Web Appendix for Building Bridges Between Structural and Reduced Form Approaches to Evaluating Policy

James J. Heckman*

University of Chicago,

University College Dublin,

Cowles Foundation,

Yale University,

and the American Bar Foundation

This draft: March 20, 2010

*Department of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637, USA; tel.: 773-702-3478; fax: 773-702-8490; e-mail: jjh@uchicago.edu. I thank Steve Durlauf, Lars Hansen and Ed Vytlacil for useful conversations over the years on the issues addressed in this paper although none have read this version. Heckman thanks the National Institutes of Health (R01-HD054702), the JB and MK Pritzker Family Foundation, the Buffett Early Childhood Fund, the American Bar Foundation, and the Committee for Economic Development with a grant from The Pew Charitable Trusts and the Partnership for America's Economic Success. Heckman also thanks the Cowles Foundation at Yale University, which supported a visit that facilitated completion of this research.

1 IV in Choice Models

We adjoin a choice equation to outcome equations (**??**) and (**??**). A standard binary threshold crossing model for D writes

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

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

A familiar case is

$$D^* = \gamma Z - V \tag{2}$$

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

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

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 in outcomes (**??**) and in costs $C = \mu_C(W) + U_C$, $Z = (X, W)$, $\mu_D(Z) = \mu_1(X) - \mu_0(X) - \mu_C(W)$, $V = -(U_1 - U_0 - U_C)$. In constructing examples, we use a special version where $U_C = 0$. We call this version the **extended Roy model**.[1] Our analysis, however, applies to more general models.

In the case where β (given X) is a constant under **??** and **??**, it is not necessary to specify the choice model to identify β. We show that in a general model with heterogenous 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

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

Thus even if Z and η are independent, they are not independent conditional on $D = \mathbf{1}[\gamma Z > V]$

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

if η $(= U_1 - U_0)$ is dependent on V (i.e., if the decision maker has partial knowledge of η and acts on it). Selection models allow for this dependence (see Ahn and Powell (1993); Heckman and Robb (1985, 1986); Powell (1994)). Keeping X implicit and assuming that

$$(U_1, U_0, V) \perp\!\!\!\perp Z \tag{3}$$

(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

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

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

Similarly,

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

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 the 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 η $(= \beta - \bar{\beta})$ is independent of V, only one instrument from vector Z needs to be used. Missing instruments play no role in identifying 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 heterogenous β. We now review some

results established in the preceding literature that form the platform on which we build.

1.1 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 (**??**) and allows for nonseparable errors:

$$\begin{aligned} Y_1 &= \mu_1(X, U_1), \\ Y_0 &= \mu_0(X, U_0), \end{aligned} \tag{4}$$

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.[2] The observed outcome is produced by equation (**??**).

Choices are generated by a standard discrete choice model. We generalize choice model (1) and (2) for D^*, a latent utility,[3]

$$D^* = \mu_D(Z) - V \text{ and } D = \mathbf{1}[D^* \geq 0]. \tag{5}$$

$\mu_D(Z) - V$ 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; $D = 0$ otherwise. Section **??** discusses the important role played by additive separability in the recent instrumental variable literature on essential heterogeneity.

In terms of the notation used in Section **??**, $\beta = Y_1 - Y_0 = \mu_1(X, U_1) - \mu_0(X, U_0)$. A special case that links our analysis to standard models in econometrics writes $Y_1 = X\beta_1 + U_1$ 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.[4]

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).***

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

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

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

(A-2) *The distribution of* $\mu_D(Z)$ *conditional on* X *is nondegenerate* ***(Rank Condition for IV)***.[5]

(A-3) *The distribution of* V *is continuous.*[6]

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

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

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

(A-1) and (A-2) generalize **??** and **??** 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 be 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 does 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 a set of outcomes corresponding to each X_d state.

