constant terms.[76] Nothing in the method intrinsically requires that X or Z be stochastically independent of U_1 or U_0, although conventional methods often assume this.

If one assumes that $(U_1,\ U_V) \perp\!\!\!\perp (X,\ Z)$ and adopts equation (22a) as the treatment choice model augmented so X and Z are determinants of treatment choice, one obtains

$$E(U_1|X,Z,D=1) = E(U_1|U_V \geq -\mu_V(X,Z)) = K_1(P(X,Z)),$$

so the control function depends only on $P(X,\ Z)$. By similar reasoning, if $(U_0,\ U_V) \perp\!\!\!\perp (X,\ Z)$,

$$E(U_0|X,Z,D=0) = E(U_0|U_V < -\mu_V(X,Z)) = K_0(P(X,Z))$$

and the control function depends only on the probability of selection ("the propensity score"). The key assumption needed to represent the control function solely as a function of $P(X,\ Z)$ is

$$(U_1, U_0, U_V) \perp\!\!\!\perp (X,Z). \tag{C-1}$$

Under this condition

$$\begin{aligned} E(Y_1|X,Z,D=1) &= \mu_1(X) + K_1(P(X,Z)) \\ E(Y_0|X,Z,D=0) &= \mu_0(X) + K_0(P(X,Z)) \end{aligned}$$

[76]Heckman and Robb (1985, 1986) introduce this general formulation of control functions. The identifiability requires that the members of the pairs $(\mu_1(X),\ E\ (U_1|X,\ Z,\ D=1))$ and $(\mu_0(X),\ E\ (U_0|X,\ Z,\ D=0))$ be "variation free" so that they can be independently varied against each other; see Heckman and Vytlacil (2006a, b) for a precise statement of these conditions.

with $\lim_{P\to 1} K_1(P) = 0$ and $\lim_{P\to 0} K_0(P) = 0$ where it is assumed that Z can be independently varied for all X, and the limits are obtained by changing Z while holding X fixed.[77] These limit results simply state that when the values of X,Z are such that the probability of being in a sample is 1, there is no selection bias. One can approximate the $K_1(P)$ and $K_0(P)$ terms by polynomials in P (Heckman 1980; Heckman and Robb 1985, 1986; Heckman and Hotz 1989).

If $K_1(P(X, Z))$ can be independently varied from $\mu_1(X)$ and $K_0(P(X, Z))$ can be independently varied from $\mu_0(X)$, one can identify $\mu_1(X)$ and $\mu_0(X)$ up to constants. If there are limit sets $\mathbb{Z}_0$ and $\mathbb{Z}_1$ such that for each X $\lim_{Z\to\mathbb{Z}_0} P(X,Z) = 0$ and $\lim_{Z\to\mathbb{Z}_1} P(X,Z) = 1$, then one can identify these constants, since in those limit sets we identify $\mu_1(X)$ and $\mu_0(X)$.[78] Under these conditions, it is possible to nonparametrically identify all three conditional treatment parameters:

$$
\begin{aligned}
ATE(X) &= \mu_1(X) - \mu_0(X) \\
TT(X, D=1) &= \mu_1(X) - \mu_0(X) + E(U_1 - U_0 | X, D=1) \\
&= \mu_1(X) - \mu_0(X) + E_{Z|X,D=1} \\
&\quad \left[K_1(P(X,Z)) + \left(\frac{1-P}{P}\right) K_0(P(X,Z)) \right],^{79}
\end{aligned}
$$

[77]More precisely, assume that $Supp\ (Z|X) = Supp\ (Z)$ and that limit sets of Z, $\mathbb{Z}_0$, and $\mathbb{Z}_1$ exist such that as $Z \to \mathbb{Z}_0$, $P(Z, X) \to 0$ and as $Z \to \mathbb{Z}_1$, $P(Z, X) \to 1$. This is also the support condition used in the generalization of matching by Heckman, Ichimura, and Todd (1997).

[78]This condition is sometimes called "identification at infinity"; see Heckman (1990) or Andrews and Schafgans (1998).

[79]Since

$$
\begin{aligned}
E(U_0) &= 0 \\
&= E(U_0|D=1,X,Z)P(X,Z) + E(U_0|D=0,X,Z)(1-P(X,Z)) \\
E(U_0|D=1,X,Z) &= -\frac{(1-P(X,Z))}{P(X,Z)} E(U_0|D=0,X,Z) = -\frac{(1-P(X,Z))}{P(X,Z)} K_0(P(X,Z))
\end{aligned}
$$

See Heckman and Robb (1986). The expression $E_{Z|X,\ D=1}$ integrates out Z for a given X, $D = 1$.

$$\begin{aligned} MTE(X, Z, V = 0) &= \mu_1(X) - \mu_1(X) + E(U_1 - U_0 \mid \mu_V(Z, X) \\ &= -U_V) \\ &= \mu_1(X) - \mu_0(X) \\ &\quad + \frac{\partial[E(U_1 - U_0|X, Z, D = 1)P(X, Z)]}{\partial(P(X, Z))}.^{80} \end{aligned}$$

Unlike the method of matching, the method of control functions allows the marginal treatment effect to be different from the average treatment effect or from the effect of treatment on the treated (i.e., the second term on the right-hand side of the first equation for $MTE(X,\ Z,\ U = 0)$ is, in general, nonzero). Although conventional practice is to derive the functional forms of $K_0(P)$ and $K_1(P)$ by making distributional assumptions (e.g., normality or other conventional distributional assumptions about $(U_0,\ U_1,\ U_V)$; see Heckman, Tobias, and Vytlacil 2001, 2003), this is not an intrinsic feature of the method and there are many non-normal and semiparametric versions of this method (see Powell 1994 or Heckman and Vytlacil 2006a,b for surveys).

Without invoking parametric assumptions, the method of control functions requires an exclusion restriction (a Z not in X) to achieve nonparametric identification.[81] Without any functional form assumptions, one cannot rule out a worst-case analysis where—for example, if $X = Z$, $K_1(P(X)) = \alpha\mu(X)$ where α is a scalar. Then, there

[80]As established in Heckman and Vytlacil (2000, 2005) and Heckman (2001), under assumption (C-1) and additional regularity conditions

$$E(U_1-U_0|X,Z,D=1)P(X,Z)=\int_{-P(X,Z)}^{1}\int_{-\infty}^{\infty}(U_1-U_0)f\big(U_1-U_0|U_V^*\big)d(U_1-U_0)dU_V^*,$$

where $U_V^*=F_V(U_V)$, so

$$\frac{\partial[E(U_1-U_0|X,Z,D=1)P(X,Z)]}{\partial P(X,Z)}=E\big(U_1-U_0|U_V^*=-P(X,Z)\big).$$

The third expression follows from algebraic manipulation. Expressions conditional on X and $V = 0$ are obtained by integrating out Z conditional on X and $V = 0$.

[81]For many common functional forms for the distributions of unobservables, no exclusion is required.

is perfect collinearity between the control function and the conditional mean of the outcome equation, and it is impossible to control for selection with this method. Even though this case is not generic, it is possible. The method of matching does not require an exclusion restriction because it makes a stronger assumption, which we clarify below. Without additional assumptions, the method of control functions requires that, for some Z values for each X, $P(X, Z) = 1$ and $P(X, Z) = 0$ to achieve full nonparametric identification.[82] The conventional method of matching excludes this case.

Both methods require that treatment parameters be defined on a common support that is the intersection of the supports of X given $D = 1$ and X given $D = 0$:

$$Supp\ (X|D = 1) \cap Supp\ (X|D = 0).$$

A similar requirement is imposed on the generalization of matching with exclusion restrictions introduced in Heckman, Ichimura, Smith, and Todd (1998). Recall that exclusion (adding a Z in the probability of treatment equation that is not in the outcome equation where $\Pr(D = 1|X, Z)$ is the choice probability), both in matching and selection models, enlarges the set of X values that satisfy this condition. If $P(X, Z)$ depends on Z, then even if $P(X, Z) = 1$ for some $Z = z$ it can be that $P(X, Z) < 1$ for $Z = z'$ if $z \neq z'$. A similar argument applies to $P(X, Z) = 0$ for $Z = z''$ but $P(X, Z) > 0$ for $Z = z'''$ if $z'' \neq z'''$. This requires the existence of such Z values in the neighborhood of all values of X, Z such that $P(X, Z) = 0$ or 1.

