Econometrica, Vol. 73, No. 3 (May, 2005), 669–738

STRUCTURAL EQUATIONS, TREATMENT EFFECTS, AND ECONOMETRIC POLICY EVALUATION[1]

BY JAMES J. HECKMAN AND EDWARD VYTLACIL

This paper uses the marginal treatment effect (MTE) to unify the nonparametric literature on treatment effects with the econometric literature on structural estimation using a nonparametric analog of a policy invariant parameter; to generate a variety of treatment effects from a common semiparametric functional form; to organize the literature on alternative estimators; and to explore what policy questions commonly used estimators in the treatment effect literature answer. A fundamental asymmetry intrinsic to the method of instrumental variables (IV) is noted. Recent advances in IV estimation allow for heterogeneity in responses but not in choices, and the method breaks down when both choice and response equations are heterogeneous in a general way.

KEYWORDS: Instrumental variables, selection models, program evaluation.

EVALUATING THE IMPACTS OF PUBLIC POLICIES, forecasting their effects in new environments, and predicting the effects of policies never tried are three central tasks of economics. The structural approach and the treatment effect approach are two competing paradigms of policy evaluation.

The structural approach emphasizes clearly articulated economic models that can be used to accomplish all three tasks under the exogeneity and parameter policy invariance assumptions presented in that literature (see Hansen and Sargent (1981), Hendry (1995)). Economic theory is used to guide the construction of models and to suggest included and excluded variables. The functional form and exogeneity assumptions invoked in this literature are sometimes controversial (see, e.g., Angrist and Krueger (1999)) and the sources of identification of parameters of these models are often not clearly articulated.

[1]This paper was presented by Heckman as the Fisher–Schultz Lecture at the Eighth World Meetings of the Econometric Society, Seattle, Washington, August 13, 2000. Because of its coauthorship, this lecture was subject to the usual refereeing practices of *Econometrica* and has been through two rounds of reviews. This paper was also presented at the seminar on Applied Price Theory at the Graduate School of Business, University of Chicago in October 2000, at a seminar at Uppsala University in December 2000, at Harvard University in April 2001, and at the Montreal Econometrics Seminar in September 2003. We thank Jaap Abbring, Richard Blundell, and two anonymous referees for helpful comments on the first round reports. We benefited from the close reading by Ricardo Avelino, Jean-Marc Robin, Sergio Urzua, and Weerachart Kilenthong on the second draft. We have benefited from a close reading by Jora Stixrud and Sergio Urzua on the third draft. We have also benefited from comments by an anonymous referee on the second draft of this paper. Sergio Urzua provided valuable research assistance in programming the simulations reported in this paper and was assisted by Hanna Lee. Urzua made valuable contributions to our understanding of the random coefficient case and cases with negative weights, and made numerous valuable comments on this draft, as did Weerachart Kilenthong. See our companion paper (Heckman, Urzua, and Vytlacil (2004)), where these topics are developed further. This research was supported by NSF 97-09-873, NSF 00-99195, NSF SES-0241858, and NICHD-40-403-000-85-261, and the American Bar Foundation.

The treatment effect literature as currently developed focuses on the first task—evaluating the impact of a policy in place—in the special case where there is a "treatment group" and a "comparison group," i.e., a group of nonparticipants. In the language of that literature, "internal validity" is the primary goal and issues of forecasting out of sample or of evaluating new policies receive little attention.[2] Because of its more limited goals, fewer explicit functional form and exogeneity assumptions are invoked. The literature on treatment effects has given rise to a new language of economic policy analysis where the link to economic theory is often obscure and the economic policy questions being addressed are not always clearly stated. Different instruments answer different economic questions that typically are not clearly stated. Relationships among the policy parameters implicitly defined by alternative choices of instruments are not articulated.

This paper unites the two approaches to policy evaluation using the marginal treatment effect (MTE) under the assumption that analysts have access to treatment and comparison groups. The MTE is the mean response of persons to treatment at a margin that is precisely defined in this paper. It is a willingness to pay measure when outcomes are values under alternative treatment regimes.

Under the conditions specified in this paper, the MTE can be used to construct and compare alternative conventional treatment effects, a new class of policy relevant treatment effects, and the probability limits produced from instrumental variable estimators and matching estimators. Using the MTE, this paper unites the selection (control function) approach, defined in a nonparametric setting, with the recent literature on instrumental variables.

A major focus in the recent microeconomic policy evaluation literature, and a major theme of this paper, is on constructing and estimating models with heterogeneity in responses to treatment among otherwise observationally identical people. This literature emphasizes that responses to treatment vary among observationally identical people and, crucially, that agents select (or are selected) into treatment at least in part based on their own idiosyncratic response to it. This emphasis is in marked contrast to the emphasis in the conventional representative-agent macro-time-series literature that ignores such heterogeneity despite ample microeconometric evidence on it.[3]

Entire classes of econometric evaluation estimators can be organized by whether or not they allow for the possibility of selection based on unobserved components of heterogeneous responses to treatment. In the presence of such heterogeneity, a variety of different mean treatment effects can be defined for

[2]Internal validity means that a treatment parameter defined in a specified environment is free of selection bias. It is defined more precisely below.

[3]Heckman (2001) summarizes the evidence on heterogeneity in responses to treatment on which agents select into treatment.

different instruments and conditioning sets. In the absence of such heterogeneity, these different treatment effects collapse to the same parameter.[4]

The dependence of estimated treatment parameters on instruments is an important and not widely understood feature of models with heterogeneous responses on which people act.[5] Instrument-dependent parameters arise in this class of models, something excluded by assumption in conventional structural econometric models that emphasize the estimation of invariant parameters. Two economists analyzing the same dataset but using different valid instruments will estimate different parameters that have different economic interpretations. Even more remarkably, two economists using the same instrument but with different notions about what variables belong in choice equations will interpret the output of an instrumental variable analysis differently. Intuitions about "identifying strategies" acquired from analyzing conventional models where responses to treatment do not vary among persons are not valid in the more general setting analyzed in this paper. The choice of an instrument defines the treatment parameter being estimated. The relevant question regarding the choice of instrumental variables in the general class of models studied in this paper is "What parameter is being identified by the instrument?" rather than the traditional question of "What is the efficient combination of instruments for a fixed parameter?"—the question that has traditionally occupied the attention of econometricians who study instrumental variables (IV). Even in the presence of least squares bias, and even assuming large samples, IV based on classical assumptions may be more biased for a given policy parameter than ordinary least squares (OLS). The cure may be worse than the disease.

We extend the method of instrumental variables to estimate economically interpretable parameters in models with heterogeneous treatment outcomes. We note a fundamental asymmetry intrinsic to the method of instrumental variables. Treatment outcomes can be heterogeneous in a general way that we make precise in this paper. Choice equations cannot be heterogeneous in the same general way. When choices and treatment outcomes are analyzed symmetrically, the method of instrumental variables and our extension of it breaks down, and more explicit structural approaches are necessary to solve policy evaluation problems.

The plan of this paper is as follows. Section 1 presents a prototypical microeconometric structural model as a benchmark to define and motivate the various treatment parameters used in the literature and to compare and contrast structural estimation approaches with those used in the literature on treatment effects. We then define our general model and assumptions in Section 2. Our model extends the treatment effect literature by introducing choice

[4]See Heckman (1997), Heckman and Robb (1985, 1986 (reprinted 2000)), and Heckman and Vytlacil (1999).

[5]This dependence was first noted by Heckman and Robb (1985, p. 196). See also Angrist, Graddy, and Imbens (2000).

theory into it and by using a weaker set of assumptions than those used in the structural literature to define and identify the marginal treatment effect. This section shows how the MTE can be used to generate and unify the various treatment parameters advocated in the recent literature and provides an economic foundation for the treatment effect literature. We derive a set of testable restrictions implied by our model, and we apply the general analysis to the special case of a parametric generalized Roy model.

The conventional treatment parameters do not, in general, answer questions of economic or policy interest. Section 3 shows how to use the MTE to define policy relevant parameters that answer well-posed economic questions. Evaluation of different policies requires different weights for the MTE. The MTE plays the role of a policy invariant structural parameter in conventional econometrics for a class of policy interventions defined in this paper.[6]

Section 4 organizes entire classes of econometric estimators on the basis of what they assume about the role of unobservables in the MTE function, conditional on X. Our analysis shows that traditional instrumental variables procedures require that the marginal treatment effect is the same for all persons of given X characteristics. When the marginal treatment effect varies over individuals with the same X, we show how the instrumental variables estimand (the probability limit of the instrumental variables estimator) can be written as a weighted average of MTE, where our general expressions nest previous results in the literature as special cases. The interpretation of the IV estimand depends not only on the choice of instrument used, but also on what other variables are included in the choice model even if they are not used as instruments. We show that it is not always possible to pick an instrument that answers a particular policy problem of interest, and we show that not all instruments answer well defined policy questions. We present necessary and sufficient conditions to construct an instrument to produce a particular policy counterfactual, and show how to construct the instrument when the conditions are satisfied. We develop necessary and sufficient conditions for a particular instrument to answer some well defined policy question, and show how to construct the policy counterfactual when the conditions are satisfied. We focus on instrumental variables in this paper, but also consider matching and ordinary least squares as special cases of our general model for IV.

Section 5 returns to the policy evaluation problem. The treatment effect literature can be used to answer certain narrowly focused questions under weaker assumptions than are required to recover conventional structural parameters that answer a broad range of questions. When we attempt to address the broader set of questions entertained in the structural econometrics literature, additional conditions are required to extrapolate existing policies to new environments and to provide accurate forecasts of new policies

[6]Hendry (1995) discusses the role of policy invariant parameters in macro-forecasting and policy evaluation.

never previously experienced. The weaker identifying assumptions invoked in the treatment effect literature are possible because of the narrower set of questions addressed by that literature. In the language of the treatment effect literature, internal validity (absence of selection bias) does not imply external validity (the ability to generalize). When the same policy forecasting questions addressed by the structural literature are asked of the treatment effect literature, the assumption sets used in the two literatures look very similar, especially for nonparametric versions of structural models. External validity requires stronger conditions.

Section 6 discusses the fundamental role played by the assumed absence of general forms of heterogeneity in choice equations invoked in the recent literature under the rubric of "monotonicity" assumptions. When both choices and treatment outcomes are modeled symmetrically, the method of instrumental variables breaks down, and a different approach to policy analysis is required. Section 7 concludes.

1. A LATENT VARIABLE FRAMEWORK

The treatment effect literature investigates a class of policies that have partial participation at a point in time so there is a "treatment" group and a "comparison" group. It is not helpful in evaluating policies that have universal participation. In contrast, the structural econometrics literature can evaluate policies with universal participation by using functional form and support conditions to substitute for lack of a comparison group (see Heckman and Vytlacil (2005)). Throughout this paper we follow the conventional practice in the literature and ignore general equilibrium effects.[7]

To link our discussion to the literature on structural econometrics, it is fruitful to compare how the two different approaches analyze a generalized Roy model for two potential outcomes (Y_0, Y_1). This model is widely used in applied econometrics (see Amemiya (1985), Heckman (2001)).

Write potential outcomes (Y_0, Y_1) for conditioning variables X as

$$Y_0 = \mu_0(X) + U_0 \tag{1a}$$

and

$$Y_1 = \mu_1(X) + U_1, \tag{1b}$$

where Y_1 is the outcome if treated and Y_0 is the outcome if not treated.[8] In a model of educational attainment, Y_1 is the present value of college earn-

[7]See, however, the studies by Heckman, Lochner, and Taber (1998), who demonstrate the empirical importance of investigating general equilibrium effects in the context of evaluating the returns to schooling.

[8]Throughout this paper, we denote random variables/random vectors by capital letters and potential realizations by the corresponding lowercase letter. For example, X denotes the random vector and x denotes a potential realization of the random vector X.

ings and Y_0 is the present value of earnings in the benchmark no-treatment state (e.g., high school). Let $D = 1$ denote receipt of treatment so that Y_1 is observed, while $D = 0$ denotes that treatment was not received so that Y_0 is observed. In the educational attainment example, $D = 1$ if the individual selects into college; $D = 0$ otherwise. The observed outcome Y is given by

$$Y = DY_1 + (1 - D)Y_0. \tag{1c}$$

Let

$$C = \mu_C(Z) + U_C \tag{1d}$$

denote the cost of receiving treatment. Net utility is $D^* = Y_1 - Y_0 - C$ and the agent selects into treatment if the net utility from doing so is positive, $D = \mathbb{1}[D^* \geq 0]$.

The original Roy (1951) model is a special case of this framework when there are zero costs of treatment, $\mu_C(Z) = 0$ and $U_C = 0$. The generalized Roy model allows for costs of treatment, both driven by observable determinants of the cost of treatment, Z, and unobservable determinants of the cost of treatment, U_C. For example, in the educational attainment example, tuition and family income operate through direct costs $\mu_C(Z)$ to determine college attendance, while U_C might include disutility from studying. The model can be generalized to incorporate uncertainty about the benefits and costs of treatment and to allow for more general decision rules. Let $\mathcal{I}$ denote the information set available to the agent at the time when the agent is deciding whether to select into treatment. If, for example, the agent selects into treatment when the expected benefit exceeds the expected cost, then the index is $D^* = E(Y_1 - Y_0 - C|\mathcal{I})$. The decision to participate is based on $\mathcal{I}$ and $D = \mathbb{1}[D^* \geq 0]$, where D^* is a random variable measurable with respect to $\mathcal{I}$.[9]

Conventional approaches used in the structural econometrics literature assume that $(X, Z) \perp\!\!\!\perp (U_0, U_1, U_C)$, where $\perp\!\!\!\perp$ denotes independence. In addition, they adopt parametric assumptions about the distributions of the error terms and functional forms of the estimating equations, and identify the full model that can then be used to construct a variety of policy counterfactuals. The most commonly used specification of this model writes $\mu_0(X) = X\beta_0$, $\mu_1(X) = X\beta_1$, $\mu_C(Z) = Z\beta_C$ and assumes $(U_0, U_1, U_C) \sim N(0, \Sigma)$. This is the normal selection model (Heckman (1976)).

The parametric normal framework can be used to answer all three policy evaluation questions. First, it can be used to evaluate existing policies by asking how policy-induced changes in X or Z affect (Y, D). Second, it can be used to extrapolate old policies to new environments by computing outcomes for the values of X, Z that characterize the new environment. Linearity and distributional assumptions make extrapolation straightforward. Third, this framework

[9]See Cunha, Heckman, and Navarro (2005) for a version of this model.

can be used to evaluate new policies if they can be expressed as some known functions of (X, Z). For example, consider the effect of charging tuition in an environment where tuition has never before been charged. If tuition can be put on the same footing as (made comparable with) another measure of cost that is measured and varies, or with returns that can be measured and vary, then we can use the estimated response to the variation in observed costs or returns to estimate the response to the new tuition policy.[10]

This paper relaxes the functional form and distributional assumptions used in the structural literature and still identifies an economically interpretable model that can be used for policy analysis. Recent semiparametric approaches relax both distributional and functional form assumptions of selection models, but typically assume exogeneity of X (see, e.g., Powell (1994)) and do not estimate treatment effects except through limit arguments (Heckman (1990), Andrews and Schafgans (1998)).[11] The treatment effect literature seeks to bypass the ad hoc assumptions used in the structural literature and estimate treatment effects under weaker conditions. The goal of this literature is to examine the effects of policies in place (i.e., to produce internally valid estimators) rather than to forecast new policies or old policies on new populations.

2. TREATMENT EFFECTS

We now present the model of treatment effects developed in Heckman and Vytlacil (1999, 2001a), which relaxes most of the controversial assumptions discussed in Section 1. It is a nonparametric selection model with testable restrictions that can be used to unify the treatment effect literature, identify different treatment effects, link the literature on treatment effects to the literature in structural econometrics, and interpret the implicit economic assumptions underlying instrumental variables and matching methods. We follow Heckman and Vytlacil (1999, 2001a) in considering binary treatments. Heckman and Vytlacil (2005) and Heckman, Urzua, and Vytlacil (2004) extend this analysis to the case of a discrete, multivalued treatment, for both ordered and unordered models, while Florens, Heckman, Meghir, and Vytlacil (2004) develop a related model with a continuum of treatments.

We use the general framework of Section 1, Equations (1a)–(1d), and define Y as the measured outcome variable. We do not impose any assumption on the support of the distribution of Y. We use the more general nonlinear

[10]For example, in a present value income maximizing model of schooling, costs and returns are on the same footing, so knowledge of how schooling responds to returns is enough to determine how schooling responds to costs. See Section 5.1.

[11]A large part of the literature is concerned with estimation of slope coefficients (e.g., Ahn and Powell (1993)) and not the counterfactuals needed for policy analysis. Heckman (1990) develops the more demanding conditions required to identify policy counterfactuals.

and nonseparable outcome model

$$(2a) \qquad Y_1 = \mu_1(X, U_1),$$

$$(2b) \qquad Y_0 = \mu_0(X, U_0).$$

Examples include conventional latent variable models: $Y_i = 1$ if $Y_i^* = \mu_i(X) + U_i \geq 0$ and $Y_i = 0$ otherwise; $i = 0, 1$. Notice that in the general case, $\mu_i(X, U_i) - E(Y_i|X) \neq U_i$, $i = 0, 1$, so even if the μ_i are structural, the $E(Y_i|X)$ are not.[12]

The individual treatment effect associated with moving an otherwise identical person from 0 to 1 is $Y_1 - Y_0 = \Delta$ and is defined as the effect on Y of a *ceteris paribus* move from 0 to 1. These *ceteris paribus* effects are called causal effects. To link this framework to the literature on structural econometrics, we characterize the decision rule for program participation by an index model

$$(3) \qquad D^* = \mu_D(Z) - U_D, \quad D = 1 \quad \text{if} \quad D^* \geq 0, \quad D = 0 \quad \text{otherwise},$$

where (Z, X) is observed and (U_1, U_0, U_D) is unobserved. The random variable U_D may be a function of (U_0, U_1). For example, in the Roy model, $U_D = U_1 - U_0$, and in the generalized Roy model, $U_D = U_1 - U_0 - U_C$. Without loss of generality, Z includes all of the elements of X. However, our analysis requires that Z contain at least one element not in X. The following assumptions are weaker than those used in the conventional literature on structural econometrics or the recent literature on semiparametric selection models and at the same time can be used both to define and to identify different treatment parameters.[13] The assumptions are the following:

(A-1) The term $\mu_D(Z)$ is a nondegenerate random variable conditional on X.

(A-2) The random vectors (U_1, U_D) and (U_0, U_D) are independent of Z conditional on X.

(A-3) The distribution of U_D is absolutely continuous with respect to Lebesgue measure.

(A-4) The values of $E|Y_1|$ and $E|Y_0|$ are finite.

(A-5) $1 > \Pr(D = 1|X) > 0$.

Assumptions (A-1) and (A-2) are "instrumental variable" assumptions that there is at least one variable that determines participation in the program that is not in X and that is independent of potential outcomes (Y_0, Y_1) given X. These are the assumptions used in the natural and social experiment literatures where randomization or pseudorandomization generates instruments.

[12] See Heckman and Vytlacil (2005) for alternative definitions of structure.

[13] As noted in Section 2.1 and Heckman and Vytlacil (2001a), a much weaker set of conditions is required to define the parameters than is required to identify them. As noted in Section 5, stronger conditions are required for policy forecasting.

Assumption (A-2) also assumes that U_D is independent of Z given X and is used below to generate counterfactuals. Assumption (A-3) is a technical assumption made primarily for expositional convenience. Assumption (A-4) guarantees that the conventional treatment parameters are well defined. Assumption (A-5) is the assumption in the population of both a treatment and a control group for each X. Observe that there are no exogeneity requirements for X. This is in contrast to the assumptions commonly made in the conventional structural literature and the semiparametric selection literature (see, e.g., Powell (1994)). A counterfactual "no feedback" condition facilitates interpretability so that conditioning on X does not mask the effects of D. Letting X_d denote a value of X if D is set to d, leads to a sufficient condition that rules out feedback from D to X:

(A-6) $X_1 = X_0$ almost everywhere.

Condition (A-6) is not strictly required to formulate an evaluation model, but it enables an analyst who conditions on X to capture the "total" or "full effect" of D on Y (see Pearl (2000)). This assumption imposes the requirement that X is an external variable determined outside the model and is not affected by counterfactual manipulations of D. However, the assumption allows for X to be freely correlated with U_1, U_0, and U_D so it can be endogenous in this sense. In this paper, we examine treatment effects conditional on X and we maintain assumption (A-6).

Define $P(Z)$ as the probability of receiving treatment given Z: $P(Z) \equiv \Pr(D = 1|Z) = F_{U_D|X}(\mu_D(Z))$, where $F_{U_D|X}(\cdot)$ denotes the distribution of U_D conditional on X.[14] We often denote $P(Z)$ by P, suppressing the Z argument. As a normalization, we impose $U_D \sim \text{Unif}[0, 1]$ and $\mu_D(Z) = P(Z)$. This normalization is innocuous given our assumptions, because if the latent variable generating choices is $D^* = \nu(Z) - V$, where V is a general continuous random variable, we can apply a probability transform to reparameterize the model so that $\mu_D(Z) = F_{V|X}(\nu(Z))$ and $U_D = F_{V|X}(V)$.[15]

Vytlacil (2002) establishes that assumptions (A-1)–(A-5) for selection model (2a), (2b), and (3) are equivalent to the assumptions used to generate the local average treatment effects (LATE) model of Imbens and Angrist (1994). Thus the nonparametric selection model for treatment effects developed in

[14]Throughout this paper, we will refer to the cumulative distribution function of a random vector A by $F_A(\cdot)$ and to the cumulative distribution function of a random vector A conditional on random vector B by $F_{A|B}(\cdot)$. We will write the cumulative distribution function of A conditional on $B = b$ by $F_{A|B}(\cdot|b)$.

[15]This representation is valid whether or not (A-2) is true. However, (A-2) imposes restrictions on counterfactual choices. For example, if a change in government policy changes the distribution of Z by an external manipulation, under (A-2) the model can be used to generate the choice probability from $P(z)$ evaluated at the new arguments, i.e., the model is invariant with respect to the distribution Z.

this paper is equivalent to an influential instrumental variable model for treatment effects. Our latent variable model satisfies their assumptions and their assumptions generate our latent variable model. Our latent variable model is a version of the standard sample selection bias model.

Our model and assumptions (A-1)–(A-5) impose two testable restrictions on the distribution of (Y, D, Z, X). First it imposes an index sufficiency restriction: for any measurable set $\mathcal{A}$ and for $j = 0, 1$,

$$\Pr(Y_j \in \mathcal{A}|X, Z, D = j) = \Pr(Y_j \in \mathcal{A}|X, P(Z), D = j).$$

This restriction has empirical content when Z contains two or more variables not in X. Second, the model also imposes a testable monotonicity restriction in $P = p$ for $E(YD|X = x, P = p)$ and $E(Y(1 - D)|X = x, P = p)$ which we develop in Appendix A.

Even though the model of treatment effects developed in this paper is not the most general possible model, it has testable implications and hence empirical content. It unites various literatures and produces a nonparametric version of the widely used selection model, and links the treatment literature to economic choice theory.

2.1. *Definitions of Treatment Effects*

The difficulty of observing the same individual in both treated and untreated states leads to the use of various population level treatment effects widely used in the biostatistics literature and applied in economics.[16] The most commonly invoked treatment effect is the average treatment effect (ATE) $\Delta^{\text{ATE}}(x) \equiv E(\Delta|X = x)$, where $\Delta = Y_1 - Y_0$. This is the effect of assigning treatment randomly to everyone of type X, assuming full compliance, and ignoring general equilibrium effects. The average impact of treatment on persons who actually take the treatment is treatment on the treated (TT): $\Delta^{\text{TT}}(x) \equiv E(\Delta|X = x, D = 1)$. This parameter can also be defined conditional on $P(Z)$: $\Delta^{\text{TT}}(x, p) \equiv E(\Delta|X = x, P(Z) = p, D = 1)$.[17]

The mean effect of treatment on those for whom $X = x$ and $U_D = u_D$, the marginal treatment effect, plays a fundamental role in our analysis:

$$\Delta^{\text{MTE}}(x, u_D) \equiv E(\Delta|X = x, U_D = u_D). \tag{4}$$

The MTE is the expected effect of treatment conditional on observed characteristics X and conditional on U_D, the unobservables from the first stage decision rule. For u_D evaluation points close to zero, $\Delta^{\text{MTE}}(x, u_D)$ is the expected

[16]Heckman, LaLonde, and Smith (1999) discussed panel data cases where it is possible to observe both Y_0 and Y_1 for the same person.

[17]These two definitions of treatment on the treated are related by integrating out the conditioning p variable: $\Delta^{\text{TT}}(x) = \int_0^1 \Delta^{\text{TT}}(x, p)\, dF_{P(Z)|X,D}(p|x, 1)$, where $F_{P(Z)|X,D}(\cdot|x, 1)$ is the distribution of $P(Z)$ given $X = x$ and $D = 1$.

effect of treatment on individuals with the value of unobservables that make them most likely to participate in treatment and who would participate even if the mean scale utility $\mu_D(Z)$ were small. If U_D is large, $\mu_D(Z)$ would have to be large to induce people to participate.

One can also interpret $E(\Delta|X = x, U_D = u_D)$ as the mean gain in terms of $Y_1 - Y_0$ for persons with observed characteristics X who would be indifferent between treatment or not if they were exogenously assigned a value of Z, say z, such that $\mu_D(z) = u_d$. When Y_1 and Y_0 are value outcomes, MTE is a mean willingness to pay measure. The MTE is a choice-theoretic building block that unites the treatment effect, selection, and matching literatures.