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 monotonicity (uniformity) condition **??** from choice equation (5). Fixing Z at two different values moves $D(Z)$ in the same direction for everyone. Vytlacil (2002) shows that under independence, rank and some regularity conditions, monotonicity **??** implies the existence of a V in representation (5). Thus the IV model for the general case and the economic choice model turn out to have *identical* representations. 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 (1) and (2), we may

[5] $\mu_D(\cdot)$ is assumed to be a measurable function of Z given X.
[6] The distribution is absolutely continuous with respect to Lebesgue measure.

write the model for D using the distribution of V, F_V, as

$$D = \mathbf{1}\left[F_V\left(\mu_D\left(Z\right)\right) > F_V\left(V\right)\right] = \mathbf{1}\left[P\left(Z\right) > U_D\right], \tag{6}$$

where $U_D = F_V\left(V\right)$ and $P\left(Z\right) = F_V\left(\mu_D(Z)\right) = \Pr(D = 1 \mid Z)$, 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 ($U_D \sim Unif[0,1]$).

1.1.1 LATE, 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 marginal treatment effect (or MTE) conditional on X and U_D:[7]

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

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. Since $P\left(Z\right)$ is a monotonic transformation of the mean net utility $\mu_D\left(Z\right)$, and U_D is a monotonic function of V, when we evaluate $\Delta^{\text{MTE}}\left(u_D\right)$ at the value $P(z) = u_D$, it is the marginal return to agents with $Z = z$ characteristics who are just indifferent between sector 1 and sector 0. In other words, at this point of evaluation, $\Delta^{\text{MTE}}\left(u_D\right)$ 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 $\mu_D\left(z\right) = v$, so $\mu_D\left(z\right) - v = 0$ which is equivalent to the event that $P\left(z\right) = F_V\left(\mu_D\left(z\right)\right) = F_V\left(v\right) = u_D$. 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

[7] 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).

parameters, matching estimators, IV estimators based on $J(Z)$, a scalar function of Z, and OLS estimators can be written as weighted averages of the MTE. Tables 1A and 1B summarize 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 at our website, where software for doing so is available.[8] Heckman and Vytlacil (2001b, 2007b) show that these weights can be constructed and the relationships among the parameters shown in Tables 1A and 1B hold even if a nonseparable choice model, instead of (5), is used and even if assumption (A-2) is weakened. We discuss this result in Section **??**.

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 one. Thus, $E(Y_1 - Y_0 \mid X = x) = \text{ATE} = E(Y_1 - Y_0 \mid X = x, D = 1) = \text{TT} = \text{MTE}$, 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 **??**, let $\mathcal{Z}(x)$ denote the support of the distribution of Z conditional on $X = x$. For any $(z, z') \in \mathcal{Z}(x) \times \mathcal{Z}(x)$ so that $P(z) > P(z')$, under **??** and independence (A-1), LATE is:

$$\Delta^{\text{LATE}}\left(z', z\right) = E\left(Y_1 - Y_0 \mid D(z) = 1, D\left(z'\right) = 0\right), \tag{7a}$$

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

$$\begin{aligned} & E\left(Y_1 - Y_0 \mid D(z) = 1, D(z') = 0\right) \\ = \; & E\left(Y_1 - Y_0 \mid u'_D < U_D < u_D\right) \\ = \; & \Delta^{\text{LATE}}(u_D, u'_D)' \end{aligned} \tag{7b}$$

where $u_D = \Pr(D(z) = 1) = \Pr(D = 1 \mid Z = z) = P(z)$, $u'_D = \Pr(D(z') = 1 \mid Z = z') = \Pr(D(z') = 1) = P(z')$.[9] In the limit, as $u'_D \to u_D$, LATE converges to MTE.

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

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

Imbens and Angrist (1994) define the LATE parameter from hypothetical manipulations of an instrument. Heckman 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 definition (7b) 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).[10] 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)

$$E(Y \mid Z = z) = E(Y \mid P(Z) = p)$$

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

As a consequence,

$$\frac{\partial}{\partial p} E(Y \mid P(Z) = p)\Big|_{P(z)=p} = E(Y_1 - Y_0 | U_D = p). \tag{8}$$

Expression (8) shows how the derivative of $E(Y \mid Z = z)$, which is the local instrumental variable (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 MTE at a value of $U_D = u_D$ that is high, i.e. 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 6). Thus high p values identify returns to persons whose unobservables make them *less* likely to participate in the program. Software for estimating MTE using local linear regression is described in Appendix **??** and is available online at jenni.uchicago.edu/underiv.