In the method of control functions, $P(X, Z)$ is a conditioning variable used to predict U_1 conditional on D, X, and Z and U_0 conditional on D, X, and Z. In the method of matching, it is used to characterize the stochastic independence between (U_0, U_1) and D. In the method of control functions, as conventionally applied, $(U_0, U_1) \perp\!\!\!\perp (X, Z)$, but this assumption is not intrinsic to the method.[83]

[82]Symmetry of the errors can be used in place of the appeal to limit sets that put $P(X, Z) = 0$ or $P(X, Z) = 1$; see Chen (1999).

[83]Relaxing it, however, requires that the analyst model the dependence of the unobservables on the observables and that certain variation-free conditions are satisfied; see Heckman and Robb (1985).

This assumption plays no role in matching if the correct conditioning set is known (i.e., one that satisfies (M-1) and (M-2)). However, as noted in Heckman and Navarro (2004), exogeneity plays a key role in devising rules to select appropriate conditioning variables. The method of control functions does not require that $(U_0, U_1) \perp\!\!\!\perp D|(X, Z)$, which is a central requirement of matching. Equivalently, the method of control functions does not require

$$(U_0, U_1) \perp\!\!\!\perp U_V|(X, Z)$$

whereas matching does. Thus matching assumes access to a richer set of conditioning variables than is assumed in the method of control functions.

The method of control functions is more robust than the method of matching, in the sense that it allows for outcome unobservables to be dependent on D even after conditioning on (X, Z), and it models this dependence, whereas the method of matching assumes no such dependence. Matching under the assumed conditions is a special case of the method of control functions[84] in which under assumptions (M-1) and (M-2),

$$\begin{aligned} E(U_1|X, Z, D = 1) &= E(U_1|X, Z) \\ E(U_0|X, Z, D = 0) &= E(U_0|X, Z). \end{aligned}$$

In the method of control functions in the case when $(X, Z) \perp\!\!\!\perp (U_0, U_1, U_V)$

$$\begin{aligned} E(Y|X, Z, D) &= E(Y_1|X, Z, D = 1)D + E(Y_0|X, Z, D = 0)(1 - D) \\ &= \mu_0(X) + (\mu_1(X) - \mu_0(X))D \\ &\quad + E(U_1|X, Z, D = 1)D + E(U_0|P(X, Z), D = 0)(1 - D) \\ &= \mu_0(X) + (\mu_1(X) - \mu_0(X))D \\ &\quad + E(U_1|P(X, Z), D = 1)D + E(U_0|P(X, Z), D = 0)(1 - D) \\ &= \mu_0(X) + [\mu_1(X) - \mu_0(X) + K_1(P(X, Z)) - K_0(P(X, Z))]D \\ &\quad + K_0(P(X, Z)). \end{aligned}$$

[84]See Aakvik et al. (2005); Carneiro et al. (2003); and Cunha et al. (2005a, 2005b) for a generalization of matching that allows for selection on unobservables by imposing a factor structure on the errors and estimating the distribution of the unobserved factors.

To identify $\mu_1(X) - \mu_0(X)$, the average treatment effect, one must isolate it from $K_1(P(X, Z))$ and $K_0(P(X, Z))$. The coefficient on D in this regression does not correspond to any one of the treatment effects presented above.

Under assumptions (M-1) and (M-2) of the method of matching, one may write expressions conditional on $P(W)$:

$$E(Y|P(W), D) = \mu_0(P(W)) + \big[(\mu_1(P(W)) - \mu_0(P(W))) + E(U_1|P(W)) - E(U_0|P(W))\big] D + \{E(U_0|P(W))\}.$$

Notice that if the analyst further invokes (C-1)

$$E(Y|P(W), D) = \mu_0(P(W)) + [\mu_1(P(W)) - \mu_0(P(W))]D,$$

since $E(U_1|P(W)) = E(U_0|P(W)) = 0$. A parallel argument can be made conditioning on X and Z instead of $P(W)$.

Under the assumptions that justify matching, treatment effects ATE or TT (conditional on $P(W)$) are identified from the coefficient on D in either of the two preceding equations. It is not necessary to invoke (C-1) in the application of matching although it simplifies expressions. One can define the parameters conditional on X, allowing the X to be endogenous. Condition (M-2) guarantees that D is not perfectly predictable by W so the variation in D identifies the treatment parameter. Thus the coefficient on D in the regression associated with the more general control function model does not correspond to any treatment parameter whereas the coefficient on D in the regression associated with matching corresponds to a treatment parameter under the assumptions of the matching model. Under (C-1), $\mu_1(P(W)) - \mu_0(P(W)) = ATE$ and $ATE = TT = MTE$, so the method of matching identifies all of the (conditional on $P(W)$) mean treatment parameters.[85] Under the assumptions justifying matching, when means of Y_1 and Y_0 are the

[85]This result also holds if (C-1) is not satisfied, but then the treatment effects include

$$E(U_1|P(W)) - E(U_0|P(W))$$

.

parameters of interest, and W satisfies (M-1) and (M-2), the bias terms defined in Section 4.3 vanish. They do not in the more general case considered in the method of control functions. The vanishing of the bias terms in matching is the mathematical counterpart of the randomization implicit in matching: conditional on W or $P(W)$, (U_1, U_0) are random with respect to D. The method of control functions allows them to be nonrandom with respect to D. In the absence of functional form assumptions, an exclusion restriction is required in the analysis of control functions to separate out $K_0(P(X, Z))$ from the coefficient on D. Matching produces identification without exclusion restrictions whereas identification with exclusion restrictions is a central feature of the control function method in the absence of functional form assumptions. The implicit randomization in matching plays the role of an exclusion restriction in the method of instrumental variables.

The work of Rosenbaum (1995) and Robins (1997) implicitly recognizes that the control function approach is more general than the matching approach. Their sensitivity analyses for matching when there are unobserved conditioning variables are, in their essence, sensitivity analyses using control functions.[86] Aakvik, Heckman, and Vytlacil (2005), Carneiro, Hansen, and Heckman (2003), and Cunha, Heckman, and Navarro (2005a) explicitly model the relationship between matching and selection models using factor structure models, treating the omitted conditioning variables as unobserved factors and estimating their distribution.

Tables 2 and 3 perform sensitivity analyses under different assumptions about the parameters of the underlying selection model. In particular, I assume that the data are generated by the model of equations (22a)–(22c), with (22c) having the explicit representation

$$
\begin{aligned}
V &= Z\gamma + U_V, \\
(U_1, U_0, U_V)' &\sim N(0, \Sigma) \\
corr\ (U_j, U_V) &= \rho_{jV} \\
var\ (U_j) &= \sigma_j^2; \quad j = \{0, 1\}.
\end{aligned}
$$

[86]See also Vijverberg (1993), who performs a sensitivity analysis in a parametric selection model with an unidentified parameter.