A third interpretation is that MTE conditions on X and the residual defined by subtracting the expectation of D^* from D^*: $\tilde{U}_D = D^* - E(D^*|Z, X)$. These three interpretations are equivalent under separability in D^*, i.e., when (3) characterizes the choice equation, but lead to three different definitions of MTE when a more general nonseparable model is developed. This point is developed further in Section 6.

The LATE parameter of Imbens and Angrist (1994) is a version of MTE. Define D_z as a counterfactual choice variable with $D_z = 1$ if D would have been chosen if Z had been set to z and with $D_z = 0$ otherwise. Let $\mathcal{Z}(x)$ denote the support of the distribution of Z conditional on $X = x$. For any $(z, z') \in \mathcal{Z}(x) \times \mathcal{Z}(x)$ such that $P(z) > P(z')$, LATE is $E(\Delta|X = x, D_z = 1, D_{z'} = 0) = E(Y_1 - Y_0|X = x, D_z = 1, D_{z'} = 0)$, the mean gain to persons who would be induced to switch from $D = 0$ to $D = 1$ if Z were manipulated externally from z' to z. From the latent index model, it follows that LATE can be written as

$$
\begin{aligned}
&E(Y_1 - Y_0|X = x, D_z = 1, D_{z'} = 0) \\
&\quad = E(Y_1 - Y_0|X = x, u'_D \leq U_D < u_D) \\
&\quad = \Delta^{\text{LATE}}(x, u_D, u'_D)
\end{aligned}
$$

for $u_D = \Pr(D_z = 1) = P(z)$, $u'_D = \Pr(D_{z'} = 1) = P(z')$, where assumption (A-2) implies that $\Pr(D_z = 1) = \Pr(D = 1|Z = z)$ and $\Pr(D_{z'} = 1) = \Pr(D = 1|Z = z')$. Imbens and Angrist define the LATE parameter as the probability limit of an estimator. Their analysis conflates issues of definition of parameters with issues of identification. Our representation of LATE allows us to separate these two conceptually distinct matters and to define the LATE parameter more generally. One can imagine evaluating the right-hand side of this equation at any u_D, u'_D points in the unit interval and not only at points in the support of the distribution of the propensity score $P(Z)$ conditional on $X = x$ where it is identified. From assumptions (A-2)–(A-4), $\Delta^{\text{LATE}}(x, u_D, u'_D)$ is continuous in u_D and u'_D, and $\lim u'_D \uparrow u_D \Delta^{\text{LATE}}(x, u_D, u'_D) = \Delta^{\text{MTE}}(x, u_D)$.[18]

[18]This follows from Lebesgue's theorem for the derivative of an integral and holds almost everywhere with respect to Lebesgue measure. The ideas of the marginal treatment effect and

TABLE IA

TREATMENT EFFECTS AND ESTIMANDS AS WEIGHTED AVERAGES OF THE MARGINAL TREATMENT EFFECT

$$\mathrm{ATE}(x) = \int_0^1 \Delta^{\mathrm{MTE}}(x, u_D)\, du_D$$

$$\mathrm{TT}(x) = \int_0^1 \Delta^{\mathrm{MTE}}(x, u_D) h_{\mathrm{TT}}(x, u_D)\, du_D$$

$$\mathrm{LATE}(x, u_D, u_D') = \frac{1}{u_D - u_D'}\left[\int_{u_D'}^{u_D} \Delta^{\mathrm{MTE}}(x, u)\, du\right]$$

$$\mathrm{TUT}(x) = \int_0^1 \Delta^{\mathrm{MTE}}(x, u_D) h_{\mathrm{TUT}}(x, u_D)\, du_D$$

$$\mathrm{PRTE}(x) = \int_0^1 \Delta^{\mathrm{MTE}}(x, u_D) h_{\mathrm{PRTE}}(x, u_D)\, du_D$$

$$\mathrm{IV}(x) = \int_0^1 \Delta^{\mathrm{MTE}}(x, u_D) h_{\mathrm{IV}}(x, u_D)\, du_D$$

$$\mathrm{OLS}(x) = \int_0^1 \Delta^{\mathrm{MTE}}(x, u_D) h_{\mathrm{OLS}}(x, u_D)\, du_D$$

Heckman and Vytlacil (1999) use assumptions (A-1)–(A-5) and the latent index structure to develop the relationship between MTE and the various treatment effect parameters shown in the first three lines of Table IA. For example, in that table $\Delta^{\mathrm{TT}}(x)$ is a weighted average of Δ^{MTE},

$$\Delta^{\mathrm{TT}}(x) = \int_0^1 \Delta^{\mathrm{MTE}}(x, u_D) h_{TT}(x, u_D)\, du_D,$$

where

$$(5) \qquad h_{\mathrm{TT}}(x, u_D) = \frac{1 - F_{P|X}(u_D|x)}{\int_0^1 (1 - F_{P|X}(t|x))\, dt} = \frac{S_{P|X}(u_D|x)}{E(P(Z)|X = x)},$$

and $S_{P|X}(u_D|x)$ is $\Pr(P(Z) > u_D|X = x)$ and $h_{\mathrm{TT}}(x, u_D)$ is a weighted distribution (see Heckman and Vytlacil (2001a)). The parameter $\Delta^{\mathrm{TT}}(x)$ oversamples $\Delta^{\mathrm{MTE}}(x, u_D)$ for those individuals with low values of u_D that make them more likely to participate in the program being evaluated. Treatment on the untreated (TUT) is defined symmetrically with TT and oversamples those least likely to participate. The various weights are displayed in Table IB. The other

the limit form of LATE were first introduced in the context of a parametric normal generalized Roy model by Björklund and Moffitt (1987), and were analyzed more generally by Heckman (1997). Angrist, Graddy, and Imbens (2000) also define and develop a limit form of LATE.

TABLE IB

WEIGHTS

$$h_{\text{ATE}}(x, u_D) = 1$$

$$h_{\text{TT}}(x, u_D) = \left[\int_{u_D}^{1} f(p|X=x)\,dp\right]\frac{1}{E(P|X=x)}$$

$$h_{\text{TUT}}(x, u_D) = \left[\int_{0}^{u_D} f(p|X=x)\,dp\right]\frac{1}{E((1-P)|X=x)}$$

$$h_{\text{PRTE}}(x, u_D) = \left[\frac{F_{P^*,X}(u_D|x) - F_{P,X}(u_D|x)}{\Delta\bar{P}(x)}\right], \text{ where } \Delta\bar{P}(x) = E(P|X=x) - E(P^*|X=x)$$

$$h_{\text{IV}}(x, u_D) = \left[\int_{u_D}^{1} (p - E(P|X=x))f(p|X=x)\,dp\right]\frac{1}{\text{Var}(P|X=x)} \text{ for } P(Z) \text{ as an instrument}$$

$$h_{\text{OLS}}(x, u_D) = 1 + \frac{E(U_1|X=x, U_D=u_D)h_1(x, u_D) - E(U_0|X=x, U_D=u_D)h_0(x, u_D)}{\Delta^{\text{MTE}}(x, u_D)},$$

$$\text{if } \Delta^{\text{MTE}}(x, u_D) \neq 0,$$

$$= 0 \text{ otherwise}$$

$$h_1(x, u_D) = \left[\int_{u_D}^{1} f(p|X=x)\,dp\right]\left[\frac{1}{E(P|X=x)}\right]$$

$$h_0(x, u_D) = \left[\int_{0}^{u_D} f(p|X=x)\,dp\right]\frac{1}{E((1-P)|X=x)}$$

weights, treatment effects, and estimands shown in this table are discussed later. A central theme of this paper is that under our assumptions all estimators and estimands can be written as weighted averages of MTE.

Observe that if $E(\Delta|X=x, U_D=u_D) = E(\Delta|X=x)$, so Δ is mean independent of U_D given $X=x$, then $\Delta^{\text{MTE}} = \Delta^{\text{ATE}} = \Delta^{\text{TT}} = \Delta^{\text{LATE}}$. Therefore, in cases where there is no heterogeneity in terms of unobservables in MTE (Δ constant conditional on $X=x$) or agents do not act on it so that U_D drops out of the conditioning set, marginal treatment effects are average treatment effects, so that all of the evaluation parameters are the same. Otherwise, they are different. Only in the case where the marginal treatment effect is the average treatment effect will the "effect" of treatment be uniquely defined.

Figure 1A plots weights for a parametric normal generalized Roy model generated from the parameters shown at the base of Figure 1B. We discuss the contents of Figure 1B in Section 4. A high u_D is associated with higher cost, relative to return, and less likelihood of choosing $D=1$. The decline of MTE in terms of higher values of u_D means that people with higher u_D have lower gross returns. TT overweights low values of u_D (i.e., it oversamples U_D that make it likely to have $D=1$). ATE samples U_D uniformly. Treatment on the untreated ($E(Y_1 - Y_0|X=x, D=0)$) oversamples the values of U_D unlikely to have $D=1$.

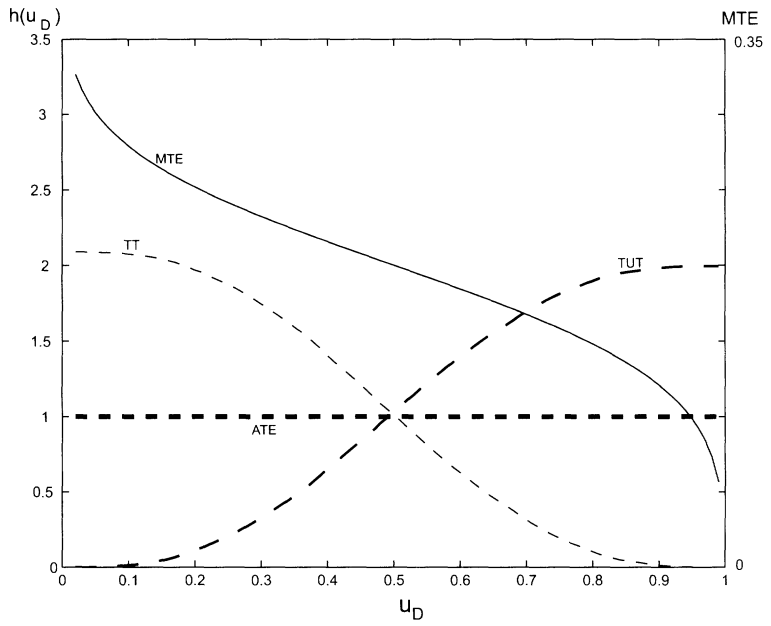

FIGURE 1A.—Weights for the marginal treatment effect for different parameters.

Table II shows the treatment parameters produced from the different weighting schemes. Given the decline of the MTE in u_D, it is not surprising that TT > ATE > TUT. The difference between TT and ATE is a sorting gain: $E(Y_1 - Y_0|X, D = 1) - E(Y_1 - Y_0|X)$, the average gain experienced by people who sort into treatment compared to what the average person would experience. Purposive selection on the basis of gains should lead to positive sorting gains of the sort found in the table. We return to this table to discuss the other numbers in it.

Heckman (2001) presents evidence on the nonconstancy of the MTE drawn from a variety of studies of schooling, job training, migration, and unionism. With the exception of studies of unionism, a common finding in the empirical literature is the nonconstancy of MTE given X.[19] The evidence from the literature suggests that different treatment parameters measure different effects and that persons participate in programs based on heterogeneity in responses to the program being studied. The phenomenon of nonconstancy of the MTE that we analyze in this paper is of substantial empirical interest.

The additively separable latent index model for D (Equation (3)) and assumptions (A-1)–(A-5) are far stronger than what is required to define the parameters in terms of the MTE. The representations of treatment effects defined in Table IA remain valid even if Z is not independent of U_D, if there

[19]However, most of the empirical evidence is based on parametric models.

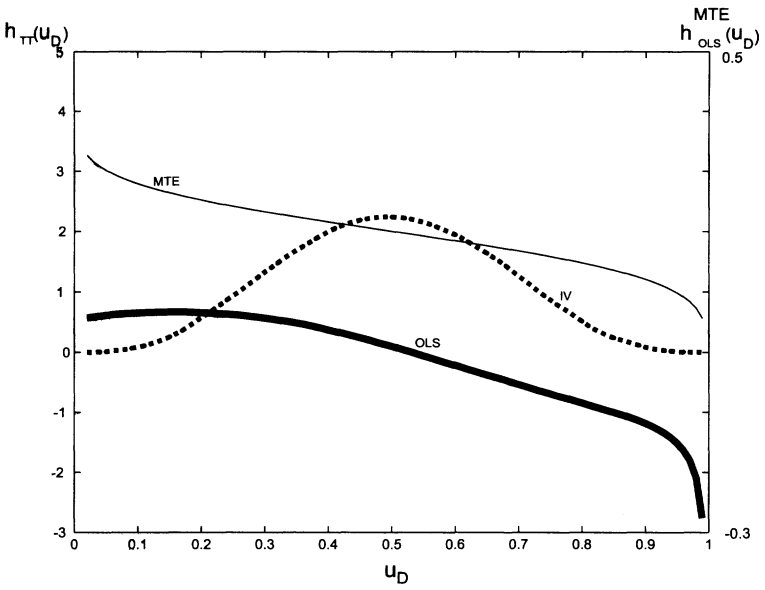

FIGURE 1B.—Marginal treatment effect vs. linear instrumental variables and ordinary least squares weights. Model used to generate Figures 1A and 1B:

$$Y_1 = \gamma + \alpha + U_1, \quad U_1 = \sigma_1 \varepsilon, \quad \gamma = 0.67, \quad \sigma_1 = 0.012,$$
$$Y_0 = \gamma + U_0, \quad U_0 = \sigma_0 \varepsilon, \quad \alpha = 0.2, \quad \sigma_0 = -0.050,$$
$$D = 1 \quad \text{if } Z - V > 0, \quad V = \sigma_V \varepsilon, \quad \varepsilon \sim N(0,1), \quad \sigma_V = -1.000,$$
$$U_D = \Phi\left(\frac{V}{\sigma_V \sigma_\varepsilon}\right), \quad Z \sim N(-0.0026, 0.2700).$$

TABLE II

TREATMENT PARAMETERS AND ESTIMANDS IN THE GENERALIZED ROY EXAMPLE

Treatment on the treated	0.2353
Treatment on the untreated	0.1574
Average treatment effect	0.2000
Sorting gain[a]	0.0353
Policy relevant treatment effect (PRTE)	0.1549
Selection bias[b]	−0.0628
Linear instrumental variables[c]	0.2013
Ordinary least squares	0.1725

[a] TT − ATE = $E(Y_1 - Y_0|D=1) - E(Y_1 - Y_0)$.

[b] OLS − TT = $E(Y_0|D=1) - E(Y_0|D=0)$.

[c] Using propensity score $P(Z)$ as the instrument.

Note: The model used to create Table II is the same as those used to create Figures 1A and 1B. The PRTE is computed using a policy t characterized as follows:
If $Z > 0$ then $D = 1$ if $Z(1+t) - V > 0$.
If $Z \leq t$ then $D = 1$ if $Z - V > 0$.
For this example t is set equal to .2.

are no variables in Z that are not also contained in X, or if a more general nonseparable choice model generates D (so $D^* = \mu_D(Z, U_D)$). No instrument is needed to define the parameters. These issues are discussed further in Section 6.

Assumptions (A-1)–(A-5) will be used to interpret what instrumental variables estimate and to relate instrumental variables to the policy relevant treatment effects. They are sufficient to identify $\Delta^{\text{MTE}}(x, u_D)$ at any u_D evaluation point that is a limit point of the support of the distribution of $P(Z)$ conditional on $X = x$.[20] As developed in Section 6, without these assumptions and representations (in particular Equation (3)) for the choice equations, the IV method and our extension of it does not identify any economically interpretable parameters.

The literature on structural econometrics is clear about the basic parameters of interest although it is not always clear about the exact combinations of parameters needed to answer specific policy problems.[21] The literature on treatment effects offers a variety of evaluation parameters. Missing from that literature is an algorithm for defining treatment effects that answer precisely formulated policy questions. The MTE provides a framework for developing such an algorithm, which we now develop.

3. POLICY RELEVANT TREATMENT PARAMETERS

The conventional treatment parameters do not always answer economically interesting questions. Their link to cost–benefit analysis and interpretable economic frameworks is often obscure.[22] Each answers a different question. Ignoring general equilibrium effects, Δ^{TT} is one ingredient for determining whether or not a given program should be shut down or retained. It is informative on the question of whether the persons participating in a program benefit from it in gross terms.[23] The parameter Δ^{MTE} estimates the gross gain from a marginal expansion of a program. Many investigators estimate a treatment effect and hope that it answers an interesting question. A more promising approach to defining parameters is to postulate a policy question or decision problem

[20]For example, if we additionally impose that the distribution of $P(Z)$ conditional on X has a density with respect to Lebesgue measure, then (A-1)–(A-5) enable us to identify $\Delta^{\text{MTE}}(x, u_D)$ at all (x, u_D) evaluation points in the support of the distribution of $(X, P(Z))$.

[21]In a fundamental paper, Marschak (1953) shows how different combinations of structural parameters are required to forecast the impacts of different policies. It is possible to answer many policy questions without identifying any of the structural parameters individually. The treatment effect literature partially embodies this vision, but typically does not define the economic question being answered, in contrast to Marschak's approach. See Heckman (2001) and Heckman and Vytlacil (2005).

[22]Heckman and Vytlacil (2005) develop the relationship between these parameters and the requirements of cost–benefit analysis.

[23]It is necessary to account for costs to conduct a proper cost–benefit analysis. See the discussion in Heckman and Vytlacil (2005) for nonparametric cost–benefit analysis.

of interest and to derive the treatment parameter that answers it. Taking this approach does not in general produce the conventional treatment parameters or the estimands produced from instrumental variables.

We consider a class of policies that affect P, the probability of participation in a program, but do not affect Δ^{MTE}. The policies analyzed in the treatment effect literature that change the Z not in X are more restrictive than the general policies that shift X and Z analyzed in the structural literature. An example from the schooling literature would be policies that change tuition or distance to school but do not directly affect the gross returns to schooling. Since we ignore general equilibrium effects in this paper, the effects on (Y_0, Y_1) from changes in the overall level of education are assumed to be negligible.

Let a and a' denote two potential policies, and let D_a and $D_{a'}$ denote the choices that would be made under policies a and a'. Let the corresponding decision rules be $D_a = \mathbb{1}[P_a(Z_a) \geq U_D]$ and $D_{a'} = \mathbb{1}[P_{a'}(Z_{a'}) \geq U_D]$, where $P_a(Z_a) = \Pr(D_a = 1|Z_a)$ and $P_{a'}(Z_{a'}) = \Pr(D_{a'} = 1|Z_{a'})$. To simplify the exposition, we will suppress the arguments of these functions and write P_a and $P_{a'}$ for $P_a(Z_a)$ and $P_{a'}(Z_{a'})$. Define $(Y_{0,a}, Y_{1,a}, U_{D,a})$ as (Y_0, Y_1, U_D) under policy a, and define $(Y_{0,a'}, Y_{1,a'}, U_{D,a'})$ correspondingly under policy a'. We assume that Z_a and $Z_{a'}$ are independent, respectively, of $(Y_{0,a}, Y_{1,a}, U_{D,a})$ and $(Y_{0,a'}, Y_{1,a'}, U_{D,a'})$ conditional on X_a and $X_{a'}$. Let $Y_a = D_a Y_{1,a} + (1 - D_a) Y_{0,a}$ and $Y_{a'} = D_{a'} Y_{1,a'} + (1 - D_{a'}) Y_{0,a'}$ denote the outcomes that would be observed under policies a and a', respectively.

We define Δ^{MTE} as policy invariant if

Policy Invariance: $E(Y_{1,a}|U_{D,a} = u, X_a = x)$ and $E(Y_{0,a}|U_{D,a} = u, X_a = x)$, are invariant to the choice of policy a.

Policy invariance can be justified by the strong assumption that the policy change does not change the counterfactual outcomes, covariates, or unobservables, i.e., $(Y_{0,a}, Y_{1,a}, X_a, U_{D,a}) = (Y_{0,a'}, Y_{1,a'}, X_{a'}, U_{D,a'})$. However, Δ^{MTE} is policy invariant if this assumption is relaxed to the weaker assumption that the policy change does not affect the distribution of these variables conditional on X:

(A-7) The distribution of $(Y_{0,a}, Y_{1,a}, U_{D,a})$ conditional on $X_a = x$ is the same as the distribution of $(Y_{0,a'}, Y_{1,a'}, U_{D,a'})$ conditional on $X_{a'} = x$.

We assume (A-7) holds and discuss invariance further in Appendix B.

For the widely used Benthamite social welfare criterion $V(Y)$, comparing policies using mean outcomes and considering the effect for individuals with a given level of $X = x$, we obtain the policy relevant treatment effect (PRTE) denoted $\Delta^{\text{PRTE}}(x)$:

$$
\begin{aligned}
(6) \qquad & E(V(Y_a)|X = x) - E(V(Y_{a'})|X = x) \\
& = \int_0^1 \Delta_V^{\text{MTE}}(x, u_D) \left\{ F_{P_{a'}|X}(u_D|x) - F_{P_a|X}(u_D|x) \right\} du_D,
\end{aligned}
$$

where $F_{P_a|X}(\cdot|x)$ and $F_{P_{a'}|X}(\cdot|x)$ are the distributions of P_a and $P_{a'}$ conditional on $X = x$, respectively, defined for the different policy regimes and $\Delta^{\text{MTE}}_V = E(V(Y_{1,a}) - V(Y_{0,a})|U_{D,a} = u, X_a = x)$.[24,25] The weights are derived in Appendix B under the assumption that the policy does not change the joint distribution of outcomes. To simplify the notation, throughout the rest of this paper, we assume that $V(Y) = Y$. Modifications of our analysis for the more general case are straightforward.

Define $\Delta\bar{P}(x) = E(P_a|X = x) - E(P_{a'}|X = x)$, the change in the proportion of people induced into the program due to the intervention. Assuming $\Delta\bar{P}(x)$ is positive, we may define per person affected weights as $h_{\text{PRTE}}(x, u_D) = (F_{P_{a'}|X}(u_D|x) - F_{P_a|X}(u_D|x))/(\Delta\bar{P}(x))$. These are the weights displayed in Table IB. As demonstrated in the next section, in general, conventional IV weights Δ^{MTE}_V differently than either the conventional treatment parameters (Δ^{ATE} or Δ^{TT}) or the policy relevant parameters, and so does not recover these parameters.

Instead of hoping that conventional treatment parameters or favorite estimators answer interesting economic questions, one approach developed in this paper is to estimate Δ^{MTE}_V and weight it by the appropriate weight determined by how the policy changes the distribution of P to construct Δ^{PRTE}. An alternative approach produces a policy weighted instrument to identify Δ^{PRTE} by standard instrumental variables. We develop both approaches in the next section. Before doing so, we first consider what conventional IV estimates and conditions for identifying Δ^{MTE}. We also consider matching methods and OLS.

4. INSTRUMENTAL VARIABLES, LOCAL INSTRUMENTAL VARIABLES, OLS, AND MATCHING

In this section, we use Δ^{MTE} to organize the literature on econometric evaluation estimators. We assume (A-7), but for simplicity suppress the a and a' subscripts. We focus primarily on instrumental variable estimators, but also briefly consider the method of matching. We present the method of local instrumental variables. Well established intuitions about instrumental variable identification strategies break down when Δ^{MTE} is nonconstant in u_D

[24]We could define policy invariance for Δ^{MTE} in terms of expectations of $V(Y_{1,a})$ and $V(Y_{0,a})$.

[25]If we assume that the marginal distributions of X_a and $X_{a'}$ are the same as the marginal distribution of a benchmark X, the weights can be integrated against the distribution of X to obtain the total effect of the policy in the population:

$$
\begin{aligned}
&E(V(Y_a)) - E(V(Y_{a'})) \\
&\quad = E_X\left\{E(V(Y_a)|X) - E(V(Y_{a'})|X)\right\} \\
&\quad = \int\left[\int_0^1 \Delta^{\text{MTE}}_V(x, u_D)\left\{F_{P_{a'}|X}(u_D|x) - F_{P_a|X}(u_D|x)\right\} du_D\right] dF_X(x).
\end{aligned}
$$

given X. Two sets of instrumental variable conditions are presented in the current literature for this more general case: those associated with conventional instrumental variable assumptions which are implied by the assumption of "no selection on heterogenous gains" and those which permit selection on heterogeneous gains. Neither set implies the other, nor does either identify the policy relevant treatment effect in the general case. Each set of conditions identifies different treatment parameters.

In place of standard instrumental variables methods, we advocate a new approach to estimating policy impacts by estimating Δ^{MTE} using local instrumental variables (LIV) to identify all of the treatment parameters from a generator Δ^{MTE}. The Δ^{MTE} can be weighted in different ways to answer different policy questions. For certain classes of policy interventions discussed in Section 5, Δ^{MTE} possesses an invariance property analogous to the invariant parameters of traditional structural econometrics.

We also consider whether it is possible to construct an instrument such that instrumental variables directly estimate Δ^{PRTE}. We establish necessary and sufficient conditions for the existence of such an instrument. We also address the inverse question of whether instrumental variable estimators always answer well-posed policy questions. In general, they do not. We present necessary and sufficient conditions for a particular instrument to answer some policy counterfactual and characterize what question is answered when an answer exists.