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

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

$$E(Y \mid Z) = a + bP(Z), \tag{9}$$

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 Tables 1A and 1B, 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.[11]

1.1.2 Understanding What IV Estimates

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

$$\Delta_J^{\text{IV}} = \int_0^1 \Delta^{\text{MTE}}(u_D)\, \omega_{\text{IV}}^J(u_D)\, du_D, \tag{10}$$

where

$$\omega_{\text{IV}}^J(u_D) = \frac{E(J(Z) - E(J(Z)) \mid P(Z) > u_D) \Pr(P(Z) > u_D)}{\text{Cov}(J(Z), D)}. \tag{11}$$

In this expression u_D is a number between zero and one. This weight depends on the choice probability $P(Z)$. For a derivation see Appendix 1.3. 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.[12]

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

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

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

case after developing the continuous case. The weights defined in (11) can be written as

$$\omega_{\mathrm{IV}}^{J}(u_D) = \frac{\int (j - E(J(Z))) \int_{u_D}^{1} f_{J,P}(j,t)\, dt\, dj}{\mathrm{Cov}(J(Z), D)}, \tag{12}$$

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,[13] 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 that $J(Z)$ is a monotonic function of 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 non-negative because from (11) $E(P(Z) \mid P(Z) > u_D) - E(P(Z)) \geq 0$. 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, weight (11) 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.[14]

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 **??**.

As is evident from Tables 1A and 1B, the weights on $\Delta^{\mathrm{MTE}}(u_D)$ generating Δ^{IV} are different from the weights on $\Delta^{\mathrm{MTE}}(u_D)$ that generate the average treatment effect which is widely regarded as an important policy parameter (see, e.g. Imbens, 2004) or from the weights associated with the policy relevant treatment parameter which answers well-posed policy questions (Heckman and

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

[14] If it is weakly monotonically increasing, the claim is evident from (11). If it is decreasing, the sign of the numerator and the denominator are both negative so the weight is nonnegative.

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. Since 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 $(p_\ell, p_{\ell+1})$ interval, and $\Pr(P(Z) \geq u_D)$ is constant in u_D for u_D within any $(p_\ell, p_{\ell+1})$ interval, and thus $\omega_{\text{IV}}^J(u_D)$ is constant in u_D over any $(p_\ell, p_{\ell+1})$ interval. Let λ_ℓ denote the weight on the LATE for the interval $(\ell, \ell+1)$. In this notation,

$$\begin{aligned}
\Delta_J^{\text{IV}} &= \int E(Y_1 - Y_0|U_D = u_D)\omega_{\text{IV}}^J(u_D)\, du_D \qquad (13)\\
&= \sum_{\ell=1}^{K-1} \lambda_\ell \int_{p_\ell}^{p_{\ell+1}} E(Y_1 - Y_0|U_D = u_D)\frac{1}{(p_{\ell+1} - p_\ell)} du_D \\
&= \sum_{\ell=1}^{K-1} \Delta^{\text{LATE}}(p_\ell, p_{\ell+1})\lambda_\ell.
\end{aligned}$$

Let j_i be the i^{th} smallest value of the support of $J(Z)$. The discrete version of (11) is

$$\lambda_\ell = \frac{\sum_{i=1}^{I}(j_i - E(J)) \sum_{t>\ell}^{K}(f(j_i, p_t))}{\text{Cov}(J(Z), D)}(p_{\ell+1} - p_\ell) \qquad (14)$$

where $f(j_i, p_t)$ is the probability frequency of (j_i, p_t): the probability that $J(Z) = j_i$ and $P(Z) = p_t$. There is no presumption 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.[15] Our expression for the weight on 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 $P(Z)$.[16] 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 regression function $E(Y \mid P(Z) = p)$ that a linear regression of Y on P estimates