TABLE 2
Mean Bias for Treatment on the Treated

ρ_{0V}	Average Bias ($\sigma_0 = 1$)	Average Bias ($\sigma_0 = 2$)
−1.00	−1.7920	−3.5839
−0.75	−1.3440	−2.6879
−0.50	−0.8960	−1.7920
−0.25	−0.4480	−0.8960
0.00	0.0000	0.0000
0.25	0.4480	0.8960
0.50	0.8960	1.7920
0.75	1.3440	2.6879
1.00	1.7920	3.5839

$$\text{BIAS}_{\text{TT}} = \rho_{0V} * \sigma_0 * M(p)$$

$$M(p) = \frac{\varphi(\Phi^{-1}(p))}{[p^*(1-p)]}$$

I assume no X and that $Z \perp\!\!\!\perp (U_1, U_0, U_V)$. Using the formulas presented in the appendix of Heckman and Navarro (2004), one can write the biases conditional on $Z = z$ as

$$\begin{aligned}
\text{Bias } TT(Z=z) &= \text{Bias } TT(P(Z)=p(z)) = \sigma_0\rho_{0V}M(p(z)) \\
\text{Bias } ATE(Z=z) &= \text{Bias } ATE(P)Z)=p(z)) \\
&= M(p(z))[\sigma_1\rho_{1V}(1-p(z)) + \sigma_0\rho_{0V}p(z)] \\
\text{Bias } MTE(Z=z) &= \text{Bias } MTE(P(Z)=p(z)) \\
&= M(p(z))[\sigma_1\rho_{1V}(1-p(z)) + \sigma_0\rho_{0V}p(z)] \\
&\quad - \Phi^{-1}(1-p(z))[\sigma_1\rho_{1V} - \sigma_0\rho_{0V}]
\end{aligned}$$

where $M(p(z)) = \frac{\phi(\Phi^{-1}(1-p(z)))}{p(z)(1-p(z))}$, $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function (pdf) and cumulative distribution function (cdf) of a standard normal random variable and $p(z)$ is the propensity score evaluated at $Z = z$. I assume that $\mu_1 = \mu_0$ so that the true average treatment effect is zero.

I simulate the mean bias for TT (Table 2) and ATE (Table 3) for different values of the ρ_{jV} and σ_j. The results in the tables show that, as one lets the variances of the outcome equations grow, the value of the mean bias that one obtains can become substantial. With larger correlations come larger biases. These

TABLE 3
Mean Bias for Average Treatment Effect

($\sigma_0 = 1$)									
ρ_{0V}	−1.00	−0.75	−0.50	−0.25	0	0.25	0.50	0.75	1.00
									continued
				ρ_{1V} ($\sigma_1 = 1$)					
−1.00	−1.7920	−1.5680	−1.3440	−1.1200	−0.8960	−0.6720	−0.4480	−0.2240	0
−0.75	−1.5680	−1.3440	−1.1200	−0.8960	−0.6720	−0.4480	−0.2240	0	0.2240
−0.50	−1.3440	−1.1200	−0.8960	−0.6720	−0.4480	−0.2240	0	0.2240	0.4480
−0.25	−1.1200	−0.8960	−0.6720	−0.4480	−0.2240	0	0.2240	0.4480	0.6720
0	−0.8960	−0.6720	−0.4480	−0.2240	0	0.2240	0.4480	0.6720	0.8960
0.25	−0.6720	−0.4480	−0.2240	0	0.2240	0.4480	0.6720	0.8960	1.1200
0.50	−0.4480	−0.2240	0	0.2240	0.4480	0.6720	0.8960	1.1200	1.3440
0.75	−0.2240	0	0.2240	0.4480	0.6720	0.8960	1.1200	1.3440	1.5680
1.00	0	0.2240	0.4480	0.6720	0.8960	1.1200	1.3440	1.5680	1.7920
				ρ_{1V} ($\sigma_1 = 2$)					
−1.00	−2.6879	−2.2399	−1.7920	−1.3440	−0.8960	−0.4480	0	0.4480	0.8960
−0.75	−2.4639	−2.0159	−1.5680	−1.1200	−0.6720	−0.2240	0.2240	0.6720	1.1200
−0.50	−2.2399	−1.7920	−1.3440	−0.8960	−0.4480	0	0.4480	0.8960	1.3440
−0.25	−2.0159	−1.5680	−1.1200	−0.6720	−0.2240	0.2240	0.6720	1.1200	1.5680
0	−1.7920	−1.3440	−0.8960	−0.4480	0	0.4480	0.8960	1.3440	1.7920
0.25	−1.5680	−1.1200	−0.6720	−0.2240	0.2240	0.6720	1.1200	1.5680	2.0159
0.50	−1.3440	−0.8960	−0.4480	0	0.4480	0.8960	1.3440	1.7920	2.2399
0.75	−1.1200	−0.6720	−0.2240	0.2240	0.6720	1.1200	1.5680	2.0159	2.4639

tables demonstrate the greater generality of the control function approach. Even if the correlation between the observables and the unobservables (ρ_{jV}) is small, so that one might think that selection on unobservables is relatively unimportant, one still obtains substantial biases if one does not control for relevant omitted conditioning variables. Only for special values of the parameters can one avoid bias by matching. These examples also demonstrate that sensitivity analyses can be conducted for analysis based on control function methods even when they are not fully identified, as noted by Vijverberg (1993).

4.4.3. *Instrumental Variables*

Both the method of matching and the method of control functions work with $E(Y|X, Z, D)$ and $\Pr(D = 1|X, Z)$. The method of instrumental variables works with $E(Y|X, Z)$ and $\Pr(D = 1|X, Z)$. There are two versions of the method of instrumental variables: (1) conventional linear instrumental variables and (2) local instrumental variables (*LIV*) (Heckman and Vytlacil 1999, 2000, 2006b; Heckman 2001). *LIV* is equivalent to a semiparametric selection model (Vytlacil 2002; Heckman and Vytlacil 2005, 2006b). It is an alternative way to implement the principle of control functions. *LATE* (Imbens and Angrist 1994) is a special case of *LIV* under the conditions I specify below.

I first consider the conventional method of instrumental variables. In this framework, $P(X, Z)$ arises less naturally than it does in the matching and control function approaches. Z is the instrument and $P(X, Z)$ is a function of the instrument.

Using the model of equations (22b) and (22c), I obtain

$$\begin{aligned} Y &= DY_1 + (1 - D)Y_0 \\ &= \mu_0(X) + (\mu_1(X) - \mu_0(X) + U_1 - U_0)D + U_0 \\ &= \mu_0(X) + \Delta(X)D + U_0, \end{aligned}$$

where $\Delta(X) = \mu_1(X) - \mu_0(X) + U_1 - U_0$. When $U_1 = U_0$, we obtain the conventional model to which *IV* is typically applied with

D correlated with U_0. Standard instrumental variable conditions apply and $P(X,Z)$ is a valid instrument if

$$E(U_0|P(X,Z),X) = E(U_0|X)^{87} \tag{IV-1}$$

and

$$\Pr(D = 1|X,Z) \tag{IV-2}$$

is a nontrivial function of Z for each X. When $U_1 \neq U_0$ but $D \perp\!\!\!\perp (U_1 - U_0)|X$ (or alternatively $U_V \perp\!\!\!\perp (U_1 - U_0)|X$), then the same two conditions identify (conditional on X):

$$\begin{aligned} ATE\ (X) &= E(Y_1 - Y_0|X) = E(\Delta(X)|X) \\ TT\ (X) &= E(Y_1 - Y_0|X, D = 1) = E(Y_1 - Y_0|X) = E(\Delta(X) \mid X) \\ &= MTE(X) \end{aligned}$$

and the marginal equals the average conditional on X and Z. The requirement that $D \perp\!\!\!\perp (U_1 - U_0)|X$ is strong and assumes that agents do not participate in the program on the basis of *any* information about unobservables in gross gains (Heckman and Robb 1985, 1986; Heckman 1997).[88]

How reasonable are the identifying assumptions of *IV*? An appeal to behavioral theory helps. Consider the use of draft lottery numbers as instruments (Z) for military service ($Z = 1$ if served in the army; $Z = 0$ otherwise). The question is how does military service affect earnings? (Angrist 1991). If agents participate in the military

[87]Observe that it is not required that $E\,(U_0|X) = 0$. We can write the *IV* estimator in the population as

$$\begin{aligned} \Delta^{IV}(x) &= \frac{E(Y|P(X=x,Z=z)=p_z,X=x) - E(Y|P(X=x,Z=z')=p_{z'},X=x)}{P(X=x,Z=z) - P(X=x,Z=z')} \\ &= \frac{[\mu_0(X) + \Delta(X)P(X=x,Z=z) + E(U_0|X) - \mu_0(X) + \Delta(X)P(X=x,Z=z) - E(U_0|X)]}{P(X=x,Z=z) - P(X=x,Z=z')} \\ &= \Delta(x) \end{aligned}$$

Thus it is not necessary to assume that $E\,(U_0 \mid X) = 0$.