4.1. *Conventional Instrumental Variables*

In the general case with $\Delta^{\text{MTE}}(x, u_D)$ nonconstant in u_D, linear IV does not estimate any of the treatment effects previously defined. Let $J(Z)$ denote an instrument written as a function of Z. We sometimes denote $J(Z)$ by J, leaving implicit that J is a function of Z. The standard conditions $J(Z) \not\perp\!\!\!\perp (U_1, U_0)$ and $\text{Cov}(J(Z), D) \neq 0$ do not, by themselves, imply that instrumental variables using $J(Z)$ as the instrument will identify conventional or policy relevant treatment effects. We must supplement the standard conditions to identify interpretable parameters. To link our analysis to conventional analyses of IV, we invoke familiar-looking representations of additive separability of outcomes in terms of (U_1, U_0) so $Y_1 = \mu_1(X) + U_1$ and $Y_0 = \mu_0(X) + U_0$, but this is not strictly required. All derivations and results in this section hold without any additive separability assumption if $\mu_1(x)$ and $\mu_0(x)$ are replaced by $E(Y_1|X = x)$ and $E(Y_0|X = x)$, respectively, and U_1 and U_0 are replaced by $Y_1 - E(Y_1|X)$ and $Y_0 - E(Y_0|X)$, respectively.

Two distinct sets of instrumental variable conditions in the literature are those due to Heckman and Robb (1985, 1986) and Heckman (1997), and those due to Imbens and Angrist (1994). In the case where Δ^{MTE} is nonconstant in u_D, linear IV estimates different parameters depending on which assumptions are maintained. To establish this point, it is useful to briefly

review the IV method in the case of a common treatment effect defined conditional on X, where $Y_1 - Y_0 = \Delta$, with Δ a deterministic function of X, and where additive separability in outcomes is assumed, as in conventional models. Using (1a) and (1b) with $U_1 = U_0 = U$, and assuming $E(U|X) = 0$, we may write $Y = \mu_0(X) + D\Delta + U$, where $\Delta = \mu_1(X) - \mu_0(X)$. By the law of iterated expectations, $E(U|X) = 0$ and $Z \perp\!\!\!\perp U|X$ imply $E(UJ(Z)|X) = 0$. The standard instrumental variables intuition is that when $E(UJ|X) = 0$ and $\mathrm{Cov}(J, D|X) \neq 0$, linear IV identifies Δ:

$$\text{(IV)} \qquad \frac{\mathrm{Cov}(J, Y|X)}{\mathrm{Cov}(J, D|X)} = \frac{\mathrm{Cov}(J, D\Delta|X)}{\mathrm{Cov}(J, D|X)} = \Delta \frac{\mathrm{Cov}(J, D|X)}{\mathrm{Cov}(J, D|X)}$$
$$= \Delta = \mu_1(X) - \mu_0(X),$$

where the second equality follows from the assumption that Δ is a deterministic function of X. This intuition breaks down in the heterogeneous response case where the outcomes are generated by different unobservables ($U_0 \neq U_1$) so $Y = \mu_0(X) + D\Delta + U_0$, where $\Delta = \mu_1(X) - \mu_0(X) + U_1 - U_0$. This is a variable response model.

There are two important cases of the variable response model. The first case arises when responses are heterogeneous, but conditional on X: people do not base their participation on these responses. In this case, the following condition holds:

(C-1) $D \perp\!\!\!\perp \Delta|X \Longrightarrow E(\Delta|X, U_D) = E(\Delta|X)$, $\Delta^{\mathrm{MTE}}(x, u_D)$ is constant in u_D and $\Delta^{\mathrm{MTE}} = \Delta^{\mathrm{ATE}} = \Delta^{\mathrm{TT}} = \Delta^{\mathrm{LATE}}$.

The second case arises when the following condition holds:

(C-2) $D \not\perp\!\!\!\perp \Delta|X$ and $E(\Delta|X, U_D) \neq E(\Delta|X)$.

In this case Δ^{MTE} is nonconstant and the treatment parameters differ among each other.

Application of the standard IV equation to the general variable coefficient model produces the first equality in IV above. Now, however, Δ is not a deterministic function of X and thus we cannot simply take Δ outside of the covariance term as in the third term of (IV). Plugging in $\Delta = \mu_1(X) - \mu_0(X) + U_1 - U_0$, we obtain

$$\frac{\mathrm{Cov}(J, D\Delta|X)}{\mathrm{Cov}(J, D|X)} = \mu_1(X) - \mu_0(X) + \frac{\mathrm{Cov}(J, D(U_1 - U_0)|X)}{\mathrm{Cov}(J, D|X)}.$$

Our independence assumptions imply that J is independent of $U_1 - U_0$ conditional on X, but do not imply that J is uncorrelated with $D(U_1 - U_0)$ conditional on X. Thus, in general, the covariance in the numerator of the second term is not zero. Knowledge of (X, Z, D) and $(X, Z, (U_0, U_1))$ dependencies is not enough to determine the covariance in the second term. We need to know joint (X, Z, D, U_0, U_1) dependencies.

A sufficient condition for producing (C-1) is the strong information condition that decisions to participate in the program are not made on the basis of $U_1 - U_0$:

(I-1) $\Pr(D = 1|Z, X, U_1 - U_0) = \Pr(D = 1|Z, X)$.

Given our assumption that $(U_1 - U_0)$ is independent of Z given X, one can use Bayes' theorem to show that (I-1) implies the weaker mean independence condition:

(I-2) $E(U_1 - U_0|Z, X, D = 1) = E(U_1 - U_0|X, D = 1)$

which is generically necessary and sufficient for linear IV to identify Δ^{TT} and Δ^{ATE}.

Case (C-2) is inconsistent with (I-2). IV estimates Δ^{LATE} under the conditions of Imbens and Angrist (1994). Δ^{LATE}, selection models, and LIV, introduced below, analyze the more general case covered by (C-2). Different assumptions define different parameters. In addition, as we establish in Section 4.3, even under the same assumptions, different instruments define different parameters and traditional intuitions about instrumental variables break down.

4.2. *Estimating the MTE Using Local Instrumental Variables*

Heckman and Vytlacil (1999, 2001a) resolve this confusion using the local instrumental variable estimator to recover Δ^{MTE} pointwise. Conditional on $X = x$, LIV is the derivative of the conditional expectation of Y with respect to $P(Z) = p$:

$$\Delta^{\text{LIV}}(x, p) \equiv \frac{\partial E(Y|X = x, P(Z) = p)}{\partial p}. \tag{7}$$

The expectation $E(Y_1 - Y_0|X, P(Z))$ exists (almost everywhere) by assumption (A-4), and $E(Y|X, P(Z))$ can be recovered over the support of $(X, P(Z))$. Assumptions (A-2)–(A-4) jointly allow one to use Lebesgue's theorem for the derivative of an integral to show that $E(Y_1 - Y_0|X = x, P(Z) = p)$ is differentiable in p. Thus we can recover $\frac{\partial}{\partial p}E(Y|X = x, P(Z) = p)$ for almost all p that are limit points of the support of distribution of $P(Z)$ conditional on $X = x$.[26] Under our assumptions, LIV identifies MTE for all limit points in the support of the distribution of $P(Z)$ conditional on X. This expression does not require additive separability of $\mu_1(X, U_1)$ or $\mu_0(X, U_0)$.[27]

[26]For example, if the distribution of $P(Z)$ conditional on X has a density with respect to Lebesgue measure, then all points in the support of the distribution of $P(Z)$ conditional on X are limit points of that support and we can identify $\Delta^{\text{LIV}}(x, p) = (\partial E(Y|X = x, P(Z) = p))/\partial p$ for p (almost everywhere).

[27]Note, however, it does require our model and assumptions, including the assumption of additive separability between U_D and Z in the latent index, for selection into treatment. See the discussion in Section 6.

Under standard regularity conditions, a variety of nonparametric methods can be used to estimate the derivative of $E(Y|X, P(Z))$ and thus to estimate Δ^{MTE}. With Δ^{MTE} in hand, if the support of the distribution of $P(Z)$ conditional on X is the full unit interval, one can generate all the treatment parameters defined in Section 2 as well as the policy relevant treatment parameter presented in Section 3 as weighted versions of Δ^{MTE}. When the support of the distribution of $P(Z)$ conditional on X is not full, it is still possible to identify some parameters. For example, Heckman and Vytlacil (2001a) show that to identify ATE under our assumptions, it is necessary and sufficient that the support of the distribution of $P(Z)$ conditional on X includes 0 and 1. Thus, identification of ATE does not require that the distribution of $P(Z)$ conditional on X be the full unit interval or that the distribution of $P(Z)$ conditional on X contain any limit points. Sharp bounds on the treatment parameters can be constructed under the same assumptions imposed in this paper without imposing full support conditions. The resulting bounds are simple and easy to apply compared with those presented in the previous literature.[28]

To establish the relationship between LIV and ordinary IV based on $P(Z)$ and to motivate how LIV identifies Δ^{MTE}, notice from the definition of Y that the conditional expectation of Y given $P(Z)$ is

$$E(Y|P(Z)=p) = E(Y_0|P(Z)=p) + E(\Delta|P(Z)=p, D=1)p,$$

where we keep the conditioning on X implicit. Our model and conditional independence assumption (A-2) imply

$$E(Y|P(Z)=p) = E(Y_0) + E(\Delta|p \geq U_D)p.$$

Applying the IV or Wald estimator for two different values of $P(Z)$, p and p', for $p \neq p'$, we obtain

$$\begin{aligned}(8)\qquad &\frac{E(Y|P(Z)=p) - E(Y|P(Z)=p')}{p-p'} \\ &= \Delta^{\text{ATE}} + \frac{E(U_1 - U_0|p \geq U_D)p - E(U_1 - U_0|p' \geq U_D)p'}{p-p'},\end{aligned}$$

where the expression is obtained under the assumption of additive separability in the outcomes so (1a) and (1b) apply. Note that exactly the same equation holds without additive separability if one replaces U_1 and U_0 with $Y_1 - E(Y_1|X)$ and $Y_0 - E(Y_0|X)$.

[28]For example, see Heckman and Vytlacil (2001b) for a comparison of sharp bounds under the nonparametric selection model with the Manski (1990) sharp bounds under a weaker mean independence condition. Heckman and Vytlacil (2005) survey and synthesize this literature and Heckman and Vytlacil (2001a) develop the bounds.

When $U_1 \equiv U_0$ or $(U_1 - U_0) \perp\!\!\!\perp U_D$ (case (C-1)), IV based on $P(Z)$ estimates Δ^{ATE} because the second term on the right-hand side of the expression (8) vanishes. Otherwise, IV estimates a difficult-to-interpret combination of MTE parameters which we analyze further below.

Another representation of $E(Y|P(Z) = p)$ that reveals the index structure under additive separability more explicitly writes (keeping the conditioning on X implicit) that

$$E(Y|P(Z) = p) = E(Y_0) + \Delta^{\text{ATE}} p + \int_0^p E(U_1 - U_0|U_D = u_D)\, du_D. \tag{9}$$

We can differentiate with respect to p and use LIV to identify Δ^{MTE}:

$$\frac{\partial E(Y|P(Z) = p)}{\partial p} = \Delta^{\text{ATE}} + E(U_1 - U_0|U_D = p) = \Delta^{\text{MTE}}(p).$$

Notice that IV estimates Δ^{ATE} when $E(Y|P(Z) = p)$ is a linear function of p. Thus a test of the linearity of $E(Y|P(Z) = p)$ in p is a test of the validity of linear IV for Δ^{ATE}, i.e., it is a test of whether or not the data are consistent with a correlated random coefficient model. The nonlinearity of $E(Y|P(Z) = p)$ in p provides a way to distinguish whether case (C-1) or case (C-2) describes the data. It is also a test of whether or not agents can at least partially anticipate future unobserved (by the econometrician) gains (the $Y_1 - Y_0$ given X) at the time they make their participation decisions. This analysis generalizes to the nonseparable outcomes case. We use separability in outcomes only to simplify the exposition and link to more traditional models. In particular, exactly the same expression holds with exactly the same derivation for the nonseparable case if we replace U_1 and U_0 with $Y_1 - E(Y_1|X)$ and $Y_0 - E(Y_0|X)$, respectively.[29]

Figure 2A plots two cases of $E(Y|P(Z) = p)$ based on the generalized Roy model used to generate the example in Figures 1A and 1B. When Δ^{MTE} does not depend on u_D, the expectation is a straight line. Figure 2B plots the derivatives of the two curves in Figure 2A. When Δ^{MTE} depends on u_D, people sort into the program being studied positively on the basis of gains from the program, and one gets the curved line depicted in Figure 2A. The levels and derivatives of $E(Y|P(Z) = p)$ and standard errors can be estimated using a variety of semiparametric methods. The derivative estimator of Δ^{MTE} is the local instrumental variable estimator of Heckman and Vytlacil (1999, 2001a). Thus it is possible to test condition (C-1) using simple econometric methods.

[29]Making the conditioning on X explicit, we obtain that $E(Y|X = x, P(Z) = p) = E(Y_0|X = x) + \Delta^{\text{ATE}}(x)p + \int_0^p E(U_1 - U_0|X = x, U_d = u_D)\, du_D$, with the derivative with respect to p given by $\Delta^{\text{MTE}}(x, p)$.

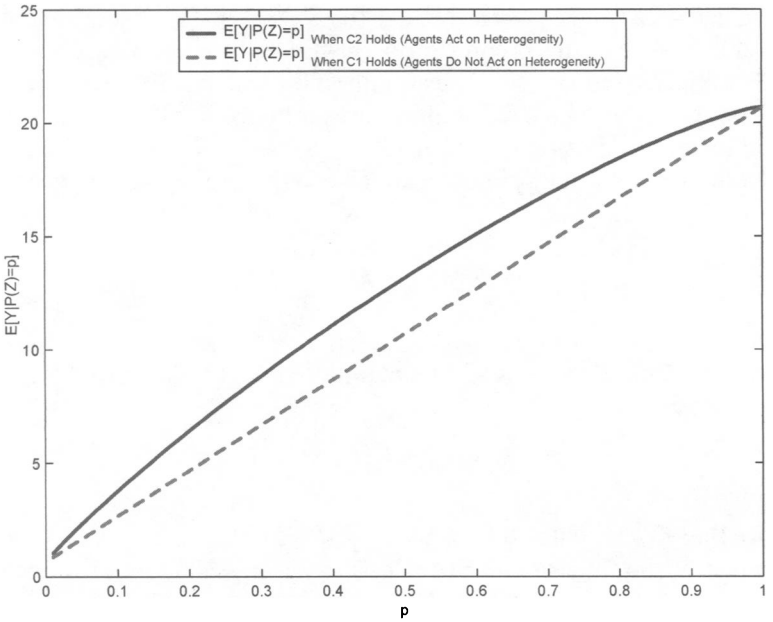

FIGURE 2A.—Plot of the $E(Y|P(Z)=p)$.

In the case without regressors, X, the null hypothesis is the parametric null of linearity.[30]

4.3. *What Does Linear IV Estimate?*

It is instructive to consider what linear IV estimates when Δ^{MTE} is nonconstant and conditions (A-1)–(A-5) hold. We consider the general nonseparable case. We consider instrumental variables conditional on $X=x$ using a general function of Z as an instrument and then specialize our result using $P(Z)$ as the instrument. Let $J(Z)$ be any function of Z such that $\text{Cov}(J(Z), D|X=x) \neq 0$. Define

$$\beta_{\text{IV}}(x;J) \equiv \big[\text{Cov}(J(Z), Y|X=x)\big]\big/\big[\text{Cov}(J(Z), D|X=x)\big].$$

[30]Thus, one can apply any one of the large number of available tests for a parametric null versus a nonparametric alternative (see, e.g., Ellison and Ellison (2000), Zheng (1996)). With regressors, the null is nonparametric, leaving $E(Y|X=x, P(Z)=p)$ unspecified except for restrictions on the partial derivatives with respect to p. In this case, the formal test is a test of a nonparametric null versus a nonparametric alternative, and a formal test of the null hypothesis can be implemented using the methodology of Chen and Fan (1999).

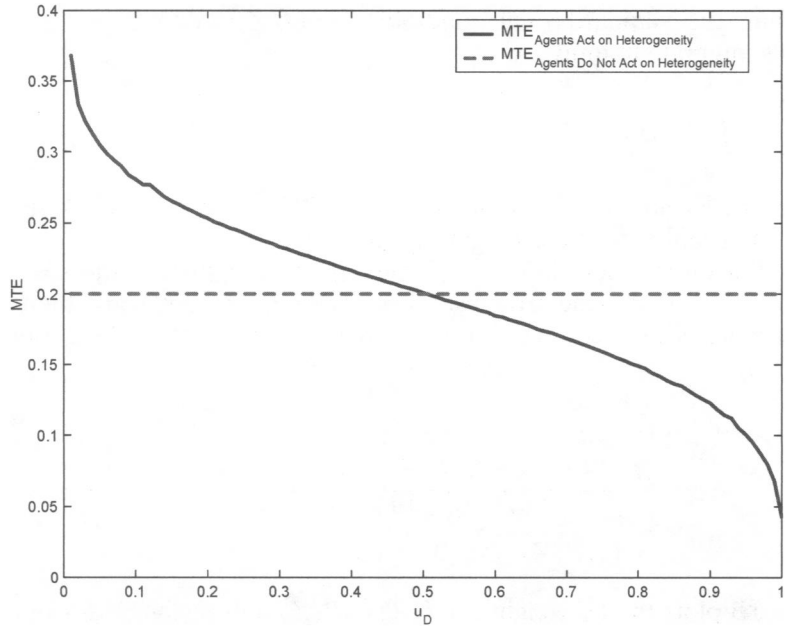

FIGURE 2B.—Plot of the identified marginal treatment effect from Figure 2A (the derivative). Note: Parameters for the general heterogeneous case are the same as those used in Figures 1A and 1B. For the homogeneous case we impose $U_1 = U_0$ ($\sigma_1 = \sigma_0 = 0.012$).

Appendix B derives an expression for the numerator of this expression, using (1c) and (A-2) and letting $\tilde{J}(Z) \equiv J(Z) - E(J(Z)|X)$:

$$\text{(10)} \qquad \begin{aligned} &\operatorname{Cov}(J(Z), Y|X) \\ &\quad = \int_0^1 \Delta^{\text{MTE}}(X, u_D) E(\tilde{J}(Z)|X, P(Z) \geq u_D) \Pr(P(Z) \geq u_D|X)\, du_D. \end{aligned}$$

The denominator follows by a similar argument. By iterated expectations, $\operatorname{Cov}(J(Z), D|X) = \operatorname{Cov}(J(Z), P(Z)|X)$. Thus

$$\beta_{\text{IV}}(x; J) = \int \Delta^{\text{MTE}}(x, u_D) h_{\text{IV}}(u_D|x; J)\, du_D,$$

where

$$\text{(11)} \qquad h_{\text{IV}}(u_D|x; J) = \frac{E(\tilde{J}(Z)|X = x, P(Z) \geq u_D) \Pr(P(Z) \geq u_D|X = x)}{\operatorname{Cov}(J(Z), P(Z)|X = x)},$$

assuming the standard rank condition $\text{Cov}(J(Z), P(Z)|X = x) \neq 0$. The weights integrate to unity,

$$\int_0^1 h_{\text{IV}}(u_D|x; J)\, du_D = 1,$$

and can be constructed from the data on X, $P(Z)$, $J(Z)$, and D. Assumptions about the properties of the weights are testable.[31]

We first discuss additional properties of the weights for the special case where $J(Z) = P(Z)$ (the propensity score is the instrument), and then analyze the properties of the weights for a general instrument $J(Z)$. From Equation (11),

$$\begin{aligned} h_{\text{IV}}(u_D|x; P(Z)) &= \frac{[E(P(Z)|X = x, P(Z) \geq u_D) - E(P(Z)|X = x)]}{\text{Var}(P(Z)|X = x)} \\ &\quad \times \Pr(P(Z) \geq u_D|X = x). \end{aligned}$$

Figure 1B plots the IV weight for $J(Z) = P(Z)$ and the MTE for our generalized Roy model example (see also Table IB). Let p_x^{Min} and p_x^{Max} denote the minimum and maximum points in the support of the distribution of $P(Z)$ conditional on $X = x$. The weights on MTE corresponding to the use of $P(Z)$ as the instrument are nonnegative for all evaluation points, are strictly positive for $u_D \in (p_x^{\text{Min}}, p_x^{\text{Max}})$, and are zero for $u_D < p_x^{\text{Min}}$ and for $u_D > p_x^{\text{Max}}$.[32]

Our expression for the weights does not impose any support conditions on the distribution of $P(Z)$ conditional on X, and thus does not require that $P(Z)$ be either continuous or discrete. To demonstrate this, consider two extreme special cases: (i) when $P(Z)$ is a continuous random variable and (ii) when $P(Z)$ is a discrete random variable.

[31] Expressions for IV and OLS as weighted averages of marginal response functions, and the properties and construction of the weights were first derived by Yitzhaki in 1989 in a paper that was eventually published in 1996 (see Yitzhaki (1996)). He does not use the MTE, however.

[32] For u_D evaluation points between p_x^{Min} and p_x^{Max}, $u_D \in (p_x^{\text{Min}}, p_x^{\text{Max}})$, we have that

$$E(P(Z)|P(Z) \geq u_D, X = x) > E(P(Z)|X = x) \quad \text{and} \quad \Pr(P(Z) \geq u_D|X = x) > 0,$$

so that $h_{\text{IV}}(u_D|x; P(Z)) > 0$ for any $u_D \in (p_x^{\text{Min}}, p_x^{\text{Max}})$. For $u_D < p_x^{\text{Min}}$,

$$E(P(Z)|P(Z) \geq u_D, X = x) = E(P(Z)|X = x).$$

For any $u_D > p_x^{\text{Max}}$, $\Pr(P(Z) \geq u_D|X = x) = 0$. Thus, $h_{\text{IV}}(u_D|x; P(Z)) = 0$ for any $u_D < p_x^{\text{Min}}$ and for any $u_D > p_x^{\text{Max}}$, $h_{\text{IV}}(u_D|x; P(Z))$ is strictly positive for $u_D \in (p_x^{\text{Min}}, p_x^{\text{Max}})$, and is zero for all $u_D < p_x^{\text{Min}}$ and all $u_D > p_x^{\text{Max}}$. Whether the weights are nonzero at the endpoints depends on the distribution of $P(Z)$. However, since the weights are defined for integration with respect to Lebesgue measure, the value taken by the weights at p_x^{Min} and p_x^{Max} does not affect the value of the integral.

First consider the case where the distribution of $P(Z)$ conditional on X has a density with respect to Lebesgue measure with nonnegative density on the interval $(p_x^{\text{Min}}, p_x^{\text{Max}})$. In this case, $\Delta^{\text{LIV}}(x, u_D)$ is well defined for all $u_D \in (p_x^{\text{Min}}, p_x^{\text{Max}})$ such that $h_{\text{IV}}(u_D|x; P(Z)) > 0$. Using the fact that $\Delta^{\text{LIV}}(x, u_D) = \Delta^{\text{MTE}}(x, u_D)$ at evaluation points where LIV is well defined, we can rewrite the expression for the IV estimator as

$$\beta_{\text{IV}}(x; P(Z)) = \int_{p_x^{\text{Min}}}^{p_x^{\text{Max}}} \Delta^{\text{LIV}}(x, u_D) h_{\text{IV}}(u_D|x; P(Z))\, du_D.^{33}$$

Next consider the case where the distribution of $P(Z)$ conditional on X has density with respect to counting measure. For simplicity, assume that the support of the distribution of $P(Z)$ conditional on X contains a finite number of values, $\{p_1, \ldots, p_K\}$ with $p_1 < p_2 < \cdots < p_K$. Then $E(P(Z)|X = x, P(Z) \geq u_D)$ is constant in u_D for u_D within any (p_j, p_{j+1}) interval, and $\Pr(P(Z) \geq u_D)$ is constant in u_D for u_D within any (p_j, p_{j+1}) interval, and thus $h_{\text{IV}}(u_D|x; P(Z))$ is constant in u_D over any (p_j, p_{j+1}) interval. Let q_j denote the value taken by $h_{\text{IV}}(u_D|x; P(Z))$ for $u_D \in (p_j, p_j + 1)$. Then, letting $\tilde{q}_j = q_j(p_{j+1} - p_j)$,

$$\begin{aligned}
&\beta_{\text{IV}}(x; P(Z)) \\
&= \int E(\Delta|X = x, U_D = u_D) h_{\text{IV}}(u_D|x; P(Z))\, du_D \\
&= \sum_{j=1}^{K-1} \int_{p_j}^{p_{j+1}} E(\Delta|X = x, U_D = u_D) q_j\, du_D \\
&= \sum_{j=1}^{K-1} q_j(p_{j+1} - p_j) \int_{p_j}^{p_{j+1}} E(\Delta|X = x, U_D = u_D) \frac{1}{(p_{j+1} - p_j)}\, du_D \\
&= \sum_{j=1}^{K-1} \Delta^{\text{LATE}}(x, p_j, p_{j+1}) \tilde{q}_j.^{34}
\end{aligned}$$

The properties of the weights for general $J(Z)$ depend critically on the relationship between $J(Z)$ and $P(Z)$. Defining $T(p|x; J) = E(J|P(Z) = p, X = x) - E(J|X = x)$,

$$h_{\text{IV}}(u_D|x; J) = \frac{\int_{u_D}^{1} T(t|x; J)\, dF_{P|X}(t|x)}{\text{Cov}(J, P|X = x)} \tag{12}$$

[33] Angrist, Graddy, and Imbens (2000) develop a special case of this expression for a scalar instrument.

[34] In this special case, our analysis is a latent variable version of the formula in Imbens and Angrist (1994).

