The Review *of* Economics *and* Statistics

VOL. LXXXVIII AUGUST 2006 NUMBER 3

UNDERSTANDING INSTRUMENTAL VARIABLES IN MODELS WITH ESSENTIAL HETEROGENEITY

James J. Heckman, Sergio Urzua, and Edward Vytlacil*

Abstract—This paper examines the properties of instrumental variables (IV) applied to models with essential heterogeneity, that is, models where responses to interventions are heterogeneous and agents adopt treatments (participate in programs) with at least partial knowledge of their idiosyncratic response. We analyze two-outcome and multiple-outcome models, including ordered and unordered choice models. We allow for transition-specific and general instruments. We generalize previous analyses by developing weights for treatment effects for general instruments. We develop a simple test for the presence of essential heterogeneity. We note the asymmetry of the model of essential heterogeneity: outcomes of choices are heterogeneous in a general way; choices are not. When both choices and outcomes are permitted to be symmetrically heterogeneous, the method of IV breaks down for estimating treatment parameters.

I. Introduction

SUPPOSE a policy is proposed for adoption in a country. It has been tried in other countries and we know outcomes there. We also know outcomes in countries where it was not adopted. From the historical record, what can we conclude about the likely effectiveness of the policy in countries that have not implemented it?

To answer this question, we build a model of counterfactuals. Let Y_0 be the outcome (for example, the GDP) of a country under a no-policy regime. Y_1 is the outcome if the policy is implemented. $Y_1 - Y_0$ is the *treatment effect* of the policy. It may vary among countries. We observe characteristics X of various countries (level of democracy, level of population literacy, and so on). It is convenient to decompose Y_1 into its mean given X, $\mu_1(X)$, and its deviation from the mean, U_1. We can make a similar decomposition for Y_0:

$$\begin{aligned} Y_1 &= \mu_1(X) + U_1, \\ Y_0 &= \mu_0(X) + U_0. \end{aligned} \tag{1}$$

It may happen that controlling for the X, $Y_1 - Y_0$ is the same for all countries. This is the case of homogeneous treatment effects given X. More likely, countries vary in their response to the policy even after controlling for X.

Figure 1 plots the distribution of $Y_1 - Y_0$ for a benchmark X. We explain the various parameters mentioned in the figure later on. The special case of homogeneity arises when the distribution collapses to its mean. It would be ideal if we could estimate the distribution of $Y_1 - Y_0$ given X, and that has sometimes been done.[1] More often, economists focus on some mean of the distribution displayed in Figure 1 and use a regression framework to interpret the data. To turn equation (1) into a regression model, it is conventional to use a switching regression framework. Define $D = 1$ if a country adopts a policy; $D = 0$ if it does not. The observed outcome, Y, is $Y = DY_1 + (1 - D)Y_0$. Substituting equation (1) into this expression, and keeping all X implicit, we obtain

$$\begin{aligned} Y &= Y_0 + (Y_1 - Y_0)D \\ &= \mu_0 + (\mu_1 - \mu_0 + U_1 - U_0)D + U_0. \end{aligned} \tag{2}$$

Using conventional regression notation,

$$Y = \alpha + \beta D + \varepsilon, \tag{3}$$

where $\alpha = \mu_0$, $\beta = (Y_1 - Y_0) = \mu_1 - \mu_0 + U_1 - U_0$, and $\varepsilon = U_0$. The coefficient on D is the treatment effect. The case where β is the same for every country is the one conventionally assumed. More elaborate versions assume that β depends on X and estimate interactions of D with X. The case where β varies even after accounting for X is called the *random-coefficient* or *heterogeneous treatment effect* case. A great deal of attention has been focused on this case in recent decades.

Received for publication December 20, 2004. Revision accepted for publication January 26, 2006.

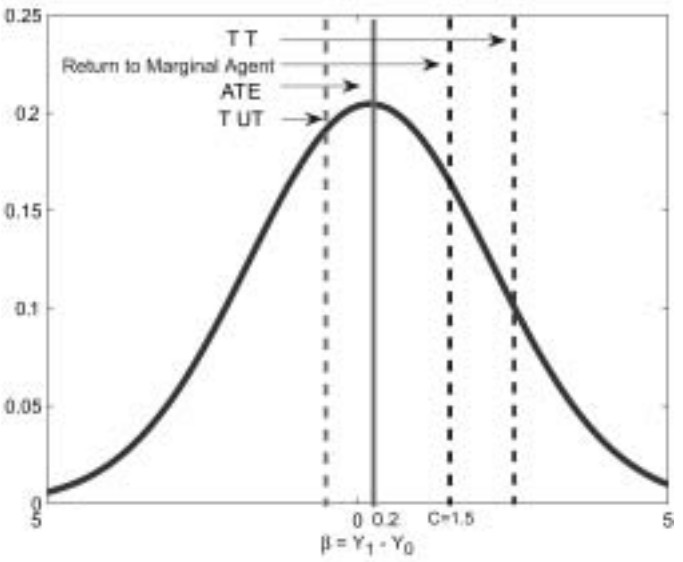
*University of Chicago, University College Dublin, and the American Bar Foundation; University of Chicago; and Columbia University, respectively.

This project was supported by NSF grants 0241858, 0099195, and 9709873; NIH grant R01-HD043411; and a grant from the American Bar Foundation. An early version of this paper was presented as the *Review of Economics and Statistics* Lecture at Harvard, April 2001; at Princeton University, December 2004; at a workshop on Causality in Economics, Oxford University, August 2005; and at the World Congress of the Econometric Society, August 2005. We thank audience participants at those seminars as well as students in Economics 350 at the University of Chicago in Winter Quarter 2005 for stimulating comments, and our discussants at the 2001 seminar, Larry Katz and Robert Moffitt, for helpful comments. We have received additional helpful comments from the editor, Jim Stock, and an anonymous referee. Bo Honoré, Derek Neal, Weerachart Kilenthong, Sergey Mityakov, Rodrigo Pinto, Jean-Marc Robin, and Jora Stixrud provided helpful comments on various drafts. Supplementary material for this paper is available at the Web site jenni.uchicago.edu/underiv.

[1] See Carneiro, Hansen, and Heckman (2001, 2003), Cunha, Heckman, and Navarro (2005, 2006), and the survey in Heckman, Lochner, and Todd (2006).

FIGURE 1.—DISTRIBUTION OF GAINS: THE ROY ECONOMY

$$U_1 - U_0 \not\perp D$$



TT= 2.666, TUT= −0.632
Return to Marginal Agent $= C = 1.5$
ATE $= \mu_1 - \mu_0 = \bar{\beta} = 0.2$

The Model

Outcomes	Choice Model
$Y_1 = \mu_1 + U_1 = \alpha + \bar{\beta} + U_1$	$D = \begin{cases} 1 \text{ if } D^* > 0 \\ 0 \text{ if } D^* \leq 0 \end{cases}$
$Y_0 = \mu_0 + U_0 = \alpha + U_0$	

General Case

$$(U_1 - U_0) \not\perp D$$
$$\text{ATE} \neq \text{TT} \neq \text{TUT}$$

The Researcher Observes (Y, D, C)

$$Y = \alpha + \beta D + U_0 \text{ where } \beta = Y_1 - Y_0$$

Parameterization

$$\alpha = 0.67 \quad (U_1, U_0) \sim N(\mathbf{0}, \mathbf{\Sigma}) \quad D^* = Y_1 - Y_0 - C$$
$$\bar{\beta} = 0.2 \quad \mathbf{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} \quad C = 1.5$$

The case where β (given X) is the same for every country is the familiar one, and we develop it first. A least squares regression of Y on D (equivalently, a mean difference in outcomes between countries with $D = 1$ and countries with $D = 0$) is possibly subject to a *selection bias.* Countries that adopt the policy may be atypical in their Y_0 ($= \alpha + \varepsilon$). Thus, if the countries that would have done well in terms of unobservable ε ($= U_0$) even in the absence of the policy are the ones that adopt the policy, then β estimated from OLS (or its nonparametric version—*matching*) is upward biased because $\text{Cov}(D, \varepsilon) > 0$.

Two main approaches have been adopted to solve this problem: (a) selection models (Gronau, 1974; Heckman, 1974, 1976a,b, 1979; Heckman and Robb, 1985, 1986; Powell, 1994) and (b) instrumental variable models (Heckman and Robb, 1985, 1986; Imbens and Angrist, 1994; Angrist and Imbens, 1995; Manski and Pepper, 2000; Heckman and Vytlacil, 1999, 2000, 2005).[2] This paper focuses on the instrumental variable (IV) approach and establishes the relationship between the selection and the IV approach using a prototypical economic model. The selection approach models levels of conditional means. The IV approach models the slopes of the conditional means. IV does not identify the constants estimated in selection models. In the general case with heterogeneity, when IV is used to identify the same level parameters that are identified by control function or selection methods, it is necessary to make the same assumptions about levels outcomes in limit sets ("identification at infinity") as are made in selection models.

For the case with homogeneous responses, if there is an instrument Z with the properties that

$$\text{Cov}(Z,D) \neq 0, \tag{4}$$

$$\text{Cov}(Z,\varepsilon) = 0, \tag{5}$$

then standard IV identifies β, at least in large samples:[3]

$$\text{plim } \hat{\beta}_{\text{IV}} = \frac{\text{Cov}(Z,Y)}{\text{Cov}(Z,D)} = \beta.$$

If other instruments exist, each identifies β. The instrument Z produces a controlled variation in D relative to ε. Randomization of assignment with full compliance with experimental protocols is an example of an instrument. From the instrumental variable estimator, we can identify the effect of adopting the policy in any country, since all countries

[2] Matching is also used. It is a form of nonparametric least squares that assumes that all relevant unobservables are accurately proxied by the observables X that the analyst happens to have at his or her disposal, so $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$ (alternatively, $(\varepsilon, \beta) \perp\!\!\!\perp D \mid X$), where $A \perp\!\!\!\perp B \mid C$ means that A is independent of B given C. See Heckman and Navarro (2004) for a discussion of matching.

[3] The proof is straightforward. Under general conditions (see, for example, White, 1984),

$$\text{plim } \hat{\beta}_{\text{IV}} = \beta + \frac{\text{Cov}(Z,\varepsilon)}{\text{Cov}(Z,D)},$$

and the second term on the right-hand side vanishes.

respond to the policy in the same way, controlling for their X.

If β $(= Y_1 - Y_0)$ varies in the population even after controlling for X, then the responses have a distribution that cannot in general be summarized by a single number. Even if we are interested in the mean of the distribution, a new phenomenon distinct from selection bias might arise. This is the problem of sorting on the gain, which is distinct from sorting on the level. If β varies even after controlling for X, there may be sorting on the gain ($\text{Cov}(\beta, D) \neq 0$). This is the model of *essential heterogeneity.*

The application of instrumental variables to this case is more problematic. Suppose that we augment the standard instrumental variable assumptions (4) and (5) by the following assumption:

$$\text{Cov}(Z,\beta) = 0. \tag{6}$$

Can we identify the mean of $Y_1 - Y_0$ using IV? In general we cannot.[4]

To see why, let $\bar{\beta} = \mu_1 - \mu_0$ be the mean treatment effect (the mean of the distribution in figure 1). Then $\beta = \bar{\beta} + \eta$, where $U_1 - U_0 = \eta$. Write equation (3) in terms of this parameter:

$$Y = \alpha + \bar{\beta}D + (\varepsilon + \eta D).$$

The error term of this equation $(\varepsilon + \eta D)$ contains two components. By assumption, Z is uncorrelated with ε and η. But to identify $\bar{\beta}$, we need the IV to be uncorrelated with $\varepsilon + \eta D$. That requires Z to be uncorrelated with ηD.

If policy adoption is made without knowledge of η $(= U_1 - U_0)$, the idiosyncratic gain to policy adoption after controlling for the observables, then η and D are statistically independent and hence uncorrelated, and IV identifies $\bar{\beta}$.[5] If, however, policy adoption is made with partial or full knowledge of η, then IV does not identify $\bar{\beta}$, because $E(\eta D|Z) = E(\eta|D = 1, Z)\Pr(D = 1|Z)$ and if sorting on the unobserved gain η occurs, the first term is not 0. Similar calculations show that IV does not identify the mean gain to the countries that adopt the policy ($E(\beta|D = 1)$) or many other summary treatment parameters.[6] Whether η $(= U_1 - U_0)$ is correlated with D depends on the quality of the data available to the empirical economist, and cannot be settled a priori. The conservative position is to allow for such correlation. However, this rules out IV as an interesting econometric strategy for identifying any of the familiar mean treatment parameters.

It is remarkable then that under certain conditions Imbens and Angrist (1994) establish that in the model with essential heterogeneity standard IV can identify an interpretable parameter. The parameter they identify is a discrete approximation to the marginal gain parameter introduced by Björklund and Moffitt (1987). Those authors demonstrate how to use a selection model to identify the marginal gain to persons induced into a treatment status by a marginal change in the cost of treatment. Imbens and Angrist (1994) show how to identify a discrete approximation to this parameter using IV.

They assume the existence of an instrument Z that takes two or more distinct values. This is implicit in equation (4). If Z assumed only one value, the covariance would be 0. Strengthening the covariance conditions of equations (5) and (6), they assume that Z is independent of β $(= Y_1 - Y_0)$ and Y_0. Let $\perp\!\!\!\perp$ denote independence. Denote by $D(z)$ the random variable indicating receipt of treatment when Z is set to z. ($D(z) = 1$ if treatment is received; $D(z) = 0$ otherwise.) The Imbens-Angrist independence assumption can be written as

IV-1: $Z \perp\!\!\!\perp (Y_0, Y_1, \{D(z)\}_{z \in \mathcal{Z}})$ (*Independence*)

where $\mathcal{Z}$ is the set of possible values of Z.

They also assume that

IV-2: $\Pr(D=1|Z)$ depends on Z (*Rank*).

This is a standard rank condition. They supplement the standard IV assumption with what they call a "monotonicity" assumption. It is a condition across persons. This assumption maintains that if Z is fixed first at one and then at the other of two distinct values, $Z = z$ and $Z = z'$, then all persons respond to the change in Z in the same way. In our policy adoption example, it states that a movement from z to z' causes all countries to move toward (or against) adoption of the policy being studied. If some adopt, others do not drop the policy in response to the same change.

More formally, letting $D_i(z)$ be the indicator $(= 1$ if adopted; $= 0$ if not) for adoption of a policy if $Z = z$ for country i, then for any distinct values z and z', Imbens and Angrist (1994) assume

IV-3: $D_i(z) \geq D_i(z')$ for all i or $D_i(z) \leq D_i(z')$ for all i, $i = 1, \ldots, I$ (*Monotonicity* or *Uniformity*).

The content in this assumption is not in the order for any person. Rather, the responses have to be *uniform* across people for a given choice of z and z'. One possibility allowed under IV-3 is the existence of three values $z < z' < z''$ such that, for all i, $D_i(z) \geq D_i(z')$ but $D_i(z') \leq D_i(z'')$. The standard usage of the term monotonicity rules out this possibility by requiring that one of the following hold for all

[4] This point was made by Heckman and Robb (1985, 1986). See also Heckman (1997).

[5] Proof:

$$\text{plim}\hat{\beta}_{\text{IV}} = \bar{\beta} + \frac{\text{Cov}(Z,\varepsilon + \eta D)}{\text{Var}(D,Z)}.$$

But $\text{Cov}(Z,\varepsilon + \eta D) = \text{Cov}(Z,\varepsilon) + \text{Cov}(Z,\eta D)$ and $\text{Cov}(Z,\eta D) = E(Z\eta D) - E(Z)\,E(\eta D)$, $E(\eta D) = 0$, by the assumed independence. Also $E(Z\eta D) = E(\eta)\,E(ZD)$ by the assumed independence, and hence $E(Z\eta D) = 0$, because $E(\eta) = 0$.

[6] See Heckman and Robb (1985, 1986), Heckman (1997), or Heckman and Vytlacil (1999).

i: (a) $z < z'$ componentwise implies $D_i(z) \geq D_i(z')$ or (b) $z < z'$ componentwise implies $D_i(z) \leq D_i(z')$. Of course, if the $D_i(z)$ are monotonic in the standard usage, they are monotonic in the sense of Imbens and Angrist.

For any value of z' in the domain of definition of Z, from IV-1 and IV-2 and the definition of $D(z)$, $(Y_0, Y_1, D(z'))$ is independent of Z. For any two values of the instrument $Z = z$ and $Z = z'$, we may write

$$\begin{aligned} &E(Y|Z = z) - E(Y|Z = z') \\ &\quad = E(Y_0 + D(Y_1 - Y_0)|Z = z) \\ &\qquad - E(Y_0 + D(Y_1 - Y_0)|Z = z') \\ &\quad = E(D(Y_1 - Y_0)|Z = z) \\ &\qquad - E(D(Y_1 - Y_0)|Z = z'). \end{aligned}$$

From the independence condition (IV-1) and the definition of $D(z)$ and $D(z')$, we may write this expression as $E((Y_1 - Y_0)\,[D(z) - D(z')])$. Using the law of iterated expectations, we have

$$\begin{aligned} &E(Y|Z = z) - E(Y|Z = z') \\ &\quad = E(Y_1 - Y_0|D(z) - D(z') = 1) \\ &\qquad \times \Pr(D(z) - D(z') = 1) \\ &\qquad - E(Y_1 - Y_0|D(z) - D(z') = -1) \\ &\qquad \times \Pr(D(z) - D(z') = -1). \end{aligned} \tag{7}$$

By the monotonicity condition (IV-3), we eliminate one or the other term in this final expression. Suppose that $\Pr(D(z) - D(z') = -1) = 0$; then

$$\begin{aligned} &E(Y|Z = z) - E(Y|Z = z') \\ &\quad = E(Y_1 - Y_0|D(z) - D(z') = 1) \\ &\qquad \times \Pr(D(z) - D(z') = 1). \end{aligned}$$

Dividing by $\Pr(D(z) - D(z') = 1) = \Pr(D = 1 \mid Z = z) - \Pr(D = 1 \mid Z = z')$ for values of z and z' that produce distinct propensity scores, we obtain the local average treatment effect,

$$\begin{aligned} \text{LATE} &= \frac{E(Y|Z = z) - E(Y|Z = z')}{\Pr(D = 1|Z = z) - \Pr(D = 1|Z = z')} \\ &= E(Y_1 - Y_0|D(z) - D(z') = 1). \end{aligned} \tag{8}$$

This is the mean gain to those induced to switch from 0 to 1 by a change in Z from z' to z.

This is not the mean of $Y_1 - Y_0$ (average treatment effect) unless Z assumes values (z, z') such that $\Pr(D(z)) = 1$ and $\Pr(D(z')) = 0$.[7] It is also not the effect of treatment on the treated ($E(Y_1 - Y_0 \mid D = 1) = E(\beta|D = 1)$) unless the analyst has access to and uses one or more values of z such that $\Pr(D(z) = 1) = 1$.

LATE depends on the particular instrument used.[8] The parameter is defined by a hypothetical manipulation of instruments. If monotonicity (uniformity) is violated, IV estimates an average response of those induced to switch into the program and those induced to switch out of the program by the change in the instrument, because both terms in (7) are present.[9]

If the analyst is interested in knowing the average response ($\bar{\beta}$), the effect of the policy on the outcomes of countries that adopt it ($E(\beta \mid D = 1)$) or the effect of the policy if a particular country adopts it, there is no guarantee that the IV estimator comes any closer to the desired target than the OLS estimator, and indeed, it may be more biased than OLS. Because different instruments define different parameters, having a wealth of different strong instruments does not improve the precision of the estimate of any particular parameter. This is in stark contrast with the traditional model with $\beta \perp\!\!\!\perp D$. In that case, all valid instruments identify $\bar{\beta}$. The Durbin (1954)–Wu (1973)–Hausman (1978) test for the validity of extra instruments applies to the traditional model. In the more general case with essential heterogeneity, because different instruments estimate different parameters, no clear inference emerges from these specification tests.

When dealing with more than two distinct values of Z, Imbens and Angrist (1994) draw on the analysis of Yitzhaki (1989), which was refined in Yitzhaki (1996) and Yitzhaki and Schechtman (2004), to produce a weighted average of pairwise LATE parameters where the scalar Z's are ordered to define the LATE parameter. In this case IV is a weighted average of LATE parameters with nonnegative weights.[10] Imbens and Angrist generalize this result to the case of vector Z, assuming that instruments are monotonic functions of the probability of selection.

This paper and our previous analysis build on the pioneering work of Yitzhaki and of Imbens and Angrist.[11] We make the following contributions to this literature:

[7] Such values of Z produce "identification at infinity," or, more accurately, limit points where $P(z) = 1$ and $P(z') = 0$.

[8] Dependence of the estimands on the choices of IV used to estimate models with essential heterogeneity was first noted in Heckman and Robb (1985, 1986).

[9] Angrist, Imbens, and Rubin (1996) consider the case of two-way flows for the special case of a scalar instrument when the monotonicity assumption is violated. Their analysis is a version of Yitzhaki's (1989, 1996) analysis. He analyzes the net effect, whereas they break the net effect into two components corresponding to the two-way flows.

[10] We have placed Yitzhaki's (1989) unpublished paper on our Web site and summarize his essential ideas in section III B and appendix C. He shows that two-stage least squares estimators of Y on $P(Z) = E(D \mid Z)$ identify weighted averages of terms like the second terms in equation (8) with positive weights. See also Yitzhaki (1996) and Yitzhaki and Schechtman (2004).

[11] See Heckman and Vytlacil (1999, 2001c, 2005).

1. We relate the LATE-IV approach to economic choice models. Using a choice-theoretic parameter (the marginal treatment effect, or MTE) introduced into the literature on selection models by Björklund and Moffitt (1987), it is possible to generate all treatment effects as different weighted averages of MTE or of LATE. Standard linear IV can also be interpreted as a weighted average of MTE or LATE. Using the economic model, MTE is a limit form of LATE, and LATE in turn is a discrete approximation to MTE, which is the marginal gain function of Björklund and Moffitt (1987). A local version of instrumental variables (LIV), distinct from the standard IV approach, identifies the MTE. These theoretical constructs can be defined independently of the data.
2. We establish the central role of the propensity score ($\Pr(D = 1 \mid Z = z) = P(z)$) in both selection and IV models.[12]
3. We show that with vector Z, and a scalar instrument constructed from Z ($J(Z)$), the weights on the LATE and MTE that are implicit in standard IV are not guaranteed to be nonnegative. Thus IV can be negative even though all pairwise LATEs and pointwise MTEs are positive. Certain instruments produce positive weights and avoid this interpretive problem. Our analysis generalizes that of Yitzhaki and of Imbens and Angrist, who analyze the case with positive weights.
4. We show the special status of $P(z)$ as an instrument. It always produces nonnegative weights for the MTE and LATE. It enables analysts to identify the MTE or LATE. With knowledge of $P(z)$, and the MTE or LATE, we can decompose any standard IV estimate into identifiable MTEs (at points) or LATEs (over intervals) and identifiable weights on MTE or LATE, where the weights can be constructed from data. This ability to decompose IV into interpretable components allows analysts to determine the response to treatment of persons at different levels of the unobserved factors that determine treatment status.
5. We present a simple test for essential heterogeneity (β dependent on D) that allows analysts to determine whether or not they can avoid the complexities that arise in the more general model with heterogeneity in response to treatments.
6. We extend the analysis of IV to models with more than two outcomes. Angrist and Imbens (1995) analyze an ordered choice model with a scalar instrument that affects choices at all margins. They show that in an ordered model with multiple outcomes, IV identifies a "causal parameter." Their causal parameter is a weighted average of parameters that are difficult to interpret as willingness-to-pay parameters or answers to well-defined choice problems. We present an economically interpretable decomposition of standard IV into willingness-to-pay components for persons at well-defined margins of choice. We show how to identify these components from data and how to construct the weights. We introduce transition-specific instruments. We generalize this analysis to an unordered choice model.
7. We show the fundamental asymmetry in the recent IV literature for models with heterogeneous outcomes. Responses to treatment are permitted to be heterogeneous in a general way. Responses of choices to instruments are not. When heterogeneity in choice is allowed for in a general way, IV and local IV do not estimate interpretable treatment parameters.

The paper is organized as follows. Section II discusses the IV approach to estimating choice models. Section III introduces a general model with essential heterogeneity and presents its implications. Section IV compares selection and IV models and shows that LIV estimates the derivative of a selection model. Section V presents theoretical and empirical examples of the model with essential heterogeneity. Section VI extends the analysis to multiple outcome models. Section VII allows choice responses to be heterogeneous in a general way. Section VIII concludes.

II. IV in Choice Models

We adjoin a choice equation to outcome equations (1) and (2). A standard binary threshold crossing model for D is

$$D = \mathbf{1}[D^* > 0], \tag{9}$$

where $\mathbf{1}[\cdot]$ is an indicator ($\mathbf{1}[A] = 1$ if A is true; 0 otherwise).

A familiar case is

$$D^* = \gamma Z - V, \tag{10}$$

where $(V \perp\!\!\!\perp Z) \mid X$ (V is independent of Z given X). The propensity score, or choice probability, is

$$P(z) = \Pr(D = 1 | Z = z) = \Pr(\gamma z > V) = F_V(\gamma z),$$

where F_V is the distribution of V, which is assumed to be continuous. In terms of the generalized Roy model where C is the cost of participation in sector 1, $D = \mathbf{1}[Y_1 - Y_0 - C > 0]$. For a separable model with outcomes (1) and costs $C = \mu_C(W) + U_C$, we have $Z = (X, W)$, $\mu_D(Z) = \mu_1(X) - \mu_0(X) - \mu_C(W)$, $V = -(U_1 - U_0 - U_C)$. In constructing examples, we use a special version where $U_C = 0$. We call this version the *extended Roy model.*[13] Our analysis, however, applies to more general models.

[12] Rosenbaum and Rubin (1983) establish the central role of the propensity score in matching models. Heckman and Robb (1985, 1986) and Heckman (1980) establish the central role of the propensity score in selection models. See also Ahn and Powell (1993) and Powell (1994).

[13] The generalized Roy model allows $U_C \neq 0$.

In the case where β (given X) is a constant under IV-1 and IV-2, it is not necessary to specify the choice model to identify β. We show that in a general model with heterogeneous responses, the specification of $P(z)$ and its relationship with the instrument play crucial roles. To see this, study the covariance between Z and ηD discussed in the introduction. By the law of iterated expectations, letting $\bar{Z}$ denote the mean of Z, we have

$$\begin{aligned}\mathrm{Cov}(Z, \eta D) &= E((Z - \bar{Z})D\eta) \\ &= E((Z - \bar{Z})\eta | D = 1)\ \Pr(D = 1) \\ &= E((Z - \bar{Z})\eta | \gamma Z > V)\ \Pr(\gamma Z > V).\end{aligned}$$

Thus, even if Z and η are independent, they are not independent conditional on $D = \mathbf{1}[\gamma Z > V]$ if $\eta\ (= U_1 - U_0)$ is dependent on V (that is, if the decision-maker has partial knowledge of η and acts on it). Selection models allow for this dependence (see Heckman and Robb, 1985, 1986; Ahn and Powell, 1993; Powell, 1994). Keeping X implicit and assuming that

$$(U_1, U_0, V) \perp\!\!\!\perp Z \tag{11}$$

(alternatively, that $(\varepsilon, \eta) \perp\!\!\!\perp Z$), we obtain $E(Y \mid D = 0, Z = z) = E(Y_0 \mid D = 0, Z = z) = \alpha + E(U_0 \mid \gamma z < V)$, which can be written as

$$E(Y | D = 0, Z = z) = \alpha + K_0(P(z)),$$

where the functional form of K_0 is produced from the distribution of (U_0, V). (This representation is derived in Heckman, 1980; Heckman and Robb, 1985, 1986; Ahn and Powell, 1993; and Powell, 1994.)

Similarly,

$$\begin{aligned}E(Y | D = 1, Z = z) &= E(Y_1 | D = 1, Z = z) \\ &= \alpha + \bar{\beta} + E(U_1 | \gamma z > V) \\ &= \alpha + \bar{\beta} + K_1(P(z)),\end{aligned}$$

where $K_0(P(z))$ and $K_1(P(z))$ are control functions in the sense of Heckman and Robb (1985, 1986). Under standard conditions, we can identify $\bar{\beta}$. Powell (1994) discusses semiparametric identification. Because we condition on $Z = z$ (or $P(z)$), correct specification of Z plays an important role in econometric selection methods. This sensitivity to the full set of instruments in Z appears to be absent from the IV method.

If β is a constant (given X), or if $\eta\ (= \beta - \bar{\beta})$ is independent of V, only one instrument from the vector Z needs to be used. Missing instruments play no role in identifying the mean responses but may affect the efficiency of the IV estimation. We establish that in a model where β is variable and not independent of V, misspecification of Z plays an important role in interpreting what IV estimates, analogous to its role in selection models. Misspecification of Z affects both approaches to identification. This is a new phenomenon in models with heterogeneous β. We now review some results established in the preceding literature that form the platform on which we build.

III. A General Model with Essential Heterogeneity in Outcomes

We now exposit the selection model developed in Heckman and Vytlacil (1999, 2001b, 2005). Their model for counterfactuals (potential outcomes) is more general than equation (1) and allows for nonseparable errors:

$$\begin{aligned}Y_1 &= \mu_1(X, U_1), \\ Y_0 &= \mu_0(X, U_0),\end{aligned} \tag{12}$$

where X are observed and (U_1, U_0) are unobserved by the analyst. The X may be dependent on U_0 and U_1 in a general way. This model is designed to evaluate policies in place and not to extrapolate to new environments characterized by X.[14] The observed outcome is produced by equation (2).

Choices are generated by a standard discrete choice model. We generalize the choice model (9) and (10) for D^*, a latent utility,[15]

$$D^* = \mu_D(Z) - V \quad \text{and} \quad D = \mathbf{1}[D^* \geq 0]. \tag{13}$$

$\mu_D(Z) - V$ can be interpreted as a net utility for a person with characteristics (Z, V). If it is positive, $D = 1$ and the person selects into treatment; otherwise $D = 0$. Section VII discusses the important role played by additive separability in the recent IV literature on essential heterogeneity.

In terms of the notation used in section I, $\beta = Y_1 - Y_0 = \mu_1(X, U_1) - \mu_0(X, U_0)$. A special case that links our analysis to standard models in econometrics is when $Y_1 = X\beta_1 + U_1$ and $Y_0 = X\beta_0 + U_0$, so $\beta = X(\beta_1 - \beta_0) + (U_1 - U_0)$. In the case of separable outcomes, heterogeneity in β arises because in general $U_1 \neq U_0$ and people differ in their X.[16]

Following Heckman and Vytlacil (2005), we assume:

A-1: (U_0, U_1, V) are independent of Z conditional on X (*Independence Condition for IV*).

A-2: The distribution of $\mu_D(Z)$ conditional on X is nondegenerate (*Rank Condition for IV*).[17]

A-3: The distribution of V is continuous.[18]

A-4: $E|Y_1| < \infty$ and $E|Y_0| < \infty$ (*Finite Means*).

[14] See Heckman and Vytlacil (2005, 2007b) for a study of exogeneity requirements on X in answering different policy questions.

[15] A large class of latent index, threshold crossing models will have this representation. See Vytlacil (2006a).

[16] In nonseparable cases, heterogeneity arises conditional on X even if $U_1 = U_0 = U$.