$$\beta_{Y,P} = \int_0^1 \left[\frac{\partial E(Y \mid P(Z) = p)}{\partial p}\right] \omega(p)\, dp, \tag{15}$$

where

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

which is exactly the weight (11) when P is the instrument. Thus we can interpret (11) as the weight on $\frac{\partial E(Y|P(Z)=p)}{\partial p}$ 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, $\left.\frac{\partial E(Y|P(Z)=p)}{\partial p}\right|_{p=u_D} = E(Y_1 - Y_0 \mid U_D = u_D) = \Delta^{\mathrm{MTE}}(u_D)$.[17] We discuss Yitzhaki's derivation which is an argument based on integration by parts in Appendix **??**. Our analysis is more general than that of Yitzhaki (1989), Imbens and Angrist (1994), or Angrist and Imbens (1995) because 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) because he does not impose uniformity (monotonicity).

[15] 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.

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

[17] Yitzhaki's weights are used by Angrist and Imbens (1995) to interpret what TSLS estimates in the model of equation (15). Yitzhaki (1989) derives the finite sample weights used by Imbens and Angrist (See his paper posted at our website). See also the refinement in Yitzhaki and Schechtman (2004).

Our simple test for the absence of general heterogeneity based on the linearity of Y in $P(Z)$ (based on equation 12) 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.[18]

1.1.3 The Central Role of the Propensity Score

Observe that both (11) and (12) (and their counterparts for LATE (13) and (14)) 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 1. In matching models, it provides a computationally convenient way to condition on Z (see, e.g. Heckman and Navarro, 2004; Rosenbaum and Rubin, 1983). For the IV weight to be correctly constructed and interpreted, we need to know the correct model for $P(Z)$, i.e., 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 with $P(Z)$, the true propensity score or choice probability and (c) the specification of the propensity score (i.e., 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.

1.1.4 Monotonicity, Uniformity and Conditional Instruments

Monotonicity or uniformity condition **??**, 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 while choice equations

[18]Note that it is possible that $E(Y \mid Z)$ 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).

cannot be. If $\mu_D(Z) = \gamma Z$, where γ is a common coefficient shared by everyone, the choice model satisfies the uniformity property. On the other hand, if γ is a random coefficient (i.e., has a nondegenerate distribution) that can take both negative and positive values, and there are two or more variables in Z with nondegenerate γ coefficients, uniformity can be violated. Different people can respond to changes in Z differently, so there is non-uniformity. 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.[19]

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.[20]

Positive weights and uniformity are distinct issues.[21] 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

[19] 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 since γ are different among persons.

[20] 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 (5) using $\mathbf{1}[\mu_D(z) \geq V]$. V assumes individual specific values which remain fixed as we set different z values. From (A-1), $\Pr(D(z) = 1) = \Pr(D = 1 \mid Z = z)$. 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 (so independent 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 Z were uncertain to the agent, this would be a rational expectations assumption. See the further discussion in Appendix II posted at the website for this paper.

[21] 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.

not necessarily satisfied for $J(Z)$. Condition **??** 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 there is 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 **??**.

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 (**??**)). If we satisfy **??** for a vector, so uniformity applies, weights for a particular instrument may be negative for certain intervals of U_D (i.e., 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, 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.[22]

In summary, nothing in the economics of choice models guarantees that if Z is changed from z to z', people respond in the same direction to the change. See the general expression (**??**). 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 "monotonicity" or "uniformity" condition **??** would be satisfied.

If responses to a common change of Z across persons are heterogenous in a general way, we

[22]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 (e.g. $E(D \mid Z) = \pi Z = J(Z)$) 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 $P(Z)$.

obtain (**??**) 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 IV equation (15) as described in Appendix **??** remains valid and the weights on $\frac{\partial E(Y|P=p)}{\partial p}$ are positive and of the same form as the weights obtained for MTE (or LATE) when the monotonicity condition holds.