From this expression, we learn that the IV estimator with $J(Z)$ as an instrument satisfies the following properties:

(i) Two instruments J and J^* weight MTE equally at all u_D evaluation points if and only if $E(J|X=x, P(Z)=p) - E(J|X=x) = E(J^*|X=x, P(Z)=p) - E(J^*|X=x)$ for all p in the support of the distribution of $P(Z)$ conditional on $X=x$.

(ii) The support of $h_{\text{IV}}(u_D|x;J)$ is contained in $(p_x^{\text{Min}}, p_x^{\text{Max}})$. Therefore, $h_{\text{IV}}(t|x;J)=0$ for $t < p_x^{\text{Min}}$ and for $t > p_x^{\text{Max}}$. Using any instrument other than $P(Z)$ leads to nonzero weights only on a subset of $(p_x^{\text{Min}}, p_x^{\text{Max}})$, and using the propensity score as an instrument leads to nonnegative weights on a larger range of evaluation points than using any other instrument.

(iii) For all u_D, $h_{\text{IV}}(u_D|x;J)$ is nonnegative if $E(J|X=x, P(Z)\geq p)$ is weakly monotonic in p. Using J as an instrument yields nonnegative weights on Δ^{MTE} if $E(J|X=x, P(Z)\geq p)$ is weakly monotonic in p. This condition is satisfied when $J(Z)=P(Z)$. More generally, if J is a monotonic function of $P(Z)$, then using J as the instrument will lead to nonnegative weights on Δ^{MTE}. There is no guarantee that the weights for a general $J(Z)$ will be nonnegative for all u_D, although the weights integrate to unity and thus must be positive over some range of evaluation points. We produce examples below where the instrument leads to negative weights for some evaluation points.

The propensity score plays a central role in determining the properties of the weights. The IV weighting formula critically depends on $T(p|x;J)$ and hence on the relationship between the instrument $J(Z)$ and the propensity score. For example, whether two instruments provide the same weights on MTE depends on their relationship with $P(Z)$ (item (i) above), the possible support of the IV weights depends on the support of $P(Z)$ (item (ii)), and whether an instrument will provide positive weights on MTE depends on the instrument's relationship with $P(Z)$ (item (iii)).

The interpretation placed on the IV estimand depends on the specification of $P(Z)$ even if only Z_1 (e.g., a coordinate of Z) is used as the instrument. This drives home the point about the difference between IV in the traditional model and IV in the more general model with heterogeneous responses analyzed in this paper. In the traditional model, the choice of any valid instrument and the specification of instruments in $P(Z)$ not used to construct a particular IV estimator does not affect the IV estimand. In the more general model analyzed in this paper, these choices matter. Two economists, using the same $J(Z)=Z_1$, will obtain the same IV point estimate, but the interpretation placed on that estimate will depend on the specification of the Z in $P(Z)$ even if $P(Z)$ is not used as an instrument. The weights can be positive for one instrument and negative for another.

Table II gives the IV estimand for the generalized Roy model used to generate Figures 1A and 1B using $P(Z)$ as the instrument. The model that generates $D=\mathbb{1}[\beta' Z > V]$ is given at the base of Figure 1B (Z is a scalar, β is 1, V is normal, $U_D=\Phi(V/\sigma_\varepsilon\sigma_V)$). We compare the IV estimand with the policy relevant treatment effect for a policy defined at the base of Table II. If $Z>0$,

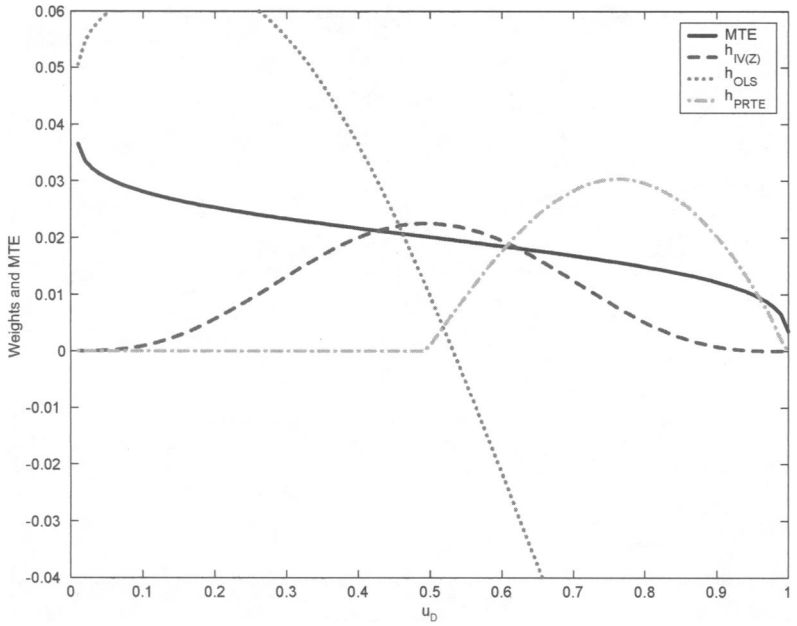

FIGURE 3A.—Marginal treatment effect vs. linear instrumental variables, ordinary least squares, and policy relevant treatment effect weights when $P(Z)$ is the instrument for the policy given at the base of Table II.

persons get a bonus Zt. Their decision rule for $Z > 0$ is $D = \mathbb{1}[Z(1+t) > V]$. People are not forced into participation in the program. Given the assumed distribution of Z, and the other parameters of the model, we obtain $h_{\text{PRTE}}(u_D)$ as plotted in Figures 3A–3C (the scales differ across the graphs). We use the per capita PRTE and consider three instruments. Table III presents estimands for three instruments in the generalized Roy models for three environments.

The first instrument we consider is $P(Z)$, which ignores the policy (t) effect on choices. It is estimated on a sample with no policy in place. Its weight is plotted in Figure 3A, which also displays the OLS weight (discussed later).

TABLE III

LINEAR INSTRUMENTAL VARIABLE ESTIMANDS AND THE POLICY RELEVANT TREATMENT EFFECT

Using propensity score $P(Z)$ as the instrument	0.2013
Using propensity score $P(Z(1+t(\mathbb{1}[Z>0])))$ as the instrument	0.1859
Using a dummy B as an instrument[a]	0.1549
Policy relevant treatment effect (PRTE)	0.1549

[a]The dummy B is such that $B = 1$ if an individual belongs to a randomly assigned eligible population and 0 otherwise.

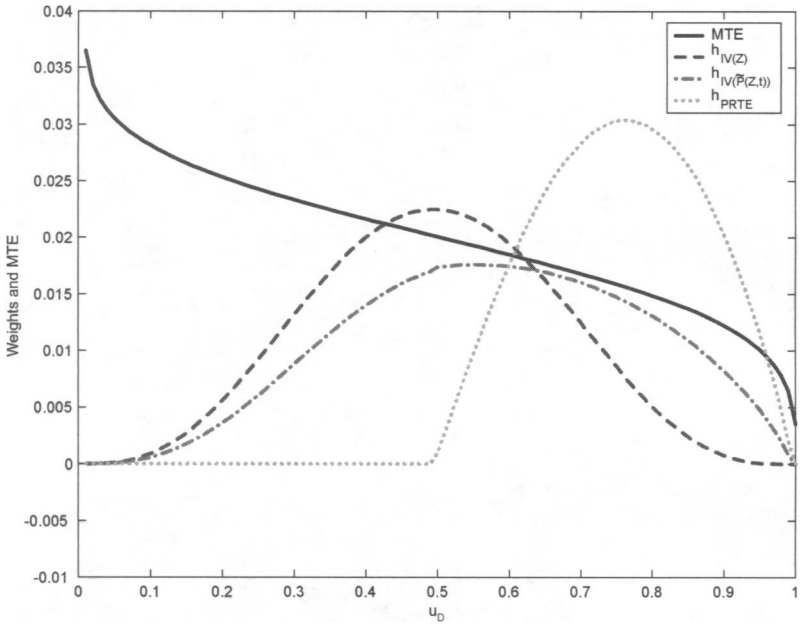

FIGURE 3B.—Marginal treatment effect vs. linear IV with Z as an instrument, linear IV with $P(Z(1+t(\mathbb{1}[Z>0]))) = \tilde{P}(Z,t)$ as an instrument, and policy relevant treatment effect weights for the policy defined at the base of Table II.

The IV weights for $P(Z)$ and the weights for Δ^{PRTE} differ. This is as it should be because Δ^{PRTE} is making a comparison across regimes but IV in this case is making a comparison within a no-policy regime. Given the shape of $\Delta^{\text{MTE}}(u_D)$, it is not surprising that the estimand for IV based on $P(Z)$ is so much above the Δ^{PRTE}, which weights a lower valued segment of $\Delta^{\text{MTE}}(u_D)$ more heavily.

The second instrument we consider exploits the variation induced by the policy in place and fits it on samples where the policy is in place. On intuitive grounds, this instrument might be thought to work well for identifying the PRTE, but in fact it does not. The instrument is $\tilde{P}(Z,t) = P(Z(1+t\mathbb{1}[Z>0]))$, which jumps in value when $Z > 0$. This is the choice probability in the regime with the policy in place. Figure 3B plots the weight for this IV along with the weight for $P(Z)$ as an IV (repeated from Figure 3A). While this weight looks a bit more like the weight for Δ^{PRTE}, it is clearly different.

Figure 3C plots the weight for an ideal instrument for PRTE: a randomization of eligibility. This compares the outcomes in a population with the policy in place with outcomes where it is not. We use an instrument B such that

$$B = \begin{cases} 1, & \text{if a person is eligible to participate in the program,} \\ 0, & \text{otherwise.} \end{cases}$$

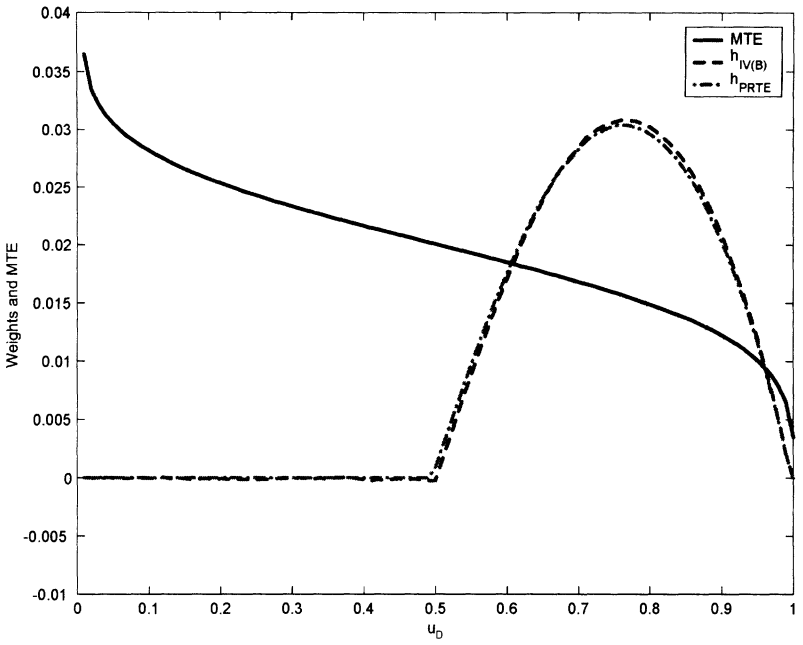

FIGURE 3C.—Marginal treatment effect vs. IV policy and policy relevant treatment effect weights for the policy defined at the base of Table II.

Persons for whom $B = 1$ make their participation choices under the policy with a jump in Z, $t\mathbb{1}(Z > 0)$ in their choice sets. If $B = 0$, persons are embargoed from the policy and there is no bonus. This is a prepolicy regime. We assume $\Pr[B = 1|Y_0, Y_1, V, Z] = \Pr[B = 1] = 0.5$, so all persons are equally likely to receive or not receive eligibility for the bonus and assignment does not depend on model unobservables in the outcome equation. The Wald estimator in this case is

$$\frac{E(Y|B=1) - E(Y|B=0)}{\Pr(D=1|B=1) - \Pr(D=1|B=0)}.$$

The IV weight for this estimator is a special case of Equation (11):

$$h_{\text{IV}}(u_D|B) = \frac{E(B - E(B)|\hat{P}(Z) \geq u_D)\Pr(\hat{P}(Z) \geq u_D)}{\text{Cov}(B, \hat{P}(Z))},$$

where $\hat{P}(Z) = P(Z(1 + t\mathbb{1}[Z > 0]))^B P(Z)^{(1-B)}$. Here, the IV is eligibility for a policy and IV is equivalent to a social experiment that identifies the mean gain per participant who switches to participation in the program. It is to be expected that this IV weight and h_{PRTE} are identical.

Monotonicity

Monotonicity property (iii) is strong. For a general $J(Z)$, there is no guarantee that it will be satisfied even if $J(Z)$ is independent of (Y_0, Y_1) given X and if $J(Z)$ is correlated with D given $X = x$ so that standard IV conditions are satisfied. Thus if Z is a K-dimensional vector and $J(Z) = Z_1$, even if conditional on $Z_2 = z_2, \ldots, Z_K = z_K$, $P(Z)$ is monotonic in Z_1, there is no guarantee that Z_1 used as an instrument for D has positive weights on the MTE.

If we redefine IV for Z_1 to be conditional on $Z_2 = z_2, \ldots, Z_K = z_K$, the weights are positive. Conditioning on instruments not used to form the primary covariance relationship is a new concept that does not appear in the conventional IV literature. In conventional cases governed by condition (C-1), any valid instrument identifies the same parameter. In the general case analyzed in this paper, different choices of instruments and the conditioning sets of other Z variables define different parameters.

Figure 4 demonstrates the possibility of negative weights for the model given at its base. In this figure, we use V rather than normalized $F_V(V) = U_D$ in order to use familiar normal algebra. This simulation is generated from a classical normal error term selection model with nonnormal instruments. The instruments are generated as mixtures of normals from two underlying populations. One can think of this example as a two-component ecological model with different $J(Z), P(Z)$ covariance relationships in the two components. An alternative way to say the same thing is that there are different $(J(Z), \beta' Z)$ covariance relationships in the two subpopulations generating $D = \mathbb{1}(\beta' Z > V)$. In the first component, the covariance between $J(Z)$ and $\beta' Z$ is 0.98. In the second, the covariance varies as shown in Table IV, where the IV is Z_1 but the choice probability depends on Z_1 and Z_2 ($\mu_D(Z) = \beta' Z$). *Ceteris paribus*, increasing Z_1 increases the probability that $D = 1$. Symmetrically, increasing Z_2 and holding Z_1 constant also increases this probability. Yet, since Z_1 and Z_2 covary, varying Z_1 implicitly varies Z_2, which may offset the *ceteris paribus* effect of Z_1 and produce nonmonotonicity and negative weights. In this example there are different covariance relationships in different normal subcomponents of the data. As Z_1 increases, $P(Z)$ increases for some people and decreases for other people, leading to two-way flows into and out of treatment for different people. IV estimates the effect of Z_1 on outcomes not controlling for the other elements of Z. For the configuration of parameters shown there (and for numerous other configurations), the IV weight is negative over a substantial range of values.

The negativity of the weights over certain regions exhibited in Figure 4 makes it clear that Z_1 (and more generally $J(Z)$) fails the monotonicity condition (iii) and does not estimate a gross treatment effect. Some agents withdraw from participation in the program when Z_1 is raised (not holding constant Z_2), while others enter, even though *ceteris paribus* a higher Z_1 raises participation (D). Thus the widely held view that IV estimates some treatment effect of a change in D induced by a change in Z_1 is in general false. It estimates a

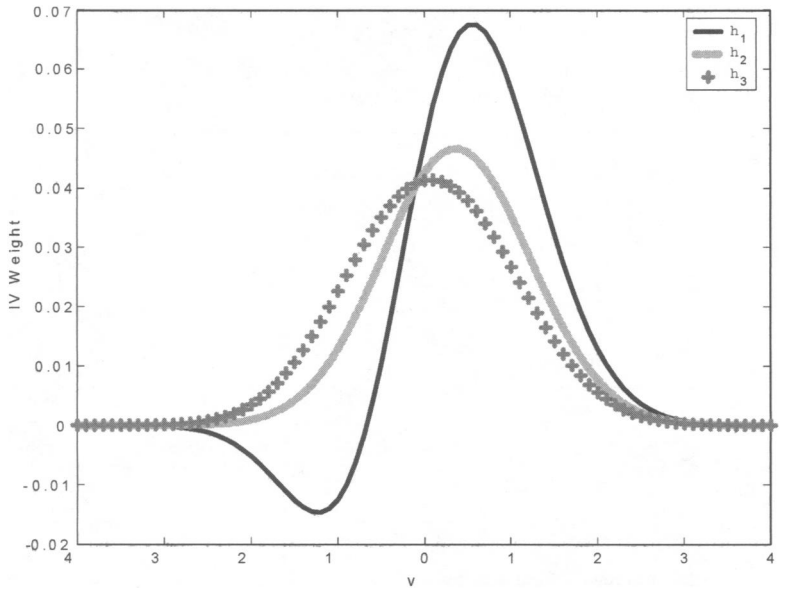

FIGURE 4.—IV weights when $Z \sim p_1 N(\mu_1, \Sigma_1) + p_2 N(\mu_2, \Sigma_2)$ for different values of Σ_2. Model used to generate Figure 4:

$$Y_1 = \gamma + \alpha + U_1 \qquad U_1 = \sigma_1 \varepsilon, \qquad \varepsilon \sim N(0,1),$$

$$Y_0 = \gamma + U_0 \qquad U_0 = \sigma_0 \varepsilon, \qquad \sigma_1 = 0.012, \quad \sigma_0 = -0.05, \quad \sigma_V = -1,$$

$$I = \beta' Z - V, \qquad V = \sigma_V \varepsilon, \qquad \gamma = 0.67, \quad \alpha = 0.2,$$

$$D = \begin{cases} 1, & \text{if } I > 0, \\ 0, & \text{if } I \le 0, \end{cases}$$

$$Z \sim p_1 N(\mu_1, \Sigma_1) + p_2 N(\mu_2, \Sigma_2),$$

$$\mu_1 = [0 \quad -1], \quad \mu_2 = [0 \quad 1], \quad \Sigma_1 = \begin{bmatrix} 1.4 & 0.5 \\ 0.5 & 1.4 \end{bmatrix},$$

$$p_1 = 0.45, \quad p_2 = 0.55, \quad \beta = [0.2 \quad 1.4],$$

$$\mathrm{Cov}(Z_1, \beta' Z) = \beta' \Sigma_1^1 = 0.98 \text{ (Group 1)},$$

$$\Delta^{\mathrm{MTE}}(v) = \alpha + \left[\frac{\mathrm{Cov}(U_1 - U_0, V)}{\mathrm{Var}(V)} \right] v,$$

$$h_{\mathrm{IV}}(v) = \frac{E(Z_1 | \beta' Z > v) \Pr(\beta' Z > v) f_V(v)}{\mathrm{Cov}(Z_1, D)},$$

$$\alpha_{\mathrm{IV}} = \int_{-\infty}^{\infty} \Delta^{\mathrm{MTE}}(v) h_{\mathrm{IV}}(v)\, dv.$$

net effect and not a treatment effect, because monotonicity may be violated. Heckman, Urzua, and Vytlacil (2004) present stark examples where MTE is

TABLE IV

THE IV ESTIMATOR AND $\mathrm{Cov}(Z_1, \beta' Z)$ ASSOCIATED WITH EACH VALUE OF Σ_2 (GROUP 2 COVARIANCE)

Weights	Σ_2	IV	$\mathrm{Cov}(Z_1, \beta' Z) = \beta' \Sigma_2^1$
h_1	$\begin{bmatrix} 0.6 & -0.3 \\ -0.3 & 0.6 \end{bmatrix}$	0.133	−0.30
h_2	$\begin{bmatrix} 0.6 & -0.1 \\ -0.1 & 0.6 \end{bmatrix}$	0.177	−0.02
h_3	$\begin{bmatrix} 0.6 & 0.1 \\ 0.1 & 0.6 \end{bmatrix}$	0.194	0.26

Weights for mixture of normals IV:

$$h_{\mathrm{IV}}(v) = \frac{\left[\frac{P_1 \beta' \Sigma_1^1}{(\beta' \Sigma_1 \beta)^{1/2}} \exp\left[-\frac{1}{2}\left(\frac{v-\beta'\mu_1}{(\beta'\Sigma_1\beta)^{1/2}}\right)^2\right] + \frac{P_2 \beta' \Sigma_2^1}{(\beta'\Sigma_2\beta)^{1/2}} \exp\left[-\frac{1}{2}\left(\frac{v-\beta'\mu_2}{(\beta'\Sigma_2\beta)^{1/2}}\right)^2\right]\right] f_V(v)}{\frac{P_1 \beta' \Sigma_1^1}{(\beta'\Sigma_1\beta+\sigma_V^2)^{1/2}} \exp\left[-\left(\frac{-\beta'\mu_1}{(\beta'\Sigma_1\beta+\sigma_V^2)^{1/2}}\right)^2\right] + \frac{P_2 \beta' \Sigma_2^1}{(\beta'\Sigma_2\beta+\sigma_V^2)^{1/2}} \exp\left[-\left(\frac{-\beta'\mu_2}{(\beta'\Sigma_2\beta+\sigma_V^2)^{1/2}}\right)^2\right]},$$

where Σ_1^1 and Σ_2^1 are the first rows of Σ_1 and Σ_2, respectively. Clearly, $h_{\mathrm{IV}}(-\infty) = 0$ and $h_{\mathrm{IV}}(\infty) = 0$. The weights clearly integrate to 1 over the support of $V = (-\infty, \infty)$. Observe that if $P_2 = 0$, the weights must be positive. Thus the structure of the covariances of the instruments is a key determinant of the positivity of the weights for any instrument. It has nothing to do with the *ceteris paribus* effect of Z_1 on $P(Z)$ in the general case (changing Z_1 holding all other components of Z fixed). Now observe that a necessary condition for $h_{\mathrm{IV}} < 0$ is that $\mathrm{sign}(\beta' \Sigma_1^1) = -\mathrm{sign}(\beta' \Sigma_2^1)$, i.e., that the covariance between Z_1 and $\beta' Z$ be of opposite signs in the two populations. Without loss of generality assume that $\beta' \Sigma_1^1 > 0$. If it equals zero, we fail the rank condition. $f_V(v)$ is the density of V.

negative, the weights are negative, and instrumental variable estimates of treatment effects are positive. Table IV shows how the IV estimand changes with the weights even though the treatment parameters are the same in all three examples.

Monotonicity condition (iii) is testable. Whether condition (iii) corresponds to positive weights on MTE depends on whether all of our assumptions hold, particularly (A-2) and representation (3). If the weights are negative, the change in $J(Z)$ induces two-way flows into and out of treatment. Since it is possible to estimate the joint density of $(J(Z), P(Z))$ given X nonparametrically, under our assumptions it is possible to test for the positivity of the weights which under our assumptions is also a test for monotonicity condition (iii). However (A-2) itself is not testable. Monotonicity condition (iii) is distinct from the condition termed "monotonicity" by Imbens and Angrist (1994). We discuss their condition in Section 6.

4.4. *Policy Relevant Instrumental Variables*

We have just analyzed what IV estimates in terms of weighting MTE. Instead of picking an instrument and hoping that it estimates something interesting, it is more natural to define an economically interesting parameter and see if in-

strumental variables identify it. Suppose that there is a parameter defined as a weighted average of Δ^{MTE} conditional on $X = x$. Can we construct a function of Z to use as an ordinary instrument so that the resulting estimand corresponds to the desired weighted average of Δ^{MTE}? This question is especially interesting if the estimand is a policy counterfactual. We also consider whether there is any policy counterfactual estimated by a given instrument. We initially consider the case where $P(Z)$ is a continuous random variable. We return at the end of this section to consider the case where the distribution of $P(Z)$ is discrete.

Suppose that we seek to recover a parameter defined by $\int \Delta^{\text{MTE}}(x, u) \times w(u|x)\,du$ by the method of linear instrumental variables. We know from Equation (12) the form of the weights corresponding to the IV estimator for any particular instrument $J(Z)$. We seek an instrument $J(Z)$ that has associated weights on MTE given by Equation (12) that are the same as those on the desired parameter

$$w(u_D|x) = \frac{\int_{u_D}^1 T(t|x; J)\,dF_{P|X}(t|x)}{\text{Cov}(J, P|X = x)} = h_{\text{IV}}(u_D|x, J).$$

Assuming that $F_{P|X}$ has a density with respect to Lebesgue measure, the second term in this expression is differentiable in u (almost everywhere). Assuming that $w(u_D|x)$ is also differentiable at all points of evaluation, it follows that

$$w'(u_D|x) = -\frac{T(u_D|x; J) f_{P|X}(u_D|x)}{\text{Cov}(J, P|X = x)}.$$

The following proposition provides conditions under which an instrument exists with the desired properties.

PROPOSITION 1: *Under the conditions*

(i) $F_{P|X}(\cdot)$ *has a density with respect to Lebesgue measure,*

(ii) $w(\cdot|x)$ *satisfies the properties* $w(u_D|x)$ *differentiable in* u_D *for all* $u_D \in [0, 1]$, $\int_0^1 w(u_D|x)\,du_D = 1$, *and* $w(1|x) = w(0|x) = 0$,

(iii) $f_{P|X}(t|x) = 0$ *implies* $w'(t|x) = 0$,

there exists an instrument $J_x(Z)$ *such that* $\text{Cov}(J_x, D|X = x) \neq 0$ *and* $w(u_D|x) = h_{\text{IV}}(u_D|x, J_x)$. *An instrument that satisfies these conditions is*[35]

$$J_x(Z) = \begin{cases} \dfrac{w'(P(Z)|x)}{f_{P|X}(P(Z)|x)}, & \text{if } f_{P|X}(P(Z)|x) > 0, \\ 0, & \text{if } f_{P|X}(P(Z)|x) = 0.^{36} \end{cases}$$

[35]When such an instrument exists, it will not be unique, since the IV estimand will be invariant to rescaling or location shifts for the instrument.