[88]We define ATE conditional on X as

$$E(Y_1 - Y_0|X = x) = \mu_1(X) - \mu_0(X) + E(U_1 - U_0|X = x).$$

based in part on the gain in the outcome measure (Y_1, Y_0) (e.g., the difference in earnings) and this is a nondegenerate random variable, then (IV-1) is violated and IV does not identify ATE. The validity of the estimator is conditional on an untestable behavioral assumption. Similar remarks apply to $LATE$ as developed by Imbens and Angrist (1994) and popularized by Angrist, Imbens, and Rubin (1996); see Heckman and Vytlacil (1999, 2000, 2005), and Vytlacil (2002) for more discussion of the implicit behavioral assumptions underlying $LATE$.

The more interesting case for many problems arises when $U_1 \neq U_0$ and D $(U_1 - U_0)$ so agents participate in a program based at least in part on factors not measured by the economist. To identify $ATE(X)$ using IV, it is required that

$$E(U_0 + D(U_1 - U_0)|P(X,Z),X) = E(U_0 + D(U_1 - U_0)|X) \quad \text{(IV-3)}$$

and condition (IV-2) (Heckman and Robb 1985, 1986; Heckman 1997). To identify $TT(X)$ using IV, it is required that

$$\begin{aligned} &E(U_0 + D(U_1 - U_0) - E(U_0 + D(U_1 - U_0)|X)|P(X,Z),X) \\ &= E(U_0 + D(U_1 - U_0) - E(U_0 + D(U_1 - U_0)|X)|X) \end{aligned} \quad \text{(IV-4)}$$

and condition (IV-2). No simple conditions exist to identify the MTE using linear instrumental variables methods in the general case where D $(U_1 - U_0)|X$, Z. Heckman and Vytlacil (2001, 2005, 2006a,b) characterize what conventional IV estimates in terms of a weighted average of MTEs.

The conditions required to identify ATE using P as an instrument may be written in the following alternative form:

$$\begin{aligned} &E(U_0|P(X,Z),X) + E(U_1 - U_0|D = 1, P(X,Z),X)P(X,Z) \\ &= E(U_0|X) + E(U_1 - U_0|D = 1, X)P(X,Z). \end{aligned}$$

If $U_1 = U_0$ (everyone with the same X responds to treatment in the same way) or $(U_1 - U_0) \perp\!\!\!\perp D|P(X,Z), X$ (people do not participate in treatment on the basis of unobserved gains), then these conditions are the standard instrumental variable conditions. In general, the conditions are not satisfied by economic choice models, except under

special cancellations. If Z is a determinant of choices, and $U_1 - U_0$ is in the agent's choice set (or is only partly correlated with information in the agent's choice set), then this condition is not satisfied generically.

These identification conditions are fundamentally different from the conditions required to justify matching and control function methods. In matching, the essential condition for means (conditioning on X and $P(X, Z)$) is

$$E(U_0|X, D = 0, P(X, Z)) = E(U_0|X, P(X, Z))$$

and

$$E(U_1|X, D = 1, P(X, Z)) = E(U_1|X, P(X, Z)).$$

These conditions require that, conditional on $P(X, Z)$ and X, U_1, and U_0 are mean independent of U_V (or D). If (C-1) is invoked, $\mu_1(W)$ and $\mu_0(W)$ are the conditional means of Y_1 and Y_0 respectively, the two preceding expressions are zero. However, as I have stressed repeatedly, (C-1) is not strictly required in matching.

The method of control functions models and estimates the dependence of U_0 and U_1 on D rather than assuming that it vanishes like the method of matching. The method of linear instrumental variables requires that the composite error term $U_0 + D(U_1 - U_0)$ be mean independent of Z (or $P(X, Z)$), given X. Essentially, these conditions require that the dependence of U_0 and $D(U_1 - U_0)$ on Z vanish through conditioning on X. Matching requires that U_1 and U_0 are independent of D given (X, Z). These conditions are logically distinct. One set of conditions does not imply the other set (Heckman and Vytlacil 2006a,b). They are justified by different *a priori* assumptions. Hence the provisional nature of causal knowledge.

Assuming finite means, local instrumental variables methods developed by Heckman and Vytlacil (1999, 2001, 2005) estimate all three treatment parameters in the general case where $(U_1 - U_0) \not\perp\!\!\!\perp D|(X, Z)$ under the following additional conditions

$$\mu_D(Z) \quad \text{is a non-degenerate random variable given } X \quad \text{(LIV-1)}$$
(exclusion restriction)

$$(U_0, U_1, U_V) \perp\!\!\!\perp Z|X \tag{LIV-2}$$

$$0 < \Pr(D = 1|X) < 1 \tag{LIV-3}$$

$$Supp\ P(D = 1|X, Z) = [0, 1]. \tag{LIV-4}$$

Under these conditions

$$\frac{\partial E(Y|X, P(X, Z))}{\partial (P(X, Z))} = MTE(X, P(X, Z), V = 0).^{89}$$

Only (LIV-1)–(LIV-3) are required to identify this parameter locally. (LIV-4) is required to use the MTE to identify the standard treatment parameters.

As demonstrated by Heckman and Vytlacil (1999, 2000, 2005) and Heckman (2001), over the support of (X, Z), MTE can be used to construct (under LIV-4) or bound (in the case of partial support of $P(Z)$) ATE and TT. Policy-relevant treatment effects can be defined.

[89]Proof: From the law of iterated expectations,

$$\begin{aligned} E(Y|X, P(Z)) =& E(Y_1|D = 1, X, P(Z))P(Z) \\ &+ E(Y_0|D = 0, X, P(Z))(1 - P(Z)) \\ =& \int_{-\infty}^{\infty} \int_{-P(Z)}^{\infty} y_1 f(y_1, U_V^*|X) dU_V^* dy_1 \\ &+ \int_{-\infty}^{\infty} \int_{-\infty}^{-P(Z)} y_0 f(y_0, U_V^*|X) dU_V^* dy_0 \end{aligned}$$

where $U_V^* = F_V(U_V)$. Thus

$$\begin{aligned} \frac{\partial E(Y|X, P(Z))}{\partial P(Z)} =& E(Y_1 - Y_0|X, U_V^* = -P(Z)) \\ =& MTE \end{aligned}.$$

LATE is a special case of this method.[90] The *LIV* approach unifies matching, control functions, and classical instrumental variables under a common set of assumptions. Table 4 summarizes the alternative assumptions used in matching, control functions, and instrumental variables to identify treatment parameters identify conditional (on X or X, Z).

4.4.4. *Directed Acyclic Graphs and the Method of* g-*Computation*

Directed acyclic graphs (DAG) (Pearl 2000) or the *g*-computation algorithm (Robins 1989) have recently been advocated as mechanisms for causal discovery. These methods improve on the method of matching by making explicit *some* of the sources of the unobservables generating the outcomes and postulating their relationships to observables. My discussion is more brief and considers only one population-level causal effect. It is based on Freedman (2001).

Figure 2, patterned after Freedman (2001), shows the essence of the method. An unobserved confounder A is a determinant of outcome F and variable B.[91] We observe (B, C, F). Unobservables are denoted by 'U'. Each of (B, C, F) is assumed to be a random variable produced in part from the variable preceding it in the triangle and from unobservables that are assumed to be mutually independent (hence the pattern of the arrows in Figure 2). Assume for simplicity that A, B, C, F are discrete random variables. Figure 2 describes a recursive model where $A = (U_A)$, C and U_F determine F; B and U_C determine C and U_B and $A = (U_A)$ determine B.

We seek to determine

$$\Pr(F = f|\text{set } B = b)$$

free of the unmeasured cofounder A, which affects both B and F. This is the probability of getting F when we set $B = b$. ("Set" is Pearl's (2000) "do" operation or Haavelmo's (1943) "fixing of the variables.") But there is confounding due to A. $A = U_A$ affects both B and F, so there may be no true causal $B - F$ relationship. How can one control for A?