1.1.5 Treatment Effects vs. Policy Effects

Even if uniformity condition **??** fails, IV may answer relevant policy questions. By Yitzhaki's result (15), 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 (**??**). 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.

1.2 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 of index sufficiency this is equivalent to

conditioning on $P(Z)$ and D:

$$\begin{aligned} E(Y \mid X, D, Z) &= \mu_0(X) + [\mu_1(X) - \mu_0(X)] D \\ &\quad + K_1(P(Z), X) D \\ &\quad + K_0(P(Z), X)(1 - D), \end{aligned}$$

where the control functions are

$$\begin{aligned} K_1(P(Z), X) &= E(U_1 \mid D = 1, X, P(Z)) \\ K_0(P(Z), X) &= E(U_0 \mid D = 0, X, P(Z)). \end{aligned}$$

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

$$\begin{aligned} E(Y \mid X, Z) &= \mu_0(X) + [\mu_1(X) - \mu_0(X)] P(Z) \\ &\quad + K_1(P(Z), X) P(Z) \\ &\quad + K_0(P(Z), X)(1 - P(Z)), \end{aligned} \tag{16}$$

the population mean outcome given X, Z.

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

$$\left.\frac{\partial E(Y \mid X, P(Z))}{\partial P(Z)}\right|_{P(Z)=p} = \text{LIV}(X, p) = \text{MTE}(X, p).^{23}$$

The distribution of $P(Z)$ and the relationship between $J(Z)$ and $P(Z)$ determine the weight on MTE.[24] Under assumptions (A-1) to (A-5), along with rank and limit conditions (Heckman, 1990; Heckman and Robb, 1985), 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

$$E(Y_1 \mid X, P(Z), D = 1) = \mu_1(X) + K_1(X, P(Z)) \tag{17a}$$

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

[24] 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 (16).

and

$$E\left(Y_0 \mid X, P(Z), D=0\right) = \mu_0\left(X\right) + K_0\left(X, P(Z)\right). \tag{17b}$$

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 $\mu_1\left(X\right)$ from $K_1\left(X, P(Z)\right)$ without invoking functional form or curvature assumptions, it is necessary to have an exclusion (a Z not in X).[25] In addition there must exist a limit set for Z given X such that $K_1\left(X, P(Z)\right) = 0$ for Z in that limit set. Otherwise, without functional form or curvature assumptions, it is not possible to disentangle $\mu_1\left(X\right)$ from $K_1\left(X, P(Z)\right)$ 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 $K_0\left(X, P(Z)\right) = 0$. Selection models operate by identifying the components of (17a) and (17b) and generating the treatment parameters from these components. Thus they work with levels of the Y.

The local IV method works with derivatives of (16) and not levels and cannot directly recover the constant terms in (17a) and (17b). 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

$$\mu_1(X) + E\left(U_1 \mid X, U_D = u_D\right)$$

and

$$\mu_0(X) + E\left(U_0 \mid X, U_D = u_D\right)$$

separately. This corresponds to what is estimated from taking the derivatives of expressions (17a)

[25] See Heckman and Navarro (2007) for use of semiparametric curvature restrictions in identification analysis that do not require functional form assumptions.

and (17b) multiplied by $P(Z)$ and $(1-P(Z))$ respectively:[26]

$$\begin{aligned} & P(Z)E\left(Y_1 \mid X, Z, D=1\right) \\ = \; & P(Z)\mu_1\left(X\right) + P(Z)K_1\left(X, P(Z)\right) \end{aligned}$$

and

$$\begin{aligned} & (1-P(Z))E\left(Y_0 \mid X, Z, D=0\right) \\ = \; & (1-P(Z))\mu_0\left(X\right) + (1-P(Z))K_0\left(X, P(Z)\right). \end{aligned}$$

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 formulae in Table 1B.