[17] $\mu_D(\cdot)$ is assumed to be a measurable function of Z given X.

[18] The distribution is absolutely continuous with respect to Lebesgue measure.

A-5: $1 > \Pr(D = 1 \mid X) > 0$ (*For each X, a treatment group and a comparison group exist*).

A-6: Let X_0 denote the counterfactual value of X that would have been observed if D were set to 0. Define X_1 analogously. Thus $X_d = X$ for $d = 0, 1$ (*The X_d are invariant to counterfactual manipulations*).

A-1 and A-2 generalize IV-1 and IV-2 respectively. A-3 is a technical condition made for convenience and is easily relaxed at some notational cost. A-4 is needed to use standard integration theorems and to have the mean treatment effect parameters well defined. A-5 is a standard requirement for any evaluation estimator: that for each value of X, there be some who are treated and some who are not. A-6 is the requirement that receipt of treatment not affect the realized value X, so we identify a full treatment effect when we condition on X instead of a treatment effect that conditions on variables affected by treatment. This assumption can be relaxed by redefining the treatment to be a set of outcomes corresponding to each state X_d.

The separability between V and $\mu_D(Z)$ in the choice equation is conventional. It plays a crucial role in justifying instrumental variable estimators in models with essential heterogeneity. It implies the monotonicity (uniformity) condition IV-3 from the choice equation (13). Fixing Z at two different values moves $D(Z)$ in the same direction for everyone. Vytlacil (2002) shows that under independence, the rank condition, and some regularity conditions, monotonicity IV-3 implies the existence of a V in the representation (13). Thus the IV model for the general case and the economic choice model turn out to have *identical* representations. The independence assumption A-1 produces the condition that everywhere Z enters the model only through $P(Z)$. This is called *index sufficiency.*

Without any loss of generality, following the same argument surrounding equations (9) and (10), we may write the model for D using the distribution of V, F_V, as

$$D = \mathbf{1}[F_V(\mu_D(Z)) > F_V(V)] = \mathbf{1}[P(Z) > U_D], \tag{14}$$

where $U_D = F_V(V)$ and $P(Z) = F_V(\mu_D(Z)) = \Pr(D = 1 \mid Z)$, the propensity score. Because F_V is assumed to be a continuous distribution, F_V is a strictly monotonic transformation that preserves the information in the original inequality. Note that U_D is uniformly distributed by construction ($U_D \sim Unif[0,1]$).

A. *The Local Average Treatment Effect, the Marginal Treatment Effect, and Instrumental Variables*

To understand what IV estimates in the model with general heterogeneity in response to treatment, we define the MTE conditional on X and U_D:[19]

$$\begin{aligned}\Delta^{\text{MTE}}(x,u_D) &= E(Y_1 - Y_0 | X = x, U_D = u_D)\\ &= E(\beta | X = x, V = v),\end{aligned}$$

for $\beta = Y_1 - Y_0$ and $v = F_V^{-1}(u_D)$, where we use both general notation and the regression-specific notation interchangeably to anchor our analysis both in the treatment effect literature and in conventional econometrics. To simplify the notation, we keep the conditioning on X implicit except when clarity of exposition dictates otherwise. Because $P(Z)$ is a monotonic transformation of the mean net utility $\mu_D(Z)$, and U_D is a monotonic function of V, when we evaluate $\Delta^{\text{MTE}}(u_D)$ at the value $P(z) = u_D$, it is the marginal return to agents with characteristics $Z = z$ who are just indifferent between sector 1 and sector 0. In other words, at this point of evaluation, $\Delta^{\text{MTE}}(u_D)$ is the gross gain of going from 0 to 1 for agents who are indifferent between the sectors when their mean utility given $Z = z$ is $\mu_D(z) = v$, so $\mu_D(z) - v = 0$, which is equivalent to the event that $P(z) = F_V(\mu_D(z)) = F_V(v) = u_D$. When Y_1 and Y_0 are denominated in value units, the MTE is a willingness-to-pay measure for persons with characteristics $Z = z$ at the specified margin.

Under assumptions A-1 to A-5, Heckman and Vytlacil (1999, 2005) show that all treatment parameters, matching estimators, IV estimators based on a scalar function $J(Z)$ of Z, and OLS estimators can be written as weighted averages of the MTE. Table 1 summarizes their results for characterizing treatment effects and estimators and the weights given data on $P(Z)$, D, and the instrument $J(Z)$. We discuss the weights for IV in the next subsection. We show how to construct these weights on our Web site, where software for doing so is available.[20] Heckman and Vytlacil (2001b, 2007b) show that these weights can be constructed and the relationships among the parameters shown in table 1 hold even if a nonseparable choice model, instead of equation (13), is used and even if assumption A-2 is weakened. We discuss this result in section VII.

Notice that when Δ^{MTE} does not depend on u_D, all of the treatment effects are the same, and that, under our assumptions, IV estimates all of them. In this case, Δ^{MTE} can be taken outside the integral and the weights all integrate to 1. Thus, $E(Y_1 - Y_0 \mid X = x) = \text{ATE}(x) = E(Y_1 - Y_0 \mid X = x, D = 1) = \text{TT}(x) = \text{MTE}(x)$, and we are back to the conventional model of homogeneous responses. This includes the case where η is nondegenerate but independent of D.

The parameters MTE and LATE are closely related. Using the definition of $D(z)$ in IV-3, 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 $P(z) > P(z')$, under IV-3 and independence (A-1), the LATE is

$$\Delta^{\text{LATE}}(z',z) = E(Y_1 - Y_0 | D(z) = 1, D(z') = 0), \tag{15a}$$

that is, the mean outcome in terms of $Y_1 - Y_0$ for persons who would be induced to switch from $D = 0$ to $D = 1$ if Z

[19] As previously noted, the concept of the marginal treatment effect and the limit form of LATE were first introduced in the literature in the context of a parametric normal generalized Roy selection model by Björklund and Moffitt (1987).

[20] See jenni.uchicago.edu/underiv/.

TABLE 1A.—TREATMENT EFFECTS AND ESTIMANDS AS WEIGHTED AVERAGES OF THE MARGINAL TREATMENT EFFECT

$$\text{ATE}(x) = E(Y_1 - Y_0 \mid X = x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, du_D$$

$$\text{TT}(x) = E(Y_1 - Y_0 \mid X = x, D = 1) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{TT}}(x, u_D)\, du_D$$

$$\text{TUT}(x) = E(Y_1 - Y_0 \mid X = x, D = 0) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{TUT}}(x, u_D)\, du_D$$

$$\text{Policy Relevant Treatment Effect } (x) = E(Y_{a'} \mid X = x) - E(Y_a \mid X = x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{PRTE}}(x, u_D)\, du_D$$

for two policies a and a' that affect the Z but not the X

$$\text{IV}_J(x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{IV}}^J(x, u_D)\, du_D, \text{ given instrument } J$$

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

TABLE 1B.—WEIGHTS

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

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

$$\omega_{\text{TUT}}(x, u_D) = \left[\int_0^{u_D} f(p \mid X = x)\, dp\right] \frac{1}{E((1 - P) \mid X = x)}$$

$$\omega_{\text{PRTE}}(x, u_D) = \left[\frac{F_{P_{a'},X}(u_D) - F_{P_a,X}(u_D)}{\Delta P}\right]$$

$$\omega_{\text{IV}}^J(x, u_D) = \left[\int_{u_D}^1 (J(Z) - E(J(Z) \mid X = x)) \int f_{J,P|X}(j, t \mid X = x)\, dt\, dj\right] \frac{1}{\text{Cov}(J(Z), D \mid X = x)}$$

$$\omega_{\text{OLS}}(x, u_D) = 1 + \frac{E(U_1 \mid X = x, U_D = u_D)\omega_1(x, u_D) - E(U_0 \mid X = x, U_D = u_D)\omega_0(x, u_D)}{\Delta^{\text{MTE}}(x, u_D)}$$

$$\omega_1(x, u_D) = \left[\int_{u_D}^1 f(p \mid X = x)\, dp\right] \left[\frac{1}{E(P \mid X = x)}\right]$$

$$\omega_0(x, u_D) = \left[\int_0^{u_D} f(p \mid X = x)\, dp\right] \frac{1}{E((1 - P) \mid X = x)}$$

Source: Heckman and Vytlacil (2005)

were manipulated externally from z' to z. As a consequence of Vytlacil's (2002) theorem, the LATE can be written as

$$\begin{aligned} &E(Y_1 - Y_0 | D(z) = 1, D(z') = 0) \\ &\quad = E(Y_1 - Y_0 | u_D' < U_D < u_D) \qquad (15b) \\ &\quad = \Delta^{\text{LATE}}(u_D, u_D'), \end{aligned}$$

where $u_D = \Pr(D(z) = 1) = \Pr(D = 1 \mid Z = z) = P(z)$, $u_D' = \Pr(D(z') = 1 \mid Z = z') = \Pr(D(z') = 1) = P(z')$.[21] In the limit, as $u_D' \to u_D$, LATE converges to MTE.

Imbens and Angrist (1994) define the LATE parameter from hypothetical manipulations of an instrument. Heck-

[21] Assumption A-1 implies that $\Pr(D(z) = 1) = \Pr(D = 1 \mid Z = z)$ and $\Pr(D(z') = 1) = \Pr(D = 1 \mid Z = z')$.

man and Vytlacil (1999, 2005) draw on choice theory and define the parameters in terms of the generalized Roy model. Their link helps to understand what IV estimates and relates IV to choice models. We work with the definition (15b) throughout the rest of this paper. It enables us to identify the margin of U_D selected by instruments, something currently not possible in results in the previous literature on IV.

The MTE can be identified by taking derivatives of $E(Y \mid Z = z)$ with respect to $P(z)$ (see Heckman and Vytlacil, 1999).[22] This derivative is called the local instrumental variable (LIV). For the model of general heterogeneity, under assumptions A-1 to A-5), we can write (keeping the conditioning on $X = x$ implicit)

$$
\begin{aligned}
&E(Y|Z = z) = E(Y|P(Z) = p),\\
&E(Y|P(Z) = p) \quad = E(DY_1 + (1 - D)Y_0|P(Z) = p)\\
&\quad = E(Y_0) + E(D(Y_1 - Y_0)|P(Z) = p)\\
&\quad = E(Y_0) + E(Y_1 - Y_0|D = 1)p\\
&\quad = E(Y_0) + \int_0^p E(Y_1 - Y_0|U_D = u_D)\, du_D.
\end{aligned}
$$

As a consequence,

$$
\frac{\partial}{\partial p} E(Y|P(Z) = p)\bigg|_{P(z)=p} = E(Y_1 - Y_0|U_D = p). \tag{16}
$$

This expression shows how the derivative of $E(Y \mid Z = z)$, which is the LIV estimand of Heckman and Vytlacil (1999), identifies the marginal treatment effect (the right-hand side of this expression) over the support of $P(Z)$. Observe that a high value of $P(Z) = p$ identifies the MTE at a value of $U_D = u_D$ that is *high*—that is associated with nonparticipation. It takes a high p to compensate for the high $U_D = u_D$ and bring the agent to indifference (see equation (14)). Thus high p values identify returns to persons whose unobservables make them *less* likely to participate in the program. Software for estimating the MTE using local linear regression is described in appendix B and is available online at jenni.uchicago.edu/underiv.

Under the special case where $\beta \perp\!\!\!\perp D$ (no essential heterogeneity), Y is linear in $P(Z)$:

$$
E(Y|Z) = a + bP(Z), \tag{17}
$$

where $b = \Delta^{\text{MTE}} = \Delta^{\text{ATE}} = \Delta^{\text{TT}}$. This representation holds whether or not Y_1 and Y_0 are separable in U_1 and U_0, respectively (see Heckman and Vytlacil, 2001b, 2007b). Thus, a test of the linearity of the conditional expectation of Y in terms of $P(Z)$ is a test of whether the conventional model or the model of essential heterogeneity generates the data. One useful empirical strategy is to test for linearity using the variety of tests developed in the literature and to determine whether the additional complexity introduced by the model of essential heterogeneity is warranted.

Using the formulae presented in table 1, all of the traditional treatment parameters as well as the IV estimator using $P(Z)$ as an instrument can be identified as weighted averages of $\Delta^{\text{MTE}}(u_D)$ if $P(Z)$ has full support. The weights can be constructed from data. If $P(Z)$ does not have full support, simple tight bounds on these parameters can be constructed.[23]

B. *Understanding What* IV *Estimates*

Standard IV based on $J(Z)$, a scalar function of a vector Z, can be written as

$$
\Delta_J^{\text{IV}} = \int_0^1 \Delta^{\text{MTE}}(u_D)\omega_{\text{IV}}^J(u_D)\, du_D, \tag{18}
$$

where

$$
\omega_{\text{IV}}^J(u_D) = \frac{E(J(Z) - E(J(Z))|P(Z) > u_D)\Pr(P(Z) > u_D)}{\text{Cov}(J(Z), D)}. \tag{19}
$$

In this expression u_D is a number between 0 and 1. This weight depends on the choice probability $P(Z)$. For a derivation see appendix A. The derivation does not impose any assumptions on the distribution of $J(Z)$ or $P(Z)$. Notice that $J(Z)$ and $P(Z)$ do not have to be continuous random variables, and that the functional forms of $P(Z)$ and $J(Z)$ are general.[24]

For ease of exposition, we initially assume that $J(Z)$ and $P(Z)$ are both continuous. This assumption plays no essential role in any of the results of this paper, and we develop the discrete case after developing the continuous case. The weights defined in equation (19) can be written as

$$
\omega_{\text{IV}}^J(u_D) = \frac{\int [j - E(J(Z))] \int_{u_D}^1 f_{J,P}(j,t)\, dt\, dj}{\text{Cov}(J(Z),D)}, \tag{20}
$$

where $f_{J,P}$ is the joint density of $J(Z)$ and $P(Z)$, and we implicitly condition on X. The weights can be negative or positive. Observe that $\omega(0) = 0$ and $\omega(1) = 0$. The weights integrate to 1,[25] so even if the weight is negative over some intervals, it must be positive over other intervals. When there is one instrument (Z is a scalar), and assumptions A-1 to A-5 are satisfied, the weights are always positive provided

[22] See also Heckman and Vytlacil (2005, 2007b).

[23] See Heckman and Vytlacil (1999, 2001a,b, 2007b).

[24] More precisely, $J(Z)$ and $P(Z)$ do not have to have distributions that are absolutely continuous with respect to Lebesgue measure.

[25] $\int\int (j - E(J(Z))) \int_{u_D}^1 f_{J,P}(j, t)\, dt\, dj\, du_D = \text{Cov}(J(Z),D)$.

that $J(Z)$ is a monotonic function of the scalar Z. In this case $J(Z)$ and $P(Z)$ have the same distribution and $f_{J,P}$ (j, t) collapses to a univariate distribution. The possibility of negative weights arises when $J(Z)$ is not a monotonic function of $P(Z)$. It can also arise when there are two or more instruments, and the analyst computes estimates with only one instrument or a combination of the Z-instruments that is not a monotonic function of $P(Z)$, so that $J(Z)$ and $P(Z)$ are not perfectly dependent. If the instrument is $P(Z)$ (so $J(Z) = P(Z)$), then the weights are everywhere nonnegative, because from equation (19) $E(P(Z) \mid P(Z) > u_D) - E(P(Z)) \geq 0$. In this case the density of $(P(Z), J(Z))$ collapses to the density of $P(Z)$. For any scalar Z we can define $J(Z)$ and $P(Z)$ so that they are perfectly dependent, provided $J(Z)$ and $P(Z)$ are monotonic in Z. More generally, the weight (19) is positive if $E(J(Z) \mid P(Z) > u_D)$ is weakly monotonic in u_D. Nonmonotonicity of this conditional expectation can produce negative weights.[26]

Observe that the weights can be constructed from data on (J, P, D). Data on $(J(Z), P(Z))$ pairs and $(J(Z), D)$ pairs (for each X value) are all that is required. We can use a smoothed sample frequency to estimate the joint density $f_{J,P}$. Thus, given our maintained assumptions, any property of the weight, including its positivity at any point (x,u_D), can be examined with data. We present examples of this approach in section V.

As is evident from table 1, the weights on $\Delta^{\text{MTE}}(u_D)$ generating Δ^{IV} are different from the weights on $\Delta^{\text{MTE}}(u_D)$ that generate the average treatment effect, which is widely regarded as an important policy parameter (see, for example, Imbens, 2004) or from the weights associated with the policy-relevant treatment parameter, which answers well-posed policy questions (Heckman and Vytlacil, 1999, 2001b, 2005, 2007b). It is not obvious why the weighted average of $\Delta^{\text{MTE}}(u_D)$ produced by IV is of any economic interest. Because the weights can be negative for some values of u_D, $\Delta^{\text{MTE}}(u_D)$ can be positive everywhere in u_D, but IV can be negative. Thus, IV may not estimate a treatment effect for any person. Therefore, a basic question is why estimate the model with IV at all, given the lack of any clear economic interpretation of the IV estimator in the general case.

Our analysis can be extended to allow for discrete instruments, $J(Z)$. Consider the case where the distribution of $P(Z)$ (conditional on X) is discrete. The support of the distribution of $P(Z)$ contains a finite number of values $p_1 < p_2 < \cdots < p_K$, and the support of the instrument $J(Z)$ is also discrete, taking I distinct values, where I and K may be distinct. $E(J(Z)|P(Z) \geq u_D)$ is constant in u_D for u_D within any interval $(p_\ell, p_{\ell+1})$ and $\Pr(P(Z) \geq u_D)$ is constant in u_D for u_D within any interval $(p_\ell, p_{\ell+1})$, and thus $\omega^J_{\text{IV}}(u_D)$ is constant in u_D over any interval $(p_\ell, p_{\ell+1})$. Let λ_ℓ denote the weight on the LATE for the interval $(\ell, \ell + 1)$. In this notation,

$$\begin{aligned}\Delta^{\text{IV}}_J &= \int E(Y_1 - Y_0|U_D = u_D)\omega^J_{\text{IV}}(u_D)\, du_D \\ &= \sum_{\ell=1}^{K-1} \lambda_\ell \int_{p_\ell}^{p_{\ell+1}} E(Y_1 - Y_0|U_D = u_D) \frac{1}{p_{\ell+1} - p_\ell}\, du_D \qquad (21) \\ &= \sum_{\ell=1}^{K-1} \Delta^{\text{LATE}}(p_\ell, p_{\ell+1})\lambda_\ell .\end{aligned}$$

Let j_i be the ith smallest value of the support of $J(Z)$. The discrete version of equation (19) is

$$\lambda_\ell = \frac{\sum_{i=1}^{I} [j_i - E(J)] \sum_{t>\ell}^{K} f(j_i, p_t)}{\text{Cov}(J(Z), D)} (p_{\ell+1} - p_\ell), \qquad (22)$$

where $f(j_i,p_t)$ is the probability frequency of (j_i,p_t): the probability that $J(Z) = j_i$ and $P(Z) = p_t$. We do not presume that high values of $J(Z)$ are associated with high values of $P(Z)$. $J(Z)$ can be one coordinate of Z that may be positively or negatively dependent on $P(Z)$, which depends on the full vector. In the case of scalar Z, as long as $J(Z)$ and $P(Z)$ are monotonic in Z, there is perfect dependence between $J(Z)$ and $P(Z)$. In this case, the joint probability density collapses to a univariate density and the weights have to be positive, exactly as in the case with continuous instruments.[27] Our expression for the weight on the LATE generalizes the expression presented by Imbens and Angrist (1994), who in their analysis of the case of vector Z only consider the case where $J(Z)$ and $P(Z)$ are perfectly dependent because $J(Z)$ is a monotonic function of $P(Z)$.[28] More generally the weights can be positive or negative for any ℓ, but they must sum to 1 over the ℓ.

Monotonicity or uniformity is a property needed with just two values of Z, $Z = z_1$ and $Z = z_2$, to guarantee that IV estimates a treatment effect. With more than two values of Z we need to weight the LATEs and MTEs. If the instrument $J(Z)$ shifts $P(Z)$ in the same way for everyone, it shifts D in the same way for everyone since $D = \mathbf{1}[P(Z) > U_D]$ and Z is independent of U_D. If $J(Z)$ is not monotonic in $P(Z)$, it may shift $P(Z)$ in different ways for different people. Negative weights are a tip-off of two-way flows.

An alternative and in some ways more illuminating way to derive the weights is to follow Yitzhaki (1989, 1996) and Yitzhaki and Schechtman (2004), who prove for a general

[26] If it is weakly monotonically increasing, the claim is evident from equation (19). If it is decreasing, the signs of the numerator and the denominator are both negative, so the weight is nonnegative.

[27] The condition for positive weights is weak monotonicity of λ_ℓ in ℓ. If λ_ℓ is monotone increasing in ℓ, the numerator and the denominator are both positive. If λ_ℓ is monotone decreasing, the numerator and the denominator are both negative and the weights are positive.

[28] In their case $I = K$ and $f(j_i, p_t) = 0\ \forall i \neq t$.

regression function $E(Y \mid P(Z) = p)$ that a linear regression of Y on P estimates

$$\beta_{Y,P} = \int_0^1 \left[\frac{\partial E(Y|P(Z) = p)}{\partial p} \right] \omega(p)\, dp, \tag{23}$$

where

$$\omega(p) = \frac{\int_p^1 [t - E(P)]\, dF_P(t)}{\mathrm{Var}(P)},$$

which is exactly the weight (19) when P is the instrument. Thus we can interpret equation (19) as the weight on $\partial E(Y|P(Z) = p)/\partial p$ when two-stage least squares (TSLS) based on $P(Z)$ as the instrument is used to estimate the "causal effect" of D on Y. Under uniformity, $\partial E(Y|P(Z) = p)/\partial p|_{p=u_D} = E(Y_1 - Y_0 \mid U_D = u_D) = \Delta^{\mathrm{MTE}}(u_D)$.[29] We discuss Yitzhaki's derivation, which is an argument based on integration by parts, in appendix C. Our analysis is more general than that of Yitzhaki (1989), Imbens and Angrist (1994), or Angrist and Imbens (1995) in that we allow for instruments that are not monotonic functions of $P(Z)$. Yitzhaki's (1989) analysis is more general than that of Imbens and Angrist (1994) in that he does not impose uniformity (monotonicity).

Our simple test for the absence of general heterogeneity, based on the linearity of Y in $P(Z)$ (based on equation (20)), applies to the case of LATE for any pair of instruments. An equivalent test is to check that all pairwise LATEs are the same over the sample support of Z.[30]

C. The Central Role of the Propensity Score

Observe that both equations (19) and (20) (and their counterparts for LATE, equations (21) and (22)) contain expressions involving the propensity score $P(Z)$, the probability of selection into treatment. Under our assumptions, it is a monotonic function of the mean utility of treatment, $\mu_D(Z)$. The propensity score plays a central role in selection models as a determinant of control functions in selection models (see Heckman and Robb, 1985, 1986), as noted in section II. In matching models, it provides a computationally convenient way to condition on Z (see, for example, Rosenbaum and Rubin, 1983; Heckman and Navarro, 2004). For the IV weight to be correctly constructed and interpreted, we need to know the correct model for $P(Z)$, that is, we need to know exactly which Z determine $P(Z)$. As previously noted, this feature is not required in the traditional model for instrumental variables based on response homogeneity. In that simpler framework, any instrument will identify $\mu_1(X) - \mu_0(X)$, and the choice of a particular instrument affects efficiency but not identifiability. One can be casual about the choice model in the traditional setup, but not in the model of choice of treatment with essential heterogeneity. Thus, unlike the application of IV to traditional models, IV applied in the model of essential heterogeneity depends on (a) the choice of the instrument $J(Z)$, (b) its dependence on $P(Z)$, the true propensity score or choice probability, and (c) the specification of the propensity score (that is, what variables go into Z). Using the propensity score, one can identify LIV and LATE and the marginal returns at values of the unobserved U_D.

D. Monotonicity, Uniformity, and Conditional Instruments

The monotonicity or uniformity condition IV-3 is a condition on counterfactuals for the same persons and is not testable. It rules out general heterogeneous responses to treatment choices in response to changes in Z. The recent literature on instrumental variables with heterogeneous responses is thus asymmetric. Outcome equations can be heterogeneous in a general way, whereas choice equations cannot be. If $\mu_D(Z) = \gamma Z$, where γ is a common coefficient shared by everyone, then the choice model satisfies the uniformity property. On the other hand, if γ is a random coefficient (that is, has a nondegenerate distribution) that can take both negative and positive values, and there are two or more variables in Z with nondegenerate coefficients γ, then uniformity can be violated. Different people can respond to changes in Z differently, so we have nonuniformity. The uniformity condition can be violated even when all components of γ are of the same sign if Z is a vector and γ is a nondegenerate random variable.[31]

Changing one coordinate of Z, holding the other coordinates at different values across people, is *not* the experiment that defines monotonicity or uniformity. Changing one component of Z, allowing the other coordinates to vary across people, does not necessarily produce uniform flows toward or against participation in the treatment status. For example, let $\mu_D(z) = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_1 z_2$, where γ_0, γ_1, γ_2, and γ_3 are constants, and consider changing z_1 from a common base state while holding z_2 fixed at different values across people. If $\gamma_3 < 0$, then $\mu_D(z)$ does not necessarily satisfy the uniformity condition. If we move (z_1, z_2) as a pair from the same base values to the same destination values z', uniformity is satisfied even if $\gamma_3 < 0$, although $\mu_D(z)$ is not a monotonic function of z.[32]

[29] Yitzhaki's weights are used by Angrist and Imbens (1995) to interpret what TSLS estimates in the model of equation (23). Yitzhaki (1989) derives the finite-sample weights used by Imbens and Angrist (see his paper posted at our Web site). See also the refinement in Yitzhaki and Schechtman (2004).

[30] Note that it is possible that $E(Y|Z)$ is linear in $P(Z)$ only over certain intervals of U_D, so there can be local dependence and local independence of (U_D, U_0, U_1).

[31] Thus if $\gamma > 0$ for each component and some components of Z are positive and others are negative, changes from z' to z can increase γZ for some and decrease γZ for others, because γ differs among persons.

[32] Associated with $Z = z$ is the counterfactual random variable $D(z)$. Associated with the scalar random variable $J(Z)$ constructed from Z is a counterfactual random variable $D(j(z))$, which is in general different from $D(z)$. The random variable $D(z)$ is constructed from equation (13) using $\mathbf{1}[\mu_D(z) \geq V]$. Here V assumes individual-specific values, which remain fixed as we set different z values. From A-1, $\Pr(D(z) = 1) = \Pr(D = 1|Z = z)$.

Positive weights and uniformity are distinct issues.[33] Under uniformity and assumptions A-1 to A-5, the weights on MTE for any particular instrument may be positive or negative. The weights for MTE using $P(Z)$ must be positive, as we have shown, so the propensity score has a special status as an instrument. Negative weights associated with the use of $J(Z)$ as an instrument do not necessarily imply failure of uniformity in Z. Even if uniformity is satisfied for Z, it is not necessarily satisfied for $J(Z)$. Condition IV-3 is an assumption about a vector. Fixing one combination of Z (when J is a function of Z) or one coordinate of Z does not guarantee uniformity in J even if one has uniformity in Z. The flow created by changing one coordinate of Z can be reversed by the flow created by other components of Z if there is negative dependence among components, even if ceteris paribus all components of Z affect D in the same direction. We present some examples in section V.

The issues of positive weights and the existence of one-way flows in response to an intervention are conceptually distinct. Even with two values for a scalar Z, flows may be two-way (see equation (7)). If we satisfy IV-3 for a vector, so uniformity applies, then weights for a particular instrument may be negative for certain intervals of U_D (that is, for some of the LATE parameters).

If we condition on $Z_2 = z_2, \ldots, Z_K = z_K$ using Z_1 as an instrument, then a uniform flow condition is satisfied. We call this *conditional uniformity*. By conditioning, we effectively convert the problem back to that of a scalar instrument where the weights must be positive. If uniformity holds for Z_1, fixing the other Z at common values, then one-dimensional LATE/MTE analysis applies. Clearly, the weights also have to be defined conditionally.

The concept of conditioning on other instruments to produce positive weights for the selected instrument is a new one, not yet appreciated in the empirical IV literature, and has no counterpart in the traditional IV model. In the conventional model, the choice of a valid instrument affects efficiency, but not the definition of the parameters as it does in the more general case.[34]

In summary, nothing in the economics of choice models guarantees that if Z is changed from z to z', people will respond in the same direction to the change. See the general expression (7). The condition that people respond to choices in the same direction for a common change in Z across people does not imply that $D(z)$ is monotonic in z for any person in the usual mathematical usage of the term monotonicity. If $D(z)$ is monotonic in the usual usage of this term, and responses are in the same direction for all people, then the "monotonicity" or "uniformity" condition IV-3 will be satisfied.

If responses to a common change of Z across persons are heterogeneous in a general way, we obtain equation (7) as the general case. Vytlacil's (2002) theorem breaks down, and IV cannot be expressed in terms of a weighted average of LATE terms. Nonetheless, Yitzhaki's characterization of the IV equation (23) as described in appendix C remains valid, and the weights on $\partial E(Y|P = p)/\partial p$ are positive and of the same form as the weights obtained for the MTE (or LATE) when the monotonicity condition holds.

E. *Treatment Effects versus Policy Effects*

Even if the uniformity condition IV-3 fails, IV may answer relevant policy questions. By Yitzhaki's result (23), IV or TSLS estimates a weighted average of marginal responses, which may be pointwise positive, zero, or negative. Policies may induce some people to switch into and others to switch out of choices, as is evident from equation (7). These net effects are of interest in many policy analyses. Thus, subsidized housing in a region supported by higher taxes may attract some to migrate to the region and cause others to leave. The net effect on earnings from the policy is all that is required to perform cost-benefit calculations of the policy on outcomes. If the housing subsidy is the instrument and the net effect of the subsidy is the parameter of interest, the issue of monotonicity is a red herring. If the subsidy is exogenously imposed, IV estimates the net effect of the policy on mean outcomes. Only if the effect of migration on earnings induced by the subsidy on outcomes is the question of interest, and not the effect of the subsidy, does uniformity emerge as an interesting condition.

IV. Comparing Selection and Local IV Models

We now show that local IV identifies the derivatives of a selection model. Making the X explicit, in the standard selection model, if the U_1 and U_0 are scalar random variables that are additively separable in the outcome equations, $Y_1 = \mu_1(X) + U_1$ and $Y_0 = \mu_0(X) + U_0$. The control function approach conditions on Z and D. As a consequence

The random variable $D(j)$ is defined by the following thought experiment. For each possible realization j of $J(Z)$, define $D(j)$ by setting $D(j) = D(Z(j))$, where $Z(j)$ is a random draw from the distribution of Z conditional on $J(Z) = j$. Set $D(j)$ equal to the choice that would be made given that draw of Z (j). Thus $D(j)$ is a function of $(Z(j), u_D)$. As long as we draw Z (j) randomly (and thus independently of Z), we have that $(Z(j), U_D) \perp\!\!\!\perp Z$, so $D(j) \perp\!\!\!\perp Z$. There are other possible constructions of the counterfactual $D(j)$, since there are different possible distributions from which Z can be drawn, apart from the actual distribution of Z. The advantage of this construction is that it equates the counterfactual probability that $D(j) = 1$ given $J(Z) = j$ with the population probability. If the value of Z were uncertain to the agent, this would be a rational expectations assumption. See the further discussion in appendix II posted at the Web site for this paper.