[36]Note that $f_{P|X}(P(Z)|x) > 0$ with probability 1 so that $J_x(Z) = w'(P(Z)|x)/f_{P|X}(P(Z)|x)$ with probability 1.

Under (i), *conditions* (ii) *and* (iii) *are necessary and sufficient for the existence of such an instrument.*

PROOF: See Appendix C.

Condition (i) is a regularity condition requiring that $P(Z)$ be a continuous random variable. We examine the case where $P(Z)$ is discrete at the end of this section. Condition (ii) requires that the desired weights on MTE be proper weights in the sense of integrating up to 1, and also that the weights satisfy the regularity conditions that the weights are differentiable and that the weights are zero when evaluated at the u_D values of 0 and 1. The first and third of these conditions mimic the properties of any IV weights, and the second condition mimics the property of any IV weights when the instrument is continuous. Condition (iii) is a strong, but natural, condition. It requires that the support of the propensity score includes the support of $w'(\cdot|x)$. Thus, the density of the propensity score has to be positive at any evaluation point where $w'(\cdot|x)$ is nonzero. This condition will always be satisfied if $f_{P|X}(t|x) > 0$ for all $t \in [0, 1]$. Given (i), if (ii) or (iii) fails, then no instrument exists that provides the desired weights. If the desired weights do not integrate to 1 or are not differentiable, then as long as $P(Z)$ is continuous (has a distribution that is absolutely continuous with respect to Lebesgue measure), there will not exist any instrument that provides the desired weights. If the weights are nonconstant over an interval outside of the support of the propensity score, then no such instrument can be constructed that would provide the desired weights.

One implication of Proposition 1 is that if $P(Z)$ is continuous, when the MTE depends on u_D in a nontrivial way, there does not exist any instrument that provides the weights corresponding to ATE or TT. To see this, recall that the weights for ATE and TT do not satisfy $w(1|x) = w(0|x) = 0$, so that the weights for ATE and TT do not satisfy condition (ii) of the proposition. Thus, under assumptions (A-1)–(A-5) and under the additional assumptions of the proposition, no instrument exists that gives the weights for ATE or TT if $P(Z)$ is a continuous random variable. This statement leaves open the question of whether instruments will exist that answer policy counterfactuals. We now specialize the previous proposition for the special case of policy weights, using the notation for policy counterfactuals from Section 3.

PROPOSITION 2: *Assume the following*:

(i) $F_{P|X}(\cdot)$, $F_{P_{a'}|X}(\cdot)$, *and* $F_{P_a|X}(\cdot)$ *have densities with respect to Lebesgue measure, where* P *is the initial* (*benchmark*) *probability, and* $P_{a'}$ *and* P_a *are the probabilities associated with two policies* (*possibly*) *distinct from the benchmark policy.*

(ii) $E(P_a|X = x) \neq E(P_{a'}|X = x)$.

(iii) *For any* t, $f_{P|X}(t|x) = 0$ *implies* $f_{P_a|X}(t|x) - f_{P_{a'}|X}(t|x) = 0$.

Define J_x to be a policy relevant instrument if it satisfies $\mathrm{Cov}(J_x, D|X = x) \neq 0$ *and*

$$h_{\mathrm{IV}}(u_D|x, J_x) = \frac{\int_{u_D}^{1} T(t|x; J_x)\, dF_{P|X}(t|x)}{\mathrm{Cov}(J_x, P|X = x)} = \frac{F_{P_{a'}|X}(t|x) - F_{P_a|X}(t|x)}{E(P_{a'}|X = x) - E(P_a|X = x)}.$$

Given conditions (i) *and* (ii), *condition* (iii) *is necessary and sufficient for the existence of such an instrument. If the instrument exists, it can be constructed as*[37]

$$J_x(Z) = \begin{cases} \dfrac{f_{P_{a'}|X}(P(Z)|x) - f_{P_a|X}(P(Z)|x)}{f_{P|X}(P(Z)|x)}, & \text{if } f_{P|X}(P(Z)|x) > 0, \\ 0, & \text{if } f_{P|X}(P(Z)|x) = 0.^{38} \end{cases}$$

PROOF: The proof follows by verifying the conditions of Proposition 1. See Appendix C. *Q.E.D.*

Condition (i) requires that the propensity score be a continuous random variable in a benchmark regime and under both alternative regimes. Condition (ii) requires that the fraction of individuals selecting into treatment under regime a is different than the fraction under regime a'. Condition (iii) imposes the requirement that the densities of the propensity score in the two regimes only differ at evaluation points in the support of the benchmark propensity score. If (iii) fails, then no policy relevant instrument can be constructed.

An immediate corollary of the proposition is that IV using the propensity score as the instrument recovers the policy relevant parameter if

$$P(Z) = \alpha(X) + \beta(X)\left[\frac{f_{P_{a'}|X}(P(Z)) - f_{P_a|X}(P(Z))}{f_{P|X}(P(Z))}\right], \tag{13}$$

where $\alpha(X) = E(P(Z)|X)$ and $\beta(X) = -\mathrm{Var}(P(Z)|X)$, i.e., only if the propensity score is linear in $\{f_{P_{a'}|X}(P(Z)) - f_{P_a|X}(P(Z))\}/f_{P|X}(P(Z))$.

A related question asks whether, given an instrument, there exists a policy counterfactual such that the given instrument is the policy relevant instrument for that counterfactual. We investigate this question for policy counterfactuals starting from a benchmark distribution of $P(Z)$ (the benchmark policy is a, so $P_a(Z_a) = P(Z)$) to some new policy characterized by $P_{a'}(Z_{a'})$. We first answer

[37]If such a $J_x(Z)$ exists, then any linear function of $J_x(Z)$ will also produce the desired set of weights.

[38]Note that $f_{P|X}(P(Z)|x) > 0$ with probability 1 so that $J_x(Z) = (f_{P_{a'}|X}(P(Z)|x) - f_{P_a|X}(P(Z)|x))/(f_{P|X}(P(Z)|x)$ with probability 1.

the question for the special case where the propensity score is the instrument. Solving for $f_{P_{a'}|X}(P(Z))$ in (13), the propensity score will be the policy relevant instrument for a policy characterized by

$$f_{P_{a'}|X}(u_D) = f_{P|X}(u_D)\left(1 - \frac{u_D - E(P(Z)|X)}{\text{Var}(P(Z)|X)}\right). \tag{14}$$

If $f_{P_{a'}|X}$ given by Equation (14) is a proper density, then instrumental variables using the propensity score directly estimate the effect of a policy intervention that changes the density of the propensity score from $f_{P|X}$ to $f_{P_{a'}|X}$, where $f_{P_{a'}|X}$ is given by Equation (14). To be a proper density, $f_{P_{a'}|X}$ must integrate to 1 and be nonnegative for all evaluation points. $f_{P_{a'}|X}(\cdot)$ integrates to 1.[39] Hence, $f_{P_{a'}|X}(\cdot)$ will be nonnegative and thus a proper density if and only if $u_D - E(P(Z)|X) \leq \text{Var}(P(Z))$ for all u_D such that $f_{P|X}(u_D) > 0$. If we let p_x^{Max} denote the maximum of the support of $P(Z)$ conditional on X, we can rewrite this condition as $p_x^{\text{Max}} - E(P(Z)|X = x) \leq \text{Var}(P(Z)|X = x)$. Nothing guarantees that this condition holds, so one cannot guarantee that an instrument produces any policy counterfactual. Not all instruments answer well-posed policy questions.

We next consider the question of whether a general instrument is the policy relevant instrument for some policy. Following the same series of steps used to establish (14), if the instrument $J(Z)$ answers a corresponding policy question, then the policy imposes the restriction that

$$f_{P_{a'}|X}(u_D) = f_{P|X}(u_D)\left(1 - \frac{E(J(Z)|X, P(Z) = u_D) - E(J(Z)|X)}{\text{Cov}(J, P(Z)|X)}\right).$$

The implied $f_{P_{a'}}(\cdot)$ integrates to 1. It is nonnegative for all evaluation points if and only if

$$\frac{E(J(Z)|X, P(Z) = u_D) - E(J(Z)|X)}{\text{Cov}(J, P(Z)|X)} \leq 1$$

for all u_D such that $f_{P|X}(u_D) > 0$. If this condition fails, the instrument is not the policy relevant instrument for any policy. Nothing in the structure of the problem imposes this requirement.

The preceding analysis conditions on X. Suppose that we wish to recover unconditional parameters, e.g., those defined by $\int [\int \Delta^{\text{MTE}}(x, u) w(u|x)\, du] \times dF_X(x)$. If the conditions of Proposition 1 hold for $X = x$ (almost everywhere), then one solution would be to construct $J_x(Z)$ for each x, estimate the parameter conditional on X for each x, and then average over x values. However,

[39] $\int f_{P|X}(u_D)(1 - (u_D - E(P(Z)|X))/(\text{Var}(P(Z)|X)))\, du_D = 1$ since $\int u_D f_{P|X}(u_D)\, du_D = E(P(Z)|X)$.

from the construction of $J_x(Z)$, one can use instrumental variables unconditional on X with the constructed $J(Z)$ as the instrument to obtain the desired parameter in one step.

PROPOSITION 3: *Assume that the conditions of Proposition* 1 *hold for almost all* X. *Construct*

$$J(Z) = \begin{cases} \dfrac{w'(P(Z)|X)}{f_{P|X}(P(Z))}, & \text{if } f_{P|X}(P(Z)) > 0, \\ 0, & \text{if } f_{P|X}(P(Z)) = 0.^{40} \end{cases}$$

Then

$$\frac{\mathrm{Cov}(J(Z), Y)}{\mathrm{Cov}(J(Z), D)} = \int \left[\int \Delta^{\mathrm{MTE}}(x, u_D) w(u_D|x)\, du \right] dF_X(x).$$

PROOF: See Appendix C.

Thus far we have considered the case where $P(Z)$ is a continuous random variable. Is it possible to construct an instrument that produces the desired weights if $P(Z)$ is discrete? The following proposition shows that instrumental variables estimators are only able to produce a very narrow range of weights if $P(Z)$ is discrete. In particular, they only produce weights given by step functions with the jumps in the weight function occurring only at the support points of $P(Z)$.

PROPOSITION 4: *Under the conditions*

(i) *The support of the distribution of* $P(Z)$ *conditional on* X *is a finite set,* $\{p_1, \ldots, p_K\}$ *with* $p_1 < p_2 < \cdots < p_K$ *and with* $\Pr[P(Z) = p_j | X = x] > 0$ *for each* $j = 1, \ldots, K$,

(ii) $w(\cdot|x)$ *satisfies the properties* $\int_0^1 w(u_D|x)\, du = 1$, *and* $w(u_D|x) = 0$ *for* $u_D \leq p_1$ *and for* $u_D > p_K$,

(iii) $w(u_D|x)$ *is constant in* u *over the interval* $(p_j, p_{j+1}]$ *for* $j = 1, \ldots, K$;

there exists an instrument $J_x(Z)$ *such that* $\mathrm{Cov}(J_x, D|X = x) \neq 0$ *and*

$$w(u_D|x) = \frac{\int_{u_D}^1 T(t|x; J_x)\, dF_{P|X}(t|x)}{\mathrm{Cov}(J_x, P|X = x)}.$$

An instrument that satisfies these conditions is

$$J_x(Z) = \sum_{j=1}^{K} \frac{1}{\Pr[P(Z) = p_j | X = x]} (w_j - w_{j+1}) \mathbb{1}[P(Z) = p_j],$$

[40]Note that $f_{P|X}(P(Z)|X) > 0$ with probability 1 so that $J(Z) = w'(P(Z)|X)/f_{P|X}(P(Z)|X)$ with probability 1.

where w_j denotes the (constant) value of $w(u_D|x)$ over the interval $(p_{j-1}, p_j]$ for $j = 2, \ldots, K-1$ and where $w_1 = w_{K+1} = 0$.[41] *Given condition* (i), *conditions* (ii) *and* (iii) *are necessary and sufficient for the existence of such an instrument.*

PROOF: See Appendix C.

Thus, if $P(Z)$ is a discrete random variable, only a very limited set of possible weights on $P(Z)$ can be captured through proper choice of the instrument, so the class of policies that can be generated by IV is very limited. We next use our framework to analyze the OLS estimator and the assumptions about Δ^{MTE} imposed in one widely used version of the method of matching.

4.5. *OLS Weights and Matching*

The OLS estimator can also be represented as a weighted average of Δ^{MTE}. The weight is given in Table IB, where U_1 and U_0 are defined as deviations from conditional expectations, $U_1 = Y_1 - E(Y_1|X)$ and $U_0 = Y_0 - E(Y_0|X)$. Unlike the weights for Δ^{TT} and Δ^{ATE}, these weights do not necessarily integrate to 1 and they are not necessarily nonnegative. The OLS weights for the generalized Roy model are plotted in Figure 1B. The negative component of the OLS weight leads to a smaller OLS treatment estimate compared to the other treatment effects in Table II.

Table II shows the estimated OLS treatment effect for the generalized Roy example. For a binary regressor D, OLS conditional on X identifies

$$\begin{aligned}\Delta^{\text{OLS}}(X) &= E(Y_1|X, D=1) - E(Y_0|X, D=0)\\ &= E(Y_1 - Y_0|X, D=1)\\ &\quad + \{E(Y_0|X, D=1) - E(Y_0|X, D=0)\},\end{aligned}$$

where the term in braces is the "selection bias" term—the difference in pretreatment outcomes between treated and untreated individuals. It is also the bias for Δ^{TT}. The large negative selection bias in this example is consistent with comparative advantage as emphasized by Roy (1951). People who are good in sector 1 (i.e., treatment) may be very poor in sector 0 (no treatment). The differences among the policy relevant treatment effects, the conventional treatment effects, and the OLS estimand are illustrated in Figure 3A and Tables II and III. As is evident from Table II, it is not at all clear that the instrumental variable estimator, with instruments that satisfy classical properties, performs better than OLS in identifying the policy relevant treatment effect.

[41] When such an instrument exists, it will not be unique, since the IV estimand will be invariant to rescaling or location shifts for the instrument.

If there is no selection conditional on covariates, $U_D \perp\!\!\!\perp (Y_1, Y_0)|X$, then $E(U_1|X, U_D) = E(U_1|X) = 0$ and $E(U_0|X, U_D) = E(U_0|X) = 0$ so that the OLS weights are unity and OLS identifies ATE. OLS is a form of matching. Furthermore, $U_D \perp\!\!\!\perp (Y_1, Y_0)|X$ implies that $\Delta^{\text{MTE}}(x, u_D)$ does not vary with u_D, i.e., $\Delta^{\text{MTE}}(X, u_D) = \Delta^{\text{MTE}}(X, u'_D)$ for u_D, u'_D (almost everywhere), so all treatment effects are the same. Observe that given the assumed conditional independence in terms of X, we can identify ATE and TT without requiring a Z that satisfies (A-2). If there is such a Z, the conditional independence condition implies under (A-1)–(A-5) that $E(Y|X, P(Z) = p)$ is linear in p. This conditional independence assumption is invoked in the method of matching and has come into widespread use. The method is based on the assumption that there is no purposeful selection into the program based on unmeasured (by the econometrician) components of gain.[42]

One can weaken the assumption that $U_D \perp\!\!\!\perp (Y_1, Y_0)|X$ to the condition that Y_1 and Y_0 are mean independent of D conditional on X.[43] However, D will be mean independent of Y_1, Y_0 conditional on X without U_D being independent of Y_1, Y_0 conditional on X only if fortuitous balancing occurs with regions of positive Y_1, Y_0 dependence on U_D and regions of negative Y_1, Y_0 dependence on U_D that just exactly offset each other. Such balancing is ruled out in the Roy model and in the generalized Roy model.[44,45] We next apply our framework to analyze policy forecasting problems.

5. OUT OF SAMPLE POLICY FORECASTING, FORECASTING THE EFFECTS OF NEW POLICIES AND STRUCTURAL MODELS BASED ON THE MTE

Section 3 introduced the concept of the Policy Relevant Treatment Effect and invoked a policy invariance assumption. In this section, we present condi-

[42]See Heckman and Navarro-Lozano (2004) and Heckman and Vytlacil (2005) for a more extensive discussion of matching estimators.

[43]See Heckman, Ichimura, Smith, and Todd (1998) and Heckman, Ichimura, and Todd (1997). If the goal of the analysis is to estimate Δ^{TT}, one can get by with the weaker assumption that only Y_0 is mean independent of D conditional on X since $E(Y_1|D = 1, X = x)$ is identified from observational data so there can be selection arising from dependence between Y_1 and D.

[44]In particular, assume $Y_j = \mu_j(X) + U_j$ for $j = 0, 1$, assume $D = \mathbb{1}[Y_1 - Y_0 \geq C(Z) + U_C]$, and let $U_D = U_C - (U_1 - U_0)$. Then if $U_C \perp\!\!\!\perp (U_1 - U_0)$ and U_C has a log concave density, then $E(Y_1 - Y_0|X, U_D = u_D)$ is decreasing in u_D, $\Delta^{\text{TT}}(x) > \Delta^{\text{ATE}}(x)$, and the matching conditions do not hold. If $U_C \perp\!\!\!\perp (U_1 - U_0)$ but U_C does not have a log concave density, then it is still the case that $(U_1 - U_0, U_D)$ is negative quadrant dependent. One can show that $(U_1 - U_0, U_D)$ being negative quadrant dependent implies that $\Delta^{\text{TT}}(x) > \Delta^{\text{ATE}}(x)$ and thus again that the matching conditions cannot hold. See Heckman and Vytlacil (2005) for further discussion.

[45]It is sometimes said that the matching assumptions are "for free" (see Gill and Robins (2001)) because one can always replace unobserved $F(Y_1|X = x, D = 0)$ with $F(Y_1|X = x, D = 1)$ and unobserved $F(Y_0|X = x, D = 1)$ with $F(Y_0|X = x, D = 0)$. This ignores the counterfactual states generated under the matching assumptions that (C-1) is true in the population. The assumed absence of selection is not a "for free" assumption, and produces fundamentally different counterfactual states for the same model under matching and selection assumptions.

tions for constructing PRTE for new environments and for new programs using historical data.

Using the terminology of Campbell and Stanley (1966), estimating the impact of a program in place in a particular environment is the problem of "internal validity." Extrapolating internally valid estimates to new environments, "external validity," or forecasting the effects of new policies are also important problems which we now address.

Let $a \in \mathcal{A}$ denote a policy characterized by random vector Z_a. Let $e \in \mathcal{E}$ denote an environment characterized by random vector X_e. A history, $\mathcal{H}$, is a collection of policy–environment (a, e) pairs that have been experienced and documented. We assume that the environment is autonomous, so the choice of a does not affect X_e. Letting $X_{e,a}$ denote the value of X_e under policy a, autonomy requires the following statement:

(A-8) For all a, e, $X_{e,a} = X_e$ (autonomy).

Autonomy is a more general notion than the concept introduced in (A-6). The concepts are the same when the policy is a treatment. General equilibrium feedback effects can cause a failure of autonomy. In this section we will assume autonomy, in accordance with the partial equilibrium tradition in the treatment effect literature.[46]

Evaluating a particular policy a' in environment e' is straightforward if $(a', e') \in \mathcal{H}$. One simply looks at the associated outcomes and treatment effects formed in that policy environment and applies the methods previously discussed to obtain internally valid estimates. The challenge comes in forecasting the impacts of policies (a') in environments (e') for (a', e') not in $\mathcal{H}$.

We show how Δ^{MTE} plays the role of a policy invariant functional that aids in creating counterfactual states never previously experienced. We focus on the problem of constructing the policy relevant treatment effect Δ^{PRTE}, but our discussion applies more generally to the other treatment parameters.

Given the assumptions invoked in Section 3, Δ^{MTE} can be used to evaluate a whole menu of policies characterized by different conditional distributions of $P_{a'}$. In addition, given our assumptions, we can focus on how policy a', which is characterized by $Z_{a'}$, produces the distribution $F_{P_{a'}|X}$, which weights an invariant Δ^{MTE} without having to conduct a new investigation of (Y, X, Z) relationships for each proposed policy.[47]

5.1. *Constructing Weights for New Policies in a Common Environment*

The problem of constructing Δ^{PRTE} for policy a' (compared to $\bar{a}$) in environment e when $(a', e) \notin \mathcal{H}$ entails constructing $E(V(Y_{a'}))$. We maintain the

[46]However, see Heckman, Lochner, and Taber (1998) for an example of a nonautonomous treatment model.

[47]Ichimura and Taber (2002) present a discussion of local policy analysis in a more general framework without the MTE structure, using a framework developed by Hurwicz (1962).

assumption that the baseline policy is observed, so $(\bar{a}, e) \in \mathcal{H}$. We assume (A-1)–(A-5), (A-7), and (A-8), and use (3) to characterize choices. The policy does not change the distribution of (Y_0, Y_1, U_D) conditional on X. Under these conditions, Equation (6) is a valid expression for PRTE, and constructing PRTE only requires identification of Δ^{MTE} and constructing $F_{P_{a'}|X_e}$ from the policy histories $\mathcal{H}_e$, defined as the elements of $\mathcal{H}$ for a particular environment e, $\mathcal{H}_e = \{a : (a, e) \in \mathcal{H}\}$.

Associated with the policy histories $a \in \mathcal{H}_e$ is a collection of policy variables $\{Z_a : a \in \mathcal{H}_e\}$. Suppose that a new policy a' can be written as $Z_{a'} = T_{a',j}(Z_j)$ for some $j \in \mathcal{H}_e$, where $T_{a',j}$ is a known deterministic transformation and $Z_{a'}$ has the same list of variables as Z_j. Examples of policies that can be characterized in this way are tax and subsidy policies on wages, prices, and incomes that affect unit costs (wages or prices) and transfers. Tuition might be shifted upward for everyone by the same amount or tuition might be shifted according to a nonlinear function of current tuition, parents' income, and other observable characteristics in Z_j.

Constructing $F_{P_{a'}|X_e}$ from data in the policy history entails two distinct steps. From the definitions, $\Pr(P_{a'} \leq t|X_e) = \Pr(Z_{a'} : \Pr(D_{a'} = 1|Z_{a'}, X_e) \leq t|X_e)$. If (i) we know the distribution of $Z_{a'}$ and (ii) we know the function $\Pr(D_{a'} = 1|Z_{a'} = z, X_e = x)$ over the appropriate support, we can then recover the distribution of $P_{a'}$ conditional on X_e. Given that $Z_{a'} = T_{a',j}(Z_j)$ for a known function $T_{a',j}(\cdot)$, step (i) is straightforward since we recover the distribution of $Z_{a'}$ from the distribution of Z_j by using the fact that $\Pr(Z_{a'} \leq t|X_e) = \Pr(Z_j : T_{a',j}(Z_j) \leq t|X_e)$. Alternatively, part of the specification of the policy a' might be the distribution $\Pr(Z_{a'} \leq t|X_e)$. We now turn to the second step, recovering the function $\Pr(D_{a'} = 1|Z_{a'} = z, X_e = x)$ over the appropriate support.

If $Z_{a'}$ and Z_j contain the same elements, though possibly with different distributions, then a natural approach to forecasting the new policy is to postulate that

$$
\begin{aligned}
(15) \qquad P_j(z) &= \Pr(D_j = 1|Z_j = z, X_e) \\
&= \Pr(D_{a'} = 1|Z_{a'} = z, X_e) = P_{a'}(z),
\end{aligned}
$$

i.e., that over a common support for Z_j and $Z_{a'}$ the known conditional probability function and the desired conditional probability function agree. Condition (15) will hold, for example, if $D_j = \mathbb{1}[\mu_D(Z_j) - U_D \geq 0]$, $D_{a'} = \mathbb{1}[\mu_D(Z_{a'}) - U_D \geq 0]$, $Z_j \perp\!\!\!\perp U_D|X_e$, and $Z_{a'} \perp\!\!\!\perp U_D|X_e$. Even if condition (15) is satisfied on a common support, the support of Z_j and $Z_{a'}$ may not be the same. If the support of the distribution of $Z_{a'}$ is not contained in the support of the distribution of Z_j, then some form of extrapolation is needed. Alternatively, if we strengthen our assumptions so that (15) holds for all $j \in \mathcal{H}_e$, we can identify $P_{a'}(z)$ for all z in $\bigcup_{j \in \mathcal{H}_e} \text{Supp}(Z_j)$. However, there is no guarantee that the support of the distribution of $Z_{a'}$ will be contained in $\bigcup_{j \in \mathcal{H}_e} \text{Supp}(Z_j)$, in which case some form of extrapolation is needed.

If extrapolation is required, then one approach is to assume a parametric functional form for $P_j(\cdot)$. Given a parametric functional form, one can use the joint distribution of (D_j, Z_j) to identify the unknown parameters of $P_j(\cdot)$ and then extrapolate the parametric functional form to evaluate $P_j(\cdot)$ for all evaluation points in the support of $Z_{a'}$. Alternatively, if there is overlap between the support of $Z_{a'}$ and Z_j,[48] so there is some overlap in the historical and policy a' supports of Z, we may use nonparametric methods presented in Matzkin (1994) with functional restrictions (e.g., homogeneity) to construct the desired probabilities on new supports or to bound them. Under the appropriate conditions, we may use analytic continuation to extend $\Pr(D_j = 1|Z_j = z, X_e = x)$ to a new support for each $X_e = x$ (Rudin (1974)).