Misspecification of $P\left(Z\right)$ (either its functional form or its arguments) and hence of $K_1\left(P\left(Z\right), X\right)$ and $K_0\left(P\left(Z\right), X\right)$ 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 terms $K_1\left(P\left(Z\right), X\right)$ and $K_0\left(P\left(Z\right), X\right)$ can be nonparametrically estimated so it is only necessary to know $P\left(Z\right)$ up to a monotonic transformation.[27] 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 identify 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.[28] 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,

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

[27] See Heckman et al. (1998).

[28] See Imbens and Angrist (1994). Heckman (1990) establishes the identification in the limit argument for ATE in selection models. See Heckman and Navarro (2007) for a generalization to multiple outcome models.

so an identification at infinity argument is implicit when IV is used to identify this parameter.[29] 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.

1.3 Deriving the IV Weights on 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 $\text{Cov}(J(Z), D \mid X = x) \neq 0$. Consider the population analog of the IV estimator,

$$[\text{Cov}(J(Z), Y \mid X = x)] \,/\, [\text{Cov}(J(Z), D \mid X = x)].$$

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

First consider the numerator of this expression,

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

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

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

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 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

$$\operatorname{Cov}\left(J\left(Z\right), D \mid X=x\right) = \operatorname{Cov}\left(J\left(Z\right), P\left(Z\right) \mid X=x\right).$$

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

$$\int \Delta^{\text{MTE}}(x, u_D)\omega\left(x, u_D\right) du_D \tag{18}$$

where

$$\omega\left(x, u_D\right) = \frac{\left(\begin{array}{c} E(\tilde{J}(Z) \mid X = x, P(Z) \geq u_D) \\ \quad \times \Pr(P(Z) \geq u_D \mid X = x) \end{array}\right)}{\operatorname{Cov}\left(J\left(Z\right), P\left(Z\right) \mid X = x\right)}. \tag{19}$$

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

If $J(Z)$ and $P(Z)$ are continuous random variables then a second interpretation of the weight can be derived from (19) by noting that

$$\begin{aligned} & \int \left(j - E\left(J\left(Z\right) \mid X = x\right)\right) \int_{u_D}^{1} f_{P,J}\left(t, j \mid X = x\right) \, dt \, dj \\ = & \int \left(j - E\left(J\left(Z\right) \mid X = x\right)\right) f_J\left(j \mid X = x\right) \\ & \quad \times \int_{u_D}^{1} f_{P|J,X}\left(t \mid J(Z) = j, X = x\right) \, dt \, dj. \end{aligned}$$

Write

$$\begin{aligned} & \int_{u_D}^{1} f_{P|J,X}\left(t \mid J(Z) = j, X = x\right) \, dt \\ = & \; 1 - F_{P|J,X}\left(u_D \mid J(Z) = j, X = x\right) \\ = & \; S_{P|J(Z),X}\left(u_D \mid J(Z) = j, X = x\right) \end{aligned}$$

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

$$\begin{aligned} & \omega\left(x, u_D\right) \\ = & \; \frac{\operatorname{Cov}\left(J\left(Z\right), S_{P|J,X}\left(u_D \mid J(Z), X = x\right) \mid X = x\right)}{\operatorname{Cov}\left(J\left(Z\right), S_{P|J,X}\left(U_D \mid J(Z), X = x\right) \mid X = x\right)}. \end{aligned}$$

For fixed u_D and x evaluation points, $S_{P|J,X}\left(u_D \mid J(Z), X = x\right)$ is a function of the random vari-

able $J(Z)$. 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 x evaluation point, $S_{P|J,X}\left(U_D \mid J(Z), X = x\right)$ 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 $J\left(Z\right)$ and the conditional probability that $\left(P\left(Z\right) > u_D\right)$ given $J\left(Z\right)$ is positive for all u_D, then the weights are positive. The condition is trivially satisfied if $J\left(Z\right) = P\left(Z\right)$, so the weights are positive and IV estimates a gross treatment effect.

If the $J\left(Z\right)$ and $P\left(Z\right)$ are discrete valued, we obtain expressions and (13) and (14) in the text.