[33] When they analyze the vector case, Imbens and Angrist (1994) analyze instruments that are monotonic functions of $P(Z)$. Our analysis is more general and recognizes that in the vector case, IV weights may be negative or positive.

[34] In the conventional model with homogeneous responses, a linear probability approximation to $P(Z)$ used as an instrument would identify the same parameter as $P(Z)$. In the general model, the parameters identified are different. Replacing $P(Z)$ by a linear probability approximation of it (for example, $E(D \mid Z) = \pi Z = J(Z)$) is not guaranteed to produce positive weights for $\Delta^{\text{MTE}}(x,u_D)$ or $\Delta^{\text{LATE}}(x,u'_D,u_D)$, or to replicate the weights based on the correctly specified $P(Z)$.

of index sufficiency this is equivalent to conditioning on $P(Z)$ and D:

$$E(Y|X,D,Z) = \mu_0(X) + [\mu_1(X) - \mu_0(X)]D + K_1(P(Z),X)\,D + K_0(P(Z),X)(1 - D),$$

where the control functions are

$$K_1(P(Z),X) = E(U_1|D = 1,X,P(Z)),$$

$$K_0(P(Z),X) = E(U_0|D = 0,X,P(Z)).$$

The IV approach does not condition on D. It works with

$$E(Y|X,Z) = \mu_0(X) + [\mu_1(X) - \mu_0(X)]P(Z) + K_1(P(Z),X)\,P(Z) + K_0(P(Z),X)[1 - P(Z)], \tag{24}$$

the population mean outcome given X,Z.

From index sufficiency, $E(Y|X,Z) = E(Y|X,P(Z))$. The MTE is the derivative of this expression with respect to $P(Z)$, which we have defined as LIV:[35]

$$\left.\frac{\partial E(Y|X,P(Z))}{\partial P(Z)}\right|_{P(Z)=p} = \text{LIV}(X,p) = \text{MTE}(X,p).$$

The distribution of $P(Z)$ and the relationship between $J(Z)$ and $P(Z)$ determine the weight on MTE.[36] Under assumptions A-1 to A-5, along with rank and limit conditions (Heckman and Robb, 1985; Heckman, 1990), one can identify $\mu_1(X)$, $\mu_0(X)$, $K_1(P(Z),X)$, and $K_0(P(Z),X)$.

The selection (control function) estimator identifies the conditional means

$$E(Y_1|X,P(Z),D = 1) = \mu_1(X) + K_1(X,P(Z)) \tag{25a}$$

and

$$E(Y_0|X,P(Z),D = 0) = \mu_0(X) + K_0(X,P(Z)). \tag{25b}$$

These can be identified from nonparametric regressions of Y_1 and Y_0 on X,Z in each population. To decompose these means and separate $\mu_1(X)$ from $K_1(X,P(Z))$ without invoking functional form or curvature assumptions, it is necessary to have an exclusion (a Z not in X).[37] In addition, there must exist a limit set for Z given X such that $K_1(X,P(Z)) = 0$ for Z in that limit set. Otherwise, without functional form or curvature assumptions, it is not possible to disentangle $\mu_1(X)$ from $K_1(X,P(Z))$, which may contain constants and functions of X that do not interact with $P(Z)$ (see Heckman, 1990). A parallel argument for Y_0 shows that we require a limit set for Z given X such that $K_0(X,P(Z)) = 0$. Selection models operate by identifying the components of equations (25a) and (25b) and generating the treatment parameters from these components. Thus they work with levels of the Y.

The local IV method works with derivatives of equation (24) and not levels, and cannot directly recover the constant terms in equations (25a) and (25b). Using our analysis of LIV but applied to $YD = Y_1D$ and $Y(1 - D) = Y_0(1 - D)$, it is straightforward to use LIV to estimate the components of the MTE separately. Thus we can identify

$$\mu_1(X) + E(U_1|X,U_D = u_D)$$

and

$$\mu_0(X) + E(U_0|X,U_D = u_D)$$

separately. This corresponds to what is estimated from taking the derivatives of the expressions (25a) and (25b) multiplied by $P(Z)$ and $1 - P(Z)$, respectively:[38]

$$P(Z)\,E(Y_1|X,Z,D = 1) = P(Z)\mu_1(X) + P(Z)\,K_1(X,P(Z))$$

and

$$[1 - P(Z)]E(Y_0|X,Z,D = 0) = [1 - P(Z)]\mu_0(X) + [1 - P(Z)]K_0(X,P(Z)).$$

Thus the control function method works with levels, whereas the LIV approach works with slopes. Constants that do not depend on $P(Z)$ disappear from the estimates of the model. The level parameters are obtained by integration using the formulas in table 1B.

Misspecification of $P(Z)$ (either its functional form or its arguments) and hence of $K_1(P(Z),X)$ and $K_0(P(Z),X)$ in general produces biased estimates of the parameters of the model under the control function approach even if semiparametric methods are used to estimate μ_0, μ_1, K_0, and K_1. To implement the method, we need to know all of the arguments of Z. The quantities $K_1(P(Z),X)$ and $K_0(P(Z),X)$ can be nonparametrically estimated, so it is only necessary to know $P(Z)$ up to a monotonic transformation.[39] The distributions of U_1, U_0, and V do not need to be specified to estimate control function models (see Powell, 1994).

These problems with control function models have their counterparts in IV models. If we use a misspecified $P(Z)$ to identify the MTE or its components, in general we do not

[35] Björklund and Moffitt (1987) analyze this marginal effect for a parametric generalized Roy model.

[36] Because LIV does not condition on D, it discards information. Lost in taking derivatives are the constants in the model that do not interact with $P(Z)$ in equation (24).

[37] See Heckman and Navarro (2006) for use of semiparametric curvature restrictions in identification analysis that do not require functional form assumptions.

[38] Björklund and Moffitt (1987) use the derivative of a selection model in levels to define the marginal treatment effect.

[39] See Heckman et al. (1998a).

identify the MTE or its components. Misspecification of $P(Z)$ plagues both approaches.

One common criticism of selection models is that without invoking functional form assumptions, identification of $\mu_1(X)$ and $\mu_0(X)$ requires that $P(Z) \to 1$ and $P(Z) \to 0$ in limit sets.[40] Identification in limit sets is sometimes called "identification at infinity." In order to identify ATE $= E(Y_1 - Y_0|X)$, IV methods also require that $P(Z) \to 1$ and $P(Z) \to 0$ in limit sets, so an identification-at-infinity argument is implicit when IV is used to identify this parameter.[41] The LATE parameter avoids this problem by moving the goal posts and redefining the parameter of interest away from a level parameter like ATE or TT to a slope parameter like LATE, which differences out the unidentified constants. Alternatively, if we define the parameter of interest to be LATE or MTE, we can use the selection model without invoking identification at infinity.

The IV estimator is model-dependent, just like the selection estimator, but in application, the model does not have to be fully specified to obtain Δ^{IV} using Z (or $J(Z)$). However, the distribution of $P(Z)$ and the relationship between $P(Z)$ and $J(Z)$ generates the weights. The interpretation placed on Δ^{IV} in terms of weights on Δ^{MTE} depends crucially on the specification of $P(Z)$. In both control function and IV approaches for the general model of heterogeneous responses, $P(Z)$ plays a central role.

Two economists using the same instrument will obtain the same point estimate using the same data. Their *interpretation* of that estimate will differ depending on how they specify the arguments in $P(Z)$, even if neither uses $P(Z)$ as an instrument. By conditioning on $P(Z)$, the control function approach makes the dependence of estimates on the specification of $P(Z)$ explicit. The IV approach is less explicit and masks the assumptions required to economically interpret the empirical output of an IV estimation. We now turn to some examples that demonstrate the main points of this paper.

V. Examples Based on Choice Theory

Return to the policy adoption example presented in section I. The cost of adopting the policy C is the same across all countries. Suppose that countries choose to adopt the policy if $D^* > 0$, where D^* is the net benefit of adoption: $D^* = Y_1 - Y_0 - C$ and ATE $= E(\beta) = E(Y_1 - Y_0) = \mu_1 - \mu_0$, while treatment on the treated is $E(\beta \mid D = 1) = E(Y_1 - Y_0|D = 1) = \mu_1 - \mu_0 + E(U_1 - U_0|D = 1)$.

In this setting, the gross return to the country at the margin is C:

$$E(Y_1 - Y_0|D^* = 0) = E(Y_1 - Y_0|Y_1 - Y_0 = C) = C.$$

Figure 1 presents the standard treatment parameters for the values of the outcome and choice parameters presented at the base of the figure. Countries that adopt the policy are above average. In a model where the cost varies (the generalized Roy model with $U_C \neq 0$), and C is negatively correlated with the gain, adopting countries could be below average.[42]

A. *Discrete Instruments and the Weights for* LATE

Consider what instrumental variables identify in the model of choice and outcomes described at the bottom of figure 2. Let the cost $C = \gamma Z$, where the instrument $Z = (Z_1, Z_2)$. Higher values of Z reduce the probability of adopting the policy if $\gamma \geq 0$, component by component. Consider the "standard" case depicted in figure 2A. Increasing both components of discrete-valued Z raises costs and hence raises the return observed for the country at the margin by eliminating adoption in low-return countries. In general, a different country is at the margin when different instruments are used.

Figure 3A plots the weights and figure 3B the components of the weights for the LATE values using $P(Z)$ as an instrument for the distribution of Z shown at the base of the figure. Figure 3C presents the LATE parameter derived using $P(Z)$ as the instrument. The weights are positive as predicted from equation (22) when $J(Z) = P(Z)$. Thus the monotonicity condition for the weights in terms of u_D is satisfied. The outcome and choice parameters are the same as those used to generate figures 1 and 2. There are four LATE values corresponding to the five distinct values of the propensity score for this example. The LATEs exhibit the declining pattern with u_D predicted by the Roy model.

A more interesting case is that depicted in figure 2B. In that graph, the same Z are used to generate choices as in figures 2A and 3. In this case, however, the analyst uses Z_1 as the instrument, Z_1 and Z_2 are negatively dependent, and $E(Z_1|P(Z) > u_D)$ is not monotonic in u_D. This nonmonotonicity is evident in figure 4B. This produces the pattern of negative weights shown in figure 4A. These are associated with two-way flows. Increasing Z_1, controlling for Z_2, reduces the probability of country policy adoption. However, we do not condition on Z_2 in constructing this figure. It is floating. Two-way flows are induced by uncontrolled variation in Z_2. For some units, the strength of the associated variation in Z_2 offsets the increase in Z_1, and for other units it does not. Observe that the LATE parameters defined using $P(Z)$ are the same in both examples. They are just weighted differently. We discuss the random-coefficient-choice model generating figure 2C in section VII.

The IV estimator does not identify ATE, TT, or TUT given at the bottom of figure 3. Conditioning on Z_2 produces positive weights, as shown in the weights in table 2 that

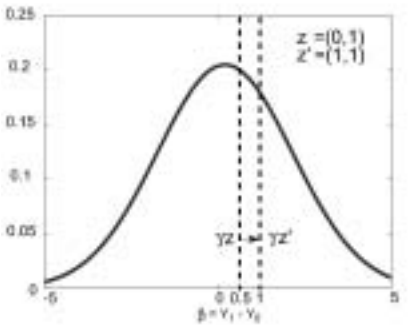
[40] See Imbens and Angrist (1994). Heckman (1990) establishes the identification in the limit argument for ATE in selection models. See Heckman and Navarro (2006) for a generalization to multiple-outcome models.

[41] Thus if the support of $P(Z)$ is not full, we cannot identify the treatment on the treated or the average treatment effect. We can construct bounds. See Heckman and Vytlacil (1999, 2001a,b, 2007b).

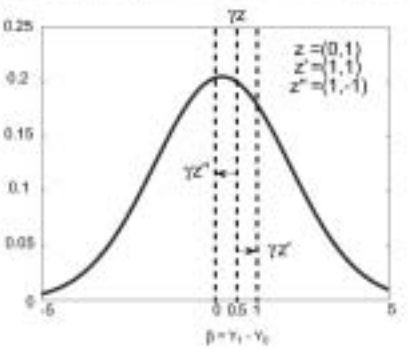
[42] See, for example, Heckman (1976a,b).

FIGURE 2.—MONOTONICITY: THE EXTENDED ROY ECONOMY

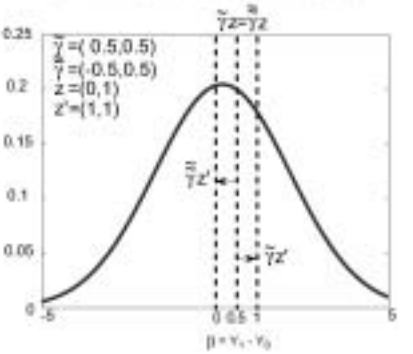
A. Standard Case



B. Changing Z_1 without Controlling for Z_2



C. Random Coefficient Case



Outcomes

$$Y_1 = \alpha + \bar{\beta} + U_1$$
$$Y_0 = \alpha + U_0$$

Choice Model

$$D = \begin{cases} 1 & \text{if } Y_1 - Y_0 - \gamma Z > 0 \\ 0 & \text{if } Y_1 - Y_0 - \gamma Z \leq 0 \end{cases}$$
$$\text{with } \gamma Z = \gamma_1 Z_1 + \gamma_2 Z_2$$

Parameterization

$$(U_1, U_0) \sim N(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}, \quad \alpha = 0.67,\ \bar{\beta} = 0.2,\ \gamma = (0.5, 0.5) \text{ (except in Case C)}$$

$$Z_1 = \{-1, 0, 1\} \text{ and } Z_2 = \{-1, 0, 1\}$$

A. Standard Case	B. Changing Z_1 without Controlling for Z_2	C. Random Coefficient Case
$z \longrightarrow z'$ $z = (0,1)$ and $z' = (1,1)$	$z \longrightarrow z'$ or $z \longrightarrow z''$ $z = (0,1)$, $z' = (1,1)$ and $z'' = (1,-1)$	$z \longrightarrow z'$ $z = (0,1)$ and $z' = (1,1)$
		γ is a random vector $\tilde{\gamma} = (0.5, 0.5)$ and $\tilde{\tilde{\gamma}} = (-0.5, 0.5)$ where $\tilde{\gamma}$ and $\tilde{\tilde{\gamma}}$ are two realizations of γ
$D(\gamma z) \geq D(\gamma z')$	$D(\gamma z) \geq D(\gamma z')$ or $D(\gamma z) < D(\gamma z'')$	$D\left(\tilde{\tilde{\gamma}} z\right) \geq D\left(\tilde{\tilde{\gamma}} z'\right)$ and $D(\tilde{\gamma} z) < D(\tilde{\gamma} z')$
For all individuals	Depending on the value of z' or z''	Depending on value of γ

condition on Z_2. Conditioning on Z_2 effectively converts the problem back into one with a scalar instrument, and the weights must be positive for that case.

By Yitzhaki's result (23), for any sample size, a regression of Y on P identifies a weighted average of slopes based on ordered regressors $[E(Y_\ell|p_\ell) - E(Y_{\ell-1}|p_{\ell-1})]/(p_\ell - p_{\ell-1})$, where $p_\ell > p_{\ell-1}$ and where the weights are the positive Yitzhaki weights derived in appendix C, in Yitzhaki (1989, 1996), or in Yitzhaki and Schechtman (2004). The weights are positive whether or not monotonicity (IV-3) holds. If monotonicity holds, IV is a weighted average of LATEs. Otherwise it is just a weighted average of ordered (by p_ℓ) estimators consistent with two-way flows.

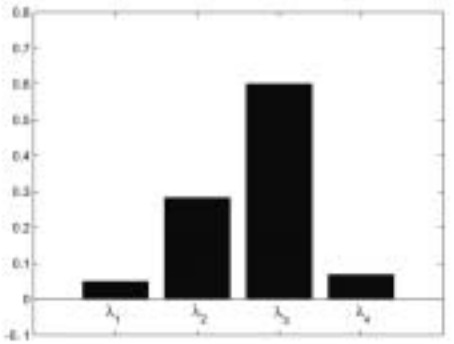
B. *Continuous Instruments*

For the case of continuous Z, we present a parallel analysis for the weights associated with the MTE. Figure 5 plots $E(Y|P(Z))$ and MTE for the models generated by the parameters displayed at the base of the figure. In cases I and II, $\beta \perp\!\!\!\perp D$. In case I, this is trivial, because β is a constant. In case II, β is random, but

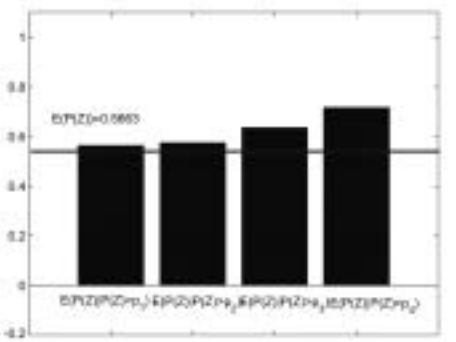
FIGURE 3.—IV WEIGHT AND ITS COMPONENTS UNDER DISCRETE INSTRUMENTS WHEN $P(Z)$ IS THE INSTRUMENT: THE EXTENDED ROY ECONOMY

The Extended Roy Economy

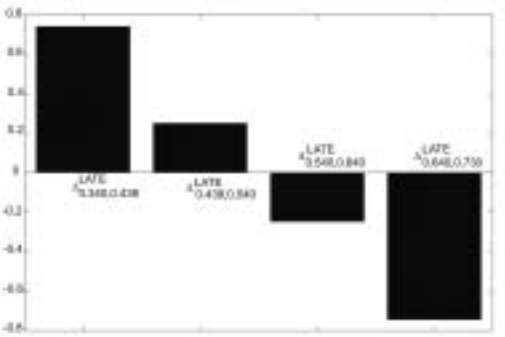
A. IV Weights



B. $E(P(Z)|P(Z) > p_\ell)$ and $E(P(Z))$



C. Local Average Treatment Effects



The model is the same as the one presented below Figure 2.

$$\text{ATE} = 0.2,\ \text{TT} = 0.5942,\ \text{TUT} = -0.4823 \text{ and } \Delta^{\text{IV}}_{P(Z)} = \sum_{\ell=1}^{K-1} \Delta^{\text{LATE}}(p_\ell, p_{\ell+1})\,\lambda_\ell = -0.09$$

$$\Delta^{\text{LATE}}(p_\ell, p_{\ell+1}) = \frac{E(Y|P(Z) = p_{\ell+1}) - E(Y|P(Z) = p_\ell)}{p_{\ell+1} - p_\ell} = \frac{\overline{\beta}(p_{\ell+1} - p_\ell) + \sigma_{U_1 - U_0}\left[\phi\left(\Phi^{-1}(1 - p_{\ell+1})\right) - \phi\left(\Phi^{-1}(1 - p_\ell)\right)\right]}{p_{\ell+1} - p_\ell}$$

$$\lambda_\ell = (p_{\ell+1} - p_\ell)\frac{\sum_{i=1}^{K}(p_i - E(P(Z)))\sum_{\ell>\ell}^{K} f(p_\ell, p_\ell)}{\text{Cov}(Z_1, D)} = (p_{\ell+1} - p_\ell)\frac{\sum_{\ell>\ell}^{K}(p_\ell - E(P(Z)))\,f(p_\ell)}{\text{Cov}(Z_1, D)}$$

Joint Probability Distribution of (Z_1, Z_2) and the Propensity Score
(joint probabilities in ordinary type $(\Pr(Z_1 = z_1, Z_2 = z_2))$; propensity score in italics $(\Pr(D = 1|Z_1 = z_1, Z_2 = z_2))$)

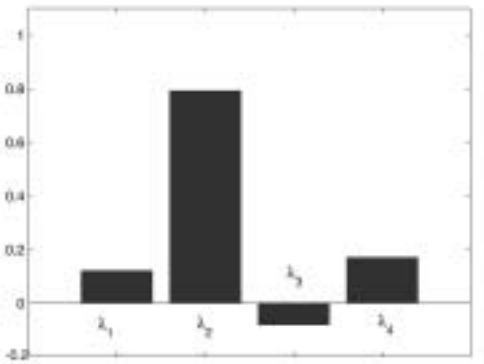
$Z_1 \backslash Z_2$	−1	0	1
−1	0.02	0.02	0.36
	0.7309	*0.6402*	*0.5409*
0	0.3	0.01	0.03
	0.6402	*0.5409*	*0.4388*
1	0.2	0.05	0.01
	0.5409	*0.4388*	*0.3408*

$\text{Cov}(Z_1, Z_2) = -0.5468$

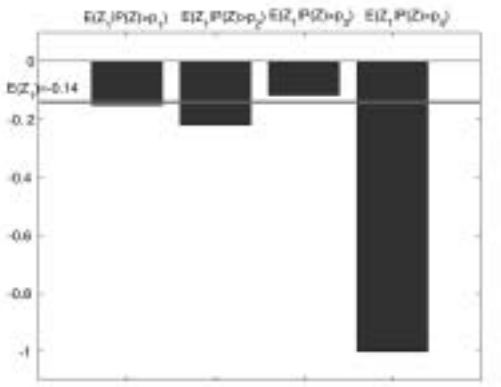
selection into D does not depend on β. Case III is the model with essential heterogeneity ($\beta \not\perp\!\!\!\perp D$). The left-hand side (figure 5A) depicts $E(Y|P(Z))$ for the three cases. Cases I and II make $E(Y|P(Z))$ linear in $P(Z)$ (see equation (17)). Case III is nonlinear in $P(Z)$. This arises when $\beta \not\perp\!\!\!\perp D$. The derivative of $E(Y|P(Z))$ is presented in the right panel (figure 5B). It is a constant for cases I and II (flat MTE), but declining in $U_D = P(Z)$ for the case with selection on the gain. A simple test for linearity in $P(Z)$ in the outcome equation reveals whether or not the analyst is in cases I and II ($\beta \perp\!\!\!\perp D$) or case III ($\beta \not\perp\!\!\!\perp D$). Recall that we keep conditioning on X implicit.

FIGURE 4.—IV WEIGHT AND ITS COMPONENTS UNDER DISCRETE INSTRUMENTS WHEN Z_1 IS THE INSTRUMENT: THE EXTENDED ROY ECONOMY

A. IV Weights



B. $E(Z_1|P(Z) > p_\ell)$ and $E(Z_1)$



The model is the same as the one presented below Figure 2. The values of the treatment parameters are the same as the ones presented below Figure 3.

$$\Delta^{\text{IV}}_{Z_1} = \sum_{\ell=1}^{K-1} \Delta^{\text{LATE}}(p_\ell, p_{\ell+1})\,\lambda_\ell = 0.1833$$

$$\lambda_\ell = (p_{\ell+1} - p_\ell)\,\frac{\sum_{i=1}^{I}(z_{1,i} - E(Z_1))\sum_{t>\ell}^{K} f(z_{1,i}, p_t)}{\text{Cov}(Z_1, D)}$$

Joint Probability Distribution of (Z_1, Z_2) and the Propensity Score
(joint probabilities in ordinary type $(\Pr(Z_1 = z_1, Z_2 = z_2))$; propensity score in italics $(\Pr(D = 1|Z_1 = z_1, Z_2 = z_2))$)

$Z_1 \backslash Z_2$	−1	0	1
−1	0.02	0.02	0.36
	0.7309	*0.6402*	*0.5409*
0	0.3	0.01	0.03
	0.6402	*0.5409*	*0.4388*
1	0.2	0.05	0.01
	0.5409	*0.4388*	*0.3408*

$\text{Cov}(Z_1, Z_2) = -0.5468$

The MTE gives the mean marginal return for persons who have utility $P(Z) = u_D$ ($P(Z) = u_D$ is the margin of indifference). Those with low u_D values have high returns. Those with high u_D values have low returns. Figure 5 highlights that the MTE (and the LATE) identify average returns for persons at the margin of indifference at different levels of the mean utility function $P(Z)$.

Figure 6A plots MTE and LATE for different intervals of u_D using the model generating figure 5. LATE is the chord of $E(Y|P(Z))$ evaluated at different points. The relationship between LATE and MTE is depicted in the right panel of figure 6. LATE is the integral under the MTE curve divided by the difference between the upper and lower limits.

The treatment parameters associated with case III are plotted in figure 7. The MTE is the same as that presented in figure 5. ATE has the same value for all p. The effect of treatment on the treated for $P(Z) = p$, $\Delta^{\text{TT}}(p) = E(Y_1 - Y_0|D = 1, P(Z) = p)$, declines in p (equivalently, it declines in u_D). The effect of treatment on the untreated given p, $\Delta^{\text{TUT}}(p) = E(Y_1 - Y_0|D = 0, P(Z) = p)$, also declines in p. Observe that

$$\text{LATE}(p,p') = \frac{\Delta^{\text{TT}}(p')p' - \Delta^{\text{TT}}(p)p}{p' - p}, \qquad p' \neq p,$$

$$\text{MTE} = \frac{\partial[\Delta^{\text{TT}}(p)p]}{\partial p}.$$

We can generate all of the treatment parameters from $\Delta^{\text{TT}}(p)$.

Matching on $P = p$ (which is equivalent to nonparametric least squares, given $P = p$) produces a biased estimator of

TABLE 2.—THE CONDITIONAL INSTRUMENTAL VARIABLE ESTIMATOR ($\Delta^{IV}_{Z_1|Z_2=z_2}$) AND CONDITIONAL LOCAL AVERAGE TREATMENT EFFECT ($\Delta^{LATE}(p_\ell, p_{\ell+1}|Z_2=z_2)$) WHEN Z_1 IS THE INSTRUMENT (GIVEN $Z_2=z_2$): THE EXTENDED ROY ECONOMY

	$Z_2=-1$	$Z_2=0$	$Z_2=1$
$P(-1,Z_2)=p_3$	0.7309	0.6402	0.5409
$P(0,Z_2)=p_2$	0.6402	0.5409	0.4388
$P(1,Z_2)=p_1$	0.5409	0.4388	0.3408
λ_1	0.8418	0.5384	0.2860
λ_2	0.1582	0.4616	0.7140
$\Delta^{LATE}(p_1,p_2)$	−0.2475	0.2497	0.7470
$\Delta^{LATE}(p_2,p_3)$	−0.7448	−0.2475	0.2497
$\Delta^{IV}_{Z_1\|Z_2=z_2}$	−0.3262	0.0202	0.3920

The model is the same as the one presented below Figure 2

$$\Delta^{IV}_{Z_1|Z_2=z_2}=\sum_{\ell=1}^{J-1}\Delta^{LATE}(p_\ell,p_{\ell+1}|Z_2=z_2)\,\lambda_{\ell|Z_2=z_2}=\sum_{\ell=1}^{I-1}\Delta^{LATE}(p_\ell,p_{\ell+1}|Z_2=z_2)\,\lambda_{\ell|Z_2=z_2}$$

$$\Delta^{LATE}(p_\ell,p_{\ell+1}|Z_2=z_2)=\frac{E(Y|P(Z)=p_{\ell+1},Z_2=z_2)-E(Y|P(Z)=p_\ell,Z_2=z_2)}{p_{\ell+1}-p_\ell}$$

$$\lambda_{\ell|Z_2=z_2}=(p_{\ell+1}-p_\ell)\frac{\sum_{i=1}^{J}(z_{1,i}-E(Z_1|Z_2=z_2))\sum_{t>\ell}^{J}f(z_{1,i},p_t|Z_2=z_2)}{\mathrm{Cov}(Z_1,D)}=(p_{\ell+1}-p_\ell)\frac{\sum_{t>\ell}^{I}(z_{1,t}-E(Z_1|Z_2=z_2))f(z_{1,t},p_t|Z_2=z_2)}{\mathrm{Cov}(Z_1,D)}$$

Probability Distribution of Z_1 Conditional on Z_2 ($\Pr(Z_1=z_1|Z_2=z_2)$)

z_1	$\Pr(Z_1=z_1\|Z_2=-1)$	$\Pr(Z_1=z_1\|Z_2=0)$	$\Pr(Z_1=z_1\|Z_2=1)$
−1	0.0385	0.25	0.9
0	0.5769	0.125	0.075
1	0.3846	0.625	0.025

TT(p). Matching assumes a flat MTE (average return equals marginal return).[43] Therefore it is systematically biased for $\Delta^{TT}(p)$ in a model with essential heterogeneity. Making observables alike makes the unobservables dissimilar. Holding p constant across treatment and control groups understates TT(p) for low p and overstates it for high p.

We now present additional examples with continuously distributed instruments. See figure 8. Instrument Z is assumed to be a random vector with a distribution function given by a mixture of two normals:

$$Z\sim P_1N(\kappa_1,\Sigma_1)+P_2N(\kappa_2,\Sigma_2),$$

where P_1 is the proportion in population 1, P_2 is the proportion in population 2, and $P_1+P_2=1$. This produces a model with continuous instruments, where $E(\tilde{J}(Z)|P(Z)>u_D)$ (with $\tilde{J}(Z)=J(Z)-E(J(Z))$) need not be monotonic in u_D. Such a data-generating process for the instrument could arise from an ecological model in which two different populations are mixed (for example, rural and urban populations).[44]

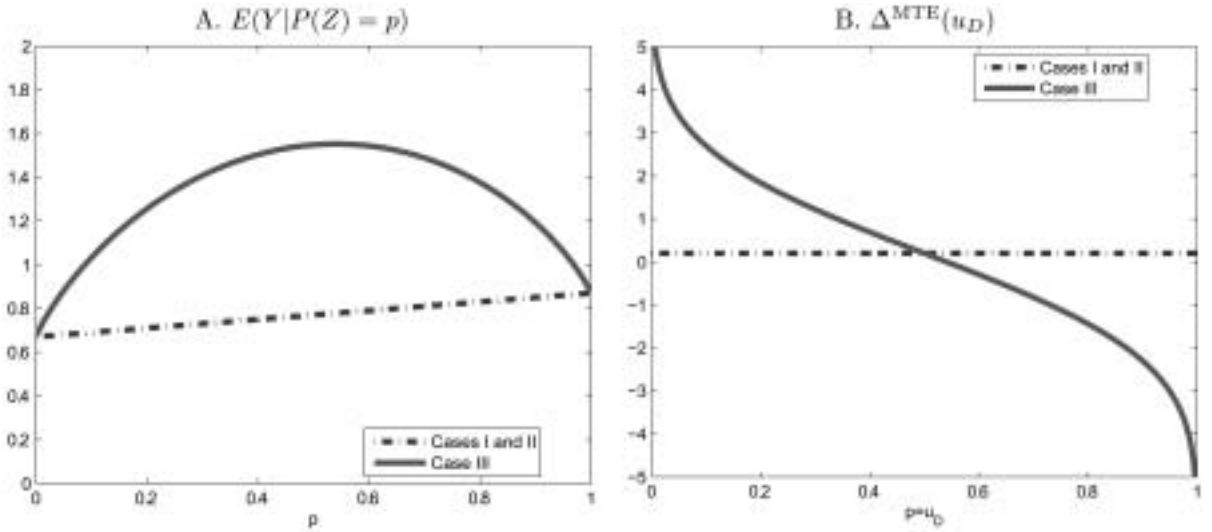
At our Web appendix, we derive explicit instrumental variable weights on Δ^{MTE} when Z_1 (the first element of Z) is used as the instrument, that is, $J(Z)=Z_1$ for this case. For simplicity we assume that there are no X regressors. The probability of selection is generated by $\mu_D(Z)=\gamma Z$. The joint distribution of $(Z_1, \gamma Z)$ is normal within each group.