The approach just presented is based on the assumption stated in Equation (15). That assumption is quite natural when $Z_{a'}$ and Z_j both contain the same elements, say they both contain tuition and parents' income. However, in some cases $Z_{a'}$ might contain additional elements not contained in Z_j. As an example, $Z_{a'}$ might include new user fees, while Z_j consists of taxes and subsidies but does not include user fees. In this case, the assumption stated in Equation (15) is not expected to hold and is not even well defined if $Z_{a'}$ and Z_j contain a different number of elements.

A more basic approach analyzes a class of policies that operate on constraints, prices, and endowments arrayed in vector C. Given the preferences and technology of the agent, a given $C = c$, however arrived at, generates the same choices for the agent. Thus a wage tax offset by a wage subsidy of the same amount produces a wage that has the same effect on choices as a no-policy wage. Policy j affects C (e.g., it affects prices paid, endowments, and constraints). Define a map $\Phi_j : Z_j \to C_j$ which maps a policy j, described by Z_j, into its consequences (C_j) for the baseline, fixed-dimensional vector C. A new policy a', characterized by $Z_{a'}$, produces $C_{a'}$ that is possibly different from C_j for all previous policies $j \in \mathcal{H}_e$.

To construct the random variable $P_{a'} = \Pr(D_{a'} = 1|Z_{a'}, X_e)$, we postulate that

$$\begin{aligned}\Pr(D_j = 1|Z_j \in \Phi_j^{-1}(c), X_e = x) &= \Pr(D_j = 1|C_j = c, X_e = x)\\ &= \Pr(D_{a'} = 1|C_{a'} = c, X_e = x)\\ &= \Pr(D_{a'} = 1|Z_{a'} \in \Phi_{a'}^{-1}(c), X_e = x),\end{aligned}$$

where $\Phi_j^{-1}(c) = \{z : \Phi_j(z) = c\}$ and $\Phi_{a'}^{-1}(c) = \{z : \Phi_{a'}(z) = c\}$. Given these assumptions, our ability to recover $\Pr(D_{a'} = 1|Z_{a'} = z, X_e = x)$ for all (z, x) in the support of $(Z_{a'}, X_e)$ depends on what Φ_j functions have been historically

[48] If we strengthen condition (15) to hold for all $j \in \mathcal{H}_e$, then the condition becomes that $\text{Supp}(Z_{a'}) \cap \bigcup_{j \in \mathcal{H}_e} \text{Supp}(Z_j)$ is not empty.

observed and how rich the histories of C_j, $j \in \mathcal{H}_e$ are. For each $z_{a'}$ evaluation point in the support of the distribution of $Z_{a'}$, there is a corresponding $c = \Phi_{a'}(z_{a'})$ evaluation point in the support of the distribution of $C_j = \Phi_j(Z_j)$. If, in the policy histories, there is at least one $j \in \mathcal{H}_e$ such that $\Phi_j(z_j) = c$ for a z_j with (z_j, x) in the support of the distribution of (Z_j, X_e), then we can construct the probability of the new policy from data in the policy histories. The methods used to extrapolate $P_{a'}(\cdot)$ over new regions, discussed previously, apply here. If the distribution of $C_{a'}$ (or $\Phi_{a'}$ and the distribution of $Z_{a'}$) is known as part of the specification of the proposed policy, the distribution of $F_{P_{a'}|X_e}$ can be constructed using the constructed $P_{a'}$. Alternatively, if we can relate $C_{a'}$ to C_j by $C_{a'} = \Psi_{a',j}(C_j)$ for a known function $\Psi_{a',j}$ or if we can relate $Z_{a'}$ to Z_j by $Z_{a'} = T_{a',j}(Z_j)$ for a known function $T_{a',j}$, and the distributions of C_j and/or Z_j are known for some $j \in \mathcal{H}_e$, we can apply the method previously discussed to derive $F_{P_{a'}|X_e}$ and hence the policy weights for the new policy.

This approach assumes that a new policy acts on components of C like a policy in $\mathcal{H}_e$, so it is possible to forecast the effect of a policy with nominally new aspects. The essential idea is to recast the new aspects of policy in terms of old aspects previously measured. Thus in a model of schooling, let $D = \mathbb{1}[Y_1 - Y_0 - B \geq 0]$, where $Y_1 - Y_0$ is the discounted gain in earnings from going to school and B is the tuition cost. Here the effect of cost is just the negative of the effect of return. Historically, we might only observe variation in $Y_1 - Y_0$ (say tuition has never previously been charged), but B is on the same footing (has the same effect on choice, except for sign) as $Y_1 - Y_0$. This identified historical variation in $Y_1 - Y_0$ can be used to nonparametrically forecast the effect of introducing B, provided that the support of $P_{a'}$ is in the historical support generated by the policy histories in $\mathcal{H}_e$. Otherwise, some functional structure (parametric or semiparametric) must be imposed to solve the support problem for $P_{a'}$.

As another example, following Marschak (1953), consider the introduction of wage taxes in a world where there has never before been a tax. Let Z_j be the wage without taxes. We seek to forecast a posttax net wage $Z_{a'} = (1-\tau)Z_j + b$, where τ is the tax rate and b is a constant shifter. Thus $Z_{a'}$ is a known linear transformation of policy Z_j. We can construct $Z_{a'}$ from Z_j. We can forecast under (A-2) using $\Pr(D_j = 1 | Z_j = z) = \Pr(D_{a'} = 1 | Z_{a'} = z)$. This assumes that the response to after tax wages is the same as the response to wages at the after tax level. The issue is whether $P_{a'|X_e}$ lies in the historical support or whether extrapolation is needed. Nonlinear versions of this example can be constructed.

As a final example, environmental economists use variation in one component of cost (e.g., travel cost) to estimate the effect of a new cost (e.g., a park registration fee). See Smith and Banzhaf (2004). Relating the costs and characteristics of new policies to the costs and characteristics of old policies is a standard, but sometimes controversial, method for forecasting the effects of new policies.

In the context of our model, extrapolation and forecasting are confined to constructing $P_{a'}$ and its distribution. If policy a', characterized by vector $Z_{a'}$, consists of new components that cannot be related to Z_j, $j \in \mathcal{H}_e$, or a base set of characteristics whose variation can be identified, the problem is intractable. Then $P_{a'}$ and its distribution cannot be formed using econometric methods applied to historical data.

When it can be applied, our approach allows us to simplify the policy forecasting problem and concentrate our attention on forecasting choice probabilities and their distribution in solving the policy forecasting problem. We can use choice theory and choice data to construct these objects to forecast the impacts of new policies by relating new policies to previously experienced policies.

5.2. *Forecasting the Effects of Policies in New Environments*

When the effects of policy a are forecast for a new environment e' from baseline environment e, and when $X_e \neq X_{e'}$, in general both $\Delta^{\text{MTE}}(x, u_D)$ and $F_{P_a|X_e}$ will change. In general, neither object is environment invariant.[49] The new $X_{e'}$ may have a different support than X_e or any other environment in $\mathcal{H}$. In addition, the new $(X_{e'}, U_D)$ stochastic relationship may be different from the historical (X_e, U_D) stochastic relationship. Constructing $F_{P_a|X_{e'}}$ from $F_{P_a|X_e}$ and $F_{Z_a|X_{e'}}$ from $F_{Z_a|X_e}$ can be done using (i) functional form (including semiparametric functional restrictions) or (ii) analytic continuation methods. Notice that the maps $T_{a,j}$ and Φ_a may depend on X_e, and so the induced changes in these transformations must also be modeled. There is a parallel discussion for $\Delta^{\text{MTE}}(x, u_D)$. The stochastic dependence between $X_{e'}$ and (U_1, U_0, U_D) may be different from the stochastic dependence between X_e and (U_1, U_0, U_D). We suppress the dependence of U_0 and U_1 on e and a only for convenience of exposition and make it explicit in the next paragraph.

Forecasting new stochastic relationships between $X_{e'}$ and (U_1, U_0, U_D) is a difficult task. It can be avoided if we invoke the traditional exogeneity assumptions of classical econometrics:

(A-9) For all e, a, $(U_{1,e,a}, U_{0,e,a}, U_{D,e,a}) \perp\!\!\!\perp (X_e, Z_a)$.

Under (A-9), we only encounter the support problems for both Δ^{MTE} and the distribution of $\Pr(D_a = 1 | Z_a, X_e)$ in constructing policy counterfactuals.

Conditions (A-7)–(A-9) are unnecessary if the only goal of the analysis is to establish internal validity, the standard objective of the treatment effect literature. Autonomy and exogeneity conditions become important issues if we seek external validity. An important lesson from this analysis is that as we try to make the treatment effect literature do the tasks of structural econometrics (i.e., make out of sample forecasts), the assumptions invoked in the two literatures come together.

[49] We suppress the dependence of U_D on a for notational convenience.

5.3. *A Comparison of Three Approaches*

Table V compares the strengths and limitations of the three approaches to policy evaluation that we have discussed: the structural approach, the conventional treatment effect approach, and the recently developed approach to treatment effects based on the MTE function developed in this paper.

The approach based on the MTE function and the structural approach share interpretability of parameters. Like the structural approach, it addresses a range of policy evaluation questions. The MTE parameter is less comparable and less easily extrapolated across environments than are structural parameters, unless nonparametric versions of invariance and exogeneity assumptions are made. However, Δ^{MTE} is comparable across populations with different distributions of P (conditional on X_e) and results from one population can be applied to another population under the conditions presented in this section. Analysts can use Δ^{MTE} to forecast a variety of policies. This invariance property is shared with conventional structural parameters. Our framework solves the problem of external validity which is ignored in the standard treatment effect approach. The price of these advantages of the structural approach is the greater range of econometric problems that must be solved. They are avoided in the conventional treatment approach at the cost of producing parameters that cannot be linked to well-posed economic models and hence do not provide building blocks for an empirically motivated general equilibrium analysis or for investigation of the impacts of new public policies. The Δ^{MTE} estimates the preferences of the agents being studied and provides a basis for integration with well-posed economic models. If the goal of a study is to examine one policy in place (the problem of internal validity), the stronger assumptions invoked in this section of the paper, and in structural econometrics, are unnecessary. Even if this is the only goal of the analysis, however, our approach allows the analyst to generate all treatment effects and IV estimands from a common parameter and provides a basis for unification of the treatment effect literature.

6. MONOTONICITY, UNIFORMITY, NONSEPARABILITY, INDEPENDENCE, AND POLICY INVARIANCE: THE LIMITS OF INSTRUMENTAL VARIABLES

The analysis of this paper and the entire recent literature on instrumental variables estimators for models with heterogeneous responses (i.e., models with outcomes of the form (2a) and (2b)) relies critically on the assumption that the treatment choice equation has a representation in the additively separable form (3). From Vytlacil (2002), we know that, under our assumptions, this assumption is equivalent to the assumption of monotonicity as defined by Imbens and Angrist (1994). Using the notation of Section 2.1, Imbens and Angrist define monotonicity as the following condition: if the Z are changed for everyone from $Z = z$ to $Z = z'$, $D_z \geq D_{z'}$ or $D_z \leq D_{z'}$ for all U_D conditional on X. A better name for this condition would be "uniformity," since it

TABLE V

COMPARISON OF ALTERNATIVE APPROACHES TO PROGRAM EVALUATION

	Structural Econometric Approach	Treatment Effect Approach	Approach Based on MTE
Interpretability	Well defined economic parameters and welfare comparisons	Link to economics and welfare comparisons obscure	Interpretable in terms of willingness to pay; weighted averages of the MTE answer well-posed economic questions
Range of questions addressed	Answers many counterfactual questions	Focuses on one treatment effect or narrow range of effects	With support conditions, generates all treatment parameters
Extrapolation to new environments	Provides ingredients for extrapolation	Evaluates one program in one environment	Can be partially extrapolated; extrapolates to new policy environments with different distributions of the probability of participation due solely to differences in distributions of Z
Comparability across studies	Policy invariant parameters comparable across studies	Not generally comparable	Partially comparable; comparable across environments with different distributions of the probability of participation due solely to differences in distributions of Z
Key econometric problems	Exogeneity, policy invariance, and selection bias	Selection bias	Selection bias: exogeneity and policy invariance if used for forecasting
Range of policies that can be evaluated	Programs with either partial or universal coverage, depending on variation in data (prices/endowments)	Programs with partial coverage (treatment and control groups)	Programs with partial coverage (treatment and control groups)
Extension to general equilibrium evaluation	Need to link to time series data; parameters compatible with general equilibrium theory	Difficult because link to economics is not precisely specified	Can be linked to nonparametric general equilibrium models under exogeneity and policy invariance

describes a condition across people rather than the shape of a function for a particular person.

This uniformity condition imparts an asymmetry to the entire instrumental variable enterprise. Responses are permitted to be heterogeneous in a general way, but choices of treatment are not. In this section, we relax the assumption of additive separability in (3). We establish that in the absence of additive separability or uniformity, the entire instrumental variable identification strategy in this paper and the entire recent literature collapses. Parameters can be defined as weighted averages of an MTE, but MTE and the derived parameters cannot be identified using any instrumental variable strategy.

One natural benchmark nonseparable model is a random coefficient model of choice $D = \mathbb{1}[Z\beta \geq 0]$, where β is a random coefficient vector and $\beta \perp\!\!\!\perp (Z, U_0, U_1)$. If β is a random coefficient with a nondegenerate distribution and with components that take both positive and negative values, uniformity is clearly violated. However, it can be violated even when all components of β are of the same sign if Z is a vector.

To consider a more general case, relax the assumption of Equation (3)

$$D^* = \mu_D(Z, U_D), \tag{16a}$$

where $\mu_D(Z, U_D)$ is not necessarily additively separable in Z and U_D, and U_D is not necessarily a scalar.[50] In the random coefficient example, $U_D = \beta$.

$$D = \mathbb{1}[D^* \geq 0]. \tag{16b}$$

We maintain assumptions (A-1)–(A-5) and (A-8).

In special cases, (16a) can be expressed in an additively separable form. For example, if D^* is weakly separable in Z and U_D, $D^* = \mu_D(\theta(Z), U_D)$ for any U_D, where $\theta(Z)$ is a scalar function, μ_D is increasing in $\theta(Z)$, and U_D is a scalar, then we can write (16b) in the same form as (3):

$$D = \mathbb{1}[\theta(Z) \geq \tilde{U}],$$

where $\tilde{U} = \mu_D^{-1}(0; U_D)$ and $\tilde{U} \perp\!\!\!\perp Z|X$, and the inverse function is expressed with respect to the first argument. Vytlacil (2002) shows that any model that does not satisfy uniformity (or "monotonicity") will not have a representation in this form.[51]

In the additively separable case, the MTE (4) has three equivalent interpretations. (i) The term U_D is the only unobservable in the first stage decision

[50]The additively separable latent index model is more general than it may at first appear. It is shown in Vytlacil (2004) that a wide class of threshold crossing models without the additive structure on the latent index will have a representation with the additively separable structure on the latent index.

[51]In the random coefficient case where $Z = (1, Z_1)$, where Z_1 is a scalar, and $\beta = (\beta_0, \beta_1)$ if $\beta_1 > 0$ for all realizations, we can write the choice rule in the form of (3): $Z_1\beta_1 > -\beta_0 \Rightarrow Z > -\beta_0/\beta_1$ and $U_D = -\beta_0/\beta_1$. This trick does not work in the general case.

rule, and MTE is the average effect of treatment given the unobserved characteristics in the decision rule ($U_D = u_D$). (ii) A person with $U_D = u_D$ would be indifferent between treatment or not if $P(Z) = u_D$, where $P(Z)$ is a mean scale utility function. Thus, the MTE is the average effect of treatment given that the individual would be indifferent between treatment or not if $P(Z) = u_D$. (iii) One can also view the additively separable form (3) as intrinsic in the way we are defining the parameter and interpret the MTE (4) as an average effect conditional on the additive error term from the first stage choice model. Under all interpretations of the MTE and under the assumptions used in the preceding sections of this paper, MTE can be identified by LIV; the MTE does not depend on Z, and hence it is policy invariant and the MTE integrates up to generate all treatment effects, all policy effects, and all IV estimands.

The three definitions are not the same in the general nonseparable case (16a). Heckman and Vytlacil (2001a) extend MTE in the nonseparable case using interpretation (i). The MTE defined this way is policy invariant to changes in Z. They show that LIV is a weighted average of the MTE with possibly negative weights and does not identify MTE. If uniformity does not hold, the definition of MTE allows one to integrate MTE to obtain all of the treatment effects, but the instrumental variables estimator breaks down.

Alternatively, one could define MTE based on (ii):

$$\Delta_B^{\text{MTE}}(x, z) = E\big(Y_1 - Y_0 | X = x, U_D \in \{u_D : \mu_D(z, u_D) = 0\}\big).$$

This is the average treatment effect for individuals who would be indifferent between treatment or not at a given value of z. Heckman and Vytlacil (2001a) show that in the nonseparable case LIV does not identify this MTE and that MTE does not change when the distribution of Z changes, provided the support of MTE does not change.[52] In general, this definition of MTE does not allow one to integrate up MTE to obtain the treatment parameters.

A third possibility is to force the index rule into an additive form by taking $\mu_D^*(Z) = E(\mu_D(Z, U_D)|Z)$, defining $U_D^* = \mu_D(Z, U) - E(\mu_D(Z, U)|Z)$, and define MTE as $E(Y_1 - Y_0 | X = x, U_D^* = u_D^*)$. Note that U_D^* is not independent of Z, is not policy invariant, and is not structural. LIV does not estimate MTE. With this definition of the MTE, it is not possible in general to integrate up MTE to obtain the various treatment effects.

For any version of the nonseparable model except those that can be transformed to separability, index sufficiency fails. To see this most directly, assume that $\mu_D(Z, U_D)$ is absolutely continuous with respect to Lebesgue measure. Define $\Omega(z) = \{u_D : \mu_D(z, u_D) \geq 0\}$. In the additively separable case, $P(z) \equiv \Pr(D = 1|Z = z) = \Pr(u_D \in \Omega(z))$ and $P(z) = P(z') \Leftrightarrow \Omega(z) = \Omega(z')$. This produces index sufficiency. In the more general case of (16a) it is possible to have (z, z') such that $P(z) = P(z')$ and $\Omega(z) \neq \Omega(z')$, so index sufficiency does not hold.

[52] If the support of Z changes, then the MTE must be extended to a new support.

6.1. *Implications of Nonseparability*

In this section, we develop generalization (i), leaving development of the other interpretations for another occasion. We focus on PRTE. The analysis of the other treatment parameters follows by parallel arguments.

For any u_D in the support of the distribution of U_D, define $\Omega_{u_D} = \{z : \mu_D(z, u_D) \geq 0\}$. For example, in the random coefficient case, with $U_D \equiv \beta$ and $D = \mathbb{1}[Z\beta \geq 0]$, we have $\Omega_b = \{z : zb \geq 0\}$, where b is a realization of β. Define $\mathbb{1}_{\mathcal{A}}(t)$ to be the indicator function for the event $t \in \mathcal{A}$. Then Appendix B shows that

$$
\begin{aligned}
(17) \qquad & E(Y_a) - E(Y_{a'}) \\
& = E[E(Y_a|X) - E(Y_{a'}|X)] \\
& = \int \left[\int E(\Delta|X = x, U_D = u_D) \right. \\
& \qquad \left. \times \begin{pmatrix} \Pr[Z_a \in \Omega_{u_D}|X = x] \\ -\Pr[Z_{a'} \in \Omega_{u_D}|X = x] \end{pmatrix} dF_{U_D|X}(u_D|x) \right] dF_X(x).
\end{aligned}
$$

Thus, without additive separability, we can still derive an expression for PRTE and by similar reasoning the other treatment parameters. However, to evaluate the expression requires knowledge of MTE, of $\Pr[Z_a \in \Omega_{u_D}|X = x]$ and $\Pr[Z_{a'} \in \Omega_{u_D}|X = x]$ for every (u_D, x) in the support of the distribution of (U_D, X), and of the distribution of U_D. In general, if no structure is placed on the μ_D function, one can normalize U_D to be unit uniform (or a vector of unit uniform random variables) so that $F_{U_D|X}$ will be known. However, in this case the $\Omega_{u_D} = \{z : \mu_D(z, u_D) \geq 0\}$ sets will not in general be identified. If structure is placed on the μ_D function, one might be able to identify the $\Omega_{u_D} = \{z : \mu_D(z, u_D) \geq 0\}$ sets, but then one needs to identify the distribution of U_D conditional on X. If structure is placed on μ_D, one cannot in general normalize the distribution of U_D to be unit uniform without undoing the structure being imposed on μ_D.

In particular, consider the random coefficient model $D = \mathbb{1}[Z\beta \geq 0]$, where $U_D = \beta$ is a random vector, so that $\Omega_\beta = \{z : z\beta \geq 0\}$. In this case, if all of the other assumptions hold, including $Z \perp\!\!\!\perp \beta|X$, and the policy change does not affect (Y_1, Y_0, X, β), the PRTE is given by

$$
\begin{aligned}
& E(Y_a) - E(Y_{a'}) \\
& = E[E(Y_a|X) - E(Y_{a'}|X)] \\
& = \int \left[\int E(\Delta|X = x, \beta = b) \right. \\
& \qquad \left. \times \begin{pmatrix} \Pr[Z_a \in \Omega_b|X = x] \\ -\Pr[Z_{a'} \in \Omega_b|X = x] \end{pmatrix} dF_{\beta|X}(b|x) \right] dF_X(x).
\end{aligned}
$$

Because structure has been placed on the $\mu_D(Z, \beta)$ function, the sets Ω_β are known. However, evaluating the function requires knowledge of the distribution of β which will not in general be identified without further assumptions.[53] Normalizing the distribution of β to be a vector of unit uniform random variables produces the distribution of β, but eliminates the assumed linear index structure on μ_D and results in Ω_β sets that are not identified.

Even if the weights are identified, Heckman and Vytlacil (2001a) show that it is not possible to use LIV to identify MTE without additive separability between Z and U_D in the selection rule index. Appendix D develops this point for the random coefficient model. Thus, without additive separability in the latent index for the selection rule, we can still create an expression for PRTE (and the other treatment parameters), but both the weights and the MTE function are no longer identified using instrumental variables.

One superficially plausible way to avoid these problems would be to define $\tilde{\mu}_D(Z) = E(\mu_D(Z, U_D)|Z)$ and $\tilde{U}_D = \mu_D(Z, U_D) - E(\mu_D(Z, U_D)|Z)$, producing the model $D = \mathbb{1}[\tilde{\mu}_D(Z) - \tilde{U}_D \geq 0]$. We keep the conditioning on X implicit. One could redefine MTE using $\tilde{U}_D$ and proceed as if the true model possessed additive separability between observables and unobservables in the latent index. This is the method pursued in approach (iii).

For two reasons, this approach does not solve the problem of providing an adequate generalization of MTE. First, with this definition, $\tilde{U}_D$ is a function of (Z, U_D), and a policy that changes Z will then also change $\tilde{U}_D$. Thus, policy invariance of the MTE no longer holds. Second, this approach generates a $\tilde{U}_D$ that is no longer statistically independent of Z, so that assumption (A-2) no longer holds when $\tilde{U}_D$ is substituted for U_D even when (A-2) is true for U_D. Lack of independence between observables and unobservables in the latent index both invalidates our expression for PRTE (and the expressions for the other treatment effects) and causes LIV to no longer identify MTE.

The nonseparable model can also restrict the support of $P(Z)$. For example, consider a standard normal random coefficient model with a scalar regressor $(Z = (1, Z_1))$. Assume $\beta_0 \sim N(0, \sigma_0^2)$, $\beta_1 \sim N(\bar{\beta}_1, \sigma_1^2)$, and $\beta_0 \perp\!\!\!\perp \beta_1$. Then

$$P(z_1) = \Phi\left(\frac{\bar{\beta}_1 z_1}{\sqrt{\sigma_0^2 + \sigma_1^2 z_1^2}}\right),$$

where in this usage Φ is the standard cumulative normal distribution. If the support of z_1 is $\Re^1$, then in the standard additive model $\sigma_1^2 = 0$ and $P(z_1)$ has support $[0, 1]$. When $\sigma_1^2 > 0$, the support is strictly within the unit interval.[54] In the special case when $\sigma_0^2 = 0$, the support is one point $(P(z) = \Phi(\bar{\beta}_1/\sigma_1))$.

[53] See, e.g., Ichimura and Thompson (1998) for conditions for identifying the distribution of β in a random coefficient discrete choice model when $Z \perp\!\!\!\perp \beta$.

[54] The interval is $[\Phi(-|\beta_1|/\sigma_1), \Phi(|\beta_1|/\sigma_1)]$.

We cannot, in general, identify ATE, TT, or any treatment effect requiring the endpoints 0 or 1.

Thus the more general case model of nonuniformity presented in this section does not satisfy the index sufficiency property, and the support of the treatment effects and estimators is, in general, less than full. The random coefficient model for choice may explain the empirical support problems for $P(Z)$ found in Heckman, Ichimura, Smith, and Todd (1998).