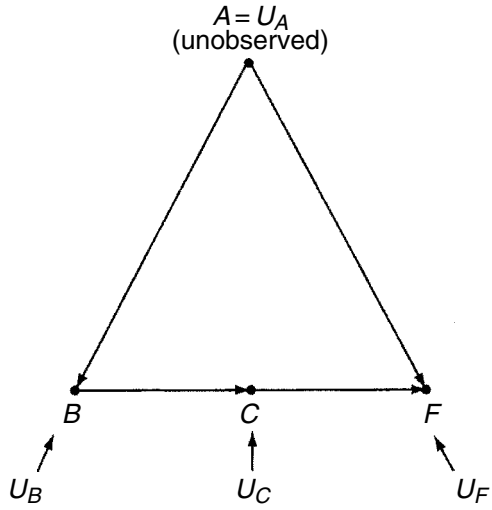
[90]Vytlacil (2002) establishes that *LATE* is a semiparametric version of a control function estimator.

[91]The symbols used in this subsection are not the same as those used in the previous sections of this paper.

TABLE 4
Identifying Assumptions and Implicit Economic Assumptions Underlying the Four Methods Discussed in this Paper
Conditional on X and Z

Method	Exclusion Required?	Separability of Observables and Unobservables in Outcome Equations?	Functional Forms Required?	Marginal = Average? (Given X, Z)	Key Identification Condition for Means (assuming separability)
Matching*	No	No	No	Yes	$E(U_1\|X, D = 1, Z) = E(U_1\|X, Z)$ $E(U_0\|X, D = 0, Z) = E(U_0\|X, Z)$
Control Function**	Yes (for nonparametric identification)	Conventional, but not required	Conventional, but not required	No	$E(U_0\|X, D = 0, Z)$ and $E(U_1\|X, D = 1, Z)$ can be varied independently of $\mu_0(X)$ and $\mu_1(X)$, respectively and intercepts can be identified through limit arguments or symmetry assumptions
IV (conventional)	Yes	Yes	No	No (Yes in standard case)	$E(U_0 + D(U_1 - U_0)\|X, Z)$ $= E(U_0 + D(U_1 - U_0)\|X)$ (ATE) $E(U_0 + D(U_1 - U_0) -$ $E(U_0 + D(U_1 - U_0)\|X)\|P(Z), X)$ $= E(U_0 + D(U_1 - U_0) - E(U_0 + D(U_1$ $- U_0)\|X)\|X)$ (TT)
LIV	Yes	No	No	No	$(U_0, U_1, U_v) \perp\!\!\!\perp Z\|X$ $\Pr(D = 1\|Z, X)$ is a nontrivial function of Z for each X.

*For propensity score matching, (X, Z) are replaced with $P(X, Z)$ in defining parameters and conditioning sets.
**Conditions for writing the control function in terms of $P(X, Z)$ are given in the text.



We know

$$\Pr(C=c \mid B=b)$$

$$\Pr(F=f \mid C=c)=\sum_a \Pr(F=f \mid A=a, C=c)\Pr(A=a)$$

and

$$\Pr(F=f \mid B=b)=\sum_c \Pr(F=f \mid C=c)\Pr(C=c \mid B=b)$$

FIGURE 2. DAG analysis. Adapted from Freedman (2001).

The *g*-computation algorithm operates by computing the following probabilities based on observables. From the data, we can compute $\Pr(C = c|B = b)$. We can also compute the left-hand side of

$$\Pr(F = f|C = c) = \sum_a \Pr(F = f|A = a, C = c) \Pr(A = a).$$

Hence we can identify the desired causal object using the following calculation:

$$\Pr(F = f|\text{set } B = b) = \sum_c \Pr(F = f|C = c) \Pr(C = c|B = b).$$

The ingredients on the right-hand side can be calculated from the available data (recall that A is not observed).

This very useful result breaks down entirely if we add an arrow like that shown in Figure 3, because in this case A also confounds C. The role of the *a priori* theory is to specify the arrows. No purely empirical algorithm can find causal effects in general models, a point emphasized by Freedman (2001). Figure 4 shows another case where the g-computation approach breaks down in nonrecursive simultaneous equations models. $F - C$ and $U_F - U_C$ interdependence create further problems ruled out in the DAG approach. These examples all illustrate the provisional nature of causal inference and the role of theory in justifying the estimators of causal effects.

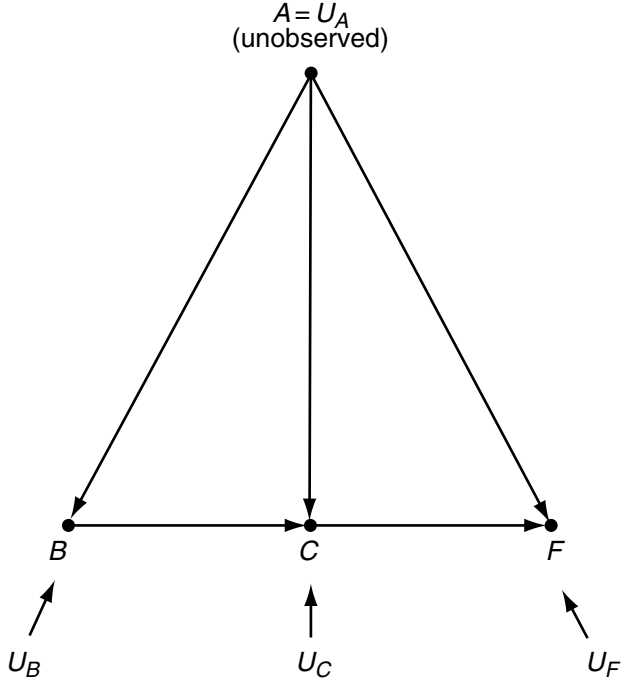


FIGURE 3. If another arrow is added to Figure 2, the argument breaks down. Where do arrows come from?

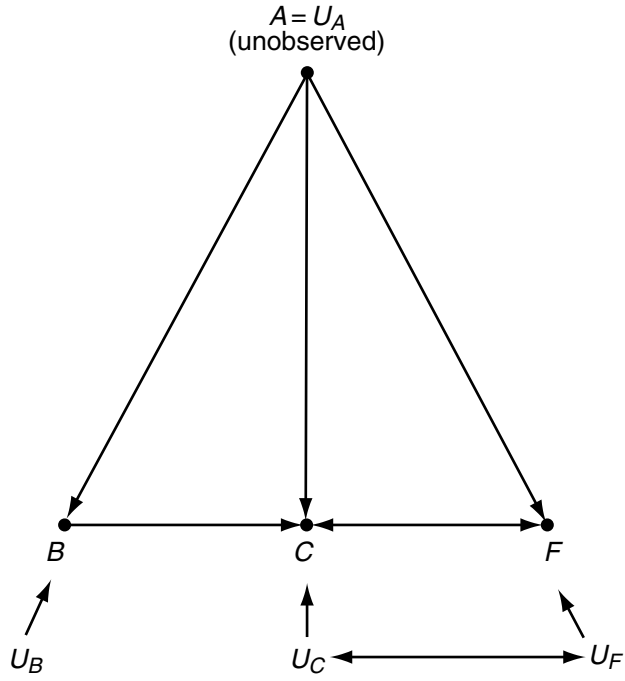


FIGURE 4. Nonrecursive. Argument breaks down. DAG is one estimation scheme for one hypothetical model, not a general algorithm for causal discovery.

5. SUMMARY AND CONCLUSIONS

This paper defines counterfactual models, causal parameters, and structural models and relates the parameters of the treatment effect literature to the parameters of structural econometrics and scientific causal models. I distinguish counterfactuals from scientific causal models. Counterfactuals are an ingredient of causal models. Scientific causal models also specify a mechanism for selecting counterfactuals. I present precise definitions of causal effects within structural models that are inclusive of the specification of a mechanism (a formal model) by which causal variables are externally manipulated (i.e., outcomes are selected). Models of causality advocated in statistics are incomplete because they do not specify the mechanisms of external variation that are central to the definition of causality, nor do they specify the sources of randomness producing outcomes and the relationship between outcomes and

selection mechanisms. By not determining the causes of effects, or modeling the relationship between potential outcomes and assignment to treatment, statistical models of causality cannot be used to provide valid answers to the numerous counterfactual questions required for policy analysis. They do not exploit relationships among potential outcomes, assignment to treatment, and the variables causing potential outcomes that can be used to devise econometric evaluation estimators. The statistical approach does not model the choice of treatment mechanism and its relationship with outcome equations, whereas the scientific approach makes the choice of treatment equation a centerpiece of identification analysis. The statistical model does not apply to nonrecursive settings, whereas the econometric model can be readily adapted to handle both recursive and nonrecursive cases.