In our example, the dependence between Z_1 and γZ ($=F_V(\gamma z)=P(Z)$) is negative in one population and

[43] See Heckman and Vytlacil (2005, 2007b).

[44] Observe that $E(Z)=P_1\kappa_1+P_2\kappa_2$.

FIGURE 5.—CONDITIONAL EXPECTATION OF Y ON $P(Z)$ AND THE MARGINAL TREATMENT EFFECT (MTE): THE EXTENDED ROY ECONOMY



Outcomes

$$Y_1 = \alpha + \bar{\beta} + U_1$$
$$Y_0 = \alpha + U_0$$

Choice Model

$$D = \begin{cases} 1 \text{ if } D^* > 0 \\ 0 \text{ if } D^* \leq 0 \end{cases}$$

Case I	Case II	Case III
$U_1 = U_0$ $\bar{\beta}$ =ATE=TT=TUT=IV	$U_1 - U_0 \perp\!\!\!\perp D$ $\bar{\beta}$ =ATE=TT=TUT=IV	$U_1 - U_0 \not\perp\!\!\!\perp D$ $\bar{\beta}$ =ATE≠TT≠TUT≠IV

Parameterization

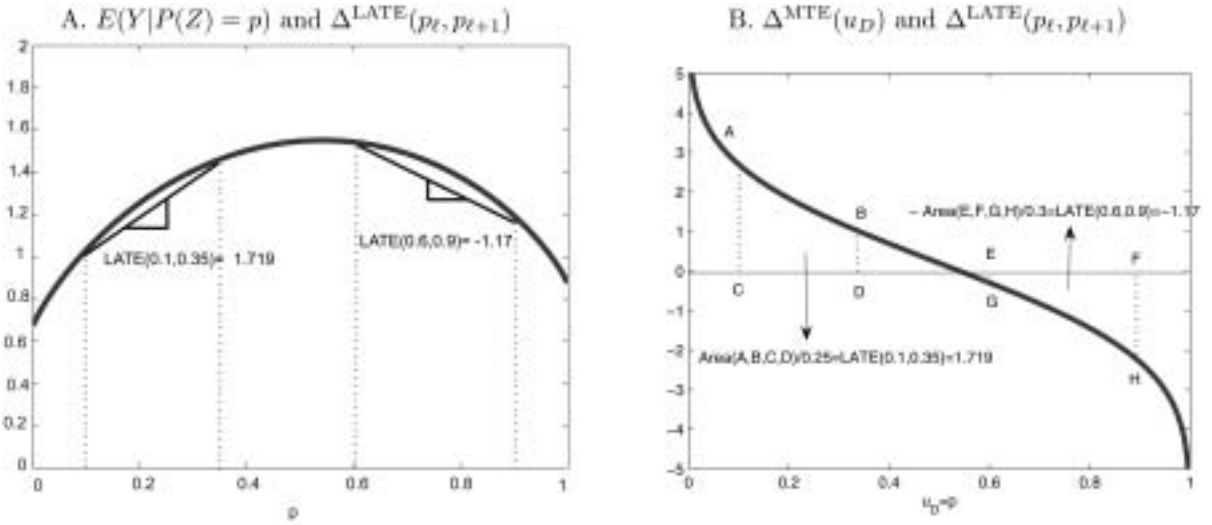
Cases I, II and III	Cases II and III	Case III
$\alpha = 0.67$ $\bar{\beta} = 0.2$	$(U_1, U_0) \sim N(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$	$D^* = Y_1 - Y_0 - \gamma Z$ $Z \sim N(\mu_Z, \mathbf{\Sigma}_Z)$ $\mu_Z = (2, -2)$ and $\mathbf{\Sigma}_Z = \begin{bmatrix} 9 & -2 \\ -2 & 9 \end{bmatrix}$ $\gamma = (0.5, 0.5)$

positive in another. Thus in one population, as Z_1 increases, $P(Z)$ increases. In the other population as Z_1 increases, $P(Z)$ decreases. If this second population is sufficiently big (P_1 is small) or the negative dependence in the second population is sufficiently big, the weights can become negative because $E(\tilde{J}(Z)|P(Z) > u_D)$ is not monotonic in u_D.

We present examples for a conventional normal outcome model generated by the parameters at the bottom of figure 8. The discrete choice equation is a conventional probit, as in the other examples. The outcome equations are linear normal equations. Thus $\Delta^{\text{MTE}}(v) = E(Y_1 - Y_0|V = v)$ is linear in v:

$$E(Y_1 - Y_0|V = v) = \mu_1 - \mu_0 + \frac{\text{Cov}(U_1 - U_0, V)}{\text{Var}(V)} v.$$

FIGURE 6.—THE LOCAL AVERAGE TREATMENT EFFECT: THE EXTENDED ROY ECONOMY



$$\Delta^{\text{LATE}}(p_\ell, p_{\ell+1}) = \frac{E(Y|P(Z)=p_{\ell+1}) - E(Y|P(Z)=p_\ell)}{p_{\ell+1}-p_\ell} = \frac{\int_{p_\ell}^{p_{\ell+1}} \Delta^{\text{MTE}}(u_D)du_D}{p_{\ell+1}-p_\ell}$$

$$\begin{aligned} \Delta^{\text{LATE}}(0.6, 0.9) &= -1.17 \\ \Delta^{\text{LATE}}(0.1, 0.35) &= 1.719 \end{aligned}$$

Outcomes	Choice Model
$Y_1 = \alpha + \bar{\beta} + U_1$	$D = \begin{cases} 1 \text{ if } D^* > 0 \\ 0 \text{ if } D^* \leq 0 \end{cases}$
$Y_0 = \alpha + U_0$	with $D^* = Y_1 - Y_0 - \gamma Z$

Parameterization

$$(U_1, U_0) \sim N(\mathbf{0}, \mathbf{\Sigma}) \text{ and } Z \sim N(\mu_Z, \mathbf{\Sigma}_Z)$$

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}, \ \mu_Z = (2, -2) \text{ and } \mathbf{\Sigma}_Z = \begin{bmatrix} 9 & -2 \\ -2 & 9 \end{bmatrix}$$

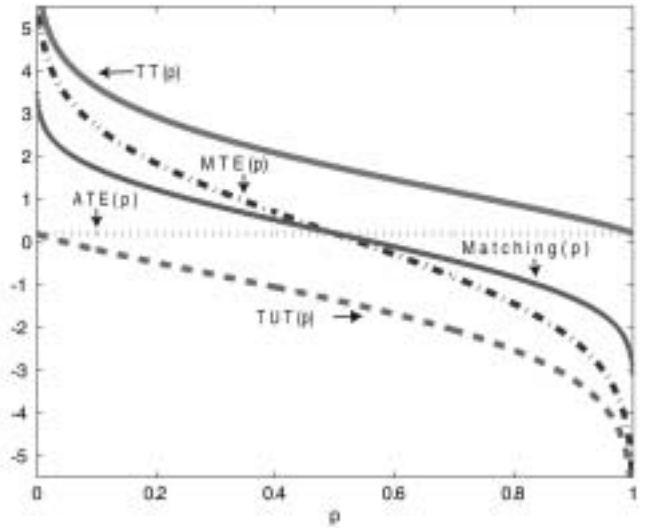
$$\alpha = 0.67, \bar{\beta} = 0.2, \gamma = (0.5, 0.5)$$

At the bottom of the figure, we define $\bar{\beta} = \mu_1 - \mu_0$ and $\alpha = \mu_0$. The average treatment effects are the same for all distributions of the Z.

In each of the following examples, we show results for models with vector Z that satisfies IV-1 and IV-2 and with $\gamma > 0$ componentwise, where γ is the coefficient on Z in the cost equation. We vary the weights and means of the instruments. Ceteris paribus, an increase in each component of Z increases $\Pr(D = 1 \mid Z = z)$. Table 3 presents the effect of treatment on the treated ($E(Y_1 - Y_0|D = 1)$), that of treatment on the untreated ($E(Y_1 - Y_0|D = 0)$), and the average treatment effect ($E(Y_1 - Y_0)$) produced by our model.

In standard IV analysis, the distribution of Z does not affect the probability limit of the IV estimator. It only affects its sampling distribution. Figure 8A shows three weights corresponding to the perturbations of the variance of the instruments in the second component population Σ_2 and the means (κ_1, κ_2) shown in table 3. The

FIGURE 7.—TREATMENT PARAMETERS AND OLS/MATCHING AS A FUNCTION OF $P(Z) = p$



Parameter	Definition	Under Assumptions (*)
Marginal Treatment Effect	$E[Y_1 - Y_0 \mid D^* = 0, P(Z) = p]$	$\overline{\beta} + \sigma_{U_1 - U_0}\Phi^{-1}(1-p)$
Average Treatment Effect	$E[Y_1 - Y_0 \mid P(Z) = p]$	$\overline{\beta}$
Treatment on the Treated	$E[Y_1 - Y_0 \mid D^* > 0, P(Z) = p]$	$\overline{\beta} + \sigma_{U_1 - U_0}\frac{\phi\left(\Phi^{-1}(1-p)\right)}{p}$
Treatment on the Untreated	$E[Y_1 - Y_0 \mid D^* \leq 0, P(Z) = p]$	$\overline{\beta} - \sigma_{U_1 - U_0}\frac{\phi\left(\Phi^{-1}(1-p)\right)}{1-p}$
OLS/Matching on $P(Z)$	$E[Y_1 \mid D^* > 0, P(Z) = p] - E[Y_0 \mid D^* \leq 0, P(Z) = p]$	$\overline{\beta} + \left(\frac{\sigma_{U_1} - \sigma_{U_1, U_0}}{\sqrt{\sigma_{U_1 - U_0}}}\right)\left(\frac{1-2p}{p(1-p)}\right)\phi\left(\Phi^{-1}(1-p)\right)$

Note: $\Phi(\cdot)$ and $\phi(\cdot)$ represent the cdf and pdf of a standard normal distribution, respectively. $\Phi^{-1}(\cdot)$ represents the inverse of $\Phi(\cdot)$.

(*): The model in this case is the same as the one presented below Figure 6.

MTE used in all of our examples is plotted in figure 8B. The MTE has the familiar shape, reported in Heckman (2001) and Heckman, Tobias, and Vytlacil (2003), that returns are highest for those with values of v that make them more likely to get treatment (that is, low values of v).

The weights ω_1 and ω_3 correspond to the case where $E(Z_1 - E(Z_1) \mid P(Z) > u_D)$ is not monotonic in u_D. In these cases the relationship between Z_1 and $P(Z)$ is not the same in the two subpopulations. The IV estimates range all over the place, even though the parameters of the outcome and choice model are the same.[45] Only the distributions of the instruments are different.

Different distributions of Z critically affect the probability limit of the IV estimator in the model of essential heterogeneity. The model of outcomes and choices is the same across all of these examples. The MTE and ATE parameters are the same. Only the distribution of the instrument differs. The IV estimand is sometimes positive and sometimes negative, and oscillates wildly in magnitude depending on the distribution of the instruments. The estimated "effect" is often way off the mark for any desired treatment parameter. These examples show how uniformity in Z does not translate into uniformity in $J(Z)$ (Z_1 in this example). This sensitivity is a phenomenon that does not appear in the conventional homogeneous response model but is a central feature of a model with essential heterogeneity.[46]

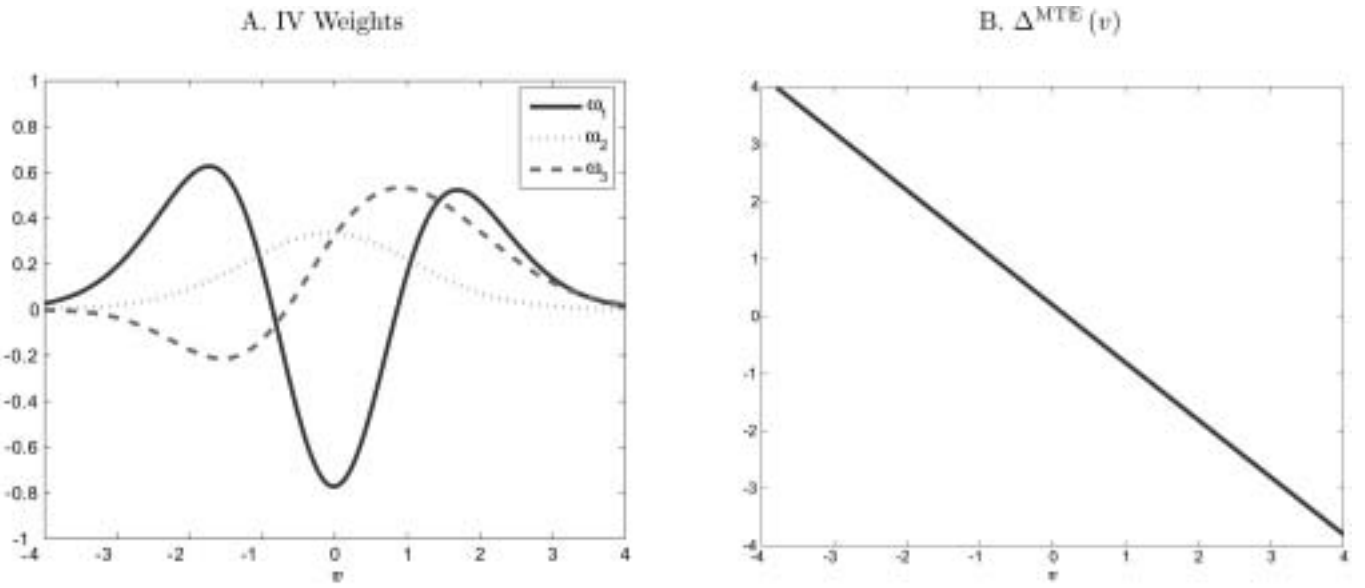
C. *Empirical Example: Using IV to Estimate the "Effect" of High School Graduation on Wages*

The previous examples demonstrate logical possibilities. This subsection shows that these logical possibilities arise in real data. We study the effects of graduating from high school on wages, using data from the National Longitudinal Survey of Youth 1979 (NLSY79). This survey gathers

[45] Because TT and TUT depend on the distribution of $P(Z)$, they are not invariant to changes is the distribution of the Z.

[46] We note parenthetically that if we assume that $P_1 = 0$ (or $P_2 = 0$), the weights are always positive even if we use only Z_1 as an instrument and Z_1 and Z_2 are negatively correlated. This follows from the monotonicity of $E(R|S > c)$ in c for vector R. See Heckman and Honoré (1990).

FIGURE 8.—MARGINAL TREATMENT EFFECT AND IV WEIGHTS USING Z_1 AS THE INSTRUMENT WHEN $Z = (Z_1, Z_2) \sim p_1 N(\kappa_1, \Sigma_1) + p_2 N(\kappa_2, \Sigma_2)$ FOR DIFFERENT VALUES OF Σ_2



Outcomes

$$Y_1 = \alpha + \beta + U_1$$
$$Y_0 = \alpha + U_0$$

Choice Model

$$D = \begin{cases} 1 & \text{if } D^* > 0 \\ 0 & \text{if } D^* \leq 0 \end{cases}$$
$$D^* = Y_1 - Y_0 - \gamma Z \text{ and } V = -(U_1 - U_0)$$

Parameterization

$$(U_1, U_0) \sim N(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}, \quad \alpha = 0.67, \beta = 0.2$$

$$Z = (Z_1, Z_2) \sim p_1 N(\kappa_1, \Sigma_1) + p_2 N(\kappa_2, \Sigma_2)$$

$$p_1 = 0.45,\ p_2 = 0.55 \quad ; \quad \Sigma_1 = \begin{bmatrix} 1.4 & 0.5 \\ 0.5 & 1.4 \end{bmatrix}$$

$$\text{Cov}(Z_1, \gamma Z) = \gamma \Sigma_1^1 = 0.98 \quad ; \quad \gamma = (0.2, 1.4)$$

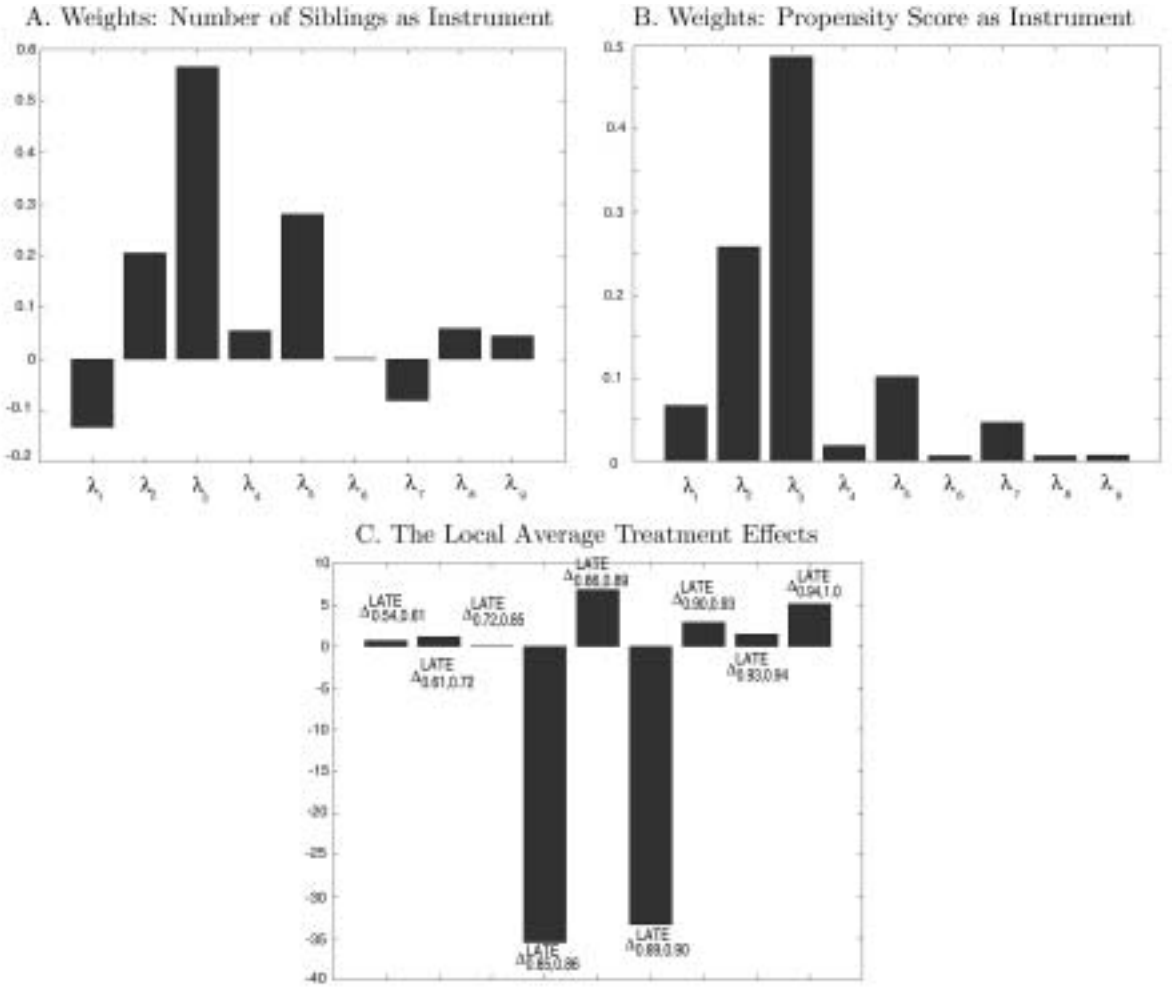
Table 3. IV estimator and $\text{Cov}(Z_2, \gamma Z)$ associated with each value of Σ_2

Weights	Σ_2	κ_1	κ_2	IV	ATE	TT	TUT	$\text{Cov}(Z_2, \gamma Z) = \gamma \Sigma_2^1$
ω_1	$\begin{bmatrix} 0.6 & -0.5 \\ -0.5 & 0.6 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \end{bmatrix}$	0.434	0.2	1.401	−1.175	−0.58
ω_2	$\begin{bmatrix} 0.6 & 0.1 \\ 0.1 & 0.6 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \end{bmatrix}$	0.078	0.2	1.378	−1.145	0.26
ω_3	$\begin{bmatrix} 0.6 & -0.3 \\ -0.3 & 0.6 \end{bmatrix}$	$\begin{bmatrix} 0 & -1 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \end{bmatrix}$	−2.261	0.2	1.310	−0.859	−0.30

information at multiple points in time on the labor market activities for men and women born in the years 1957–1964.

We estimate LATE using log hourly wages at age 30 as the outcome measure. Following a large body of research (see Mare, 1980), we use the number of siblings and mother's graduation status as instruments. Figure 9 plots the weights on LATE using the estimated $P(Z)$. The weights are based on equation (22). The LATE parameters are both positive and negative. The weights using siblings as an instrument are both positive and negative. The weights

FIGURE 9.—IV WEIGHTS: THE EFFECT OF GRADUATING FROM HIGH SCHOOL: SAMPLE OF HIGH SCHOOL DROPOUTS AND HIGH SCHOOL GRADUATES: WHITE MALES—NLSY79



Y = Log per-hour wage at age 30, Z_1 = Number of Siblings in 1979, Z_2 = Mother is a High School Graduate

$$D = \begin{cases} 1 & \text{if High School Graduate} \\ 0 & \text{if High School Dropout} \end{cases}$$

IV Estimates
(bootstrap std. errors in parenthesis - 100 replications)

Instrument	Value
Number of Siblings in 1979	0.115
	(0.695)
Propensity Score	0.316
	(0.110)

Joint Probability Distribution of (Z_1, Z_2) and the Propensity Score
(joint probabilities $\Pr(Z_1 = z_1, Z_2 = z_2)$ in ordinary type; propensity score $\Pr(D = 1 | Z_1 = z_1, Z_2 = z_2)$ in italics)

$Z_2 \backslash Z_1$	0	1	2	3	4
0	0.07	0.03	0.47	0.121	0.06
	1.0	*0.54*	*0.86*	*0.72*	*0.61*
1	0.039	0.139	0.165	0.266	0.121
	0.94	*0.89*	*0.90*	*0.85*	*0.93*

$\text{Cov}(Z_1, Z_2) = -0.066$ - Number of Observations $= 1,702$

using $P(Z)$ as an instrument are positive, as they must be, following the analysis of Yitzhaki. The two IV estimates differ from each other because the weights are different. The overall IV estimate is a crude summary of the underlying component LATEs, which are often large and positive or large and negative.

We next turn to an extension of our model to multiple outcomes.

VI. Extensions to More than Two Outcomes

Angrist and Imbens (1995) extend their analysis of LATE to an ordered choice model with outcomes generated by a scalar instrument that can assume multiple values. From their analysis of the effect of schooling on earnings, it is unclear, even under a strengthened "monotonicity" condition, whether IV estimates the effect of a change of schooling on earnings for a well-defined margin of choice. To summarize their analysis, let $\bar{S}$ be the number of possible outcome states with associated outcomes Y_s and choice indicators D_s, $s = 1, \ldots, \bar{S}$. The s in their analysis correspond to different levels of schooling. For any two instrument values $Z = z_i$ and $Z = z_j$ with $z_i > z_j$, we can define associated indicators $\{D_s(z_i)\}_{s=1}^{\bar{S}}$ and $\{D_s(z_j)\}_{s=1}^{\bar{S}}$, where $D_s(z_i) = 1$ if a person assigned instrument value z_i chooses state s. As in the two outcome model, the instrument Z is assumed to be independent of the potential outcomes $\{Y_s\}_{s=1}^{\bar{S}}$ as well as the associated indicator functions defined by fixing Z at z_i and z_j. Observed schooling for the instrument z_j is $S(z_j) = \sum_{s=1}^{\bar{S}} sD_s(z_j)$. Observed outcomes with this instrument are $Y(z_j) = \sum_{s=1}^{\bar{S}} Y_s D_s(z_j)$. Angrist and Imbens show that IV (with $Z = z_i$ and z_j) applied to S in a two-stage least squares regression of Y on S identifies a "causal parameter"

$$\Delta^{\text{IV}} = \sum_{s=2}^{\bar{S}} \{E(Y_s - Y_{s-1} | S(z_i) \geq s > S(z_j))\} \times \frac{Pr(S(z_i) \geq s > S(z_j))}{\sum_{s=2}^{\bar{S}} \Pr(S(z_i) \geq s > S(z_j))}. \tag{26}$$

This parameter is a weighted average of the gross return from going from $s - 1$ to s for persons induced by the change in the instrument to move from *any* schooling level below s to *any* schooling level s or above. Thus the conditioning set defining the s component of IV includes people who have schooling below $s - 1$ at instrument value $Z = z_j$ and people who have schooling above level s at instrument value $Z = z_i$. In this sum, the average return experienced by some of the people in the conditioning set for each component conditional expectation does not correspond to the average outcome corresponding to the gain in the argument of the expectation. In the case where $\bar{S} = 2$, agents face only two choices and the margin of choice is well defined. Agents in each conditioning set are at different margins of choice. The weights are positive, but, as noted by Angrist and Imbens, persons can be counted multiple times in forming the weights. When they generalize their analysis to multiple-valued instruments, they use the Yitzhaki (1989) weights.

Whereas the weights in equation (26) can be constructed empirically, the quantities in braces cannot be identified by any standard IV procedure. We present decompositions with components that are recoverable, whose weights can be estimated from the data and that are economically interpretable.

We generalize LATE to a multiple-outcome case where we can identify agents at different well-defined margins of choice. Specifically, we (1) analyze both ordered and unordered choice models, (2) analyze outcomes associated with choices at various well-defined margins, and (3) develop models with multiple instruments that can affect different margins of choice differently. With our methods, we can define and estimate a variety of economically interpretable parameters, whereas the Angrist-Imbens analysis produces a single "causal parameter" (26) that does not answer any well-defined policy question. We first consider an explicit ordered choice model and decompose the IV into policy-useful, identifiable components.

A. Analysis of an Ordered Choice Model

Ordered choice models arise in many settings. In schooling models, there are multiple grades. One has to complete grade $s - 1$ to proceed to grade s. The ordered choice model has been widely used to fit data on schooling transitions (Harmon and Walker, 1999; Cameron and Heckman, 1998). Its nonparametric identifiability has been studied (Carneiro, Hansen, and Heckman, 2003; Cunha, Heckman, and Navarro, 2007). It can also be used as a duration model for dynamic treatment effects with associated outcomes, as in Cunha, Heckman, and Navarro (2007). It also represents the "vertical" model of the choice of product quality (Prescott and Visscher, 1977; Shaked and Sutton, 1982; Bresnahan, 1987).

Our analysis generalizes the preceding analysis for the binary model in a parallel way. Write potential outcomes as

$$Y_s = \mu_s(X, U_s), \qquad s = 1, \ldots, \bar{S}.$$

The $\bar{S}$ could be different schooling levels or product qualities. We define latent variables $D_S^* = \mu_D(Z) - V$, where

$$D_s = \mathbf{1}[C_{s-1}(W_{s-1}) < \mu_D(Z) - V \leq C_s(W_s)],$$

$$s = 1, \ldots, \bar{S},$$

and the cutoff values satisfy

$$C_{s-1}(W_{s-1}) \leq C_s(W_s), \qquad C_0(W_0) = -\infty,$$

$$\text{and} \quad C_{\bar{S}}(W_{\bar{S}}) = \infty.$$

The cutoffs used to define the intervals are allowed to depend on observed (by the economist) regressors W_s. In appendix D we extend the analysis to allow the cutoffs to depend on unobserved regressors as well, following structural analysis along these lines by Carneiro, Hansen, and Heckman (2003) and Cunha, Heckman, and Navarro (2007). The observed outcomes are $Y = \sum_{s=1}^{\bar{S}} Y_s D_s$. The components of Z shift the index generally; those of W_s affect s-specific transitions. Thus, in a schooling example, Z could include family background variables, while W_s could include college tuition or opportunity wages for unskilled labor.[47] Collect the W_s into $W = (W_1, \ldots, W_{\bar{S}})$, and the U_s into $U = (U_1, \ldots, U_{\bar{S}})$. Larger values of $C_s(W_s)$ make it more likely that $D_s = 1$. The inequality restrictions on the functions $C_s(W_s)$ play a critical role in defining the model and producing its statistical implications.

Analogous to the assumptions made for the binary outcome model, we assume

OC-1: $(U_s, V) \perp\!\!\!\perp (Z,W)|X$, $s = 1, \ldots, \bar{S}$ (*conditional independence of the instruments*).

OC-2: $\mu_D(Z)$ is a nondegenerate random variable conditional on X and W (*rank condition*).

OC-3: The distribution of V is continuous.[48]

OC-4: $E(|Y_s|) < \infty$, $s = 1, \ldots, \bar{S}$ (*finite means*).

OC-5: $0 < \Pr(D_s = 1|X) < 1$ for $s = 1, \ldots, \bar{S}$ for all X (*In large samples, there are some persons in each treatment state*).

OC-6: For $s = 1, \ldots, \bar{S} - 1$, the distribution of $C_s(W_s)$ conditional on X, Z and the other $C_j(W_j)$, $j = 1, \ldots, \bar{S}$, $j \neq s$, is nondegenerate and continuous.[49]

Assumptions OC-1 to OC-5 play roles analogous to their counterparts in the two-outcome model, A-1 to A-5. OC-6 is a new condition that is key to identification of the Δ^{MTE} defined below for each transition. It assumes that we can vary the choice sets of agents at different margins of schooling choice without affecting other margins of choice. A necessary condition for OC-6 to hold is that at least one element of W_s is nondegenerate and continuous conditional on X, Z, and $C_j(W_j)$ for $j \neq s$. Intuitively, one needs an instrument (or source of variability) for each transition. The continuity of the regressor allows us to differentiate with respect to $C_s(W_s)$, as we differentiated with respect to $P(Z)$ to estimate the MTE in the analysis of the two-outcome model.

The analysis of Angrist and Imbens (1995) discussed in the introduction to this section makes independence and monotonicity assumptions that generalize their earlier work. They do not consider estimation of transition-specific parameters as we do, or even transition-specific LATE. We present a different decomposition of the IV estimator where each component can be recovered from the data, and where the transition-specific MTEs answer well-defined and economically interpretable policy evaluation questions.[50]

The probability of $D_s = 1$ given X, Z, and W is generated by an ordered choice model:

$$\Pr(D_s = 1|W,Z,X) \equiv P_s(Z,W,X) = \Pr(C_{s-1}(W_{s-1}) < \mu_D(Z) - V \leq C_s(W_s)|X).$$

Analogous to the binary case, we can define $U_D = F_V(V|X = x)$, so $U_D \sim \text{Unif}[0,1]$ under our assumption that the distribution of V is absolutely continuous with respect to Lebesgue measure. The probability integral transformation used extensively in the binary choice model is somewhat less useful for analyzing ordered choices, so we work with both U_D and V in this section of the paper. Monotonic transformations of V induce monotonic transformations of $\mu_D(Z) - C_s(W_s)$, but one is not free to form arbitrary monotonic transformations of $\mu_D(Z)$ and $C_s(W_s)$ separately. Using the probability integral transformation, the expression for choice s is $D_s = \mathbf{1}[F_V(\mu_D(Z) - C_{s-1}(W_{s-1})) > U_D \geq F_V(\mu_D(Z) - C_s(W_s))]$. Keeping the conditioning on X implicit, we define $P_s(Z,W) = F_V(\mu_D(Z) - C_{s-1}(W_{s-1})) - F_V(\mu_D(Z) - C_s(W_s))$. It is convenient to work with the probability that $S > s$, $\pi_s(Z,W_s) = F_V(\mu_D(Z) - C_s(W_s)) = \Pr(\sum_{j=s+1}^{\bar{S}} D_j = 1|Z,W_s)$, $\pi_{\bar{S}}(Z,W_{\bar{S}}) = 0$, $\pi_0(Z,W_0) = 1$, *and* $P_s(Z,W) = \pi_{s-1}(Z,W_{s-1}) - \pi_s(Z,W_s)$.