6.2. *Implications of Dependence*

We next consider relaxing the independence assumption (A-2) to allow $Z \not\perp\!\!\!\perp U_D|X$ while maintaining the assumption that $Z \perp\!\!\!\perp (Y_1, Y_0)|(X, U_D)$. We maintain the other assumptions, including additive separability between Z and U_D in the latent index for the selection rule (Equation (3)) and the assumption that the policy changes Z but does not change (U_D, Y_0, Y_1, X). Thus we assume that the policy change does not change the MTE function (policy invariance). Given these assumptions, we derive in Appendix B the following expression for PRTE in the nonindependent case:

$$
\begin{aligned}
(18) \qquad & E(Y_a) - E(Y_{a'}) \\
& = E[E(Y_a|X) - E(Y_{a'}|X)] \\
& = \int \Bigg[\int E(\Delta|X = x, U_D = u_D) \\
& \qquad \times \begin{pmatrix} \Pr[\mu_D(Z_{a'}) < u_D|X = x, U_D = u_D] \\ - \Pr[\mu_D(Z_a) < u_D|X = x, U_D = u_D] \end{pmatrix} \\
& \qquad \times dF_{U_D|X}(u_D|x) \Bigg] dF_X(x).
\end{aligned}
$$

Although we can derive an expression for PRTE without requiring independence between Z and U_D, to evaluate this expression requires knowledge of MTE, of $\Pr[\mu_D(Z_{a'}) < u_D|X = x, U_D = u_D]$, and of $\Pr[\mu_D(Z_a) < u_D|X = x, U_D = u_D]$ for every (x, u_D) in the support of the distribution of (X, U_D). This requirement is stronger than in the case of independence, since the weights no longer depend only on the distribution of $P_a(Z_a)$ and $P_{a'}(Z_{a'})$ conditional on X. To evaluate these weights requires knowledge of the function μ_D and of the joint distribution of (U_D, Z_a) and $(U_D, Z_{a'})$ conditional on X, and these will in general not be identified without further assumptions.

Even if the weights are identified, Heckman and Vytlacil (2001a) show that it is not possible to use LIV to identify MTE without independence between Z and U_D conditional on X. Thus, without conditional independence between Z and U_D in the latent index for the decision rule, we can still create an expression for PRTE, but both the weights and the MTE function are no longer identified without invoking further assumptions.

One superficially appealing way to avoid these problems is to define $\tilde{U}_D = F_{U_D|X,Z}(U_D)$ and $\tilde{\mu}_D(Z) = F_{U_D|X,Z}(\mu_D(Z))$, so $D = \mathbb{1}[\mu_D(Z) - U_D \geq 0] = \mathbb{1}[\tilde{\mu}_D(Z) - \tilde{U}_D \geq 0]$ with $\tilde{U}_D \sim \text{Unif}[0, 1]$ conditional on X and Z, and so $\tilde{U}_D$ is independent of X and Z. It might seem that the previous analysis would carry over. However, by defining $\tilde{U}_D = F_{U_D|X,Z}(U_D)$, we have defined $\tilde{U}_D$ in a way that depends functionally on Z and X, and hence we violate invariance of the MTE with respect to the shifts in the distribution of Z given X.

6.3. *Do We Need Uniformity?*

The monotonicity or uniformity condition and the additional condition of positive weights for MTE are both required to obtain gross treatment effects using IV. If these conditions are violated, changes in Z induce two-way flows, with some people changing into treatment and others leaving it. Thus IV does not identify the "gross effect" of treatment. Recall from our discussion in Section 4.3 that even if we have monotonicity or uniformity as defined in this section (a necessary and sufficient condition for the existence of representation (4)), the discussion in Section 4.3 reveals that in a model with multiple instruments we may still obtain negative IV weights unless we condition on the other instruments.

Monotonicity and independence are invoked when the treatment (indicated by D) is the policy being evaluated. However, treatments are only a subset of all possible policies of interest, and if the goal is to evaluate the effects of a policy on aggregate outcomes, as in Δ^{PRTE}, the monotonicity requirement may not be needed. In that case, one is interested in the net impact of the policy and not the impact of treatment operating through a particular mechanism or treatment. The policy of interest may entail two-way flows.

Consider the case where D indicates schooling, which is the treatment. Define $D = 1$ if the person goes to college and $D = 0$ otherwise. Suppose that the policy being studied is the introduction of a physical education (PE) requirement in colleges along with mandatory augmented athletics facilities. The policy has no effect on (Y_1, Y_0) (e.g., potential earnings), but it affects the choice of college, so it is a valid Z. Some people hate PE while others love it and are attracted by colleges with good gyms, so monotonicity (uniformity) is violated. If $Z_a = z$ is the policy with PE and $Z_{a'} = z'$ is the policy without PE, $E(Y|Z_a = z) - E(Y|Z_{a'} = z')$ is a perfectly valid policy parameter—the effect of the policy on aggregate outcomes—even if uniformity is violated and Δ^{MTE} is not policy invariant. One only needs uniformity, policy invariance, and the other assumptions only when the policy is a treatment.

6.4. *The Limits of Instrumental Variable Estimators*

The treatment effect literature focuses on a class of policies that move treatment choices in the same direction for everyone. General instruments do not

have universally positive weights on Δ^{MTE}. They are not guaranteed to shift everyone in the same direction. They do not necessarily estimate gross treatment effects. However, the effect of treatment is not always the parameter of policy interest. Thus, in the example just presented, schooling is the vehicle through which policy operates. One might be interested in the effect of schooling (the treatment effect) or the effect of the policy. These are separate issues unless the policy is the treatment.

Generalizing the MTE to the case of a nonseparable choice equation that violates the monotonicity condition, we can define but cannot identify the policy parameters of interest. If we make the model symmetrically heterogeneous in outcome and choice equations, the method of instrumental variables and our extensions of it break down in terms of estimating economically interpretable parameters. This case is beyond the outer limits of an entire literature, although it captures intuitively plausible phenomena. More general structural methods are required.[55]

7. SUMMARY AND PROPOSED EXTENSIONS

This paper develops an approach to policy evaluation based on the marginal treatment effect (Δ^{MTE}), which provides a choice-theoretic foundation for organizing the treatment effect literature. All of the conventional treatment effect parameters can be expressed as different weighted averages of Δ^{MTE}. These conventional treatment effect parameters do not, in general, answer economically interesting questions. We define the policy relevant treatment effect as the solution to a Benthamite policy criterion for policies operating on decisions to participate, but not on potential outcomes. The policy relevant treatment effect can be represented as a weighted average of Δ^{MTE}, where the weights differ in general from the weights used to generate conventional treatment effects. Thus the conventional treatment effects are not guaranteed to answer policy relevant questions.

Instrumental variable estimators and OLS estimators converge to expressions that can be represented as weighted averages of Δ^{MTE} parameters, with the weights in general different from those used to define the various treatment effects and the weights not necessarily positive, so they do not identify a gross treatment effect. We show how to check whether the weights are positive. Conventional IV and matching assumptions impose a strong condition on the Δ^{MTE}—that selection into programs is not made in terms of any unobservable gain from program participation.

We present methods for estimating Δ^{MTE} based on local instrumental variables and we develop a new instrumental variable for recovering policy relevant

[55]The framework of Carneiro, Hansen, and Heckman (2003) can be generalized to allow for random coefficients models in choice equations and lack of policy invariance in the sense of (A-7). However, a fully semiparametric analysis does not appear to be possible. This generalization is being prepared for publication.

treatment effects using standard instrumental variable methods. We present conditions for using IV to estimate well-posed policy questions. We show that IV need not generate any interesting policy counterfactual and that there are policy counterfactuals for which no IV can be generated. In a model of heterogeneous responses, there is no natural superiority of conventional IV over OLS in estimating policy relevant parameters. We develop the conditions required to forecast the effects of old policies on new environments and the effects of new policies. These issues are typically ignored in the treatment effect literature, but are central to the structural policy evaluation literature.

The model presented in this paper and the models presented in the recent literature on instrumental variables in heterogeneous response models are fundamentally asymmetric. Responses to treatment are allowed to be heterogeneous in a general way, but choices of treatment are not. When we develop a symmetrically heterogeneous model, the method of instrumental variables breaks down entirely and a different approach to econometric policy analysis is required.

Dept. of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637, U.S.A.; jjh@uchicago.edu

and

Dept. of Economics, Stanford University, 579 Serra Mall, Palo Alto, CA 94305, U.S.A.; vytlacil@stanford.edu.

Manuscript received June, 2001; final revision received October, 2004.

APPENDIX A: TESTABLE MONOTONICITY IN P RESTRICTION

THEOREM 1: *Assume* Y_0, Y_1, *and* D *are determined by Equations* (2a), (2b), *and* (3), *respectively. Assume conditions* (A-1) *through* (A-5) *hold.*

(i) *Let* g_0, g_1 *be any real valued functions such that* $g_0(Y_0, X), g_1(Y_1, X) \geq 0$ *with probability* 1. *Then* $E((1-D)g_0(Y,X)|X, P(Z)=p)$ *is weakly decreasing in* p *and* $E(Dg_1(Y,X)|X, P(Z)=p)$ *is weakly increasing in* p.

(ii) *Let* g_0, g_1 *be any real valued functions such that* $g_0(Y_0, X), g_1(Y_1, X) > 0$ *with probability* 1. *Then* $E((1-D)g_0(Y,X)|X, P(Z)=p)$ *is strictly decreasing in* p *and* $E(Dg_1(Y,X)|X, P(Z)=p)$ *is strictly increasing in* p.

PROOF: Consider assertion (i). Consider $E(Dg_1(Y,X)|X=x, P(Z)=p)$ for some x. Let p_1, p_0 denote any two points in the support of the distribution of $P(Z)$ conditional on $X=x$ such that $p_1 > p_0$. Then

$$\begin{aligned}
&E(Dg_1(Y,X)|X=x, P(Z)=p_1) \\
&\quad - E(Dg_1(Y,X)|X=x, P(Z)=p_0) \\
&= E\big(\mathbb{1}[U_D \leq P(Z)]g_1(Y_1,X)|X=x, P(Z)=p_1\big) \\
&\quad - E\big(\mathbb{1}[U_D \leq P(Z)]g_1(Y_1,X)|X=x, P(Z)=p_0\big)
\end{aligned}$$

$$
\begin{aligned}
&= E\big(\mathbb{1}[U_D \le p_1] g_1(Y_1, X) | X = x\big) \\
&\quad - E\big(\mathbb{1}[U_D \le p_0] g_1(Y_1, X) | X = x\big) \\
&= E\big(\{\mathbb{1}[U_D \le p_0] + \mathbb{1}[p_0 < U_D \le p_1]\} g_1(Y_1, X) | X = x\big) \\
&\quad - E\big(\mathbb{1}[U_D \le p_0] g_1(Y_1, X) | X = x\big) \\
&= E\big(\mathbb{1}[p_0 < U_D \le p_1] g_1(Y_1, X) | X = x\big) \\
&\ge 0,
\end{aligned}
$$

where the first equality follows from the definition of D and uses the fact that $Dg_1(Y, X) = Dg_1(Y_1, X)$; the second equality uses independence condition (A-2); the third equality uses the fact that $p_0 < p_1$ and thus that $\mathbb{1}[U_D \le p_1] = \mathbb{1}[U_D \le p_0] + \mathbb{1}[p_0 < U_D \le p_1]$; the fourth equality follows from linearity of expectations; and the final inequality follows from $g_1(Y_1, X) \ge 0$ with probability 1. The proof that $E((1-D)g_0(Y, X)|X, P(Z) = p)$ is decreasing in p follows from a similar argument. Assertion (ii) follows from a trivial modification of the last line of the above proof. *Q.E.D.*

Consider the following examples of g_0 and g_1:

(i) If Y_1, Y_0 are known to be nonnegative (for example, Y_1, Y_0 are indicator variables or Y_1, Y_0 are wages), then we may choose $g_j(Y, X) = Y$. In this example, Theorem 1 implies that $E((1 - D)Y|X, P(Z) = p)$ is weakly decreasing in p and $E(DY|X, P(Z) = p)$ is weakly increasing in p. More generally, if Y_1, Y_0 are known to be bounded from below by a function of X so that $Y_1 \ge l_1(X)$, $Y_0 \ge l_0(X)$ with probability 1, then we may choose $g_j(Y, X) = Y - l_j(X)$ so that Theorem 1 implies that $E((1-D)(Y - l_0(X))|X, P(Z) = p)$ is weakly decreasing in p and $E(D(Y - l_1(X))|X, P(Z) = p)$ is weakly increasing in p.

(ii) Without any assumptions on the support of the distribution of Y_1, Y_0, let t denote a real number and take $g_j(Y, X) = \mathbb{1}[Y \le t]$ for $j = 0, 1$. Then Theorem 1 implies that $\Pr(D = 0, Y \le t|X, P(Z) = p)$ is weakly decreasing in p and $\Pr(D = 1, Y \le t|X, P(Z) = p)$ is weakly increasing in p. More generally, let $\mathcal{A}$ denote any measurable subset of the real line and take $g_j(Y, X) = \mathbb{1}[Y \in \mathcal{A}]$ for $j = 0, 1$. Then the conclusion of the proposition can be rewritten as $\Pr(D = 0, Y \in \mathcal{A}|X, P(Z) = p)$ is weakly decreasing in p and $\Pr(D = 1, Y \in \mathcal{A}|X, P(Z) = p)$ is weakly increasing in p.

For any choice of g_0, g_1, the restriction of Theorem 1 leads to the prediction that regression functions $E((1-D)g_0(Y, X)|X, P(Z) = p)$ and $E(Dg_1(Y, X)|X, P(Z) = p)$ satisfy the monotonicity conditions. This is an example of a nonparametric null with shape restrictions versus a nonparametric alternative. A formal test of the null hypothesis can be implemented using the methodology of Ghosal, Sen, and van der Vaart (2000).

The restrictions of Theorem 1 nest the Imbens–Rubin (1997) restrictions on IV as a special case. They assume a binary Z, and obtain the density of

Y_1 and Y_0 from the observed data and derive the testable restriction that these densities be nonnegative.[56] Our analysis is more general.[57]

APPENDIX B: DERIVATION OF PRTE AND IV WEIGHTS

PROOF OF EQUATION (6): To simplify the notation, assume that $V(Y) = Y$. Modifications required for the more general case are obvious. Define $\mathbb{1}_{\mathcal{A}}(t)$ to be the indicator function for the event $t \in \mathcal{A}$. Then

$$
\begin{aligned}
E(Y_a|X) &= \int_0^1 E(Y_a|X, P_a(Z_a) = p)\, dF_{P_a|X}(p) \\
&= \int_0^1 \Bigg[\int_0^1 \mathbb{1}_{[0,p]}(u_D) E(Y_{1,a}|X, U_D = u_D) \\
&\qquad + \mathbb{1}_{(p,1]}(u_D) E(Y_{0,a}|X, U_D = u_D)\, du \Bigg] dF_{P_a|X}(p) \\
&= \int_0^1 \Bigg[\int_0^1 \big[\mathbb{1}_{[u_D,1]}(p) E(Y_{1,a}|X, U_D = u_D) \\
&\qquad + \mathbb{1}_{(0,u_D]}(p) E(Y_{0,a}|X, U_D = u_D) \big] dF_{P_a|X}(p) \Bigg] du_D \\
&= \int_0^1 \big[(1 - F_{P_a|X}(u_D)) E(Y_{1,a}|X, U_D = u_D) \\
&\qquad + F_{P_a|X}(u_D) E(Y_{0,a}|X, U_D = u_D) \big] du_D.
\end{aligned}
$$

This derivation involves changing the order of integration. Note that from (A-4),

$$
\begin{aligned}
&E\big| \mathbb{1}_{[0,p]}(u_D) E(Y_{1,a}|X, U_D = u_D) + \mathbb{1}_{(p,1]}(u_D) E(Y_{0,a}|X, U_D = u_D) \big| \\
&\quad \leq E(|Y_1| + |Y_0|) < \infty,
\end{aligned}
$$

[56] See also Heckman, Smith, and Taber (1998) for a closely related test.

[57] For ease of exposition, suppress conditioning on X. Take the case where $Z = 0, 1$ and $P(1) > P(0)$. Consider the Y_1 outcome; the analysis for Y_0 is completely symmetric. For binary Z with $P(1) > P(0)$, our restriction can be rewritten as $E(Dg_1(Y)|Z = 1) \geq E(Dg_1(Y)|Z = 0)$. Take $g_1(Y) = \mathbb{1}[Y \in \mathcal{A}]$ for any prespecified set $\mathcal{A}$, for example, the intervals examined in the histogram analyzed in the Imbens–Rubin paper. In this special case, our monotonicity restriction is that $\Pr(D = 1, Y \in \mathcal{A}|Z = 1) - \Pr(D = 0, Y \in \mathcal{A}|Z = 0) > 0$, and the restriction is the same as the Imbens and Rubin restriction of a nonnegative density. Our analysis replaces their densities with the probability that Y lies in any given set. Thus, their restriction is a very special case of the general monotonicity restriction developed in this paper.

so the change in the order of integration is valid by Fubini's theorem. Comparing policy a to policy a',

$$E(Y_a|X) - E(Y_{a'}|X) = \int_0^1 E(\Delta|X, U_D = u_D)\left(F_{P_{a'}|X}(u_D) - F_{P_a|X}(u_D)\right) du_D,$$

which gives the required weights. (Recall $\Delta = Y_1 - Y_0$ and from (A-7) we can drop the a, a' subscripts on outcomes and errors.) *Q.E.D.*

Relaxing (A-7): *Implications of Noninvariance for PRTE*

Suppose that all of the assumptions invoked up to Section 3 are satisfied, including additive separability in the latent index choice Equation (3) (equivalently, the monotonicity or uniformity condition). Impose the normalization that the distribution of U_D is unit uniform. Suppose however, contrary to (A-7), that the distribution of (Y_1, Y_0, U_D, X) is different under the two regimes a and a'. Thus, let $(Y_{1,a}, Y_{0,a}, U_{D,a}, X_a)$ and $(Y_{1,a'}, Y_{0,a'}, U_{D,a'}, X_{a'})$ denote the random vectors under regimes a and a', respectively. Following the same analysis used to derive Equation (6), the PRTE conditional on X is given by

$$\begin{aligned}
&E(Y_a|X_a = x) - E(Y_{a'}|X_{a'} = x) \\
\text{(I)} \qquad &= \int_0^1 E(Y_{1,a} - Y_{0,a}|X_a = x, U_{D,a'} = u) \\
&\qquad \times \left[F_{P_{a'}|X_{a'}}(u|x) - F_{P_a|X_a}(u|x)\right] du \\
\text{(II)} \qquad &+ \int_0^1 \left[E(Y_{0,a}|X_a = x, U_{D,a} = u)\right. \\
&\qquad \left. - E(Y_{0,a'}|X_{a'} = x, U_{D,a'} = u)\right] du \\
\text{(III)} \qquad &+ \int_0^1 \left[(1 - F_{P_{a'}|X}(u|x))\right. \\
&\qquad \times \left(E(Y_{1,a} - Y_{0,a}|X_a = x, U_{D,a} = u)\right. \\
&\qquad \left.\left. - E(Y_{1,a'} - Y_{0,a'}|X_{a'}, U_{D,a'} = u)\right)\right] du.
\end{aligned}$$

Thus, when the policy affects the distribution of (Y_1, Y_0, U_D, X), the PRTE is given by the sum of three terms: (I) the value of PRTE if the policy did not affect (Y_1, Y_0, X, U_D), (II) the weighted effect of the policy change on $E(Y_0|X, U_D)$, and (III) the weighted effect of the policy change on MTE. Evaluating the PRTE requires knowledge of the MTE function in both regimes and knowledge of $E(Y_0|X = x, U_D = u)$ in both regimes, as well as knowledge of the distribution of $P(Z)$ in both regimes. Note, however, that if we assume

that the distribution of $(Y_{1,a}, Y_{0,a}, U_{D,a})$ conditional on $X_a = x$ equals the distribution of $(Y_{1,a'}, Y_{0,a'}, U_{D,a'})$ conditional on $X_{a'} = x$, then $E(Y_{1,a}|U_{D,a} = u, X_a = x) = E(Y_{1,a'}|U_{D,a'} = u, X_{a'} = x)$, $E(Y_{0,a}|U_{D,a} = u, X_a = x) = E(Y_{0,a'}|U_{D,a'} = u, X_{a'} = x)$, and thus terms (II) and (III) are zero and the expression for PRTE simplifies to the expression of Equation (6).

PROOF OF EQUATION (10): We have

$$
\begin{aligned}
&\operatorname{Cov}(J(Z), Y|X) \\
&\quad = E\big([J(Z) - E(J(Z)|X)]Y|X\big) \\
&\quad = E\big(\big(J(Z) - E(J(Z)|X)\big)(Y_0 + D(Y_1 - Y_0))|X\big) \\
&\quad = E\big(\big(J(Z) - E(J(Z)|X)\big)D(Y_1 - Y_0)|X\big) \\
&\quad = E\big(\tilde{J}(Z)\mathbb{1}[U_D \le P(Z)](Y_1 - Y_0)|X\big) \\
&\quad = E\big(\tilde{J}(Z)\mathbb{1}[U_D \le P(Z)]E(Y_1 - Y_0|X, Z, U_D)|X\big) \\
&\quad = E\big(\tilde{J}(Z)\mathbb{1}[U_D \le P(Z)]E(Y_1 - Y_0|X, U_D)|X\big) \\
&\quad = E\big(E\big[\tilde{J}(Z)\mathbb{1}[U_D \le P(Z)]|X = x, U_D\big]E(Y_1 - Y_0|X, U_D)|X\big) \\
&\quad = \int_0^1 \big[E(\tilde{J}(Z)|X, P(Z) \ge u_D) \\
&\qquad\qquad \times \Pr(P(Z) \ge u_D)E(Y_1 - Y_0|X, U_D = u_D)\big]\, du_D \\
&\quad = \int_0^1 \Delta^{\mathrm{MTE}}(X, u_D)E(\tilde{J}(Z)|X, P(Z) \ge u_D)\Pr(P(Z) \ge u_D)\, du_D.
\end{aligned}
$$

The third equality follows from (3); the fourth equality follows from the law of iterated expectations with the inside expectation conditional on (X, Z, U_D); the fifth equality follows from assumption (A-2); the sixth equality follows from the law of iterated expectations with the inside expectation conditional on $(X = x, U_D)$; the seventh equality follows from Fubini's Theorem and the normalization that U_D is distributed unit uniform conditional on X; and the final equality follows from plugging in the definition of Δ^{MTE}. Yitzhaki (1996, 1999) was the first to develop the interpretation of IV as a weighted average, although he did not develop the MTE. *Q.E.D.*

PROOF OF EQUATION (17): We have

$$
\begin{aligned}
&E(Y_a|X) \\
&\quad = \int E(Y_a|X, U_D = u_D, Z_a = z)\, dF_{U_D, Z_a|X}(u_D, z)
\end{aligned}
$$

$$
\begin{aligned}
&= \int \big[\mathbb{1}_{\Omega_{u_D}}(z)E(Y_1|X, U_D = u_D, Z_a = z) \\
&\qquad + \mathbb{1}_{\Omega^c_{u_D}}(z)E(Y_0|X, U_D = u_D, Z_a = z)\big]\, dF_{U_D,Z_a|X}(u_D, z) \\
&= \int \big[\mathbb{1}_{\Omega_{u_D}}(z)E(Y_1|X, U_D = u_D) \\
&\qquad + \mathbb{1}_{\Omega^c_{u_D}}(z)E(Y_0|X, U_D = u_D)\big]\, dF_{U_D,Z_a|X}(u_D, z) \\
&= \int \Bigg[\int \big[\mathbb{1}_{\Omega_{u_D}}(z)E(Y_1|X, U_D = u_D) \\
&\qquad\qquad + \mathbb{1}_{\Omega^c_{u_D}}(z)E(Y_0|X, U_D = u_D)\big]\, dF_{Z_a|X}(z) \Bigg]\, dF_{U_D|X}(u_D) \\
&= \int \big[\Pr[Z_a \in \Omega_{u_D}|X]E(Y_1|X, U_D = u_D) \\
&\qquad + (1 - \Pr[Z_a \in \Omega_{u_D}|X])E(Y_0|X, U_D = u_D)\big]\, dF_{U_D|X}(u_D),
\end{aligned}
$$

where $\Omega^c_{u_D}$ denotes the complement of Ω_{u_D} and where the first equality follows from the law of iterated expectations; the second equality follows by plugging in our threshold crossing model for D; the third equality follows from independence $Z \perp\!\!\!\perp (Y_1, Y_0, U_D)|X$; and the fourth and fifth equalities follow by an application of Fubini's Theorem and a rearrangement of terms. Fubini's Theorem may be applied by assumption (A-4). Thus comparing policy a to policy a', we obtain (17):

$$
\begin{aligned}
&E(Y_a|X) - E(Y_{a'}|X) \\
&\quad = \int E(\Delta|X, U_D = u_D) \\
&\qquad \times \big(\Pr[Z_a \in \Omega_{u_D}|X] - \Pr[Z_{a'} \in \Omega_{u_D}|X]\big)\, dF_{U_D|X}(u_D). \quad \textit{Q.E.D.}
\end{aligned}
$$

PROOF OF EQUATION (18): We have

$$
\begin{aligned}
&E(Y_a|X) \\
&\quad = \int E(Y_a|X, U_D = u_D, Z_a = z)\, dF_{U_D,Z_a|X}(u_D, z) \\
&\quad = \int \begin{bmatrix} \mathbb{1}_{[0,\mu_D(z)]}(u_D)E(Y_1|X, Z = z, U_D = u_D) \\ + \mathbb{1}_{(\mu_D(z),1]}(u_D)E(Y_0|X, Z = z, U_D = u_D) \end{bmatrix} dF_{U_D,Z_a|X}(u_D, z) \\
&\quad = \int \begin{bmatrix} \mathbb{1}_{[0,\mu_D(z)]}(u_D)E(Y_1|X, U_D = u_D) \\ + \mathbb{1}_{(\mu_D(z),1]}(u_D)E(Y_0|X, U_D = u_D) \end{bmatrix} dF_{U_D,Z_a|X}(u_D, z)
\end{aligned}
$$

$$= \int \left[\int \left(\begin{array}{l} \mathbb{1}_{[0,\mu_D(z)]}(u_D) E(Y_1|X, U_D = u_D) \\ + \mathbb{1}_{(\mu_D(z),1]}(u_D) E(Y_0|X, U_D = u_D) \end{array} \right) dF_{Z_a|U_D}(z|u_D) \right]$$
$$\times dF_{U_D|X}(u_D)$$
$$= \int \left[\begin{array}{l} (1 - \Pr[\mu_D(Z_a) < u_D|U_D = u_D]) E(Y_1|X, U_D = u_D) \\ + \Pr[\mu_D(Z_a) < u_D|U_D = u_D] E(Y_0|X, U_D = u_D) \end{array} \right]$$
$$\times dF_{U_D|X}(u_D),$$

where the first equality follows from the law of iterated expectations; the second equality follows by plugging in our model for D; the third equality follows from independence $Z \perp\!\!\!\perp (Y_1, Y_0)|X, U_D$; the fourth equality follows by an application of Fubini's Theorem; and the final equality follows immediately. Thus comparing policy a to policy a', we obtain (18) in the text. *Q.E.D.*

APPENDIX C: PROOFS OF PROPOSITIONS

PROOF OF PROPOSITION 1: We first show that, given (i), conditions (ii) and (iii) are sufficient for the instrument $J_x(Z)$ defined in the proposition to have the desired properties. As a preliminary step, note that

$$E(J_x|X = x) = \int_0^1 \mathbb{1}[f_{P|X}(p|x) > 0] w'(p|x)\, dp = \int_0^1 w'(p|x)\, dp = 0,$$

where the first equality comes from plugging in the proposed J_x and using condition (i); the second equality follows from condition (iii); and the final equality follows from condition (ii). We now check that the proposed J_x is correlated with D under conditions (i)–(iii):

$$\begin{aligned} \mathrm{Cov}(J_x(Z), D|X = x) &= \mathrm{Cov}(J_x(Z), P(Z)|X = x) \\ &= \int_0^1 \mathbb{1}[f_{P|X}(p|x) > 0] w'(p|x)\, p\, dp \\ &= \int_0^1 w'(p|x)\, p\, dp = -1, \end{aligned}$$

where the first equality follows from the law of iterated expectations; the second equality comes from plugging in the proposed J_x and using $E(J_x|X = x) = 0$; and the third equality uses condition (iii) and the final equality follows from integration by parts using condition (ii). We now check that the proposed instrument J_x implies the desired weights on Δ^{MTE}. With the proposed J_x, for u such that $f_{P|X}(u|x) > 0$,

$$-\frac{T(u|x; J_x) f_{P|X}(u|x)}{\mathrm{Cov}(J_x, P|X = x)} = w'(u|x),$$

where the equality comes from plugging in the proposed J_x and using $E(J_x|X=x)=0$ and $\text{Cov}(J_x, P|X=x)=-1$. Thus, for u such that $f_{P|X}(u|x)>0$, we have that $-T(u|x;J_x)f_{P|X}(u|x)/\text{Cov}(J_x,P|X=x)=w'(u|x)$ as desired. For u such that $f_{P|X}(u|x)=0$, condition (iii) implies that $w'(u|x)=0$ and thus trivially $-T(u|x;J_x)f_{P|X}(u|x)/\text{Cov}(J_x,P|X=x)=w'(u|x)$ for u such that $f_{P|X}(u|x)=0$.