2 The Choice Model and Assumptions

Following Heckman, Urzua, and Vytlacil (2006, 2009) and Heckman and Vytlacil (2007b), consider the following model with multiple choices and associated multiple outcome states. Let $\mathcal{J}$ denote the agent's choice set, where $\mathcal{J}$ contains a finite number of elements. For example, $\mathcal{J}$ enumerates possible schooling states (e.g., GED, high school dropout, high school graduate). The value to the agent of choosing $j \in \mathcal{J}$ is

$$R_j(Z_j) = \vartheta_j(Z_j) - V_j, \tag{20}$$

where Z_j are the agent's observed characteristics that affect the utility from choosing j, and V_j is the unobserved shock to the agent's utility from choice j. We sometimes write R_j for $R_j(Z_j)$ to simplify notation. Let Z denote the random vector containing all unique elements of $\{Z_j\}_{j\in\mathcal{J}}$. We write $R_j(Z)$ for $R_j(Z_j)$, leaving implicit the condition that $R_j(\cdot)$ only depends on the elements of Z that are contained in Z_j. Let D_j be a variable indicating whether the agent would choose j if

confronted with choice set $\mathcal{J}$:[30]

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

Array the D_j into a vector D. Let Y be the outcome that would be observed if the agent faced choice set $\mathcal{J}$, defined as

$$Y = \sum_{j \in \mathcal{J}} D_j Y_j,$$

where Y_j is a potential outcome observed only if option j is chosen. 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 $\{X_j\}_{j \in \mathcal{J}}$. (Z, X, D, Y) is assumed to be observed by the analyst.[31]

Define $R_{\mathcal{J}}$ as the maximum obtainable value given choice set $\mathcal{J}$:

$$\begin{aligned} R_{\mathcal{J}} &= \max_{j \in \mathcal{J}} \{R_j\} \\ &= \textstyle\sum_{j \in \mathcal{J}} D_j R_j. \end{aligned} \tag{21}$$

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

$$D_j = 1 \Longleftrightarrow R_j \geq R_{\mathcal{J} \setminus j}.$$

Heckman, Urzua, and Vytlacil (2006, 2009) and Heckman and Vytlacil (2007b) show that this simple, well-known, representation is the key intuition for understanding how instrumental variables estimate the effect of a given choice versus the "next best" alternative. IV is a weighted average

[30] Below, we invoke conditions so that ties, $R_j = R_k$ for $j \neq k$, occur with probability zero.

[31] Depending on the choice model, Z may or may not include the X. For example, in a Roy model of schooling under perfect certainty (e.g. Willis and Rosen, 1979), X would be among the Z. In models of schooling under uncertainty (e.g. Cunha, Heckman, and Navarro, 2005, Cunha and Heckman, 2007 and Urzua, 2008) innovations in X unknown at the time schooling decisions are made would not be in Z. The key condition on Z is given in Assumption (A-2) below.

of the effects for people induced into a choice from different margins. Analogous to the definition of $R_{\mathcal{J}}$, we define $R_{\mathcal{J}}(z)$ to be the maximum obtainable value given choice set $\mathcal{J}$ when instruments are fixed at $Z = z$,

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

Following the analysis in Heckman et al. (2006, 2009) and Heckman and Vytlacil (2007b), we assume:

(A-1) The distribution of $(\{V_j\}_{j \in \mathcal{J}})$ is continuous.[32]

(A-2) $\{(V_j, U_j)\}_{j \in \mathcal{J}}$ is independent of Z conditional on X.

(A-3) $E \mid Y_j| < \infty$ for all $j \in \mathcal{J}$.

(A-4) $\Pr(D_j = 1 \mid X) > 0$ for all $j \in \mathcal{J}$.

In addition, we assume an exclusion restriction that requires some additional notation.[33] Let $Z^{[-l]}$ denote all elements of Z except for the lth component. We assume

(A-5) For each $j \in \mathcal{J}$, their exists at least one element of Z, say $Z^{[l]}$, such that the distribution of $\vartheta_j(Z_j)$ conditional on $(X, Z^{[-l]})$ is continuous.