The transition-specific Δ^{MTE} for the transition from s to $s + 1$ is defined in terms of U_D:

$$\Delta^{\text{MTE}}_{s,s+1}(x, u_D) = E(Y_{s+1} - Y_s|X = x, U_D = u_D), \quad s = 1, \ldots, \bar{S} - 1.$$

Alternatively, one can condition on V. Analogously to the analysis of the earlier sections of this paper, when we set $u_D = \pi_s(Z,W_s)$, we obtain the mean return to persons indifferent between s and $s + 1$ at mean level of utility $\pi_s(Z,W_s)$.

In this notation, keeping X implicit, the mean outcome Y, conditional on (Z,W), is the sum of the mean outcomes conditional on each state weighted by the probability of being in each state summed over all states:

[47] Many of the instruments studied by Harmon and Walker (1999) and Card (2001) are transition-specific. Card's model of schooling is not sufficiently rich to make the distinction between Z and W. See Heckman and Navarro (2006) and Cunha, Heckman, and Navarro (2007) for more general models of schooling that make these distinctions explicit.

[48] Absolutely continuous with respect to Lebesgue measure.

[49] Absolutely continuous with respect to Lebesgue measure.

[50] Vytlacil (2006b) shows that their monotonicity and independence conditions imply (and are implied by) a more general version of the ordered choice model with stochastic thresholds, which appears in Heckman, LaLonde, and Smith (1999), Carneiro, Hansen, and Heckman (2003), and Cunha, Heckman, and Navarro (2007) and is analyzed in appendix D.

$$E(Y|Z,W) = \sum_{s=1}^{\bar{S}} E(Y_s|D_s = 1,Z,W)\Pr(D_s = 1|Z,W)$$

$$= \sum_{s=1}^{\bar{S}} \int_{\pi_s(Z,W_s)}^{\pi_{s-1}(Z,W_{s-1})} E(Y|U_D = u_D)\, du_D, \tag{27}$$

where we use the conditional independence assumption OC-1 to obtain the final expression. Analogous to the result for the binary outcome model, we obtain the index sufficiency restriction $E(Y|Z,W) = E(Y|\pi(Z,W))$, where $\pi(Z,W) = [\pi_1(Z,W_1), \ldots, \pi_{\bar{S}-1}(Z,W_{\bar{S}-1})]$. The choice probabilities encode all of the influence of (Z,W) on outcomes.

We can identify $\pi_s(z,w_s)$ for (z,w_s) in the support of the distribution of (Z,W_s) from the relationship $\pi_s(z,w_s) = \Pr(\sum_{j=s+1}^{\bar{S}} D_j = 1|Z = z,\ W_s = w_s)$. Thus $E(Y \mid \pi(Z,W) = \pi)$ is identified for all π in the support of $\pi(Z,W)$. Assumptions OC-1, OC-3, and OC-4 imply that $E(Y|\pi(Z,W) = \pi)$ is differentiable in π. So $\partial E(Y|\pi(Z,\ W) = \pi)/\partial\pi$ is well defined.[51] Thus, analogous to the result obtained in the binary case, we have

$$\frac{\partial E(Y|\pi(Z,W) = \pi)}{\partial \pi_s} = \Delta^{\text{MTE}}_{s,s+1}(U_D = \pi_s) = E(Y_{s+1} - Y_s|U_D = \pi_s). \tag{28}$$

Equation (28) is the basis for identification of the transition-specific MTE from data on (Y,Z,X).

From index sufficiency, we can express equation (27) as

$$E(Y|\pi(Z,W) = \pi) = \sum_{s=1}^{\bar{S}} E(Y_s|\pi_s \le U_D < \pi_{s-1})(\pi_{s-1} - \pi_s)$$

$$= \sum_{s=1}^{\bar{S}-1} [E(Y_{s+1}|\pi_{s+1} \le U_D < \pi_s) - E(Y_s|\pi_s \le U_D < \pi_{s-1})]\pi_s + E(Y_1|\pi_1 \le U_D < 1) \tag{29}$$

$$= \sum_{s=1}^{\bar{S}-1} [m_{s+1}(\pi_{s+1},\pi_s) - m_s(\pi_s,\pi_{s-1})]\pi_s + E(Y_1|\pi_1 \le U_D < 1),$$

where $m_s(\pi_s, \pi_{s-1}) = E(Y_s|\pi_s \le U_D < \pi_{s-1})$. In general this expression is a nonlinear function of (π_s,π_{s-1}). This model has a testable restriction of index sufficiency in the general case: $E(Y|\pi(Z,W) = \pi)$ is a nonlinear function that is additive in functions of (π_s, π_{s-1}), so there are no interactions between π_s and $\pi_{s'}$ if $|s - s'| > 1$, that is,

$$\frac{\partial^2 E(Y|\pi(Z,W) = \pi)}{\partial \pi_S \partial \pi_{s'}} = 0 \qquad \text{if} \quad |s - s'| > 1.$$

Observe that if $U_D \perp\!\!\!\perp U_s$ for $s = 1, \ldots, \bar{S}$,

$$E(Y|\pi(Z,W) = \pi) = \sum_{s=1}^{\bar{S}} E(Y_s)(\pi_{s-1} - \pi_s) = \sum_{s=1}^{\bar{S}-1} [E(Y_{s+1}) - E(Y_s)]\pi_s + E(Y_1).$$

Defining $E(Y_{s+1}) - E(Y_s) = \Delta^{\text{ATE}}_{s,s+1}$, we have $E(Y|\pi(Z,W)=\pi) = \sum_{s=1}^{\bar{S}-1} \Delta^{\text{ATE}}_{s,s+1}\, \pi_s + E(Y_1)$. Thus, under full independence, we obtain linearity of the conditional mean of Y in the π_s's. This result generalizes the test for the presence of essential heterogeneity presented in section IIIA to the ordered case. We can ignore the complexity induced by the model of essential heterogeneity if $E(Y|\pi(Z,W)= \pi)$ is linear in the π's and can use conventional IV estimators to identify well-defined treatment effects.[52]

What Do Instruments Identify in the Ordered Choice Model? We now characterize what scalar instrument $J(Z,W)$ identifies. When Y is log earnings, it is common practice to regress Y on D, the completed years of schooling, and call the coefficient on D a rate of return.[53] We seek an expression for the IV estimator of the effect of D on Y in the ordered choice model:

$$\frac{\text{Cov}(J(Z,W),Y)}{\text{Cov}(J(Z,W),D)}, \tag{30}$$

where $D = \sum_{s=1}^{\bar{S}} sD_s$ the number of years of schooling attainment. We keep the conditioning on X implicit. We now present the weights for IV. Their full derivation is presented in appendix E.

Define $K_s(v) = E(\tilde{J}(Z,W)|\mu_D(Z) - c_s(W_s) > v) \cdot \Pr(\mu_D(Z) - C_s(W) > v)$, where $\tilde{J}(Z,W) = J(Z,W) - E(J(Z,W))$. Thus,

$$\Delta^{\text{IV}}_J = \frac{\text{Cov}(J,Y)}{\text{Cov}(J,D)} = \sum_{s=1}^{\bar{S}-1} \int E(Y_{s+1} - Y_s|V = v)\omega(s,v) f_V(v)\, dv, \tag{31}$$

where

[51] For almost all π that are limit points of the support of distribution of $\pi(Z,W)$. We use the Lebesgue theorem for the derivative of an integral. Under assumption OC-6, all points in the support of the distribution of $\pi(Z,W)$ will be limit points of that support, and we thus have that $\partial E(Y|\pi(Z,\ W) = \pi)/\partial\pi$ is well defined and is identified for (a.e.) π.

[52] Notice that if $U_D \not\perp\!\!\!\perp U_s$ for some s, then we obtain an expression with nonlinearities in π_s, π_{s-1} in equation (29).

[53] Heckman, Lochner, and Todd (2006) present conditions under which this economic interpretation is valid.

$$\omega(s,v) = \frac{K_s(v)}{\sum_{s=1}^{\bar{S}} s \int [K_{s-1}(v) - K_s(v)] f_V(v)\, dv}$$
$$= \frac{K_s(v)}{\sum_{s=1}^{\bar{S}-1} \int K_s(v) f_V(v)\, dv},$$

and clearly $\sum_{s=1}^{\bar{S}-1} \int \omega(s,v) f_V(v)\, dv = 1$, $\omega(0,v) = 0$, and $\omega(\bar{S}, v) = 0$. We can rewrite this result in terms of the MTE, expressed in terms of u_D:

$$\Delta^{\text{MTE}}_{s,s+1}(u_D) = E(Y_{s+1} - Y_s | U_D = u_D),$$

so that

$$\frac{\text{Cov}(J,Y)}{\text{Cov}(J,D)} = \sum_{s=1}^{\bar{S}-1} \int \Delta^{\text{MTE}}_{s,s+1}(u_D)\tilde{\omega}(s,u_D)\, du_D,$$

where

$$\tilde{\omega}(s,u_D) = \frac{\tilde{K}_s(u_D)}{\sum_{s=1}^{\bar{S}} s \int_0^1 [\tilde{K}_{s-1}(u_D) - \tilde{K}_s(u_D)] du_D} \quad (32)$$
$$= \frac{\tilde{K}_s(u_D)}{\sum_{s=1}^{\bar{S}-1} \int_0^1 \tilde{K}_s(u_D) du_D}$$

and

$$\tilde{K}_s(u_D) = E(\tilde{J}(Z,W) | \pi_s(Z,W_s) > u_D) \times \Pr(\pi_s(Z,W_s) \geq u_D). \quad (33)$$

Compare equations (32) and (33) for the ordered choice model with equations (19) and (20) for the binary choice model. The numerator of the weights for the Δ^{MTE} for a particular transition in the ordered choice model is exactly the numerator of the weights implied for the binary choice model, substituting $\pi_s(Z,W_s) = \Pr(D > s | Z,W_s)$ for $P(Z) = \Pr(D = 1|Z)$. The numerator for the weights for IV in the binary choice model is driven by the connection between the instrument and $P(Z)$. The numerator for the weights for IV in the ordered choice model for a particular transition is driven by the connection between the instrument and $\pi_s(Z,W_s)$. The denominator of the weights is the covariance between the instrument and D for both the binary and ordered cases. However, in the binary case the covariance between the instrument and D is completely determined by the covariance between the instrument and $P(Z)$, whereas in the ordered choice case the covariance depends on the relationship between the instrument and the full vector $[\pi_1(Z,W_1), \ldots, \pi_{\bar{S}-1}(Z,W_{\bar{S}-1})]$. Comparing our decomposition of Δ^{IV} with the decomposition (26), we see that ours corresponds to weighting up marginal outcomes across well-defined and adjacent boundary values experienced by agents having their instruments manipulated, whereas the Angrist-Imbens decomposition corresponds to outcomes not experienced by some of the persons whose instruments are being manipulated.

From equation (33), the IV estimator using $J(Z,W)$ as an instrument satisfies the following properties: (a) The numerator of the weights on $\Delta^{\text{MTE}}_{s,s+1}(u_D)$ is nonnegative for all u_D if $E(J(Z,W_s)|\pi_s((Z,W_s) \geq \pi_s)$ is weakly monotonic in π_s. For example, if $\text{Cov}(\pi_s(Z,W_s), D) > 0$, setting $J(Z,W) = \pi_s(Z,W_s)$ will lead to nonnegative weights on $\Delta^{\text{MTE}}_{s,s+1}(u_D)$, though it may lead to negative weights on other transitions. A second property (b) is that the support of the weights on $\Delta^{\text{MTE}}_{s,s+1}$ using $\pi_s(Z,W_s)$ as the instrument is $(\pi_s^{\text{Min}}, \pi_s^{\text{Max}})$, where π_s^{Min} and π_s^{Max} are the minimum and maximum values in the support of $\pi_s(Z,W_s)$, respectively, and the support of the weights on $\Delta^{\text{MTE}}_{s,s+1}$ using any other instrument is a subset of $(\pi_s^{\text{Min}}, \pi_s^{\text{Max}})$. A third property (c) is that the weights on $\Delta^{\text{MTE}}_{s,s+1}$ implied by using $J(Z,W)$ as an instrument are the same as the weights on $\Delta^{\text{MTE}}_{s,s+1}$ implied by using $E(J(Z,W)|\pi_s(Z,W))$ as the instrument.

Suppose that the distributions of W_s, $s = 1, \ldots, \bar{S}$, are degenerate, so that the C_s are constants satisfying $C_1 < \cdots < C_{\bar{S}-1}$. This is the classical ordered choice model. In this case, $\pi_s(Z,W_s) = F_V(\mu_D(Z) - C_s)$ for any $s = 1, \ldots, \bar{S}$. For this special case, using J as an instrument will lead to nonnegative weights on all transitions if $J(Z,W_s)$ is a monotonic function of $\mu_D(Z)$. For example, note that $\mu_D(Z) - C_s > v$ can be written as $\mu_D(Z) > C_s + F_V^{-1}(u_D)$. Using $\mu_D(Z)$ as the instrument leads to weights on $\Delta^{\text{MTE}}_{s,s+1}$ (u_D) of the form specified above with $\tilde{K}_s(u_D) = [E(\mu_D(Z) \mid \mu_D(Z) > F_V^{-1}(u_D) + C_s) - E(\mu_D(Z))] \times \Pr(\mu_D(Z) > F_V^{-1}(u_D) + C_s)$. Clearly, these weights will be nonnegative for all points of evaluation and will be strictly positive for any evaluation point u_D such that $1 > \Pr(\mu_D(Z) > F_V^{-1}(u_D) + C_s) > 0$. We now present some examples of the weights for IV.

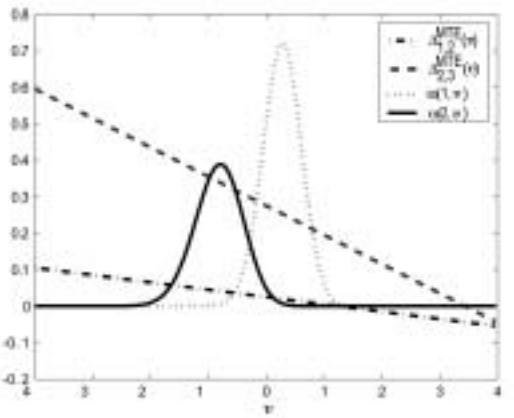
Examples of Weights for IV. Figures 10 and 11 plot the transition-specific MTEs and the IV weights for the models and distributions of the data at the bottom of each of the figures. We work with a normal V and U_s, so we get MTEs linear in V from standard normal regression theory. The IV estimates using Z and W_1 as instruments are reported transition by transition, along with the overall decomposition of the IV representation (31) into its transition-specific components.[54]

[54] In particular, when $J(Z)$ is used as the instrument, we decompose $\Delta^{\text{IV}_{J(Z)}}$ as

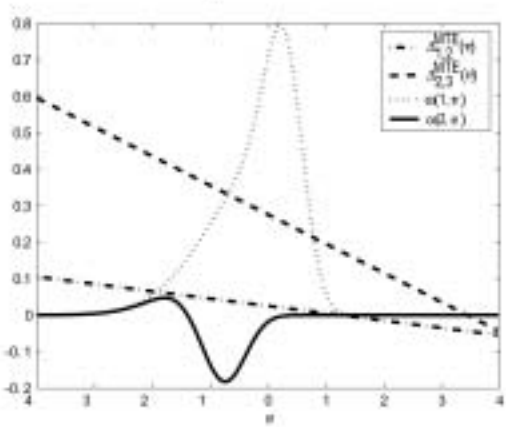
$$\Delta^{\text{IV}_{J(Z)}} = \sum_{s=1}^{\bar{S}-1} \int E(Y_{s+1} - Y_s | V = v)\omega^{J(Z)}(s,v) f_V(v) dv$$
$$= \sum_{s=1}^{\bar{S}-1} \int \Delta^{\text{MTE}}_{s,s+1}(v)\omega^{J(Z)}(s,v) f_V(v)\, dv$$
$$= \sum_{s=1}^{\bar{S}-1} \Delta^{\text{IV}_{J(Z)}}_{s,s+1}.$$

FIGURE 10.—TREATMENT PARAMETERS AND IV:
THE GENERALIZED ORDERED CHOICE ROY MODEL UNDER NORMALITY: CASE I

A. Z as Instrument



B. W_1 as Instrument



Outcomes

$$Y_1 = \alpha + \beta_1 + U_1$$
$$Y_2 = \alpha + \beta_2 + U_2$$
$$Y_3 = \alpha + \beta_3 + U_3$$

Choice Model

$$D_s = \mathbf{1}[W_{s-1} < \gamma Z - V \leqslant W_s]$$
$$s = 1, 2, 3$$

Parameterization

$$(U_1, U_2, U_3, V) \sim N(\mathbf{0}, \mathbf{\Sigma}_{UV}),\ (Z, W_1, W_2) \sim N(\mu_{ZW}, \mathbf{\Sigma}_{ZW}) \text{ and } W_0 = -\infty; W_3 = \infty.$$

$$\mathbf{\Sigma}_{UV} = \begin{bmatrix} 1 & 0.16 & 0.2 & -0.3 \\ 0.16 & 0.64 & 0.16 & -0.32 \\ 0.2 & 0.16 & 1 & -0.4 \\ -0.3 & -0.32 & -0.4 & 1 \end{bmatrix},\ \mu_{ZW} = (-0.6, -1.08, 0.08) \text{ and } \mathbf{\Sigma}_{ZW} = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.1 & -0.09 \\ 0 & -0.09 & 0.25 \end{bmatrix}$$

$$Cov(U_2 - U_1, V) = -0.02 \quad \text{Cov}(U_3 - U_2, V) = -0.08$$
$$\beta_1 = 0;\ \beta_2 = 0.025;\ \beta_3 = 0.3;\ \gamma = 1$$

IV Estimates and Their Components*

Parameter	Value
Δ^{IVz}	0.1489
Δ_{12}^{IVz}	*0.0117*
Δ_{23}^{IVz}	*0.1372*
Δ^{IVw_1}	0.0017
$\Delta_{12}^{IVw_1}$	*0.0325*
$\Delta_{23}^{IVw_1}$	*−0.0308*

Treatment Parameters and Their Values

Parameter	Value
$\text{ATE}_{12} = E(Y_2 - Y_1)$	0.025
$\text{ATE}_{23} = E(Y_3 - Y_2)$	0.275
$\text{TT}_{12} = E(Y_2 - Y_1 \mid D_2 = 1)$	0.0271
$\text{TT}_{23} = E(Y_3 - Y_2 \mid D_3 = 1)$	0.1871
$\text{TUT}_{12} = E(Y_2 - Y_1 \mid D_1 = 1)$	0.0047
$\text{TUT}_{23} = E(Y_3 - Y_2 \mid D_2 = 1)$	0.2854

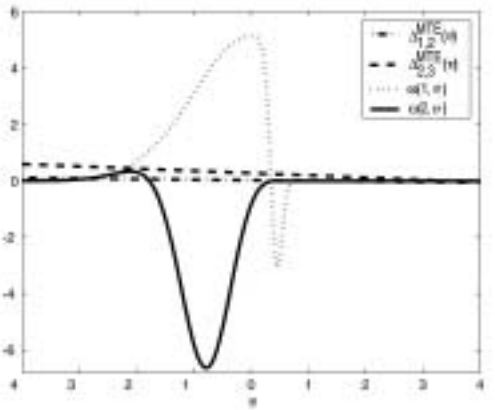
*Δ^{IVz} is decomposed as:

$$\Delta^{IVz} = \int E(Y_2 - Y_1 | V = v)\omega^Z(1, v) f_V(v)\, dv + \int E(Y_3 - Y_2 | V = v)\omega^Z(2, v) f_V(v)\, dv = \Delta_{12}^{IVz} + \Delta_{23}^{IVz}$$

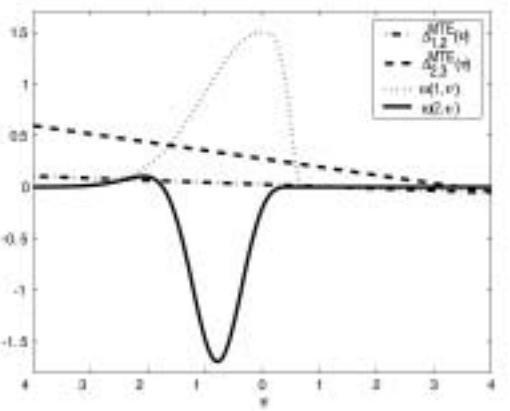
An analogous decomposition applies to Δ^{IVw_1}.

FIGURE 11.—TREATMENT PARAMETERS AND IV:
THE GENERALIZED ORDERED CHOICE ROY MODEL UNDER NORMALITY: CASE II

A. Z as Instrument



B. W_1 as Instrument



Outcomes
$Y_1 = \alpha + \beta_1 + U_1$
$Y_2 = \alpha + \beta_2 + U_2$
$Y_3 = \alpha + \beta_3 + U_3$

Choice Model
$D_s = \mathbf{1}[W_{s-1} < \gamma Z - V \leqslant W_s]$
$s = 1, 2, 3$

Parameterization

$$(U_1, U_2, U_3, V) \sim N(\mathbf{0}, \mathbf{\Sigma}_{UV}), \quad (Z, W_1, W_2) \sim N(\boldsymbol{\mu}_{ZW}, \mathbf{\Sigma}_{ZW}) \quad \text{and} \quad W_0 = -\infty;\ W_3 = \infty.$$

$$\mathbf{\Sigma}_{UV} = \begin{bmatrix} 1 & 0.16 & 0.2 & -0.3 \\ 0.16 & 0.64 & 0.16 & -0.32 \\ 0.2 & 0.16 & 1 & -0.4 \\ -0.3 & -0.32 & -0.4 & 1 \end{bmatrix}, \ \boldsymbol{\mu}_{ZW} = (-0.6, -1.08, 0.08) \text{ and } \mathbf{\Sigma}_{ZW} = \begin{bmatrix} 0.1 & 0.092 & -0.036 \\ 0.092 & 0.1 & -0.09 \\ -0.036 & -0.09 & 0.25 \end{bmatrix}$$

$$\text{Cov}(U_2 - U_1, V) = -0.02 \quad \text{Cov}(U_3 - U_2, V) = -0.08$$
$$\beta_1 = 0;\ \beta_2 = 0.025;\ \beta_3 = 0.3;\ \gamma = 1$$

IV Estimates and Their Components[†]

Parameter	Value
Δ^{IVz}	−1.8091
Δ^{IVz}_{12}	*0.2866*
Δ^{IVz}_{23}	*-2.0957*
Δ^{IVw_1}	−0.4284
$\Delta^{IVw_1}_{12}$	*0.0909*
$\Delta^{IVw_1}_{23}$	*-0.5193*

Treatment Parameters and Their Values

Parameter	Value
$\text{ATE}_{12} = E(Y_2 - Y_1)$	0.025
$\text{ATE}_{23} = E(Y_3 - Y_2)$	0.275
$\text{TT}_{12} = E(Y_2 - Y_1 \mid D_2 = 1)$	0.0283
$\text{TT}_{23} = E(Y_3 - Y_2 \mid D_3 = 1)$	0.1754
$\text{TUT}_{12} = E(Y_2 - Y_1 \mid D_1 = 1)$	0.0025
$\text{TUT}_{23} = E(Y_3 - Y_2 \mid D_2 = 1)$	0.2898

[†]See the footnote below Figure 10 for details of the decomposition of Δ^{IVz}. An analogous decomposition is used for Δ^{IVw_1}.

The IV weights are defined by equations (32) and (33). The bottom table presents the transition-specific treatment parameters.

In figure 10, the IV weights based on Z and W_1 are very different. So, correspondingly, are the IV estimates produced from each instrument, which are far off the mark of the standard treatment parameters shown at the bottom of the table. Observe that the IV weight for W_1 in the second transition is negative for an interval of values. This accounts for the dramatically lower IV estimate based on W_1 as the instrument. Figure 11 shows a different configuration of (Z,W_1,W_2). This produces negative weights for Z for both transitions and a negative weight for W_1 in the second transition. For both instruments, the IV is negative even though both MTEs are positive throughout most of their range. IV provides a misleading summary of the underlying marginal treatment effects. In digesting figures 10 and 11, it is important to recall that both are based on the same structural model. Both have the same MTE and average treatment effects. But the IV estimates are very different, solely as a consequence of the differences in the distributions of instruments across examples.

These simulations show a rich variety of shapes and signs for the weights. They illustrate a main point of this paper—that standard IV methods are not guaranteed to weight marginal treatment effects positively or to produce estimates close to any of the standard treatment effects. Estimators based on LIV and its extension to the ordered model (28) identify Δ^{MTE} for each transition and answer policy-relevant questions. We now turn to development of a more general unordered model.

B. Extension to Multiple Treatments That Are Unordered

In this subsection, we develop a framework for multiple treatments with a choice equation that is based on a nonparametric version of the classical multinomial choice model.[55] Within this framework, treatment effects can be defined as the difference in the counterfactual outcomes that would have been observed if the agent had faced different choice sets, that is, the effect of the individual being forced to choose from one choice set instead of another.

We analyze the return to the agent of choosing between option j and the next best option. The analysis of this case is very similar to the analysis presented in section III in that it converts a multiple choice problem to a binary choice problem. Exclusion restrictions allow analysts to identify generalizations of the LATE parameter and MTE parameters corresponding to the effect of one choice versus the next best alternative. This identification analysis does not require large support assumptions.

Consider the following model with multiple outcome states. Let $\mathcal{J}$ denote the agent's choice set, containing a finite number of elements. The reward (psychic and monetary) of choosing $j \in \mathcal{J}$ is

$$R_j(Z_j) = \vartheta_j(Z_j) - V_j, \tag{34}$$

where Z_j are the agent's observed characteristics that affect the utility from choosing choice j, and V_j is the unobserved shock to the agent's utility from choice j.[56] Let Z denote the random vector containing all unique elements of $\{Z_j\}_{j\in\mathcal{J}}$, that is, Z is the union of $\{Z_j\}_{j\in\mathcal{J}}$. We write $R_j(Z)$ for $R_j(Z_j)$, leaving implicit that $R_j(\cdot)$ only depends on those elements of Z that are contained in Z_j. Let $D_{\mathcal{J},j}$ be an indicator variable for whether the agent would choose option j if confronted with choice set $\mathcal{J}$:[57]

$$D_{\mathcal{J},j} = \begin{cases} 1 & \text{if} \quad R_j \geq R_k \ \ \forall\, k \in \mathcal{J}, \\ 0 & \text{otherwise.} \end{cases}$$

Let $I_{\mathcal{J}}$ denote the choice that would be made by the agent if confronted with the choice set $\mathcal{J}$: $I_{\mathcal{J}} = j \Leftrightarrow D_{\mathcal{J},j} = 1$. Let $Y_{\mathcal{J}}$ be the outcome variable that would be observed if the agent faced the choice set $\mathcal{J}$. It is

$$Y_{\mathcal{J}} = \sum_{j\in\mathcal{J}} D_{\mathcal{J},j} Y_j, \tag{35}$$

where Y_j is the potential outcome, observed only if option j is chosen. This expression generalizes equation (2). We assume that Y_j is determined by $Y_j = \mu_j(X_j, U_j)$, where X_j is a vector of the agent's observed characteristics and U_j is an unobserved random vector. Let X denote the random vector containing all unique elements of $\{X_j\}_{j\in\mathcal{J}}$, that is, X is the union of $\{X_j\}_{j\in\mathcal{J}}$. We assume that $(Z, X, I_{\mathcal{J}}, Y_{\mathcal{J}})$ is observed.[58] Define $R_{\mathcal{J}}$ as the maximum obtainable value given choice set $\mathcal{J}$:

$$R_{\mathcal{J}} = \max_{j\in\mathcal{J}}\{R_j\} = \sum_{j\in\mathcal{J}} D_{\mathcal{J},j} R_j.$$

We obtain the traditional representation of the decision process that if choice j is optimal, choice j is better than the next best option:

$$I_{\mathcal{J}} = j \Leftrightarrow R_j \geq R_{\mathcal{J}\setminus j},$$

where $\mathcal{J}\setminus j$ means $\mathcal{J}$ with the jth element removed. More generally, a choice with $\mathcal{K}$ optimal means that the highest value obtainable from choices in $\mathcal{K}$ is higher than the highest value that can be obtained from choices outside that set:

[55] Heckman and Navarro (2006) and Heckman and Vytlacil (2007a) present a semiparametric analysis of identification for the multinomial choice model.

[56] More consistently with the notation used in the previous section, we could define $R_j(Z_j) = D_j^*$. A more precise, but tedious notation would use $R_j(Z_j, V_j)$, but we use the simpler notation.

[57] We will impose conditions such that ties ($R_j = R_k$ for $j \neq k$) occur with probability 0.

[58] One possible extension is to the case where one does not observe which choice was made, but only whether one particular choice was made, that is, one observes $D_{\mathcal{J},0}$ but not $I_{\mathcal{J}}$. The analysis of Thompson (1989) suggests that this extension should be possible.

$$I_{\mathcal{J}} \in K \Leftrightarrow R_K \geq R_{\mathcal{J}\setminus\mathcal{K}}.$$

As we will show, this well-known representation, used by Lee (1983), Dahl (2002) and others, is key for understanding how nonparametric instrumental variables estimates the effect of a given choice versus the next best alternative.

Analogously to our definition of $R_{\mathcal{J}}$, we define $R_{\mathcal{J}}(z)$ to be the maximum attainable value given the choice set $\mathcal{J}$ when instruments are fixed at $Z = z$,

$$R_{\mathcal{J}}(z) = \max_{j\in\mathcal{J}}\{R_j(z)\}.$$

Thus, for example, a choice from $\mathcal{K}$ is optimal when instruments are fixed at $Z = z$ if $R_{\mathcal{K}}(z) \geq R_{\mathcal{J}\setminus\mathcal{K}}(z)$.

We make the following assumptions, which generalize assumptions A-1 to A-5 for the multiple-treatment case and are presented in a parallel fashion (B-2 is stated below):

B-1: $\{(V_j, U_j)\}_{j\in\mathcal{J}}$ is independent of Z conditional on X.

B-3: The distribution of $(\{V_j\}_{j\in\mathcal{J}})$ is absolutely continuous with respect to Lebesgue measure on $\Pi_{j\in J}\mathbb{R}$.

B-4: $E|Y_j| < \infty$ for all $j \in \mathcal{J}$.

B-5: $\Pr(I_{\mathcal{J}} = j|X) > 0$ for all $j \in \mathcal{J}$.

Assumptions B-1 and B-3 imply that $R_j \neq R_k$ w.p. 1 for $j \neq k$, so that argmax $\{R_j\}$ is unique w.p. 1. Assumption B-4 is required for the mean treatment parameters to be well defined.[59] Assumption B-5 requires that at least some individuals participate in each program for all X.