Statistical treatment effects are typically proposed to answer a more limited set of questions than are addressed by structural equation models and it is not surprising that they can do so under weaker conditions than are required to identify structural equations. At the same time, if treatment effects are used structurally—that is, to forecast the effect of a program on new populations or to forecast the effects of new programs—stronger assumptions are required of the sort used in standard structural econometrics (see Heckman 2001; Heckman and Vytlacil 2005, 2006b).

Table 5 compares scientific models with statistical "causal" models. Statistical causal models, in their current state, are not fully articulated models. Crucial assumptions about sources of randomness are kept implicit. The assumptions required to project treatment parameters to different populations are not specified. The scientific approach has no substitute for making out-of-sample predictions—that is, for answering policy questions P2 and P3. The scientific approach distinguishes derivation of a model as an abstract theoretical activity from the problem of identifying models from data.

APPENDIX: THE VALUE OF STRUCTURAL EQUATIONS IN MAKING POLICY FORECASTS

Structural equations are useful for three different purposes. First, the derivatives of such functions or finite changes generate the

TABLE 5
Econometric Versus Statistical Causal Models

	Statistical Causal Models	Econometric Models
Sources of randomness	Implicit	Explicit
Models of conditional counterfactuals	Implicit	Explicit
Mechanism of intervention for defining counterfactuals	Hypothetical randomization	Many mechanisms of hypothetical interventions including randomization; mechanism is explicitly modeled
Treatment of interdependence	Recursive	Recursive or simultaneous systems
Social/market interactions	Ignored	Modeled in general equilibrium frameworks
Projections to different populations?	Does not project	Projects
Parametric?	Nonparametric	Becoming nonparametric
Range of questions answered	One focused treatment effect	In principle, answers many possible questions

continued

comparative statics *ceteris paribus* variations produced by scientific theory. For example, tests of economic theory and measurements of economic parameters (price elasticities, measurements of consumer surplus, etc.) are often based on structural equations.

Second, structural equations can be used to forecast the effects of policies evaluated in one population in other populations, provided that the parameters are invariant across populations and that support conditions are satisfied. However, a purely nonparametric structural equation determined on one support cannot be extrapolated to other populations with different supports.

Third, as emphasized by Marschak (1953), Marshallian causal functions and structural equations are one ingredient required to forecast the effect of a new policy, never previously implemented.

The problem of forecasting the effects of a policy evaluated on one population but applied to another population can be formulated in the following way. Let $Y(\omega) = \varphi(X(\omega), U(\omega))$, where $\varphi : \mathcal{D} \rightarrow \mathcal{Y}$, $\mathcal{D} \subseteq \mathbb{R}^J$, $\mathcal{Y} \subseteq \mathbb{R}$. φ is a structural equation determining outcome Y, and we assume that it is known only over $Supp(X(\omega), U(\omega)) = \mathcal{X} \times \mathcal{U}$. $X(\omega)$ and $U(\omega)$ are random input variables. The mean outcome conditional on $X(\omega) = x$ is

$$E_H(Y|X = x) = \int_{\mathcal{U}} \varphi(X = x, u) dF_H(u|X = x),$$

where $F_H(u|X)$ is the distribution of U in the historical data. We seek to forecast the outcome in a target population that may have a different support. The average outcome in the target population (T) is

$$E_T(Y|X = x) = \int_{\mathcal{U}^T} \varphi(X = x, u) dF_T(u|X = x)$$

where $\mathcal{U}^T$ is the support of U in the target population. Provided the support of (X, U) is the same in the source and the target populations, from knowledge of F_T it is possible to produce a correct value of $E_T(Y|X = x)$ on the target population. Otherwise, it is possible to evaluate this expectation only over the intersection set $Supp_T(X) \cap Supp_H(X)$, where $Supp_A(X)$ is the support of X in the A population. In order to extrapolate over the whole set $Supp_T(X)$, it is necessary to adopt some form of parametric or functional structure. Additive

separability in φ simplifies the extrapolation problem. If φ is additively separable

$$Y = \varphi(X) + U,$$

$\varphi(X)$ applies to all populations for which we can condition on X. However, some structure may have to be imposed to extrapolate from $Supp_H(X)$ to $Supp_T(X)$ if $\varphi(X)$ on T is not determined nonparametrically from H.

The problem of forecasting the effect of a new policy, never previously experienced, is similar in character to the policy forecasting problem just discussed. It shares many elements in common with the problem of forecasting the demand for a new good, never previously consumed.[92] Without imposing some structure on this problem, it is impossible to solve. The literature in structural econometrics associated with the work of the Cowles Commission adopts the following five-step approach to this problem.

1. Structural functions are determined (e.g., $\varphi(X)$).
2. The new policy is characterized by an invertible mapping from observed random variables to the characteristics associated with the policy: $C = q(X)$, where c is the set of characteristics associated with the policy and q, $q{:}\mathbb{R}^J \to \mathbb{R}^J$, is a *known* invertible mapping.
3. $X = q^{-1}(C)$ is solved to associate characteristics that in principle can be observed with the policy. This places the characteristics of the new policy on the same footing as those of the old.
4. It is assumed that, in the historical data, $Supp(q^{-1}(C) \subseteq Supp(X))$. This ensures that the support of the new characteristics mapped into X space is contained in the support of X. If this condition is not met, some functional structure must be used to forecast the effects of the new policy, to extend it beyond the support of the source population.
5. The forecast effect of the policy on Y is $Y(C) = \varphi(q^{-1}(C))$.

[92]Quandt and Baumol (1966), Lancaster (1971), Gorman (1980), McFadden (1974), and Domencich and McFadden (1975) consider the problem of forecasting the demand for a new good. Marschak (1953) is the classic reference for evaluating the effect of a new policy; see Heckman (2001).

The leading example of this approach is Lancaster's method for estimating the demand for a new good (Lancaster 1971). New goods are viewed as bundles of old characteristics. McFadden's conditional logit scheme (1974) is based on a similar idea.[93]

Marschak's analysis of the effect of a new commodity tax is another example. Let $P(\omega)$ be the random variable denoting the price facing consumer ω. The tax changes the product price from $P(\omega)$ to $P(\omega)(1 + t)$, where t is the tax. With sufficient price variation so that the assumption in step 4 is satisfied so that the support of the price after tax, $Supp_{\text{post tax}}(P(\omega)(1 + t)) \subseteq Supp_{\text{pretax}}(P(\omega))$, it is possible to use reduced form demand functions fit on a pretax sample to forecast the effect of a tax never previously put in place. Marschak uses a linear structural equation to solve the problem of limited support. From linearity, determination of the structural equations over a small region determines it everywhere.

Marshallian or structural causal functions are an essential ingredient in constructing such forecasts because they explicitly model the relationship between U and X. The treatment effect approach does not explicitly model this relationship so that treatment parameters cannot be extrapolated in this fashion, unless the dependence of potential outcomes on U and X is specified, and the required support conditions are satisfied. The Rubin (1978)–Holland (1986) model does not specify the required relationships.

REFERENCES

Aakvik, A., J. J. Heckman, and E. J. Vytlacil. 1999. "Training Effects on Employment When the Training Effects are Heterogeneous: An Application

[93]McFadden's stochastic specification is different from Lancaster's specification. See Heckman and Snyder (1997) for a comparison of these two approaches. Lancaster assumes that the $U(\omega)$ are the same for each consumer in all choice settings. (They are preference parameters in his setting.) McFadden allows for $U(\omega)$ to be different for the same consumer across different choice settings but assumes that the $U(\omega)$ in each choice setting are draws from a common distribution that can be determined from the demand for old goods.

to Norwegian Vocational Rehabilitation Programs." University of Bergen Working Paper 0599.