We now show that, given condition (i), conditions (ii) and (iii) are necessary. First, consider condition (ii). We have previously established that the weights corresponding to any instrument must integrate to 1, and that the weights corresponding to any instrument must satisfy $w(0|x)=0$. One can also directly verify that $w(1|x)=0$ unless the conditional distribution of $P(Z)$ has a mass point at 1. The conditional distribution of $P(Z)$ does not have a mass point at 1 by condition (i). Using condition (i), one can apply Lebesgue's theorem for the derivative of an integral to show that the weights corresponding to any instrument will be differentiable. Thus, given condition (i), the weights corresponding to any instrument will satisfy condition (ii), and thus condition (ii) is a necessary condition for there to exist an instrument that corresponds to the desired weights. Now assume conditions (i) and (ii), and consider condition (iii). Suppose that (iii) does not hold, so that there exists a set of t values such that $f_{P|X}(t|x)=0$ but $w'(t|x)>0$. Then, for such values of t,

$$-\frac{T(t|x;J_x)f_{P|X}(t|x)}{\text{Cov}(J_x,P|X=x)}=0$$

for any potential instrument J_x while $w'(t|x)>0$, and thus trivially there cannot exist an instrument J_x such that

$$-\frac{T(t|x;J_x)f_{P|X}(t|x)}{\text{Cov}(J_x,P|X=x)}=w'(t|x)\quad\text{for all } t.$$

Thus, given condition (i), conditions (ii) and (iii) are necessary for the existence of an instrument with the desired properties. *Q.E.D.*

PROOF OF PROPOSITION 2: Define

$$w(\cdot|x)\equiv\frac{F_{P_{a'}|X}(\cdot|x)-F_{P_a|X}(\cdot|x)}{E(P_a|X=x)-E(P_{a'}|X=x)}.$$

We now show that conditions (i) and (ii) of Proposition 2 imply conditions (i) and (ii) of Proposition 1 when $w(\cdot|x)$ is defined in this manner. Note that condition (i) of Proposition 2 immediately implies that condition (i) of Proposition 1 holds. Condition (i) of Proposition 2 implies that $w(\cdot|x)$ is differentiable for all evaluation points with

$$w'(\cdot|x)=\frac{f_{P_{a'}|X}(\cdot|x)-f_{P_a|X}(\cdot|x)}{E(P_a|X=x)-E(P_{a'}|X=x)}.$$

Since $F_{P_a|X}(\cdot)$ and $F_{P_{a'}|X}(\cdot)$ are distribution functions, one can directly verify that $\int_0^1 w(u|x)\,du = 1$. Since the propensity score is always bounded by 0 and 1, using condition (i), $F_{P_a|X}(1|x) = F_{P_{a'}|X}(1|x) = 1$ and $F_{P_a|X}(0|x) = F_{P_{a'}|X}(0|x) = 0$, and thus $w(1|x) = w(0|x) = 0$. Defining $w(\cdot|x)$ in this manner, we have that conditions (i) and (ii) of Proposition 2 imply conditions (i) and (ii) of Proposition 1. Given

$$w'(\cdot|x) = \frac{f_{P_{a'}|X}(\cdot|x) - f_{P_a|X}(\cdot|x)}{E(P_a|X=x) - E(P_{a'}|X=x)},$$

we have that condition (iii) of Proposition 2 is equivalent to condition (iii) of Proposition 1 for this choice of $w(\cdot|x)$. The result now follows directly from Proposition 1. *Q.E.D.*

PROOF OF PROPOSITION 3: Assume that the conditions of Proposition 1 hold almost everywhere with respect to X. From the proof of Proposition 1, under the stated conditions, $E(J(Z)|X) = 0$, $\mathrm{Cov}(J(Z), D|X) = -1$, and

$$\frac{\mathrm{Cov}(J(Z), Y|X)}{\mathrm{Cov}(J(Z), D|X)} = \int \Delta^{\mathrm{MTE}}(X, u) w(u|X)\,du.$$

It follows that $\mathrm{Cov}(J(Z), D|x) = \mathrm{Cov}(J(Z), D|X) = -1$, that $\mathrm{Cov}(J(Z), Y) = E(J(Z)Y) = E[E(J(Z)Y|X)]$, and thus that

$$\begin{aligned}\frac{\mathrm{Cov}(J(Z), Y)}{\mathrm{Cov}(J(Z), D)} &= E\big[-E(J(Z)Y|X)\big] \\ &= E\left[\frac{\mathrm{Cov}(J(Z), Y|X)}{\mathrm{Cov}(J(Z), D|X)}\right] \\ &= \int\left[\int_0^1 \Delta^{\mathrm{MTE}}(x, u) w(u|x)\,du\right] dF_X(x). \qquad \textit{Q.E.D.}\end{aligned}$$

PROOF OF PROPOSITION 4: We first show that, given condition (i), conditions (ii) and (iii) are sufficient for the instrument $J_x(Z)$ defined by the proposition to have the desired properties. For notational convenience, define $\pi_l = \Pr[P(Z) = p_l]$, for $l = 1, \ldots, K$. As a preliminary step, note that with this definition of $J_x(Z)$,

$$\begin{aligned}E(J_x|X=x) &= \sum_{l=1}^{K} \frac{1}{\pi_l}[w_l - w_{l+1}]\pi_l \\ &= \sum_{l=1}^{K}[w_l - w_{l+1}] = w_1 - w_{K+1} = 0,\end{aligned}$$

where we use the fact that $w_1 = w_{K+1} = 0$. We now check that the proposed $J_x(Z)$ is correlated with D under conditions (i) to (iii):

$$\begin{aligned}\operatorname{Cov}(J_x(Z), D|X=x) &= \operatorname{Cov}(J_x(Z), P(Z)|X=x)\\ &= \sum_{l=1}^{K} \frac{1}{\pi_l}[w_l - w_{l+1}]p_l\pi_l = \sum_{l=2}^{K}[p_l - p_{l-1}]w_l\\ &= \sum_{l=1}^{K}\int_{p_{l-1}}^{p_l} w(t|x)\,dt = \int_0^1 w(t|x)\,dt = 1,\end{aligned}$$

where the first equality follows from the law of iterated expectations; the second equality comes from $E(J_x(Z)|X=x)=0$ and plugging in the proposed $J_x(Z)$; the third equality rearranges terms in the sum using $w_1 = w_{K+1} = 0$; the fourth equality uses condition (iii) and the definition of w_l; the fifth equality uses linearity of integration; and the final equality uses condition (ii). We now check that the proposed instrument J_x implies the desired weights on Δ^{MTE}. Using $\operatorname{Cov}(J_x(Z), D|X=x) = 1$ and that $E(J_x(Z)|X=x) = 0$, the IV weights corresponding to the proposed $J_x(Z)$ as given by $h_{\text{IV}}(u) = E(J_x(Z)\mathbb{1}[P(Z) \geq u])$. We immediately have $h_{\text{IV}}(u) = w(u|x) = 0$ for $u \in (p_K, 1]$ and for $u \in [0, p_1]$. For $u \in (p_{j-1}, p_j]$, $j = 2, \ldots, K$, we have

$$\begin{aligned}E\big(J_x(Z)\mathbb{1}[P(Z) \geq u]\big) &= \sum_{l=1}^{K} \frac{1}{\pi_l}[w_l - w_{l+1}]\mathbb{1}[p_l \geq u]\pi_l\\ &= \sum_{l=j}^{K}[w_l - w_{l+1}] = w_j - w_{K+1} = w_j = w(u|x),\end{aligned}$$

where the first equality comes from plugging in the proposed $J_x(Z)$; the second equality follows from $u \in (p_{j-1}, p_j]$; the third equality follows by rearranging terms; the fourth equality follows from $w_{K+1} = 0$; and the final equality follows by the definition of w_j and the fact that $u \in (p_{j-1}, p_j]$.

We now show that, given condition (i), conditions (ii) and (iii) are necessary. First, consider condition (ii). One can verify that the weights corresponding to any instrument must integrate to 1. One can also verify that the weights corresponding to any instrument must satisfy $w(u|x) = 0$ for $u \leq p_x^{\text{Min}}$ and for $u > p_x^{\text{Max}}$, where p_x^{Min} and p_x^{Max} are the minimum and maximum values in the support of the conditional distribution of $P(Z)$. Given condition (i), one can immediately verify that the IV weights for any instrument must satisfy condition (iii). Thus, given condition (i), the weights corresponding to any instrument will satisfy conditions (ii) and (iii), and thus, given condition (i), conditions (ii) and (iii) are necessary conditions for there to exist an instrument that corresponds to the desired weights. *Q.E.D.*

APPENDIX D: LOCAL INSTRUMENTAL VARIABLES FOR THE RANDOM COEFFICIENT MODEL

Consider the model

$$D = \mathbb{1}[Z\beta \geq 0],$$

where β is a random variable. For ease of exposition, we leave implicit the conditioning on X covariates. Assume that $(Y_0, Y_1, \beta) \perp\!\!\!\perp Z$. Assume that β has a density that is absolutely continuous with respect to Lebesgue measure on $\Re^K$. We have

$$E(Y|Z=z) = E(DY_1|Z=z) + E((1-D)Y_0|Z=z).$$

To simplify the exposition, first consider the first term, $E(DY_1|Z=z)$. Using the model, the independence assumption, and the law of iterated expectations, we have

$$\begin{aligned} E(DY|Z=z) &= E(\mathbb{1}[z\beta \geq 0]Y_1) = E\big(\mathbb{1}[z\beta \geq 0]E(Y_1|\beta)\big) \\ &= E\big(\mathbb{1}\{z^{[K]}\beta^{[K]} \geq -z^{[-K]}\beta^{[-K]}\}E(Y_1|\beta)\big), \end{aligned}$$

where the final outer expectation is over β. Consider taking the derivative with respect to the Kth element of Z assumed to be continuous. Let $Z^{[K]}$ denote the Kth element of Z and let $Z^{[-K]}$ denote all other elements of Z, and write $Z = (Z^{[-K]}, Z^{[K]})$. Likewise, partition z, β, and b as $z = (z^{[-K]}, z^{[K]})$, $\beta = (\beta^{[-K]}, \beta^{[K]})$, and $b = (b^{[-K]}, b^{[K]})$, where z is a realization of Z and b is a realization of β. For simplicity, suppose that the Kth element of z is positive, $z^{[K]} > 0$. We obtain

$$\begin{aligned} E(DY|Z=z) &= E\big[E\big(\mathbb{1}\{z^{[K]}\beta^{[K]} \geq -z^{[-K]}\beta^{[-K]}\}E(Y_1|\beta)|\beta^{[-K]}\big)\big] \\ &= E\left[E\left(\mathbb{1}\left\{\beta^{[K]} \geq \frac{-z^{[-K]}\beta^{[-K]}}{z^{[K]}}\right\}E(Y_1|\beta)\middle|\beta^{[-K]}\right)\right], \end{aligned}$$

where the inside expectation is over $\beta^{[K]}$ conditional on $\beta^{[-K]}$, i.e., is over the Kth element of β conditional on all other components of β. Thus,

$$\frac{\partial}{\partial z^{[K]}}E(DY|Z=z) = \int E\big(Y_1|\beta = M(b^{[-K]})\big)\tilde{w}(b^{[-K]})\,db^{[-K]},$$

where

$$M(b^{[-K]}) = \left((b^{[-K]})', \frac{-z^{[-K]}b^{[-K]}}{z^{[K]}}\right)' \quad \text{and}$$

$$\tilde{w}(b^{[-K]}) = \frac{z^{[-K]}b^{[-K]}}{(z^{[K]})^2} f\left(b^{[-K]}, \frac{-z^{[-K]}b^{[-K]}}{z^{[K]}}\right)$$

with $f(\cdot)$ the density of β (with respect to Lebesgue measure), and where for notational simplicity we suppress the dependence of the function $M(\cdot)$ and the weights $\tilde{w}(\cdot)$ on the z evaluation point. In this expression, we are averaging over $E(Y_1|\beta = b)$, but only over b evaluation points such that $zb = 0$. In particular, the expression averages over the $K-1$ space of $b^{[-K]}$, while for each potential realization of $b^{[-K]}$ it is filling in the value of $b^{[K]}$ such that $z^{[K]}b^{[K]} = -z^{[-K]}b^{[-K]}$, so that $z^{[K]}b^{[K]} + z^{[-K]}b^{[-K]} = 0$. Note that the weights $\tilde{w}(b^{[-K]})$ will be zero for any $b^{[-K]}$ such that $f(b^{[-K]}, (-z^{[-K]}b^{[-K]})/z^{[K]}) = 0$, i.e., the weights will be zero for any $b^{[-K]}$ such that there does not exist $b^{[K]}$ in the conditional support of $\beta^{[K]}$ with $z^{[K]}b^{[K]} = -z^{[-K]}b^{[-K]}$.

Following the same logic for $E((1-D)Y_0|Z = z)$ we obtain

$$\frac{\partial}{\partial z^{[K]}} E((1-D)Y|Z = z) = -\int E\big(Y_0|\beta = M(b^{[-K]})\big)\tilde{w}(b^{[-K]})\,db^{[-K]}$$

and likewise have

$$\frac{\partial}{\partial z^{[K]}} \Pr(D = 1|Z = z) = \int \tilde{w}(b^{[-K]})\,db^{[-K]}$$

so that

$$\frac{(\partial/\partial z^{[K]})E(Y|Z = z)}{(\partial/\partial z^{[K]})\Pr(D = 1|Z = z)} = \int E\big(Y_1 - Y_0|\beta = M(b^{[-K]})\big)w(b^{[-K]})\,db^{[-K]},$$

where

$$w(b^{[-K]}) = \tilde{w}(b^{[-K]}) \Big/ \int \tilde{w}(b^{[-K]})\,db^{[-K]}.$$

Now consider the question of whether this expression will include both positive and negative weights. Recall that

$$\tilde{w}(b^{[-K]}) = \frac{z^{[-K]}b^{[-K]}}{(z^{[K]})^2} f\left(b^{[-K]}, \frac{-z^{[-K]}b^{[-K]}}{z^{[K]}}\right).$$

Thus,

$$\tilde{w}(b^{[-K]}) \begin{cases} \geq 0, & \text{if } z^{[-K]}b^{[-K]} > 0, \\ \leq 0, & \text{if } z^{[-K]}b^{[-K]} < 0, \end{cases}$$

and will be nonzero if $z^{[-K]}b^{[-K]} \neq 0$ and there exists $b^{[K]}$ in the conditional support of $\beta^{[K]}$ with $z^{[K]}b^{[K]} = z^{[-K]}b^{[-K]}$, i.e., with $zb = 0$. We thus have that there will be both positive and negative weights on the MTE if there exist values of b

in the support of β with both $z^{[-K]}b^{[-K]} > 0$ and $zb = 0$, and there exist other values of b in the support of β with $z^{[-K]}b^{[-K]} < 0$ and $zb = 0$.

REFERENCES

AHN, H., AND J. L. POWELL (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58, 3–29.

AMEMIYA, T. (1985): *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

ANDREWS, D. W. K., AND M. M. A. SCHAFGANS (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies*, 65, 497–517.

ANGRIST, J., K. GRADDY, AND G. IMBENS (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *Review of Economic Studies*, 67, 499–527.

ANGRIST, J., AND A. KRUEGER (1999): "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics*, Vol. 3A, ed. by O. Ashenfelter and D. Card. Amsterdam: Elsevier Science, 1277–1366.

BJÖRKLUND, A., AND R. MOFFITT (1987): "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *Review of Economics and Statistics*, 69, 42–49.

CAMPBELL, D. T., AND J. C. STANLEY (1966): *Experimental and Quasi-Experimental Designs for Research*. Skokie, IL: Rand McNally.

CARNEIRO, P., K. HANSEN, AND J. HECKMAN (2003): "Estimating Distributions of Counterfactuals with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on Schooling Choice," *International Economic Review*, 44, 361–422.

CHEN, X., AND Y. FAN (1999): "Consistent Hypothesis Testing in Semiparametric and Nonparametric Models for Econometric Time Series," *Journal of Econometrics*, 91, 373–401.

CUNHA, F., J. HECKMAN, AND S. NAVARRO (2005): "Separating Heterogeneity from Uncertainty in Modeling Schooling Choices," Hicks Lecture, Oxford University, April 2004; *Oxford Economic Papers*, 57, 191–261.

ELLISON, G., AND S. F. ELLISON (2000): "A Simple Framework for Nonparametric Specification Testing," *Journal of Econometrics*, 96, 1–23.

FLORENS, J.-P., J. HECKMAN, C. MEGHIR, AND E. VYTLACIL (2004): "Instrumental Variables, Local Instrumental Variables, and Control Functions," Unpublished Working Paper, University of Chicago.

GHOSAL, S., A. SEN, AND A. VAN DER VAART (2000): "Testing Monotonicity of Regression," *The Annals of Statistics*, 28, 1054–1082.

GILL, R. D., AND J. M. ROBINS (2001): "Causal Inference for Complex Longitudinal Data: The Continuous Case," *The Annals of Statistics*, 29, 1–27.

HANSEN, L., AND T. SARGENT (1981): "Linear Rational Expectations Models of Dynamically Interrelated Variables," in *Rational Expectations and Econometric Practice*, ed. by R. Lucas and T. Sargent. Minneapolis: University of Minnesota Press, 127–156.

HECKMAN, J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables," *Annals of Economic and Social Measurement*, 5, 475–492.

——— (1990): "Varieties of Selection Bias," *American Economic Review*, 80, 313–318.

——— (1997): "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources*, 32, 441–462.

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

HECKMAN, J., H. ICHIMURA, AND P. TODD (1997): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

HECKMAN, J., H. ICHIMURA, J. SMITH, AND P. TODD (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

HECKMAN, J., R. LALONDE, AND J. SMITH (1999): "The Economics and Econometrics of Active Labor Market Programs," in *Handbook of Labor Economics*, Vol. 3A, ed. by O. Ashenfelter and D. Card. Amsterdam: Elsevier Science, 1865–2097.

HECKMAN, J., L. LOCHNER, AND C. TABER (1998): "General Equilibrium Treatment Effects: A Study of Tuition Policy," *American Economic Review*, 88, 381–386.

HECKMAN, J., AND S. NAVARRO-LOZANO (2004): "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models," *Review of Economics and Statistics*, 86, 30–57.

HECKMAN, J., AND R. ROBB (1985): "Alternative Methods for Estimating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data,* ed. by J. Heckman and B. Singer. New York: Cambridge University Press, 156–245.

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

HECKMAN, J., J. SMITH, AND C. TABER (1998): "Accounting for Dropouts in Evaluations of Social Programs," *Review of Economics and Statistics*, 80, 1–14.

HECKMAN, J., S. URZUA, AND E. VYTLACIL (2004): "Understanding Instrumental Variables," *Review of Economics and Statistics Lecture*, Harvard University, April 2001; revised 2004.

HECKMAN, J., AND E. VYTLACIL (1999): "Local Instrumental Variable and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 96, 4730–4734.

——— (2001a): "Local Instrumental Variables," in *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, ed. by C. Hsiao, K. Morimune, and J. Powell. Cambridge, U.K.: Cambridge University Press, 1–46.

——— (2001b): "Instrumental Variables, Selection Models, and Tight Bounds on the Average Treatment Effect," in *Econometric Evaluations of Active Labor Market Policies in Europe*, ed. by M. Lechner and F. Pfeiffer, Heidelberg and Berlin: Physica, 1–23.

——— (2005): "Econometric Evaluation of Social Programs," in *Handbook of Econometrics*, Vol. 6, ed. by J. Heckman and E. Leamer. Amsterdam: Elsevier Science, forthcoming.

HENDRY, D. (1995): *Dynamic Econometrics*. Oxford, U.K.: Oxford University Press.

HURWICZ, L. (1962): "On the Structural Form of Interdependent Systems," in *Logic, Methodology and Philosophy of Science*, ed. by E. Nagel, P. Suppes, and A. Tarski. Stanford, CA: Stanford University Press, 232–239.

ICHIMURA, H., AND C. TABER (2002): "Direct Estimation of Policy Impacts," Unpublished Working Paper, Department of Economics, University College London.

ICHIMURA, H., AND T. THOMPSON (1998): "Maximum Likelihood Estimation of a Binary Choice Model with Random Coefficients of Unknown Distribution," *Journal of Econometrics*, 86, 269–295.

IMBENS, G., AND J. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

IMBENS, G., AND D. RUBIN (1997): "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," *Review of Economic Studies*, 64, 555–574.

MANSKI, C. (1990): "Nonparametric Bounds on Treatment Effects," *American Economic Review*, 80, 319–323.

MARSCHAK, J. (1953): "Economic Measurements for Policy and Predictions," in *Studies in Econometric Method*, ed. by W. C. Hood and T. C. Koopmans, Cowles Commission for Research in Economics Monograph 14. New York: Wiley, 1–26.

MATZKIN, R. (1994): "Restrictions of Economic Theory in Nonparametric Methods," in *Handbook of Econometrics*, Vol. 4, ed. by R. F. Engle and D. McFadden. New York: North-Holland, 2523–2558.

PEARL, J. (2000): *Causality.* Cambridge, U.K.: Cambridge University Press.

POWELL, J. L. (1994): "Estimation of Semiparametric Models," in *Handbook of Econometrics*, Vol. 4, ed. by R. F. Engle and D. McFadden. New York: North-Holland, 2443–2521.

ROY, A. (1951): "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3, 135–146.

RUDIN, W. (1974): *Real and Complex Analysis*. New York: McGraw–Hill.

SMITH, V. K., AND H. S. BANZHAF (2004): "A Diagrammatic Exposition of Weak Complementarity and the Willig Condition," *American Journal of Agricultural Economics*, 86, 455–466.

VYTLACIL, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341.

——— (2004): "A Note on Additive Separability and Latent Index Models of Binary Choice: Representation Results," Unpublished Manuscript, Stanford University.

YITZHAKI, S. (1996): "On Using Linear Regression in Welfare Economics," *Journal of Business and Economic Statistics*, 14, 478–486.

——— (1999): "The Gini Instrumental Variable, or 'The Double IV Estimator'," Unpublished Manuscript, Department of Economics, Hebrew University.

ZHENG, J. (1996): "A Consistent Test of Functional Form via Nonparametric Estimation Techniques," *Journal of Econometrics*, 75, 263–289.