Definitions of the treatment parameters only require assumptions B-1 and B-3 to B-5. However, we use exclusion restrictions to secure identification. Let $Z^{[j]}$ denote the jth component of Z. Let $Z^{[-j]}$ denote all elements of Z except for the jth. We will work with two alternative assumptions for the exclusion restriction:[60]

B-2a: For each $j \in \mathcal{J}$, their exists at least one element of Z, say $Z^{[j]}$, such that $Z^{[j]}$ is not an element of Z_k, $k \neq j$, and such that the distribution of $\vartheta_j(Z_j)$ conditional on $(X, Z^{[-j]})$ is nondegenerate.

B-2b: For each $j \in \mathcal{J}$, their exists at least one element of Z, say $Z^{[j]}$, such that $Z^{[j]}$ is not an element of Z_k, $k \neq j$, and such that the distribution of $\vartheta_j(Z_j)$ conditional on $(X, Z^{[-j]})$ is absolutely continuous with respect to Lebesgue measure.

Assumption B-2a requires that the analyst be able to independently vary the index for the given value function. It imposes an exclusion restriction, that for any $j \in \mathcal{J}$, Z contains an element such that (i) it is contained in Z_j; (ii) it is not contained in any Z_k for $k \neq j$, and (iii) $\vartheta_j(\cdot)$ is a nontrivial function of that element conditional on all other regressors. Assumption B-2b strengthens B-2a by adding a smoothness assumption. A necessary condition for B-2b is that the excluded variable has a density with respect to Lebesgue measure conditional on all other regressors and for $\vartheta_j(\cdot)$ to be a continuous and nontrivial function of the excluded variable.[61] Assumption B-2a is used to identify a generalization of the LATE parameter. Assumption B-2b will be used to identify a generalization of the MTE parameter. We will strengthen B-2b to a large-support assumption to identify ATE, though that strengthening will not be required for most of our analysis. Assumptions B-2a and B-2b mirror A-2 and are analogous to OC-2 and OC-6 in an ordered choice setting.

Definition of Treatment. Treatment effects are defined as the difference in the counterfactual outcomes that would have been observed if the agent faced different choice sets. For any two choice sets $\mathcal{K}$, $\mathcal{L} \subset \mathcal{J}$, define $\Delta_{\mathcal{K},\mathcal{L}} = Y_{\mathcal{K}} - Y_{\mathcal{L}}$, the effect of the individual being forced to choose from choice set $\mathcal{K}$ rather than choice set $\mathcal{L}$. The conventional treatment effect is defined as the difference in potential outcomes between two specified states,

$$\Delta_{k,\ell} = Y_k - Y_\ell,$$

which is nested within this framework by taking $\mathcal{K} = \{k\}$, $\mathcal{L} = \{\ell\}$. It is the effect for the individual of having no choice except to choose state k rather than having no choice except to choose state ℓ.

$\Delta_{\mathcal{K},\mathcal{L}}$ will be 0 for agents who make the same choice when confronted with choice set $\mathcal{K}$ and choice set $\mathcal{L}$. Thus, $I_{\mathcal{K}} = I_{\mathcal{L}}$ implies $\Delta_{\mathcal{K},\mathcal{L}} = 0$, and we have

$$\begin{aligned}\Delta_{\mathcal{K},\mathcal{L}} &= \mathbf{1}(I_{\mathcal{L}} \neq I_{\mathcal{K}})\Delta_{\mathcal{K}\setminus\mathcal{L},\mathcal{L}} \\ &= \mathbf{1}(I_{\mathcal{L}} \neq I_{\mathcal{K}})\left(\sum_{j\in\mathcal{K}\setminus\mathcal{L}} D_{\mathcal{K},j}\Delta_{j,\mathcal{L}}\right). \qquad (36)\end{aligned}$$

Two cases will be of particular importance for our analysis. First, consider the choice set $\mathcal{K} = \{k\}$ versus choice set $\mathcal{L} = \mathcal{J}\setminus\{k\}$. In this case, $\Delta_{k,\mathcal{J}k}$ is the difference between the agent's potential outcome in state k versus the outcome that would have been observed if he or she had not been allowed to choose state k. If $I_{\mathcal{J}} = k$, then $\Delta_{k,\mathcal{J}k}$ is the difference between the outcome in the agent's preferred state and the outcome in the agent's next best state. Second,

[59] It allows us to integrate to the limit.

[60] We work here with exclusion restrictions for ease of exposition. By adapting the analysis of Cameron and Heckman (1998) and Heckman and Navarro (2006), one can modify our analysis to encompass the case of no exclusion restrictions if Z contains a sufficient number of continuous variables and there is sufficient variation in the ϑ_k function across k.

[61] B-2b can be easily relaxed to the weaker assumption that the support of $\vartheta_j(Z_j)$ conditional on $(X, Z^{[-\ell]})$ contains an open interval, or further weakened to the assumption that the conditional support contains at least one limit point. In these cases, the analysis of this section goes through without change for points within the open interval or more generally for any limit point.

consider the choice set $\mathcal{K} = \mathcal{J}$ versus the choice set $\mathcal{L} = \mathcal{J}\backslash\{k\}$. In this case, $\Delta_{\mathcal{J},\mathcal{J}\backslash k}$ is the difference between the agent's observed outcome and what his or her outcome would have been if state k had not been available. Note that $\Delta_{\mathcal{J},\mathcal{J}\backslash k} = D_{\mathcal{J},k}\Delta_{k,\mathcal{J}\backslash k}$. Thus, there is a trivial connection between the two parameters $\Delta_{\mathcal{J},\mathcal{J}\backslash k}$ and $\Delta_{k,\mathcal{J}\backslash k}$. This paper focuses on $\Delta_{k,\mathcal{J}\backslash k}$, the effect of being forced to choose option k versus being denied option k. However, one can exploit equation (36) to use the results for $\Delta_{k,\mathcal{J}\backslash k}$ to obtain results for $\Delta_{\mathcal{J},\mathcal{J}\backslash k}$.

Treatment Parameters. The conventional definition of the ATE parameter is $\Delta^{\text{ATE}}_{k,\ell}(x,z) = E(\Delta_{k,\ell}|X = x, Z = z)$, which immediately generalizes to the class of parameters just discussed: $\Delta^{\text{ATE}}_{\mathcal{K},\mathcal{L}}(x,z) = E(\Delta_{\mathcal{K},\mathcal{L}}|X = x, Z = z)$. The conventional definition of the parameter for the effect of the treatment on the treated (TT) is $\Delta^{\text{TT}}_{k,\ell}(x,z) = E(\Delta_{k,\ell}|X = x, Z = z, I_{\mathcal{J}} = k)$, which generalizes to $\Delta^{\text{TT}}_{\mathcal{K},\mathcal{J}}(x,z) = E(\Delta_{\mathcal{K},\mathcal{J}}|X = x, Z = z, I_{\mathcal{J}} \in \mathcal{K})$.

We generalize the MTE parameter to be the average effect conditional on being indifferent between the best option among choice set $\mathcal{K}$ versus the best option among choice set $\mathcal{L}$ at some fixed value of the instruments, $Z = z$:

$$\Delta^{\text{MTE}}_{\mathcal{K},\mathcal{L}}(x,z) = E(\Delta_{\mathcal{K},\mathcal{L}}|X = x, Z = z, R_{\mathcal{K}}(z) = R_{\mathcal{L}}(z)). \quad (37)$$

We generalize the LATE parameter to be the average effect for someone for whom the optimal choice in choice set $\mathcal{K}$ is preferred to the optimal choice in choice set $\mathcal{L}$ at $Z = \tilde{z}$, but who prefers the optimal choice in choice set $\mathcal{L}$ to the optimal choice in choice set $\mathcal{K}$ at $Z = z$:

$$\Delta^{\text{LATE}}_{\mathcal{K},\mathcal{L}}(x,z,\tilde{z}) = E(\Delta_{\mathcal{K},\mathcal{L}} \mid X = x, Z \in \{z,\tilde{z}\}, R_{\mathcal{K}}(\tilde{z}) \geq R_{\mathcal{L}}(\tilde{z}), R_{\mathcal{L}}(z) \geq R_{\mathcal{K}}(z)). \quad (38)$$

An important special case of this parameter arises when $z = \tilde{z}$ except for elements that enter the index functions only for choices in $\mathcal{K}$ and not for any choice in $\mathcal{L}$. In that special case, the expression (38) simplifies to

$$\Delta^{\text{LATE}}_{\mathcal{K},\mathcal{L}}(x,z,\tilde{z}) = E(\Delta_{\mathcal{K},\mathcal{L}} \mid X = x, Z \in \{z,\tilde{z}\}, R_{\mathcal{K}}(\tilde{z}) \geq R_{\mathcal{L}}(z) \geq R_{\mathcal{K}}(z))$$

because $R_{\mathcal{L}}(z) = R_{\mathcal{L}}(\tilde{z})$ in this special case.

We have defined each of these parameters as conditional not only on X but also on the "instruments" Z. In general, the parameters will depend on the Z evaluation point. For example, $\Delta^{\text{ATE}}_{\mathcal{K},\mathcal{L}}(x, z)$ will in general depend on the z evaluation point. To see this, note that $Y_{\mathcal{K}} = \sum_{k\in\mathcal{K}} D_{\mathcal{K},k}Y_k$ and $Y_{\mathcal{L}} = \sum_{\ell\in\mathcal{L}} D_{\mathcal{L},\ell}Y_\ell$. By the independence assumption B-1, we have that $Z \perp\!\!\!\perp \{Y_j\}_{j\in\mathcal{J}} \mid X$, but $D_{\mathcal{K},k}$ and $D_{\mathcal{L},\ell}$ will be dependent on Z conditional on X, and thus $Y_{\mathcal{K}} - Y_{\mathcal{L}}$ will in general be dependent on Z conditional on X.[62] In other words, even though Z is conditionally independent of each individual potential outcome, it is correlated with which choice is optimal within the sets $\mathcal{K}$ and $\mathcal{L}$ and thus is related to $Y_{\mathcal{K}} - Y_{\mathcal{L}}$.

Identification: Effect of Option j versus Next Best Alternative. We now establish identification of treatment parameters corresponding to averages of $\Delta_{j,\mathcal{J}\backslash j}$, the effect of choosing option j versus the preferred option in $\mathcal{J}$ if j were not available.[63] Recall that $Z^{[j]}$ is the vector of elements of Z_j that do not enter any other choice index, and that $Z^{[-j]}$ is a vector of all elements of Z not in $Z^{[j]}$. The $Z^{[j]}$ thus act as shifters attracting people into or out of j, but not affecting the valuations in the arguments of the other choice functions. We can develop a parallel analysis to the binary case developed earlier in this paper if we condition on $Z^{[-j]}$. We obtain monotonicity or uniformity in this model if the movements among states induced by $Z^{[j]}$ are the same for all persons conditional on $Z^{[-j]} = z^{[-j]}$ and $X = x$. For example, ceteris paribus, if $Z^{[j]} = z^{[j]}$ increases, then $R_j(Z_j)$ increases but the $R_k(Z_k)$ are not affected, so the flow is toward state j.

Let $D_{\mathcal{J},j}$ be an indicator variable denoting whether option j is selected:

$$\begin{aligned} D_{\mathcal{J},j} &= \mathbf{1}(R_j(Z_j) \geq \max_{\ell\neq j}\{R_\ell(Z_\ell)\}) \\ &= \mathbf{1}(\vartheta_j(Z_j) \geq V_j + \max_{\ell\neq j}\{R_\ell(Z_\ell)\}) \\ &= \mathbf{1}(\vartheta_j(Z_j) \geq \tilde{V}_j), \end{aligned} \quad (39)$$

where $\tilde{V}_j = V_j + \max_{\ell\neq j}\{R_\ell(Z_\ell)\}$. Thus we obtain $D_{\mathcal{J},j} = \mathbf{1}(P_j(Z_j) \geq U_{D_j})$, where $U_{D_j} = F_{\tilde{V}_j}(V_j + \max_{\ell\neq j}\{R_\ell(Z_\ell)\} \mid Z^{[-j]} = z^{[-j]}$, where $F_{\tilde{V}_j}$ is the cdf of $\tilde{V}_j$ given $Z^{[-j]} = z^{[-j]}$. In a format parallel to the binary model, we write

$$Y = D_{\mathcal{J},j}Y_j + (1 - D_{\mathcal{J},j})Y_{\mathcal{J}\backslash j}, \quad (40)$$

where $Y_{\mathcal{J}\backslash j}$ is the outcome that would be observed if option j were not available. This case is just a version of the binary case developed in previous sections of the paper. We can define the MTE as

$$E(Y_j - Y_{\mathcal{J}\backslash j}|X = x, Z = z, \vartheta_j(z_j) - V_j = R_{\mathcal{J}\backslash j}(z)).$$

Recall that we have to condition on $Z = z$ because the choice sets are defined over the maximum of elements in $\mathcal{J}\backslash j$ (see equation (39)).

We now show that our identification strategies presented in the preceding part of this paper extend naturally to the identification of treatment parameters for $\Delta_{j,\mathcal{J}\backslash j}$. In particular, it is possible to recover LATE and MTE parameters for $\Delta_{j,\mathcal{J}\backslash j}$ by use of discrete-change IV methods and LIV methods, respec-

[62] An exception is if $\mathcal{K} = \{k\}$, $\mathcal{L} = \{\ell\}$, that is, both sets are singletons.
[63] Heckman and Vytlacil (2007b) consider the identification of other parameters in the general unordered case.

tively. Averages of the effect of option j versus the next best alternative are the easiest effects to study using IV methods and are natural generalizations of our two-outcome analysis.[64]

Consider identification of treatment parameters corresponding to averages of $\Delta_{j,\mathcal{J}\setminus j}$, either using a discrete-change, Wald form for the IV estimand or using the LIV estimand.[65] The discrete-change IV estimand will allow us to recover a version of the LATE parameter.[66] Let $Z^{[-j]}$ denote the excluded variable for option j with properties assumed in B-2a. We let $z = [z^{[-j]}, z^{[j]}]$ and $\tilde{z} = [\tilde{z}^{[-j]}, \tilde{z}^{[j]}]$ be two values of Z, and we only manipulate $Z^{[j]}$. Define

$$\Delta_j^{\text{Wald}}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]}) = \frac{E(Y|X = x, Z = \tilde{z}) - E(Y|X = x, Z = z)}{\left(\begin{array}{l}\Pr(D_{\mathcal{J},j} = 1|X = x, Z = \tilde{z}) \\ \qquad - \Pr(D_{\mathcal{J},j} = 1|X = x, Z = z)\end{array}\right)},$$

where for notational convenience we assume that $Z^{[j]}$ is the last component of Z. Without loss of generality, we assume that $\vartheta_j(\tilde{z}) > \vartheta_j(z)$. The LIV estimand introduced in Heckman (1997) and developed further in Heckman and Vytlacil (1999, 2001b) allows us to recover a version of the MTE parameter. Impose B-2b, and let $Z^{[j]}$ denote the excluded variable for option j with properties assumed in B-2b. Our results are invariant to which particular variable satisfying B-2b is used if there are more than one variable with those properties. Define

$$\Delta_j^{\text{LIV}}(x,z) \equiv \frac{\dfrac{\partial}{\partial z^{[j]}} E(Y|X = x, Z = z)}{\dfrac{\partial}{\partial z^{[j]}} \Pr(D_{\mathcal{J},j} = 1|X = x, Z = z)}. \tag{41}$$

$\Delta_j^{\text{LIV}}(x, z)$ is thus the limit form of $\Delta_j^{\text{Wald}}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]})$ as $\tilde{z}^{[j]}$ approaches $z^{[j]}$. Given our previous assumptions, one can easily show that this limit exists w.p. 1. We prove the following identification theorem.

Theorem 1.

1. Assume B-1, B-3 to B-5, and B-2a. Then $\Delta_j^{\text{Wald}}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]}) = \Delta_{j,\mathcal{J}\setminus j}^{\text{LATE}}(x, z, \tilde{z})$, where $\tilde{z} = (z^{[-j]}, \tilde{z}^{[j]})$.
2. Assume B-1, B-3 to B-5, and B-2b. Then $\Delta_j^{\text{LIV}}(x, z) = \Delta_{j,\mathcal{J}\setminus j}^{\text{MTE}}(x, z)$.

Proof: See appendix F.

The intuition underlying the proof is simple. Under B-1, B-3 to B-5, and B-2a we can convert the problem of comparing the outcome under j with the outcome under the next best option. This is an IV version of the selection modeling analysis of Dahl (2002). $\Delta_{j,\mathcal{J}\setminus j}^{\text{LATE}}(x, z, \tilde{z})$ is the average effect of switching to state j from state $I_{\mathcal{J}\setminus j}$ for individuals who would choose $I_{\mathcal{J}\setminus j}$ at $Z = z$ but would choose j at $Z = \tilde{z}$. $\Delta_{j,\mathcal{J}\setminus j}^{\text{MTE}}(x, z)$ is the average effect of switching to state j from state $I_{\mathcal{J}\setminus j}$ (the best option besides state j) for individuals who are indifferent between state j and $I_{\mathcal{J}\setminus j}$ at the given values of the selection indices (at $Z = z$, that is, at $\{\vartheta_k(Z_k) = \vartheta_k(z_k)\}_{k \in \mathcal{J}}$).

The mean outcome in state j versus state $I_{\mathcal{J}\setminus j}$ (the next best option) is a weighted average over $k \in \mathcal{J}\setminus j$ of the effect of state j versus state k, conditional on k being the next best option, weighted by the probability that k is the next best option. For example, for the LATE parameter we have

$$\Delta_{j,\mathcal{J}\setminus j}^{\text{LATE}}(x, z, ,\tilde{z}) = E\left(\Delta_{j,\mathcal{J}\setminus j} \middle| \begin{array}{l} X = x, \quad Z \in \{z, \tilde{z}\}, \\ R_j(\tilde{z}) \geq R_{\mathcal{J}\setminus j}(z) \geq R_j(z) \end{array}\right)$$

$$= \sum_{k \in \mathcal{J}\setminus j} \left[\begin{array}{l} \Pr\left(I_{\mathcal{J}\setminus j} = k \middle| \begin{array}{l} Z \in \{z,\tilde{z}\}, \\ R_j(\tilde{z}) \geq R_{\mathcal{J}\setminus j}(z) \geq R_j(z) \end{array}\right) \\ \times E\left(\Delta_{j,k} \middle| \begin{array}{l} X = x, Z \in \{z,\tilde{z}\}, \\ R_j(\tilde{z}) \geq R_{\mathcal{J}\setminus j}(z) \geq R_j(z), \\ I_{\mathcal{J}\setminus j} = k \end{array}\right) \end{array}\right]$$

where we use the fact that $R_{\mathcal{J}\setminus j}(z) = R_{\mathcal{J}\setminus j}(\tilde{z})$ because $z = \tilde{z}$ except for one component, which only enters the index for the jth option. How heavily each option is weighted in this average depends on

$$\Pr(I_{\mathcal{J}\setminus j} = k|Z \in \{z, \tilde{z}\}, R_j(\tilde{z}_j) \geq R_k(z_k) \geq R_j(z_j)),$$

which in turn depends on $\{\vartheta_k(z_k)\}_{k \in \mathcal{J}\setminus j}$. The higher $\vartheta_k(z_k)$, holding the other indices constant, the larger the weight given to state k as the base state.

The LIV and Wald estimands depend on the z evaluation point. Alternatively, one can define averaged versions of the LIV and Wald estimands that will recover averaged versions of the MTE and LATE parameters,[67]

[64] Heckman and Navarro (2006) consider identification of other parameters, but they use identification-at-infinity arguments not required for standard IV. See the comprehensive discussion in Heckman and Vytlacil (2007b).

[65] The estimand is the population version of the estimator.

[66] We use Z directly in the following manipulations instead of manipulating the $\{\vartheta_j(Z_j)\}$ indices. One can modify the following analysis to directly use $\{\vartheta_j(Z_j)\}$, with the disadvantage of requiring identification of $\{\vartheta_j(Z_j)\}$ (for example, by an identification-at-infinity argument) but with the advantage of being able to follow the analysis of Cameron and Heckman (1998), Chen, Heckman, and Vytlacil (1998, 1999), and Heckman and Navarro (2006) in not requiring an exclusion restriction if Z contains a sufficient number of continuous variables and there is sufficient variation in the function ϑ_k across k. See Heckman and Vytlacil (2007b) for a more general analysis.

[67] We assume that the support of $Z^{[-j]}$ conditional on $(\tilde{Z}^{[j]}, X)$ is the same as the support of $Z^{[-j]}$ conditional on $(Z^{[j]}, X)$.

$$\int \Delta_j^{\text{Wald}}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]})\, dF_{Z^{[-j]}}(z^{[-j]})$$

$$= \int \Delta_{j,\mathcal{J}\setminus j}^{\text{LATE}}(x, z, \tilde{z})\, dF_{Z^{[-j]}}(z^{[-j]})$$

$$= E\left(\Delta_{j,\mathcal{J}\setminus j} \,\middle|\, \begin{array}{l} X = x, \\ R_j(Z^{[-j]}, \tilde{z}^{[j]}) \\ \geq R_{\mathcal{J}\setminus j}(Z^{[-j]}) \;\geq\; R_j(Z^{[-j]}, z^{[j]}) \end{array} \right)$$

and

$$\int \Delta_j^{\text{LIV}}(x,z)\, dF_Z(\mathrm{z}) = \int \Delta_{j,\mathcal{J}\setminus j}^{\text{MTE}}(x,z)\, dF_Z(z)$$

$$= E(\Delta_{j,\mathcal{J}\setminus j} | X = x, R_j(Z) = R_{\mathcal{J}\setminus j}(Z)).$$

Thus far, we have only considered identification of LATE and MTE, and not of the more standard treatment parameters ATE and TT. However, following Heckman and Vytlacil (1999), LATE can approximate ATE or TT arbitrarily well given the appropriate support conditions. Theorem 1 shows that we can use Wald estimands to identify LATE for $\Delta_{j,\mathcal{J}\setminus j}$ and we can thus adapt Heckman and Vytlacil (1999) to identify ATE or TT for $\Delta_{j,\mathcal{J}\setminus j}$. With suitable modification of the weights, their analysis, summarized in section III, goes through as before. Suppose that $Z^{[j]}$ satisfies the properties assumed in B-2a, and suppose that: (i) the support of the distribution of $Z^{[j]}$ conditional on all other elements of Z is the full real line; (ii) $\vartheta_j(z_j) \to \infty$ as $z^{[j]} \to \infty$, and $\vartheta_j(z_j) \to -\infty$ as $z^{[j]} \to -\infty$. Then $\Delta_{j,\mathcal{J}\setminus j}^{\text{ATE}}(x, z)$ and $\Delta_j^{\text{LATE}}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]})$ are arbitrarily close when evaluated at a sufficiently large value of $\tilde{z}^{[j]}$ and a sufficiently small value of $z^{[j]}$. Following Heckman and Vytlacil (1999), $\Delta_{j,\mathcal{J}\setminus j}^{\text{TT}}(x, z)$ and $\Delta_j^{\text{LATE}}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]})$ are arbitrarily close for sufficiently small $z^{[j]}$. Our discussion has focused on the Wald estimands. Alternatively we could also follow Heckman and Vytlacil (1999, 2001b, 2005) in expressing ATE and TT as integrated versions of MTE. By theorem 1, we can use LIV to identify MTE and can thus express ATE and TT as integrated versions of the LIV estimand.

For a general instrument $J(Z^{[j]}, Z^{[-j]})$ constructed from $(Z^{[j]}, Z^{[-j]})$, which we denote as $J^{[j]}$, we can obtain a parallel construction to the characterization of standard IV given in equation (18):

$$\Delta_{J^{[j]}}^{\text{IV}} = \int_0^1 \Delta^{\text{MTE}}(x, z, u_{D_j}) \omega_{\text{IV}}^{J^{[j]}}(u_{D_j})\, du_{D_j}, \tag{42}$$

where

$$\omega_{\text{IV}}^{J^{[j]}} = \frac{\begin{array}{c} E[J^{[j]} - E(J^{[j]}) | P_j(Z) \geq u_{D_j}] \\ \times \Pr(P_j(Z) \geq u_{D_j} | Z^{[-j]} = z^{[-j]}) \end{array}}{\text{Cov}(Z^{[j]}, D_{\mathcal{J},j})}, \tag{43}$$

where u_{D_j} is defined at the beginning of this section and where we keep the conditioning on $X = x$ implicit.

Note that from theorem 1 we obtain that

$$\frac{\dfrac{\partial}{\partial z^{[j]}} E[Y | X = x, Z = z]}{\dfrac{\partial P_j(z)}{\partial z^{[j]}}} = \frac{\partial E[Y | X = x, Z = z]}{\partial P_j(z)}$$

$$= E[Y_j - Y_{\mathcal{J}\setminus j} | X = x, Z = z, \vartheta_j(Z_j) - V_j = R_{\mathcal{J}\setminus j}(Z)],$$

so we obtain that LIV identifies MTE and linear IV is a weighted average of LIV with the weights summing to 1. These results mirror the results established in the binary case.

In the literature on the effects of schooling ($S = \sum_{j \in \mathcal{J}} j D_{\mathcal{J},j}$) on earnings ($Y_{\mathcal{J}}$), it is conventional to instrument S. Our Web site presents an analysis of this case. For the general unordered case,

$$\Delta_{J^{[j]}}^{\text{IV}} = \frac{\text{Cov}(J^{[j]}, Y_{\mathcal{J}})}{\text{Cov}(J^{[j]}, S)}$$

can be decomposed into economically interpretable components where the weights can be identified but the objects being weighted cannot be identified using local instrumental variables or LATE without making large support assumptions. However, the components can be identified using a structural model.[68]

The trick we have used in this section, comparing outcomes in j with the next best option, converts a general unordered multiple-outcome model into a two-outcome setup. This effectively partitions $Y_{\mathcal{J}}$ into two components, as in equation (40). Thus we write

$$Y_{\mathcal{J}} = D_{\mathcal{J},j} Y_j + (1 - D_{\mathcal{J},j}) Y_{\mathcal{J}\setminus j}$$

where

$$Y_{\mathcal{J}\setminus j} = \sum_{\substack{\ell \neq j \\ \ell \in \mathcal{J}}} \frac{D_{\mathcal{J},l}}{1 - D_{\mathcal{J},j}} Y_l \times \mathbf{1}(D_{\mathcal{J},j} \neq 1).$$

In the more general unordered case with three or more choices, to analyze IV estimates of the effect of S on $Y_{\mathcal{J}}$, we must work with $Y_{\mathcal{J}} = \sum_{k \in \mathcal{J}} D_{\mathcal{J},k} Y_k$, and make multiple comparisons across potential outcomes. This requires us to

[68] See Heckman and Vytlacil (2007a) and Heckman and Navarro (2006) for analyses of semiparametric identification of structural models that can identify all treatment effects and the components of the IV decompositions. See Heckman and Vytlacil (2007b) and Heckman and Urzua (2006) for further analyses of this case.

move outside the LATE/LIV framework, which is inherently based on binary comparisons.[69] We now consider models that do not impose additive separability in choice equation (13). This includes a general random-coefficient model.

VII. Relaxing Additive Separability in the Choice Equation and Allowing for Random-Coefficient Choice Models

The analysis of this paper and of the entire recent literature on IV estimators for models with essential heterogeneity relies critically on the assumption that the treatment choice equation can be represented in the additively separable form (13). The implied uniformity condition imparts an asymmetry to the entire IV enterprise. Uniformity also underlies conventional selection models.

Responses are permitted to be heterogeneous in a general way, but choices of treatment are not. In the absence of additive separability, or uniformity, the IV identification strategy breaks down. Parameters can be defined as weighted averages of an MTE, but the MTE and the derived parameters cannot be identified using any IV strategy (see Heckman and Vytlacil, 2001b, 2005, 2007b). This point applies to models with two or more potential outcomes. For simplicity of exposition, we only analyze the two-outcome case.

One natural benchmark nonseparable model is a random-coefficient model of choice $D = \mathbf{1}[\gamma Z \geq 0]$, where γ is a random coefficient vector and $\gamma \perp\!\!\!\perp (Z, U_0, U_1)$. If γ is a random coefficient with a nondegenerate distribution and with components that take both positive and negative values, uniformity ("monotonicity") can be violated. Figure 2C illustrates this violation. Uniformity can also be violated if we change one coordinate of Z but fail to control for movements in the other coordinates. See figure 2B.

To consider a more general case, relax the separability assumption of equation (13) to consider the latent choice index

$$D^* = \mu_D(Z, V), \qquad D = \mathbf{1}[D^* \geq 0], \tag{44}$$

where $\mu_D(Z, V)$ is not necessarily additively separable in Z and V, and V is not necessarily a scalar. In the random-coefficient example, $V = \gamma$. We maintain assumptions A-1 to A-5, with A-3 suitably modified for the random-coefficient case.[70]

In the additively separable case, the MTE has three equivalent interpretations: (i) U_D $(= F_V(V))$ is the only unobservable in the first-stage decision rule, and the MTE is the average effect of treatment given the unobserved characteristics in the decision rule $(U_D = u_D)$; (ii) the MTE is the average effect of treatment given that the individual would be indifferent between treatment and no treatment if $P(Z) = u_D$, where $P(Z)$ is a mean utility function; (iii) the MTE is 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 A-1 to A-5, the MTE can be identified by LIV. The MTE does not depend on Z, and hence it is invariant to policies that shift Z. The MTE integrates up to generate all treatment effects, policy effects, and IV estimands.

The three definitions are not the same in the general nonseparable case (44). Heckman and Vytlacil (2001b, 2005, 2007b) extend the MTE to the nonseparable case. LIV is a weighted average of the MTE with possibly negative weights and does not identify the MTE. Thus, if uniformity does not hold, the definition of the MTE allows one to integrate it up to obtain all of the treatment effects. However, the IV estimator does not identify LATE or MTE.

A. Failure of Index Sufficiency in General Nonseparable Models

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 $\mu_D(Z,V)$ is a continuous random variable.[71] Define $\Omega(z) = \{v : \mu_D(z,v) \geq 0\}$. In the additively separable case, $P(z) \equiv \Pr(D = 1 \mid Z = z) = \Pr(V \in \Omega(z))$, $P(z) = P(z') \Leftrightarrow \Omega(z) = \Omega(z')$. This produces index sufficiency, so the propensity score orders the unobservables generating choices. In the more general case (44), it is possible to have (z, z') values such that $P(z) = P(z')$ and $\Omega(z) \neq \Omega(z')$, so index sufficiency does not hold. The Z's enter the model more generally, and the propensity score no longer plays the central role it plays in separable models.

B. The Support of the Propensity Score

The nonseparable model can also restrict the support of $P(Z)$. For example, consider a normal random-coefficient

[69] If we partition $Y_{\mathcal{J}}$ into two components based on general sets $\mathcal{K}$, $\mathcal{L}$, each with two or more elements, the choice equation in general is no longer characterized by the assumption of additive separability in the error, discussed in Heckman and Vytlacil (2005) and in the next section, that is required to justify application of LATE and LIV to identify the MTE. The ordered case previously analyzed has a local property that compares adjacent choices and effectively makes binary comparisons.