———. 2005. "Estimating Treatment Effects for Discrete Outcomes When Responses to Treatment Vary: An Application to Norwegian Vocational Rehabilitation Programs." *Journal of Econometrics* 125(1–2):15–51.

Abadie, A. 2003. "Semiparametric Differences-in-Differences Estimators." Department of Economics, Harvard University, Unpublished manuscript.

Abadie, A., J. D. Angrist, and G. Imbens. 2002. "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings." *Econometrica* 70(1):91–117.

Abbring, J. H., and G. J. Van Den Berg. 2003. "The Nonparametric Identification of Treatment Effects in Duration Models." *Econometrica* 71(5):1491–517.

Andrews, D. W., and M. M. Schafgans. 1998. "Semiparametric Estimation of the Intercept of a Sample Selection Model." *Review of Economic Studies* 65(3):497–517.

Angrist, J. D. 1991. "The Draft Lottery and Voluntary Enlistment in the Vietnam Era." *Journal of the American Statistical Association* 86(415):584–95.

Angrist, J. D., G. W. Imbens, and D. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91:444–55.

Björklund, A., and R. Moffitt. 1987. "The Estimation of Wage Gains and Welfare Gains in Self-selection." *Review of Economics and Statistics* 69(1):42–49.

Boadway, R. W., and N. Bruce. 1984. *Welfare Economics*. New York: Blackwell Publishers.

Brock, W. A., and S. N. Durlauf 2001. "Interactions-based models." Pp. 3463–68 in *Handbook of Econometrics*, Vol. 5, edited by J. J. Heckman and E. Leamer. New York: North-Holland.

Cameron, S. V., and J. J. Heckman. 1998. "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males." *Journal of Political Economy* 106(2):262–333.

Campbell, D. T., and J. C. Stanley. 1963. *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand McNally.

Carneiro, P. 2002. "Heterogeneity in the Returns to Schooling: Implications for Policy Evaluation." Ph. D. dissertation, University of Chicago.

Carneiro, P., K. Hansen, and J. J. Heckman. 2001. "Removing the Veil of Ignorance in Assessing the Distributional Impacts of Social Policies." *Swedish Economic Policy Review* 8(2):273–301.

———. 2003. "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice." 2001 Lawrence R. Klein Lecture. *International Economic Review* 44(2):361–422.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil. 2005. "Understanding What Instrumental Variables Estimate: Estimating Marginal and Average Returns to Education." Department of Economics, University of Chicago. Unpublished manuscript.

Chen, S. 1999. "Distribution-free Estimation of the Random Coefficient Dummy Endogenous Variable Model." *Journal of Econometrics* 91(1):171–99.

Cox, D. 1958. *Planning of Experiments*. New York: Wiley.

———. 1992. "Causality: Some Statistical Aspects." *Journal of the Royal Statistical Society*, Series A, 155:291–301.

Cox, D., and N. Wermuth. 1996. *Multivariate Dependencies: Models, Analysis and Interpretation*. New York: Chapman and Hall.

Cunha, F., J. Heckman, and S. Navarro. 2005a. "Counterfactual Analysis of Inequality and Social Mobility." In *Income Inequality*, edited by M. Gretzky. Palo Alto: Stanford University Press. Forthcoming.

———. 2005b. "Separating Heterogeneity from Uncertainty in Modeling Schooling Choices." *Oxford Economic Papers* 57(2):191–261.

Dawid, A. 2000. "Causal Inference Without Counterfactuals." *Journal of the American Statistical Association* 95(450):407–24.

Domencich, T., and D. L. McFadden. 1975. *Urban Travel Demand: A Behavioral Analysis*. Amsterdam: North-Holland.

Fisher, R. A. 1966. *The Design of Experiments*. New York: Hafner.

Florens, J.-P., and J. J. Heckman. 2003. "Causality and Econometrics." Department of Economics, University of Chicago. Unpublished working paper.

Foster, J. E., and A. K. Sen. 1998. *On Economic Inequality*. New York: Oxford University Press.

Freedman, D. 2001. "On Specifying Graphical Models for Causation and the Identification Problem." Department of Statistics, University of California at Berkeley. Unpublished manuscript.

Gill, R. D., and J. M. Robins. 2001. "Causal Inference for Complex Longitudinal Data: The Continuous Case." *Annals of Statistics* 29(6):1785–1811.

Gorman, W. M. 1980. "A Possible Procedure for Analysing Quality Differentials in the Egg Market." *Review of Economic Studies* 47(5):843–56.

Haavelmo, T. 1943. "The Statistical Implications of a System of Simultaneous Equations." *Econometrica* 11(1):1–12.

———. 1944. "The Probability Approach in Econometrics." *Econometrica* 12(suppl.):iii–vi; 1–115.

Hahn, J. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica* 66(2):315–31.

Harsanyi, J. C. 1955. "Cardinal Welfare, Individualistic Ethics and Interpersonal Comparisons of Utility." *Journal of Political Economy* 63(4):309–21.

———. 1975. "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory." *American Political Science Review* 69(2):594–606.

Heckman, J. J. 1976. "Simultaneous Equation Models with Both Continuous and Discrete Endogenous Variables with and Without Structural Shift in the Equations." Pp. 235–72 in *Studies in Nonlinear Estimation*, edited by S. Goldfeld and R. Quandt. Cambridge, MA: Ballinger.

———. 1978. "Dummy Endogenous Variables in a Simultaneous Equation System." *Econometrica* 46(4):931–59.

———. 1980. "Sample Selection Bias as a Specification Error with an Application to the Estimation of Labor Supply Functions." Pp. 206–48 in *Female Labor Supply: Theory and Estimation*, edited by J. P. Smith. Princeton, NJ: Princeton University Press.

———. 1990. "Varieties of Selection Bias." *American Economic Review* 80(2), 313–18.

———. 1992. "Randomization and Social Policy Evaluation." Pp. 201–30 in *Evaluating Welfare and Training Programs*, edited by C. Manski and I. Garfinkel. Cambridge, MA: Harvard University Press.

———. 1997. "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations." *Journal of Human Resources* 32(3):441–62.

———. 2000. "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective." *Quarterly Journal of Economics* 115(1):45–97.

———. 2001. "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture." *Journal of Political Economy* 109(4):673–748.

———. 2006. *Evaluating Economic Policy*. Princeton, NJ: Princeton University Press.

Heckman, J. J., and B. E. Honoré. 1990. "The Empirical Content of the Roy Model." *Econometrica* 58(5):1121–49.

Heckman, J. J., and V. J. Hotz. 1989. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association* 84(408):862–74.

Heckman, J. J., H. Ichimura, J. Smith, and P. E. Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66(5):1017–98.

Heckman, J. J., H. Ichimura, and P. E. Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies* 64(4):605–54.

———. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65(223):261–94.

Heckman, J. J., R. J. LaLonde, and J. A. Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." Pp. 1865–2097 in *Handbook of Labor Economics*, Vol. 3A, edited by O. Ashenfelter and D. Card. New York: North-Holland.

Heckman, J. J., and T. E. MaCurdy. 1985. "A Simultaneous Equations Linear Probability Model." *Canadian Journal of Economics* 18(1):28–37.

Heckman, J. J., and S. Navarro. 2004. "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models." *Review of Economics and Statistics* 86(1):30–57.

———. 2006. "Dynamic Discrete Choice and Dynamic Treatment Effects." *Journal of Econometrics*. Forthcoming.

Heckman, J. J., and R. Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." Pp. 156–245 in *Longitudinal Analysis of Labor*

Market Data, Vol. 10, edited by J. Heckman and B. Singer. New York: Cambridge University Press.

———. 1986. "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes." Pp. 63–107 in *Drawing Inferences from Self-Selected Samples*, edited by H. Wainer. New York: Springer-Verlag.

Heckman, J. J., and J. A. Smith. 1998. "Evaluating the Welfare State." Pp. 241–318 in *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*, edited by S. Strom. New York: Cambridge University Press.

Heckman, J. 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(221):487–536.

Heckman, J. J., and J. M. Snyder Jr. 1997. "Linear Probability Models of the Demand for Attributes with an Empirical Application to Estimating the Preferences of Legislators" (Special issue). *RAND Journal of Economics* 28:S142.

Heckman, J. J., J. L. Tobias, and E. J. Vytlacil. 2001. "Four Parameters of Interest in the Evaluation of Social Programs." *Southern Economic Journal* 68(2):210–23.

———. 2003. "Simple Estimators for Treatment Parameters in a Latent Variable Framework." *Review of Economics and Statistics* 85(3):748–54.

Heckman, J. J., and E. J. Vytlacil. 1999. "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." *Proceedings of the National Academy of Sciences* 96:4730–34.

———. 2000. "The Relationship Between Treatment Parameters Within a Latent Variable Framework." *Economics Letters* 66(1):33–39.

———. 2001. "Local Instrumental Variables." Pp. 1–46 in *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, edited by C. Hsiao, K. Morimue, and J. L. Powell. New York: Cambridge University Press.

———. 2005. "Structural Equations, Treatment Effects and Econometric Policy Evaluation." *Econometrica* 73(3):669–738.

———. 2006a. "Econometric Evaluation of Social Programs," "Part I: Causal Models, Structural Models and Econometric Policy Evaluation." In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics, Volume 6*. Amsterdam: Elsevier, forthcoming.

———. 2006b. "Econometric Evaluation of Social Programs," "Part II: Using Economic Choice Theory and the Marginal Treatment Effect to Organize Alternative Econometric Estimators." In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics, Volume 6*. Amsterdam: Elsevier, forthcoming.

Holland, P. W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945–60.

———. 1988. "Causal Inference, Path Analysis, and Recursive Structural Equation Models." Pp. 449–84 in *Sociological Methodology*, Vol. 18, edited by C. Clogg and G. Arminger. Washington, DC: American Sociological Association.

Hurwicz, L. 1962. "On the Structural Form of Interdependent Systems." Pp. 232–39 in *Logic, Methodology and Philosophy of Science*, edited by E. Nagel, P. Suppes, and A. Tarski. Stanford, CA: Stanford University Press.

Imbens, G. W., and J. D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2):467–75.

Katz, D., A. Gutek, R. Kahn, and E. Barton. 1975. *Bureaucratic Encounters: A Pilot Study in the Evaluation of Government Services*. Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan.

Knight, F. 1921. *Risk, Uncertainty and Profit*. New York: Houghton Mifflin.

Lancaster, K. J. 1971. *Consumer Demand: A New Approach*. New York: Columbia University Press.

Leamer, E. E. 1985. "Vector Autoregressions for Causal Inference?" *Carnegie-Rochester Conference Series on Public Policy* 22:255–303.

Lechner, M. 2004. "Sequential Matching Estimation of Dynamic Causal Models." Technical Report 2004, IZA Institute for the Study of Labor Discussion Paper.

Lewis, H. G. 1974. "Comments on Selectivity Biases in Wage Comparisons." *Journal of Political Economy* 82(6):1145–55.

Lucas, R. E., and T. J. Sargent. 1981. *Rational Expectations and Econometric Practice*. Minneapolis: University of Minnesota Press.

Marschak, J. 1953. "Economic Measurements for Policy and Prediction." Pp. 1–26 in *Studies in Econometric Method*, edited by W. Hood and T. Koopmans. New York: Wiley.

Marshall, A. 1890. *Principles of Economics*. New York: Macmillan.

Matzkin, R. 2003. "Nonparametric Estimation of Nonadditive Random Functions." *Econometrica* 71(5):1339–75.

———. 2004. "Unobserved Instruments." Department of Economics, Northwestern University, Evanston, IL. Unpublished manuscript.

———. 2006. "Nonparametric Identification." In *Handbook of Econometrics*, Vol. 6, edited by J. Heckman and E. Leamer. Amsterdam: Elsevier.

McFadden, D. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers in Econometrics*, edited by P. Zarembka. New York: Academic Press.

———. 1981. "Econometric Models of Probabilistic Choice." In *Structural Analysis of Discrete Data with Econometric Applications*, edited by C. Manski and D. McFadden. Cambridge, MA: MIT Press.

Mill, J. S. 1848. *Principles of Political Economy with Some of Their Applications to Social Philosophy*. London: J. W. Parker.

Moulin, H. 1983. *The Strategy of Social Choice*. New York: North-Holland.

Neyman, J. 1923. "Statistical Problems in Agricultural Experiments." *Journal of the Royal Statistical Society* Series B (suppl.) (2):107–80.

Pearl, J. 2000. *Causality*. Cambridge, England: Cambridge University Press.
Powell, J. L. 1994. "Estimation of Semiparametric Models." Pp. 2443–521 in *Handbook of Econometrics*, Vol. 4, edited by R. Engle and D. McFadden. Amsterdam: Elsevier.
Quandt, R. E. 1958. "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes." *Journal of the American Statistical Association* 53(284):873–80.
———. 1972. "A New Approach to Estimating Switching Regressions." *Journal of the American Statistical Association* 67(338):306–10.
———. 1974. "A Comparison of Methods for Testing Nonnested Hypotheses." *Review of Economics and Statistics* 56(1):92–99.
Quandt, R. E., and W. Baumol. 1966. "The Demand for Abstract Transport Modes: Theory and Measurement." *Journal of Regional Science* 6:13–26.
Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Belknap.
Robins, J. M. 1989. "The Analysis of Randomized and Non-randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies." Pp. 113–59 in *Health Services Research Methodology: A Focus on AIDS*, edited by L. Sechrest, H. Freeman, and A. Mulley. Rockville, MD: U.S. Department of Health and Human Services, National Center for Health Services Research and Health Care Technology Assessment.
———. (1997). "Causal Inference from Complex Longitudinal Data." Pp. 69–117 in *Latent Variable Modeling and Applications to Causality*. *Lecture Notes in Statistics*, edited by M. Berkane. New York: Springer-Verlag.
Rosenbaum, P. R. 1987. "Model-Based Direct Adjustment." *Journal of the American Statistical Association* 82(398):387–94.
———. 1995. *Observational Studies*. New York: Springer-Verlag.
Rosenbaum, P. R., and D. B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41–55.
Roy, A. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* 3(2):135–46.
Rubin, D. B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6(1):34–58.
———. 1986. "Statistics and Casual Inference: Comment: Which Ifs Have Casual Answers." *Journal of the American Statistical Association* 81(396):961–62.
Rubin, D. B., and N. Thomas. 1992. "Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions." *Biometrika* 79(4):797–809.
Ruud, P. A. 2000. *An Introduction to Classical Econometric Theory*. New York: Oxford University Press.
Sen, A. K. 1999. "The Possibility of Social Choice." *American Economic Review* 89(3):349–78.
Sims, C. A. 1977. "Exogeneity and Casual Orderings in Macroeconomic Models." Pp. 23–43 in *New Methods in Business Cycle Research*. Minneapolis, MN: Federal Reserve Bank of Minneapolis.

Tamer, E. 2003. "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria." *Review of Economic Studies* 70(1):147–65.

Thurstone, L. 1930. *The Fundamentals of Statistics*. New York: Macmillan.

Tukey, J. 1986. "Comments on Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes." Pp. 108–10 in *Drawing Inferences from Self-Selected Samples*, edited by H. Wainer. New York: Springer-Verlag.

Vickrey, W. 1945. "Measuring Marginal Utility by Reactions to Risk." *Econometrica* 13(4):319–33.

———. 1960. "Utility, Strategy, and Social Decision Rules." *Quarterly Journal of Economics* 74(4):507–35.

Vijverberg, W. P. M. 1993. "Measuring the Unidentified Parameter of the Extended Roy Model of Selectivity." *Journal of Econometrics* 57(1–3):69–89.

Vytlacil, E. J. 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica* 70(1):331–41.

Wainer, H. (Ed.) 1986. *Drawing Inferences from Self-Selected Samples*. New York: Springer-Verlag (Reprinted in 2000, Mahwah, NJ: Lawrence Erlbaum Associates).