[70] In special cases, equation (44) can be expressed in additively separable form. Suppose, for example, that D^* is weakly separable in Z and V; $D^* = \mu_D(\theta(Z), V)$, where $\theta(Z)$ is a scalar function and $\mu_D(\theta(Z), V)$ is strictly increasing in its first argument; and V is a scalar. For any V, then, we can write equation (44) in the same form as equation (13):

$$D = \mathbf{1}(\theta(Z) \geq \tilde{V}),$$

where $\tilde{V} = \mu_D^{-1}(0, V)$ and $\tilde{V} \perp\!\!\!\perp Z \mid X$, and the inverse function is expressed with respect to the first argument. See Vytlacil (2006a), who considers the vector V case. Vytlacil (2002) shows that any model that does not satisfy uniformity (or "monotonicity") will not have a representation in this form. In the random-coefficient case where $Z = (1, Z_1)$ with Z_1 a scalar, and $\gamma = (\gamma_0, \gamma_1)$ if $\gamma_1 > 0$ for all realizations, we can write the choice rule in the form of equation (13): $\gamma_1 Z_1 > -\gamma_0 \Rightarrow Z > -\gamma_0/\gamma_1$ and $V = -\gamma_0/\gamma_1$. However, this trick does not work in the general case.

[71] Absolutely continuous with respect to Lebesgue measure.

choice model with a scalar regressor ($Z = (1, Z_1)$). Assume $\gamma_0 \sim N(0, \sigma_0^2)$, $\gamma_1 \sim N(\bar{\gamma}_1, \sigma_1^2)$, and $\gamma_0 \perp\!\!\!\perp \gamma_1$. Then

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

where Φ is the cumulative distribution of a standard normal. If the support of Z_1 is $\mathbb{R}$, in the standard additive model ($\sigma_1^2 = 0$), $P(z_1)$ has support [0, 1]. When $\sigma_1^2 > 0$, the support is strictly within the unit interval.[72] In the special case when $\sigma_0^2 = 0$, the support is one point ($P(z) = \Phi(\bar{\gamma}_1/\sigma_1)$). We cannot, in general, identify ATE, TT, or any treatment effect requiring the endpoints 0 or 1 using IV or control-function strategies.[73] In addition, the IV weights presented in section III no longer apply. IV now fails as a method for estimating interpretable causal effects and treatment effects. Other approaches to estimation must be adopted if a fully symmetric model of heterogeneity is entertained.

C. Violations of Uniformity

The uniformity or monotonicity assumption can be violated for any vector *Z.* One source of violations is nonseparability between Z and V in equation (44). The random-coefficient model is one intuitive model where separability fails. Even if equation (44) is separable in Z and V, uniformity may fail in the case of vector *Z,* where we use only one function of Z as the instrument and do not condition on the remaining sources of variation in Z—as we demonstrated by examples in section V. If we condition appropriately, we retain monotonicity but get a new form of IV estimator that is sensitive to the specification of the Z not used as an instrument.

VIII. Summary and Conclusions

This paper considers the application of the method of instrumental variables to models where responses to treatment are heterogeneous, agents make treatment choices based in part on this heterogeneity, and some components of heterogeneity are unobserved by the economist. We call this a model with *essential heterogeneity.* Intuitions about IV that are valid for the homogeneous model are often applied inappropriately to the model of essential heterogeneity. In a model with essential heterogeneity, different instruments satisfying the traditional definition of an IV define different economic parameters. This is not the case in the classical IV literature that assumes that responses to treatment are homogeneous. Because different instruments identify different parameters, the traditional emphasis in the econometric theory literature of efficiently combining instruments, or using Durbin-Wu-Hausman tests to check for endogeneity by comparing estimates from different instruments, is inappropriate.

In the model with essential heterogeneity, the specification of the choice equation ($\Pr(D = 1 \mid Z)$) affects the interpretation of any IV estimator. This feature is absent in the classical model, where specification of the full instrument list and choice model is irrelevant to the interpretation of what IV estimates. Two economists using the same valid instrument and the same outcome equations but maintaining different models of economic choice will interpret the same point estimate differently. So will two economists using the same instrument and the same Z variables in $P(Z)$ but using distributions of Z that are different. The agnostic and robust features of IV in its classical setting disappear in a model with essential heterogeneity. We develop a simple procedure that can be applied to test whether, in a given data set, the analyst has to worry about the complications resulting from essential heterogeneity or whether they can be ignored in identifying treatment parameters.

We clarify the concept of monotonicity introduced by Imbens and Angrist (1994) and note that uniformity is a better term for their concept. Additionally, we show that this concept is not the same as that of "monotonicity" used in the literature to define positive IV weights on treatment effects. IV weights can be nonpositive even when uniformity is satisfied for a vector *Z,* if an instrument other than $P(Z)$ (or a function of $P(Z)$) is used. Uniformity plus conditioning on unused instruments is required to produce positive weights in the case of vector *Z.* We demonstrate these points with both theoretical and empirical examples.

Positivity of weights is required to interpret IV estimates as treatment effects. We argue, however, that many interesting policy questions do not require treatment effects. Policy effects and treatment effects are distinct. We develop new software for estimating the MTE and the weights for the two-outcome model.

We also compare the method of IV with the method of control functions. In the more general setting studied here, the method of control functions is explicit in formulating its identifying assumptions and recovers interpretable parameters. We establish a strong relationship between LIV, LATE, and control function models. LIV and LATE estimate the derivatives (differences) of the level functions identified by the control function approach. When we use IV and its extensions to answer the traditional questions addressed by the control function method, the same large-sample support assumptions are required to identify model intercepts.

We highlight the central role of the propensity score in IV and control function methods. Using the propensity score and the distributions of X and Z, we can generate instrument-invariant parameters and weights for any instrument from a common set of parameters. The propensity score or choice probability is more than a computational device, as it is in matching. It shows up as a fundamental feature of both IV

[72] The interval is $[\Phi(-|\gamma_1|/\sigma_1), \Phi(|\gamma_1|/\sigma_1)]$.

[73] The random-coefficient model for choice may explain the support problems for $P(Z)$ noted by many analysts. See Heckman et al. (1998a).

and control function models in the presence of essential heterogeneity.

We develop both ordered and unordered choice models with associated outcomes that extend the binary choice model for essential heterogeneity. The unordered model extends the two-outcome model in a natural way. The ordered model places some special structure on it. In the context of the ordered model, we define transition-specific treatment parameters $\Delta^{\text{MTE}}_{s,s+1}(u)$. We show how to estimate these parameters using transition-specific instruments. These instruments identify parameters that can be linked to specific choice models.

We explain why the model of essential heterogeneity as currently formulated in the recent literature on instrumental variables is asymmetric. It features heterogeneity (nonuniformity) of responses to treatment, but assumes uniformity in response to the variables generating the choice of treatment. We present new results for a random-coefficient model that allows for nonuniformity in responses of choices to instruments and responses of outcomes to treatment.

REFERENCES

Ahn, H., and J. Powell, "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics* 58:1–2 (1993), 3–29.

Angrist, J. D., and G. W. Imbens, "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association* 90:430 (1995), 431–442.

Angrist, J. D., G. W. Imbens, and D. Rubin, "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association* 91:434 (1996), 444–455.

Bjorklund, A., and R. Moffitt, "The Estimation of Wage Gains and Welfare Gains in Self-Selection," this REVIEW, 69:1 (1987), 42–49.

Bresnahan, T. F., "Competition and Collusion in the American Automobile Industry: The 1955 Price War," *Journal of Industrial Economics* 35:4 (1987), 457–482.

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

Card, D., "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica* 69:5 (2001), 1127–1160.

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

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

Chen, X., J. J. Heckman, and E. J. Vytlacil, "Non/Semiparametric Identification and Estimation of a Dynamic Discrete-Time Discrete-Choice Models with Unobserved Heterogeneity," University of Chicago, Department of Economics, working paper (1998).

——— "Identification and Square-Root-*n* Efficient Estimation of Semiparametric Panel Data Models with Binary Dependent Variables and a Latent Factor," University of Chicago, Department of Economics, working paper (1999).

Cunha, F., J. J. Heckman, and S. Navarro, "Separating Uncertainty from Heterogeneity in Life Cycle Earnings, the 2004 Hicks Lecture," *Oxford Economic Papers* 57:2 (2005), 191–261.

——— "Counterfactual Analysis of Inequality and Social Mobility" (chapter 4), in S. L. Morgan, D. B. Grusky, and G. S. Fields (Eds.), *Mobility and Inequality: Frontiers of Research in Sociology and Economics* (Palo Alto: Stanford University Press, 2006), pp. 290–348.

——— "The Identification and Economic Content of Ordered Choice Models with Stochastic Cutoffs," *International Economic Review,* forthcoming (2007).

Dahl, G. B., "Mobility and the Return to Education: Testing a Roy Model with Multiple Markets," *Econometrica* 70:6 (2002), 2367–2420.

Durbin, J., "Errors in Variables," *Review of the International Statistical Institute* 22, (1954) 23–32.

Fan, J., and I. Gijbels, *Local Polynomial Modelling and Its Applications* (New York: Chapman and Hall, 1996).

Gronau, R., "Wage Comparisons—a Selectivity Bias," *Journal of Political Economy* 82:6 (1974), 1119–1943.

Harmon, C., and I. Walker, "The Marginal and Average Returns to Schooling in the UK, *European Economic Review* 43:4–6 (1999), 879–887.

Hausman, J. A., "Specification Tests in Econometrics," *Econometrica* 46:6 (November 1978), 1251–1272.

Heckman, J. J., "Shadow Prices, Market Wages, and Labor Supply," *Econometrica* 42:4 (1974), 679–694.

——— "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement* 5:4 (1976a), 475–492.

——— "Simultaneous Equation Models with Both Continuous and Discrete Endogenous Variables with and without Structural Shift in the Equations," (pp. 235–272) in S. Goldfeld and R. Quandt (Eds.), *Studies in Nonlinear Estimation* (Cambridge, MA: Ballinger Publishing Company, 1976b).

——— "Sample Selection Bias as a Specification Error," *Econometrica* 47:1 (1979), 153–162.

——— "Addendum to Sample Selection Bias as a Specification Error," in E. Stromsdorfer and G. Farkas (Eds.), *Evaluation Studies Review Annual,* Volume 5 (Beverly Hills, CA: Sage Publications, 1980).

——— "Varieties of Selection Bias," *American Economic Review* 80:2 (1990), 313–318.

——— "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources* 32:3 (1997), 441–462. Addendum, 33:1 (1998).

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

Heckman, J. J., and B. E. Honoré, "The Empirical Content of the Roy Model," *Econometrica* 58:5 (1990), 1121–1149.

Heckman, J. J., H. Ichimura, J. Smith, and P. E. Todd, "Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evidence on the Effectiveness of Matching as a Program Evaluation Method," *Proceedings of the National Academy of Sciences* 93:23 (1996), 13416–13420.

——— "Characterizing Selection Bias Using Experimental Data," *Econometrica* 66:5 (1998a), 1017–1098.

Heckman, J. J., H. Ichimura, and P. E. Todd, "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65:223 (1998b), 261–294.

Heckman, J. J., R. J. LaLonde, and J. A. Smith, "The Economics and Econometrics of Active Labor Market Programs" (chapter 31, pp. 1865–2097), in O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics,* Volume 3A (New York: North-Holland, 1999).

Heckman, J. J., L. J. Lochner, and P. E. Todd, "Earnings Equations and Rates of Return: The Mincer Equation and Beyond," in E. A. Hanushek and F. Welch (Eds.), *Handbook of the Economics of Education* (Amsterdam: North-Holland, 2006).

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

——— "Dynamic Discrete Choice and Dynamic Treatment Effects," *Journal of Econometrics,* in press (2006).

Heckman, J. J., and R. Robb, "Alternative Methods for Evaluating the Impact of Interventions" (pp. 156–245), in J. Heckman and B. Singer (Eds.), *Longitudinal Analysis of Labor Market Data,* Volume 10 (New York: Cambridge University Press, 1985).

——— "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes" (pp. 63–107), in H. Wainer (Ed.), *Drawing Inferences from Self-Selected Samples* (New York: Springer-Verlag, 1986; Mahwah, NJ: Lawrence Erlbaum Associates, 2000).

Heckman, J. J., J. L. Tobias, and E. J. Vytlacil, "Simple Estimators for Treatment Parameters in a Latent Variable Framework," this REVIEW, 85:3 (2003), 748–754.
Heckman, J. J., and S. Urzua, "Interpreting IV Estimates of the Effect of Schooling on Earnings," University of Chicago, Department of Economics, manuscript (2006).
Heckman, J. J., and E. J. Vytlacil, "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences* 96 (1999), 4730–4734.
——— "The Relationship between Treatment Parameters within a Latent Variable Framework," *Economics Letters* 66:1 (2000), 33–39.
——— "Instrumental Variables, Selection Models, and Tight Bounds on the Average Treatment Effect" (pp. 1–15), in M. Lechner and F. Pfeiffer (Eds.), *Econometric Evaluation of Labour Market Policies* (New York: Center for European Economic Research, 2001a).
——— "Local Instrumental Variables" (pp. 1–46), in C. Hsiao, K. Morimune, and J. L. Powell (Eds.), *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya* (New York: Cambridge University Press, 2001b).
——— "Policy-Relevant Treatment Effects," *American Economic Review* 91:2 (2001c), 107–111.
——— "Structural Equations, Treatment Effects and Econometric Policy Evaluation," *Econometrica* 73:3 (2005), 669–738.
——— "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation, in J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics,* Volume 6 (Amsterdam: Elsevier, 2007a).
——— "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Economic Estimators to Evaluate Social Programs and to Forecast Their Effects in New Environments," In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics,* Volume 6 (Amsterdam: Elsevier, 2007b).
Imbens, G. W., "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," this REVIEW, 86:1 (2004), 4–29.
Imbens, G. W., and J. D. Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62:2 (1994), 467–475.
Lee, L.-F., "Generalized Econometric Models with Selectivity," *Econometrica* 51:2 (1983), 507–512.
Manski, C. F., and J. V. Pepper, "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica* 68:4, (2000) 997–1010.
Mare, R. D., "Social Background and School Continuation Decisions," *Journal of the American Statistical Association* 75:370 (1980), 295–305.
Powell, J. L., "Estimation of Semiparametric Models" (pp. 2443–2521), in R. Engle and D. McFadden (Eds.), *Handbook of Econometrics,* Volume (Amsterdam: Elsevier, 1994).
Prescott, E. C., and M. Visscher, "Sequential Location among Firms with Foresight," *Bell Journal of Economics* 8:2 (1977), 378–893.
Rosenbaum, P. R., and D. B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 10:1, (1983), 41–55.
Shaked, A., and J. Sutton, "Relaxing Price Competition through Product Differentiation," *Review of Economic Studies* 49:1 (1982), 3–13.
Thompson, T. S., "Identification of Semiparametric Discrete Choice Models," University of Minnesota Center for Economic Research discussion paper, no. 249 (1989).
Vytlacil, E. J., "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica* 70:1 (2002), 331–341.
——— "A Note on Additive Separability and Latent Index Models of Binary Choice: Representation Results," *Oxford Bulletin of Economics and Statistics,* forthcoming (2006a).
——— "Ordered Discrete Choice Selection Models: Equivalence, Nonequivalence, and Representation Results," this REVIEW (2006b).
White, H., *Asymptotic Theory for Econometricians* (Orlando, FL: Academic Press, 1984).
Wu, D., "Alternative Tests of Independence between Stochastic Regressors and Disturbances," *Econometrica* 41:4 (1973), 733–750.
Yitzhaki, S., "On Using Linear Regression in Welfare Economics," Department of Economics, Hebrew University, working paper, no. 217 (1989).
——— "On Using Linear Regressions in Welfare Economics," *Journal of Business and Economic Statistics* 14:4 (1996), 478–486.
Yitzhaki, S., and E. Schechtman, "The Gini Instrumental Variable, or the 'Double Instrumental Variable' estimator," *Metron* 62:3 (2004), 287–313.

APPENDIX A

Deriving the IV Weights on the MTE

We consider instrumental variables conditional on $X = x$ using a general function of Z as an instrument. Let $J(Z)$ be any function of Z such that $\mathrm{Cov}(J(Z), D \mid X = x) \neq 0$. Consider the population analog of the IV estimator,

$$\frac{\mathrm{Cov}(J(Z), Y|X = x)}{\mathrm{Cov}(J(Z), D|X = x)}.$$

First consider the numerator of this expression,

$$\begin{aligned}\mathrm{Cov}(J(Z),Y|X = x) &= E([J(Z) - E(J(Z)|X = x)]Y|X = x)\\ &= E([J(Z) - E(J(Z)|X = x)][Y_0 + D(Y_1 - Y_0)]|X = x)\\ &= E([J(Z) - E(J(Z)|X = x)]D(Y_1 - Y_0)|X = x),\end{aligned}$$

where the second equality comes from substituting in the definition of Y and the third equality follows from the assumption of conditional independence A-2. Define $\tilde{J}(Z) \equiv J(Z) - E(J(Z) \mid X = x)$. Then

$$\begin{aligned}&\mathrm{Cov}(J(Z),Y|X = x)\\ &= E(\tilde{J}(Z)\mathbf{1}[U_D \leq P(Z)](Y_1 - Y_0)|X = x)\\ &= E(\tilde{J}(Z)\mathbf{1}[U_D \leq P(Z)]E(Y_1 - Y_0|X = x, Z, U_D)|X = x)\\ &= E(\tilde{J}(Z)\mathbf{1}[U_D \leq P(Z)]E(Y_1 - Y_0|X = x, U_D)|X = x)\\ &= E(E[\tilde{J}(Z)\mathbf{1}[U_D \leq P(Z)] \mid X = x, U_D]E(Y_1 - Y_0 \mid X = x, U_D) \mid X = x)\\ &= \int\{E(\tilde{J}(Z)|X = x, P(Z) \geq u_D)\ \Pr(P(Z) \geq u_D|X = x)\\ &\quad \times E(Y_1 - Y_0|X = x, U_D = u_D)\}\, du_D\\ &= \int\Delta^{\mathrm{MTE}}(x, u_D)\, E(\tilde{J}(Z) \mid X = x, P(Z) \geq u_D)\ \Pr(P(Z) \geq u_D|X = x)\, du_D,\end{aligned}$$

where the first equality follows from plugging in the model for D; the second equality follows from the law of iterated expectations with the inside expectation conditional on $(X = x, Z, U_D)$; the third equality follows from the conditional independence assumption A-2; the fourth equality follows from Fubini's theorem and the law of iterated expectations with the inside expectation conditional on $(X = x, U_D = u_D)$; the fifth equality follows from the normalization that U_D is distributed uniformly [0, 1] conditional on X; and the final equality follows from plugging in the definition of Δ^{MTE}. Next consider the denominator of the IV estimand. Observe that by iterated expectations

$$\mathrm{Cov}(J(Z),D|X = x) = \mathrm{Cov}(J(Z),P(Z)|X = x).$$

Thus, the population analog of the IV estimator is given by

$$\int \Delta^{\mathrm{MTE}}(x, u_D)\omega(x, u_D)\, du_D, \tag{A1}$$

where

$$\omega(x,u_D) = \frac{E(\tilde{J}(Z)|X = x, P(Z) \geq u_D)\ \Pr(P(Z) \geq u_D|X = x)}{\mathrm{Cov}(J(Z), P(Z)|X = x)}, \tag{A2}$$

where by assumption $\mathrm{Cov}(J(Z), P(Z) \mid X = x) \neq 0$.

If $J(Z)$ and $P(Z)$ are continuous random variables, then a second interpretation of the weight can be derived from equation (A2) by noting that

$$\int [j - E(J(Z)|X = x)] \int_{u_D}^{1} f_{P,J}(t, j|X = x)dt\, dj$$

$$= \int [j - E(J(Z)|X = x)] f_J(j|X = x) \times \int_{u_D}^{1} f_{P|J,X}(t|J(Z) = j, X = x)\, dt\, dj.$$

Write

$$\int_{u_D}^{1} f_{P|J,X}(t|J(Z) = j, X = x)dt = 1 - F_{P|J,X}(u_D|J(Z) = j, X = x) = S_{P|J(Z),X}(u_D|J(Z) = j, X = x),$$

where $S_{P|J,X}(u_D \mid J(Z) = j, X = x)$ is the probability of $P(Z) \geq u_D$ given $J(Z) = j$ and $X = x$. Likewise, $\Pr(P(Z) > U_D \mid J(Z), X) = S_{P|J,X}(U_D \mid J(Z), X)$. Using these results, we may write the weight as

$$\omega(x,u_D) = \frac{\mathrm{Cov}(J(Z), S_{P|J,X}(u_D|J(Z), X = x)|X = x)}{\mathrm{Cov}(J(Z), S_{P|J,X}(U_D|J(Z), X = x)|X = x)}.$$

For fixed u_D and x evaluation points, $S_{P|J,X}(u_D \mid J(Z), X = x)$ is a function of the random variable $J(Z)$. The numerator of the preceding expression is the covariance between $J(Z)$ and the probability that the random variable $P(Z)$ is greater than the evaluation point u_D conditional on $J(Z)$.

For a fixed evaluation point x, $S_{P|J,X}(U_D \mid J(Z), X = x)$ is a given function of the random variables U_D and $J(Z)$. The denominator of the above expression is the covariance between $J(Z)$ and the probability that the random variable $P(Z)$ is greater than the random variable U_D conditional on $J(Z)$ and $X = x$.

Thus, it is clear that if the covariance between $J(Z)$ and the conditional probability that $(P(Z) > u_D)$ given $J(Z)$ is positive for all u_D, then the weights are positive. The condition is trivially satisfied if $J(Z) = P(Z)$, so the weights are positive and IV estimates a gross treatment effect.

If the $J(Z)$ and $P(Z)$ are discrete valued, we obtain the expressions (21) and (22) in the main text.

APPENDIX B

Computational Aspects: Estimating the MTE, the Treatment Parameters, and the Weights

We illustrate the computational aspects of this paper using the linear and separable version of the model of essential heterogeneity introduced in section III. More precisely, we consider the following framework:

$$Y_1 = \alpha + \varphi + \beta_1 X + U_1,$$

$$Y_0 = \alpha + \beta_0 X + U_0,$$

$$I = \gamma Z - V, \tag{B1}$$

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

where (U_0, U_1, V) are independent of Z conditional on X, but U_0, U_1, and V are not independent (even conditioning on X).

Using the same arguments presented in section III, we can show that

$$E(Y|X = x, P(Z) = p) = \alpha + \beta_0 x + [(\beta_1 - \beta_0)x]p + K(p), \tag{B2}$$

where $P(Z)$ represents the propensity score or probability of selection $(\Pr(D = 1|Z))$, p is a particular evaluation value of the propensity score, and

$$K(p) = \varphi p + E(U_1 - U_0|D = 1, P(Z) = p)p. \tag{B3}$$

Equations (B2) and (B3) are closely related to the control function approach (see section IV).

1. The Estimation of the Propensity Score and the Identification of the Relevant Support

The first step in the computation of the MTE is to estimate the probability of participation, or propensity score, $\Pr(D = 1|Z = z) = P(z)$. This probability can be estimated using different methods. In this appendix, we assume $V \sim N(0, 1)$ and thus estimate $P(z)$ using a probit model. Let $\hat{\gamma}$ denote the estimated value of γ in equation (B1). The predicted value of the propensity score (conditional on $Z = z$), $\hat{P}(z)$, is then computed as $\hat{P}(z) = \Pr(\hat{\gamma}Z > V|Z = z) = \Phi(\hat{\gamma}z)$, where Φ represents the cumulative distribution function of a standard normal random variable.

The predicted values of the propensity score allow us to define the values of u_D over which the MTE can be identified. In particular, as shown by Heckman and Vytlacil (2001b), identification of the MTE depends critically on the support of the propensity score.[74] The larger the support of the propensity score, the bigger the set over which the MTE can be identified.

In order to define the relevant support, we first estimate the frequencies of the predicted propensity scores in the samples of treated $(D = 1)$ and untreated $(D = 0)$ individuals. These frequencies can be computed using smoothed sample histograms. In both subsamples the grid Γ of values of $\hat{P}(z)$ specifies the number of points at which the histogram is to be evaluated.

Let $\mathcal{P}_\ell$ denote the set of evaluation points (coming from the grid) such that

$$\mathcal{P}_\ell = \{p \in \Gamma | \epsilon < \Pr(\hat{P}(z) = p|D = \ell)\} \qquad \text{with } \ell = 0, 1 \text{ and } \epsilon > 0,$$

so $\mathcal{P}_\ell$ represents the set of values of p for which we compute frequencies in the range $(\epsilon, 1]$ using the subsample of individuals reporting $D = \ell$ $(\ell = 0, 1)$. Notice that the extreme value 0 is excluded from $\mathcal{P}_\ell$. Finally, if we denote by $\mathcal{P}$ the set of evaluation points used to define the relevant support of the propensity score, we have that

$$\mathcal{P} = \mathcal{P}_0 \cap \mathcal{P}_1 = \{p \in \Gamma | \epsilon < \min(\Pr(\hat{P}(z) = p \mid D = 0), \Pr(\hat{P}(z) = p \mid D = 1))\}$$

for $\epsilon > 0$. Therefore, the MTE is defined only for those evaluations of $\hat{P}(z)$ for which we obtain positive frequencies for both subsamples.

In practice, after identifying the relevant or common support of the propensity score, it is necessary to adjust the sample. In particular, the observations for which $\hat{P}(z)$ is contained in the common support are kept. The rest of the sample is dropped. From this point on, our analysis refers to the resulting sample.

2. Semiparametric Estimation of the Marginal Treatment Effect in Practice

Before presenting the steps used in computing the semiparametric estimate of the MTE, recall equation (16) and make the conditioning on X explicit:

$$\Delta^{\mathrm{LIV}}(x, u_D) = \left.\frac{\partial E(Y|X = x, P(Z) = p)}{\partial p}\right|_{p=u_D} = \Delta^{\mathrm{MTE}}(x, u_D).$$

[74] Heckman et al. (1998a, 1996), and Heckman, Ichimura, and Todd (1998b) also discuss the importance of the propensity score. They present empirical evidence that failure of the full-support condition is a major source of evaluation bias.

This expression indicates that in general the computation of the MTE involves the estimation of the partial derivative of the expectation of the outcome Y (conditional on $X = x$ and $P(Z) = p$) with respect to p. This is the method of local instrumental variables introduced in Heckman and Vytlacil (2001b). However, because we are considering the linear and separable version of the model of essential heterogeneity, we can use equations (B2) and (B3) to show that

$$\left.\frac{\partial E(Y|X = x,\, P(Z) = p)}{\partial p}\right|_{p=u_D} = (\beta_1 - \beta_0)\, x + \left.\frac{\partial K(p)}{\partial p}\right|_{p=u_D}. \qquad \text{(B4)}$$

Thus, in order to compute the MTE we need to estimate values for $\beta_1 - \beta_0$ and $\partial K(p)/\partial p$. Notice that without additional assumptions, the estimation of this last quantity requires the utilization of nonparametric techniques.

Different approaches can be used in the estimation of equation (B4). The following steps describe a semiparametric one.[75]

Step 1. We first estimate the coefficients β_0 and $\beta_1 - \beta_0$ in equation (B2), using a nonparametric version of the double residual regression procedure.[76] In order to do so, we start by fitting a local linear regression (LLR) of each regressor in equation (B2) on the predicted propensity score $\hat{P}(z)$. Notice that if n_X represents the number of variables in X, this step involves the estimation of $2n_X$ local linear regressions. This is because equation (B2) also contains terms of the form $X_k\hat{P}(z)$ for $k = 1, \ldots, n_X$. We use the kth regressor in equation (B2), X_k, to illustrate the LLR procedure. Let $X_k(j)$ and $\hat{P}(z(j))$ denote the values of the kth regressor and the predicted propensity score for the jth individual, respectively, the latter evaluated at the $Z(j)$ that is observed for the individual. The estimation of the LLR of X_k on $\hat{P}(z)$ requires obtaining the values of $\{\theta_0(p), \theta_1(p)\}$ for a set of values of p contained in the support of $\hat{P}(z)$ such that

$$\{\theta_0(p),\theta_1(p)\} = \underset{\{\theta_0,\theta_1\}}{\operatorname{argmin}} \sum_{j=1}^{N} \left\{ \begin{array}{c} (X_k(j) - \theta_0 - \theta_1(\hat{P}(z(j)) - p))^2 \\ \times \Psi((\hat{P}(z(j)) - p)/h) \end{array} \right\},$$

where $\Psi(\cdot)$ and h represent the kernel function and the bandwidth, respectively, and where θ_0 and θ_1 are parameters.[77] In practice, we can use the set of all values of $\hat{P}(z)$ to define the set of evaluation points p in the LLR. This allows us to estimate the predicted value of X_k for each individual in the sample.[78] Let $\hat{X}_k(j)$ denote the predicted value of X_k for the jth individual. This procedure is repeated for each of the $2n_X$ regressors in the outcome equations.

Step 2. Given the predicted values of the $2n_X$ regressors $\hat{X}_k$ ($k = 1, \ldots, 2n_X$), we now generate the residual for each regressor k and person j,

$$\hat{e}_{X_k}(j) = X_k(j) - \hat{X}_k(j) \qquad \text{with } k = 1, \ldots, 2n_X.$$

We denote by $\hat{e}_{X_k}$ the vector of residuals $(\hat{e}_{X_k}(1), \hat{e}_{X_k}(2), \ldots, \hat{e}_{X_k}(N))'$, and by $\hat{e}_X$ the matrix of residuals such that its kth column contains the vector $\hat{e}_{X_k}$.

Step 3. As in the standard double residual regression procedure, we also need to estimate a LLR of Y on $\hat{P}(z)$. The same procedure as the one described in step 1 is used in this case. Let $\hat{Y}(j)$ denote the resulting predicted value of outcome Y for the jth individual.

Step 4. With $\hat{Y}(j)$ in hand, we generate the residual associated with outcome Y for each person j,

$$\hat{e}_Y(j) = Y(j) - \hat{Y}(j).$$

Following the notation used before, we denote by $\hat{e}_Y$ the vector of residuals $(\hat{e}_Y(1), \ldots, \hat{e}_Y(N))'$.

Step 5. Finally, we can estimate the values of β_0 and $\beta_1 - \beta_0$ in equation (B2) from a regression of $\hat{e}_Y$ on $\hat{e}_X$. Specifically,

$$[\hat{\beta}_0, \widehat{\beta_1 - \beta_0}] = [\hat{e}_X'\hat{e}_X]^{-1}[\hat{e}_X'\hat{e}_Y].$$

Heckman et al. (1998a) use a similar double residual regression argument to characterize the selection bias in a semiparametric setup that arises from using nonexperimental data.

Step 6. From equation (B4) we observe that after obtaining the estimated value of $\beta_1 - \beta_0$, only $\partial K(p)/\partial p$ remains to be estimated. However, with the estimated values of β_0 and $\beta_1 - \beta_0$ in hand, this quantity can be estimated using standard nonparametric techniques. To see why, notice that we can write

$$\tilde{Y} = K(\hat{P}(Z)) + \tilde{v}, \qquad \text{(B5)}$$

where $\tilde{Y} = Y - \hat{\beta}_0 X - \widehat{(\beta_1 - \beta_0)} X\, \hat{P}(Z)$ and, as before, we assume $E(\tilde{v}|\hat{P}(z), X) = 0$. Then, it is clear from equation (B5) that the problem reduces to the estimation of $\partial K(\hat{P}(z)) / \partial \hat{P}(z)$, where $K(\hat{P}(z))$ can be interpreted as the conditional expectation $E(\tilde{Y}|P(Z) = \hat{P}(z))$. Let $\vartheta_1(p)$ denote the nonparametric estimator of $\partial K(p)/\partial p$. Notice that we define this estimator as a function of p instead of $\hat{P}(z)$. This is because, unlike the case of the LLR estimators described in step 1, we now use a subset of values of $\hat{P}(z)$ to define the set of points p on which our estimator is evaluated. In particular, we use the set $\mathcal{P}$ to define this set of evaluation points. As shown above, $\mathcal{P}$ contains the values of $\hat{P}(z)$ for which we obtain positive frequencies in both the $D = 0$ and $D = 1$ samples. Thus, $\vartheta_1(p)$ is computed as

$$\{\vartheta_0(p),\vartheta_1(p)\} = \underset{\{\vartheta_0,\vartheta_1\}}{\operatorname{argmin}} \sum_{j=1}^{N} \left\{ \begin{array}{c} (\tilde{Y}(j) - \vartheta_0 - \vartheta_1(\hat{P}(z(j)) - p))^2 \\ \times \Psi((\hat{P}(z(j)) - p)/h) \end{array} \right\}$$

where, as before, $\Psi(\cdot)$ and h represent the kernel function and the bandwidth, respectively.[79]

Step 7. The LIV estimator of the MTE is finally computed as

$$\Delta^{\text{LIV}}(x, u_D) = \widehat{(\beta_1 - \beta_0)}'x + \left.\widehat{\frac{\partial K(p)}{\partial p}}\right|_{p\,=\,u_D} = \widehat{\text{MTE}}(x, u_D)$$

and is evaluated over the set of p's contained in $\mathcal{P}$.

[75] A Fortran code implementing this routine is available at jenni.uchicago.edu/underiv.

[76] In the textbook case $Y = \lambda_1 X_1 + \lambda_2 X_2 + \epsilon$, where ϵ is assumed independent of X_1 and X_2, a double residual regression procedure estimates λ_2 using two stages. In the first stage, the estimated residuals of regressions of Y on X_2 and X_1 on X_2 are computed. Let ϵ_Y and ϵ_{X_1} denote these estimated residuals. In the second stage, λ_2 is estimated from the regression of ϵ_Y on ϵ_{X_1}.

[77] The selection of optimal bandwidth is extensively studied in the nonparametric literature. In the code available on our Web site, two procedures computing optimal bandwidth in the context of local regressions are implemented. The first one is the standard leave-one-out cross-validation procedure. The second is the refined bandwidth selector described in section 4.6 of Fan and Gijbels (1996). Our code allows the utilization of three different kernel functions: Epanechnikov, Gaussian, and biweight.

[78] An alternative could be to use P as the set of evaluation points. In this case, in order to compute the predicted value of X_k for each individual, it would be necessary to replace its value of the predicted propensity scores by the closest value in P.

[79] The code posted on our Web site allows the utilization of local polynomials of higher order to approximate $K(p)$, and so the derivative is computed according to the selected order. It also includes the alternative of using $E(Y|P(Z) = \hat{P}(z))$ to compute a discrete version of the derivative. Furthermore, it allows the estimation of the MTE under the assumption of joint normality of the error terms.

3. The IV Weights

Let J be the instrument. For simplicity we assume that J is a scalar. The extension to the vector case is trivial. Then, as we have shown in equation (19) in the main text, the IV weight is

$$\omega_J(x,u_D) = \frac{\begin{pmatrix} E(J|P(Z) > u_D, X = x) \\ - E(J|X = x) \end{pmatrix} \Pr(P(Z) > u_D|X = x)}{\mathrm{Cov}(J,D|X = x)} \quad \text{(B6)}$$

In order to compute the weight:

Step 1. We approximate $\hat{E}(J|X = x)$ using a linear projection, that is, we assume $J = \lambda' X + V$, where $E(V|X = x) = 0$, so $\hat{E}(J|X = x) = \hat{\lambda}' x$.

Step 2. For each value of u_D we generate the auxiliary indicator function $\mathbf{1}(P(Z) > u_D)$, which is equal to 1 if the argument of the function is true and 0 otherwise.

Step 3. We use linear projections to estimate $E(J|X = x, P(Z) > u_D)$. More precisely, we use OLS to estimate the equation $J(u_D) = \lambda'_{J(u_D)} X + V$ using only the observations for which $\mathbf{1}(P(Z) > u_D) = 1$. Because we assume $E(V|X = x, P(Z) > u_D) = 0$, then $\hat{E}(J|X = x, P(Z) > u_D) = \hat{\lambda}'_{J(u_D)} x$.

Step 4. Because $\Pr(P(Z) > u_D|X = x) = \Pr(\mathbf{1}(P(Z) > u_D) = 1|X = x)$, we use a probit model (for each value of u_D) to estimate this probability. Let $\hat{\Pr}(P(Z) > u_D|X = x)$ denote the estimated probability.

Step 5. We repeat steps 2, 3, and 4 for each value of u_D.

Step 6. With $\hat{E}(J|X = x)$, $\hat{E}(J|X = x, P(Z) > u_D)$, and $\hat{\Pr}(P(Z) > u_D|X = x)$ in hand, we can compute the numerator of equation (B6). In order to get the denominator, we use the fact that

$$\int \omega^J_{\mathrm{IV}}(x, u_D)\, du_D$$

$$= \frac{1}{\mathrm{Cov}(J,D|X = x)}$$

$$\times \int \begin{pmatrix} (E(J|P(Z) > u_D, X = x) - E(J|X = x)) \\ \times \Pr(P(Z) > u_D|X = x) \end{pmatrix} du_D = 1,$$

so with the numerator in hand, it is straightforward to obtain the value of the covariance (conditional on X).

4. The Treatment Parameter Weights

We use the treatment-on-the-treated (TT) parameter to illustrate the computation of the treatment parameter weights. The TT weight is

$$\omega_{\mathrm{TT}}(x,u_D) = \frac{\Pr(P(Z) > u_D|X = x)}{\int \Pr(P(Z) > u_D|X = x)\, du_D},$$

and consequently, we can use $\hat{\Pr}(P(Z) > u_D|X = x)$ to estimate $\omega_{\mathrm{TT}}(x,u_D)$. As in the case of $\omega^J_{\mathrm{IV}}(x, u_D)$, with the estimated value of $\hat{\Pr}(P(Z) > u_D|X = x)$ in hand, we can directly obtain the value for $\int \Pr(P(Z) > u_D|X = x)\, du_D$, using the fact $\int \omega_{\mathrm{TT}}(x,u_D)\, du_D = 1$.

5. The IV and Treatment Parameter Estimators

The MTE and the weights can be used to construct the different estimators. In particular, if $\Delta^{\mathrm{IV}}_J(x)$ denotes the IV estimator obtained by using the instrument J, we know that

$$\Delta^{\mathrm{IV}}_J(x) = \int \mathrm{MTE}(x, u_D)\omega^J_{\mathrm{IV}},(x, u_D)\, du_D.$$

Likewise,

$$\Delta^{\mathrm{TT}}(x) = \int \mathrm{MTE}(x,u_D)\omega_{\mathrm{TT}}(x,u_D)\, du_D,$$

where $\Delta^{\mathrm{TT}}(x)$ represents the TT estimator conditional on $X = x$. Similar expressions exist for the other treatment parameters. Therefore, provided with $\Delta^{\mathrm{LIV}}(x,u_D)$ and the estimated values for the weights, we can compute $\hat{\Delta}^{\mathrm{IV}}_J(x)$ and $\hat{\Delta}^{\mathrm{TT}}(x)$. These estimators depend on the particular value of X considered. In order to compute their unconditional estimated values, we need to integrate X out. More precisely, we need to compute

$$\Delta^{\mathrm{IV}}_J = \int \Delta^{\mathrm{IV}}_J(x)\, dF_X(X)$$

and

$$\Delta^{\mathrm{TT}} = \int \Delta^{\mathrm{TT}}(x)\, dF_{X|D=1}(x).$$

In practice we replace $F_X(\cdot)$ and $F_{X|D=1}(\cdot)$ by their empirical analogs $\hat{F}_X(\cdot)$ and $\hat{F}_{X|D=1}(\cdot)$, leading to

$$\Delta^{\mathrm{IV}}_J = \int \Delta^{\mathrm{IV}}_J(x)\, d\hat{F}_X(x),$$

$$\Delta^{\mathrm{TT}} = \int \Delta^{\mathrm{TT}}(x)\, d\hat{F}_{X|D=1}(x).$$

APPENDIX C

Yitzhaki's Theorem (Yitzhaki, 1989)

Assume (Y, X) i.i.d., $E(|Y|) < \infty$, $E(|X|) < \infty$, $E(Y|X) = g(X)$, $g'(X)$ exists and $E(|g'(x)|) < \infty$.

Let $\mu_Y = E(Y)$ and $\mu_X = E(X)$.

Then

$$\frac{\mathrm{Cov}(Y, X)}{\mathrm{Var}(X)} = \int_{-\infty}^{\infty} g'(t)\omega(t)dt,$$

where

$$\omega(t) = \frac{1}{\mathrm{Var}(X)} \int_t^{\infty} (x - \mu_X) f_X(x)dx$$

$$= \frac{1}{\mathrm{Var}(X)} E(X - \mu_X|X > t) \Pr(X > t).$$

Proof:

$$\mathrm{Cov}(Y, X) = \mathrm{Cov}(E(Y|X), X) = \mathrm{Cov}(g(X), X)$$

$$= \int_{-\infty}^{\infty} g(t)(t - \mu_X) f_X(t)dt.$$

Integration by parts implies that

$$= g(t)\int_{-\infty}^{t}(x-\mu_X)f_X(x)dx\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty}g'(t)\int_{-\infty}^{t}(x-\mu_X)f_X(x)dx\,dt$$

$$= \int_{-\infty}^{\infty}g'(t)\int_{t}^{\infty}(x-\mu_X)f_X(x)dx\,dt,$$

because $E(X-\mu_X)=0$ and the first term in the first expression vanishes.

Therefore,

$$\mathrm{Cov}(Y,X)=\int_{-\infty}^{\infty}g'(t)E(X-\mu_X|X>t)\Pr(X>t)dt.$$

Thus

$$\omega(t)=\frac{1}{\mathrm{Var}(X)}E(X-\mu_X|X>t)\Pr(X>t). \quad \blacksquare$$

Notice that:

(i) The weights are nonnegative ($\omega(t)\geq 0$).
(ii) They integrate to 1 (use integration by parts).
(iii) They are $=0$ at $t=-\infty,\infty$.

We get the formula in the text when in place of X we use $P(Z)$ and the domain of $P(Z)$ is suitably defined.

We apply Yitzhaki's result to the treatment effect model

$$Y=\alpha+\beta D+\epsilon,$$

$$\begin{aligned}E(Y|P(Z)) &= \alpha+E(\beta|D=1,P(Z))P(Z)\\ &= \alpha+E(\beta|P(Z)>u_D,P(Z))P(Z)\\ &= g(P(Z)).\end{aligned}$$

By the law of iterated expectations, we eliminate the conditioning on $D=1$. Using our previous results for OLS, we have

$$\mathrm{IV}=\frac{\mathrm{Cov}(Y,P(Z))}{\mathrm{Cov}(D,P(Z))}=\int g'(t)\omega(t)dt,$$

$$g'(t)=\frac{\partial[E(\beta|D=1,P(Z))]P(Z)}{\partial P(Z)}\Big|_{P(Z)=t},$$

$$\omega(t)=\frac{\int_t^1[\varphi-E(P(Z))]f_P(\varphi)d\varphi}{\mathrm{Cov}(P(Z),D)}.$$

Under assumptions A-2 to A-5 and separability, we have $g'(t)=\Delta^{\mathrm{MTE}}(t)$ but $g'(t)=\mathrm{LIV}$, for $P(Z)$ as an instrument.

APPENDIX D

Generalized Ordered Choice Model with Stochastic Thresholds

The ordered choice model presented in the text with parameterized, but nonstochastic, thresholds is analyzed by Cameron and Heckman (1998), who establish its nonparametric identifiability under the conditions they specify. Treating the W_s (or components of it) as unobservables, we obtain the generalized ordered choice model analyzed in Carneiro, Hansen, and Heckman (2003) and Cunha, Heckman, and Navarro (2007). In this appendix, we present the main properties of this more general model.

The thresholds are now written as $Q_s+C_s(W_s)$ in place of $C_s(W_s)$, where Q_s is a random variable. In addition to the order on the $C_s(W_s)$ in the text, we impose the order $Q_s+C_s(W_s)\geq Q_{s-1}+C_{s-1}(W_{s-1})$, $s=2,\ldots,\bar{S}-1$. We impose the requirement that $Q_{\bar{S}}=\infty$ and $Q_0=-\infty$. The latent index D_s^* is as defined in the text, but now

$$\begin{aligned}D_s &= \mathbf{1}[C_{s-1}(W_{s-1})+Q_{s-1}<\mu_D(Z)-V\leq C_s(W_s)+Q_s]\\ &= \mathbf{1}[\ell_{s-1}(Z,W_{s-1})-Q_{s-1}>V\geq \ell_s(Z,W_s)-Q_s],\end{aligned}$$

where $\ell_s(Z,W_s)=\mu_D(Z)-C_s(W_s)$. Using the fact that $\ell_s(Z,W_s)-Q_s<\ell_{s-1}(Z,W_{s-1})-Q_{s-1}$, we obtain

$$\begin{aligned}&\mathbf{1}[\ell_{s-1}(Z,W_{s-1})-Q_{s-1}>V\geq\ell_s(Z,W_s)-Q_s]\\ &= \mathbf{1}[V+Q_{s-1}<\ell_{s-1}(Z,W_{s-1})]\\ &\quad -\mathbf{1}[V+Q_s\leq\ell_s(Z,W_s)].\end{aligned}$$

The nonparametric identifiability of this choice model is established in Carneiro, Hansen, and Heckman (2003) and Cunha, Heckman, and Navarro (2007). We retain assumptions OC-2 to OC-6, but alter OC-1 to

OC-1′: $(Q_s,U_s,V)\perp\!\!\!\perp(Z,W)|X,\ s=1,\ldots,\bar{S}$.

Vytlacil (2006b) shows that this model with no transition-specific instruments (with W_s degenerate for each s) implies and is implied by the independence and monotonicity conditions of Angrist and Imbens (1995) for an ordered model. Define $Q=(Q_1,\ldots,Q_{\bar{S}})$. Redefine $\pi_s(Z,W_s)=F_{V+Q_s}(\mu_D(Z)+C_s(W_s))$, and define $\pi(Z,W)=[\pi_1(Z,W_1),\ldots,\pi_{\bar{S}-1}(Z,W_{\bar{S}-1})]$. Redefine $U_{D,s}=F_{V+Q_s}(V+Q_s)$. We have that

$$E(Y|Z,W)=E\left(\sum_{s=1}^{\bar{S}}\mathbf{1}\left[\begin{array}{l}\ell_{s-1}(Z,W_{s-1})-Q_{s-1}\\ >V\geq\ell_s(Z,W_s)-Q_s\end{array}\right]Y_s\,\middle|\,Z,W\right)$$

$$= \sum_{s=1}^{\bar{S}}\left(\begin{array}{l}E(\mathbf{1}[V+Q_{s-1}<\ell_{s-1}(Z,W_{s-1})]Y_s|Z,W)\\ -E(\mathbf{1}[V+Q_s\leq\ell_s(Z,W_s)]Y_s|Z,W)\end{array}\right)$$

$$= \sum_{s=1}^{\bar{S}}\left(\begin{array}{l}\displaystyle\int_{-\infty}^{\ell_{s-1}(Z,W_{s-1})}E(Y_s|V+Q_{s-1}=t)\quad dF_{V+Q_{s-1}}(t)\\ \displaystyle-\int_{-\infty}^{\ell_s(Z,W_s)}E(Y_s|V+Q_s=t)\,dF_{V+Q_s}(t)\end{array}\right)$$

$$= \sum_{s=1}^{\bar{S}}\left(\begin{array}{l}\displaystyle\int_{0}^{\pi_{s-1}(Z,W_{s-1})}E(Y_s|U_{D,s-1}=t)\,dt\\ \displaystyle-\int_{0}^{\pi_s(Z,W_s)}E(Y_s|U_{D,s}=t)dt\end{array}\right).$$

We thus have the index sufficiency restriction that $E(Y|Z,W)=E(Y|\pi(Z,W))$, and in the general case $\partial/\partial\pi_s\,[E(Y\mid\pi(Z,W)=\pi)]=E(Y_{s+1}-Y_s\mid U_{D,s}=\pi_s)$. Also, notice that we have the restriction that $\partial^2/\partial\pi_s\,\partial\pi_{s'}\,[E(Y\mid\pi(Z,W)=\pi)]=0$ if $|s-s'|>1$. Under full independence between U_s and $V+Q_s$, $s=1,\ldots,\bar{S}$, we can test full independence for the more general choice model by testing for linearity of $E(Y\mid\pi(Z,W)=\pi)$ in π.

Define

$$\Delta^{\text{MTE}}_{s,s+1}(x,u) = E(Y_{s+1} - Y_s | X = x, U_{D,s} = u),$$

so that our result above can be rewritten as

$$\frac{\partial}{\partial \pi_s} E(Y | \pi(Z,W) = \pi) = \Delta^{\text{MTE}}_{s,s+1}(x, \pi_s).$$

Because $\pi(Z,W_s)$ can be nonparametrically identified immediately from $\pi_s(Z,W_s) = \Pr(\sum_{j=s+1}^{\bar{S}} D_j = 1 | Z, W_s)$ we have that the above offset equality immediately implies identification of MTE for all evaluation points within the appropriate support.

The policy relevant treatment effect is defined analogously. Recall that H^a_s is defined as the cumulative distribution function of $\mu_D(Z) - C_s(W_s)$. We have that

$$\begin{aligned}
E_a(Y_a) &= E_a(E(Y|V,Q,Z,W)) \\
&= E_a\left(\sum_{s=1}^{\bar{S}} \mathbf{1}\left[\begin{matrix} \ell_{s-1}(Z,W_{s-1}) - Q_{s-1} \\ > V \geq \ell_s(Z,W_s) - Q_s \end{matrix} \right] E(Y_s|V,Q,Z,W) \right) \\
&= E_a\left(\sum_{s=1}^{\bar{S}} \mathbf{1}\left[\begin{matrix} \ell_{s-1}(Z,W_{s-1}) - Q_{s-1} \\ > V \geq \ell_s(Z,W_s) - Q_s \end{matrix} \right] E(Y_s|V,Q) \right) \\
&= \sum_{s=1}^{\bar{S}} \left(E_a(E(Y_s|V,Q)\{H^a_s(V+Q_s) - H^a_{s-1}(V+Q_{s-1})\}) \right) \\
&= \sum_{s=1}^{\bar{S}} \int \left(\begin{matrix} E(Y_s|V=v, Q=q) \\ \cdot \{H^a_s(v+q_s) - H^a_{s-1}(v+q_{s-1})\} \end{matrix} \right) dF_{V,Q}(v,q) \\
&= \sum_{s=1}^{\bar{S}} \left(\begin{matrix} \int E(Y_s|V+Q_s=t) H^a_s(t)\, dF_{V+Q_s}(t) \\ - \int E(Y_s|V+Q_{s-1}=t) H^a_{s-1}(t)\, dF_{V+Q_{s-1}}(t) \end{matrix} \right)
\end{aligned}$$

where V, Q_s enter additively, and

$$\begin{aligned}
\Delta^{\text{PRTE}}_{a,a'} &= E_{a'}(Y) - E_a(Y) \\
&= \sum_{s=1}^{\bar{S}-1} \int \left(\begin{matrix} E(Y_{s+1} - Y_s | V + Q_s = t) \\ \cdot \{H^a_s(t) - H^{a'}_s(t)\} \end{matrix} \right) dF_{V+Q_s}(t).
\end{aligned}$$

Alternatively, we can express this result in terms of the MTE,

$$E_a(Y_a) = \sum_{s=1}^{\bar{S}} \left(\begin{matrix} \int E(Y_s|U_{D,s} = t) \tilde{H}^a_s(t)\, dt \\ - \int E(Y_s|U_{D,s-1} = t) \tilde{H}^a_{s-1}(t)\, dt \end{matrix} \right)$$

so that

$$\begin{aligned}
\Delta^{\text{PRTE}}_{a,a'} &= E_{a'}(Y) - E_a(Y) \\
&= \sum_{s=1}^{\bar{S}-1} \int \left(E(Y_{s+1} - Y_s | U_{D,s} = t)\{\tilde{H}^a_s(t) - \tilde{H}^{a'}_s(t)\} \right) dt
\end{aligned}$$

where $\tilde{H}^a_s$ is the cumulative distribution function of the random variable $F_{U_{D,s}}(\mu_D(Z) - C_s(W_s))$.

APPENDIX E

Derivation of the Weights for IV in the Ordered Choice Model

We first derive Cov $(J(Z,W),Y)$. Its derivation is typical of the other terms needed to form equation (30) in the main text. Defining $\tilde{J}(Z,W) = J(Z,W) - E(J(Z,W))$, we obtain, because $\text{Cov}(J(Z,W),Y) = E(\tilde{J}(Z,W)Y)$,

$$\begin{aligned}
&E(\tilde{J}(Z,W)Y) \\
&\quad = E\left[\tilde{J}(Z,W) \sum_{s=1}^{\bar{S}} \mathbf{1}\left[\begin{matrix} \ell_s(Z,W_s) \\ \leq V < \ell_{s-1}(Z,W_{s-1}) \end{matrix} \right] E(Y_s|V,Z,W) \right] \\
&\quad = \sum_{s=1}^{\bar{S}} E\left[\tilde{J}(Z,W) \mathbf{1}\left[\begin{matrix} \ell_s(Z,W_s) \\ \leq V < \ell_{s-1}(Z,W_{s-1}) \end{matrix} \right] E(Y_s|V) \right]
\end{aligned}$$

where the first equality comes from the definition of Y and the law of iterated expectations, and the second equality follows from linearity of expectations and the independence assumption OC-1. Let $H_s(\cdot)$ equal $H^a_s(\cdot)$ for a equal to the policy that characterizes the observed data, that is, $H_s(\cdot)$ is the cumulative distribution function of $\ell_s(Z,W_s)$,

$$\begin{aligned}
H^a_s(t) &= \Pr(\ell_s(Z,W_s) \leq t) \\
&= \Pr(\mu_D(Z) - C_s(W_s) \leq t).
\end{aligned}$$

Using the law of iterated expectations, we obtain

$$\begin{aligned}
E(\tilde{J}(Z,W)Y) &= \sum_{s=1}^{\bar{S}} E\Big[E\Big(\tilde{J}(Z,W)\big(\mathbf{1}[V < \ell_{s-1}(Z,W_{s-1})] \\
&\qquad - \mathbf{1}[V \leq \ell_s(Z,W_s)]\big) | V \Big) E(Y_s|V) \Big] \\
&= \sum_{s=1}^{\bar{S}} \int [E(Y_s|V=v)\{K_{s-1}(v) - K_s(v)\}] f_V(v)\, dv \\
&= \sum_{s=1}^{\bar{S}-1} \int [E(Y_{s+1} - Y_s | V = v) K_s(v)\}] f_V(v)\, dv
\end{aligned}$$

where $K_s(v) = E(\tilde{J}(Z,W) \mid \ell_s(Z,W_s) > v)[1 - H_s(v)]$ and we use the fact that $K_{\bar{S}}(v) = K_0(v) = 0$.

Now consider the denominator of the IV estimand,

$$\begin{aligned}
E(D\tilde{J}(Z,W)) &= E\left(\tilde{J}(Z,W) \sum_{s=1}^{\bar{S}} s\mathbf{1}[\ell_s(Z,W_s) \leq V < \ell_{s-1}(Z,W_{s-1})] \right) \\
&= \sum_{s=1}^{\bar{S}} sE\Big(\tilde{J}(Z,W)\mathbf{1}[\ell_s(Z,W_s) \leq V < \ell_{s-1}(Z,W_{s-1})] \Big) \\
&= \sum_{s=1}^{\bar{S}} sE\Big(E(\tilde{J}(Z,W)\{\mathbf{1}[V < \ell_{s-1}(Z,W_{s-1})] \\
&\qquad - \mathbf{1}[V \leq \ell_s(Z,W_s)]\} \Big| V \Big) \\
&= \sum_{s=1}^{\bar{S}} s \int [K_{s-1}(v) - K_s(v)] f_V(v)\, dv \\
&= \sum_{s=1}^{\bar{S}-1} \int K_s(v) f_V(v)\, dv.
\end{aligned}$$

Collecting results, we obtain an expression for the IV estimand (30):

$$\frac{\mathrm{Cov}(J,Y)}{\mathrm{Cov}(J,D)} = \sum_{s=1}^{\bar{S}-1} \int E(Y_{s+1} - Y_s | V = v)\omega(s,v) f_V(v)\, dv,$$

where

$$\omega(s,v) = \frac{K_s(v)}{\sum_{s=1}^{\bar{S}} \int [K_{s-1}(v) - K_s(v)] f_V(v)\, dv} = \frac{K_s(v)}{\sum_{s=1}^{\bar{S}-1} \int K_s(v) f_V(v)\, dv}$$

and clearly

$$\sum_{s=1}^{\bar{S}-1} \int \omega(s,v) f_V(v)\, dv = 1, \qquad \omega(0,v) = 0, \quad \text{and} \quad \omega(\bar{S},v) = 0.$$

APPENDIX F

Proof of Theorem 1

PROOF: The basic idea is that we can bring the model back to a two-choice setup of j versus the next best option. We prove the result for the second assertion, that $\Delta_j^{\mathrm{LIV}}(x,z)$ recovers the marginal treatment effect parameter. The first assertion, that $\Delta_j^{\mathrm{Wald}}(x, z^{[-j]}, z^{[j]}, \tilde{z}^{[j]})$ recovers a LATE parameter, follows from a trivial modification to the same proof strategy. Recall that $R_{\mathcal{J}\setminus j}(z) = \max_{i \in \mathcal{J}\setminus j} \{R_i(z)\}$ and that $I_{\mathcal{J}\setminus j} = \mathrm{argmax}_{i \in \mathcal{J}\setminus j}(R_i(Z))$. We may write $Y = Y_{\mathcal{J}\setminus j} + D_{\mathcal{J},j}(Y_j - Y_{\mathcal{J}\setminus j})$. We have

$$\begin{aligned}\Pr(D_{\mathcal{J},j} = 1 | X = x, Z = z) &= \Pr(R_j(z_j) > R_{\mathcal{J}\setminus j}(z) | X = x, Z = z) \\ &= \Pr(\vartheta_j(z_j) \geq R_{\mathcal{J}\setminus j}(z) - V_j | X = x, Z = z).\end{aligned}$$

Using the independence assumption B-1, $R_{\mathcal{J}\setminus j}(z) - V_j$ is independent of Z conditional on X, so that

$$\Pr(D_{\mathcal{J},j} = 1 | X = x, Z = z) = \Pr(\vartheta_j(z_j) \geq R_{\mathcal{J}\setminus j}(z) - V_j | X = x).$$

$\vartheta_k(\cdot)$ does not depend on $z^{[j]}$ for $k \neq j$ by assumption B-2b, and thus $R_{\mathcal{J}\setminus j}(z)$ does not depend on $z^{[j]}$, and we will therefore (with an abuse of notation) write $R_{\mathcal{J}\setminus j}(z^{[-j]})$ for $R_{\mathcal{J}\setminus j}(z)$. Write $F(\cdot\,; x, z^{[-j]})$ for the distribution function of $R_{\mathcal{J}\setminus j}(z^{[-j]}) - V_j$ conditional on $X = x$. Then

$$\Pr(D_{\mathcal{J},j} = 1 | X = x, Z = z) = F(\vartheta_j(z_j); x, z^{[-j]})$$

and

$$\frac{\partial}{\partial z^{[j]}} \Pr(D_{\mathcal{J},j} = 1 | X = x, Z = z) = \left[\frac{\partial}{\partial z^{[j]}} \vartheta_j(z_j)\right] f(\vartheta_j(z_j); x, z^{[-j]}),$$

where $f(\cdot\,; x, z^{[-j]})$ is the density of $R_{\mathcal{J}\setminus j}(z^{[-j]}) - V_j$ conditional on $X = x$. Consider

$$\begin{aligned}&E(Y | X = x, Z = z) \\ &\quad = E(Y_{\mathcal{J}\setminus j} | X = x, Z = z) + E(D_{\mathcal{J},j}(Y_j - Y_{\mathcal{J}\setminus j}) | X = x, Z = z).\end{aligned}$$

As a consequence of B-1, B-2b, B-3, and B-4, we have that $E(Y_{\mathcal{J}\setminus j} | X = x, Z = z)$ does not depend on $z^{[j]}$. Using the assumptions and the law of iterated expectations, we may write

$$\begin{aligned}&E(D_{\mathcal{J},j}(Y_j - Y_{\mathcal{J}\setminus j}) | X = x, Z = z) \\ &\quad = \int_{-\infty}^{\vartheta_j(z)} E\left(Y_j - Y_{\mathcal{J}\setminus j} \,\middle|\, \begin{matrix} X = x, Z = z, \\ R_{\mathcal{J}\setminus j}(z^{[-j]} - V_j = t) \end{matrix}\right) f(t; x, z^{[-j]})\, dt \\ &\quad = \int_{-\infty}^{\vartheta_j(z)} E\left(Y_j - Y_{\mathcal{J}\setminus j} \,\middle|\, \begin{matrix} X = x, Z^{[-j]} = z^{[-j]}, \\ R_{\mathcal{J}\setminus j}(z^{[-j]}) - V_j = t \end{matrix}\right) f(t; x, z^{[-j]})\, dt.\end{aligned}$$

Thus,

$$\begin{aligned}&\frac{\partial}{\partial z^{[j]}} E(Y | X = x, Z = z) \\ &= E\left(Y_j - Y_{\mathcal{J}\setminus j} \,\middle|\, \begin{matrix} X = x, Z^{[-j]} = z^{[-j]}, \\ R_j(z) = R_{\mathcal{J}\setminus j}(z) \end{matrix}\right) \\ &\quad \times \left[\frac{\partial}{\partial z_j^{[j]}} \vartheta_j(z_j)\right] f(\vartheta_j(z_j)).\end{aligned}$$

Combining results, we have

$$\begin{aligned}&\frac{\dfrac{\partial}{\partial z^{[j]}} E(Y | X = x, Z = z)}{\dfrac{\partial}{\partial z^{[j]}} \Pr(D_{\mathcal{J},j} = 1 | X = x, Z = z)} \\ &\quad = E(Y_j - Y_{\mathcal{J}\setminus j} | X = x, Z^{[-j]} = z^{[-j]}, R_j(z) = R_{\mathcal{J}\setminus j}(z)).\end{aligned}$$

Finally, noting that

$$\begin{aligned}&E(Y_j - Y_{\mathcal{J}\setminus j} | X = x, Z^{[-j]} = z^{[-j]}, R_j(z) = R_{\mathcal{J}\setminus j}(z)) \\ &\quad = E(Y_j - Y_{\mathcal{J}\setminus j} | X = x, Z = z, R_j(z) = R_{\mathcal{J}\setminus j}(z))\end{aligned}$$

provides the stated result. The proof for the LATE result follows from a parallel argument using discrete changes in the instrument. ■