With these assumptions, one can generalize the analysis of Heckman and Vytlacil (1999, 2001c, 2005) to the unordered case. Assumptions (A-1) and (A-2) imply that $R_j \neq R_k$ (with probability 1) for $j \neq k$, so that $\text{argmax}_{j \in \mathcal{J}}\{R_j\}$ is unique (with probability 1). Assumption (A-2) assures the existence of an instrument. Assumption (A-3) is required for mean treatment parameters to be well defined. It also allows one to integrate to the limit and to produce well-defined means. Assumption (A-4) requires that at least some individuals participate in each choice for all X. Assumption (A-5) imposes the requirement that one be able to independently vary the index for the given value function. It imposes a type of 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.[34]

[32] Absolutely continuous with respect to Lebesgue measure on $\prod_{j \in \mathcal{J}} \Re$.

[33] We work here with exclusion restrictions in part for ease of exposition. By adapting the analysis of Cameron and Heckman (1998) and Heckman and Navarro (2007), one can modify our analysis for 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.

[34] See Heckman and Vytlacil (2007b) for additional discussion.

In a series of papers, Heckman and Vytlacil (1999, 2001c, 2005, 2007b), develop the method of local instrumental variables (LIV) to estimate the marginal treatment effect (MTE) for the case of binary choices. We now define and interpret the MTE and LIV in the case of general unordered choices.

2.1 Interpreting Local Instrumental Variables in the Unordered Case

We define local instrumental variables (LIV) using a variable that shifts people toward (or against) choice j by operating only on $R_j(Z_j)$. LIV identifies an average marginal return to j vs. the next best alternative across persons.[35] However, without further assumptions, LIV will not decompose the average marginal return into its component parts corresponding to the effects for persons induced into j from each of the possible origin states.

To see this, consider a three outcome case, $\mathcal{J} = \{1, 2, 3\}$. For concreteness, we pursue the education example previously stated and let 1 be GED, 2 be high school dropout, and 3 be high school graduate. Our results are more general but the three outcome case is easy to exposit.

In this section, we assume that Z_1, Z_2, Z_3 are disjoint sets of regressors so $Z = (Z_1, Z_2, Z_3)$ but they are not necessarily statistically independent. We can easily relax this assumption but making it simplifies the notation. We condition on X and keep it implicit throughout the analysis of this paper.[36] In this notation,

$$\begin{aligned} E(Y \mid Z) &= E\left[\sum_{j=1}^{3} Y_j D_j \,\middle|\, Z\right] \\ &= E(Y_1 D_1 \mid Z) + E(Y_2 D_2 \mid Z) + E(Y_3 D_3 \mid Z). \end{aligned} \tag{22}$$

$E(Y|Z)$ and its components can be estimated from data on (Y, Z). IV is based on (22). From (21), choices are generated by the following inequalities:

$$\begin{aligned} D_1 &= \mathbf{1}(R_1 \geq R_2, R_1 \geq R_3) \\ D_2 &= \mathbf{1}(R_2 \geq R_1, R_2 \geq R_3) \\ D_3 &= \mathbf{1}(R_3 \geq R_1, R_3 \geq R_2). \end{aligned}$$

[35] See Heckman, Urzua, and Vytlacil (2006) and Heckman and Vytlacil (2007b).
[36] See Heckman and Vytlacil (2007b) for a more general analysis.

We define the marginal change in Y with respect to Z_1. IV methods are based on such types of variation. The local instrumental variable estimator using Z_1 as an instrument is the sample analogue of

$$\left.\frac{\frac{\partial E(Y|Z)}{\partial Z_1}}{\frac{\partial \Pr(D_1=1|Z)}{\partial Z_1}}\right|_{Z=z} = \text{LIV}(z),$$

where LIV is a function of z. In the case of three choices, there are two margins from which persons can be attracted into or out of choice 1 by Z_1.[37]

From local variations in Z_1, one can recover the following combinations of parameters from the data on $Y_1 D_1$:

$$\begin{aligned}
&\frac{\partial E\left(Y_1 D_1 \mid Z=z\right)}{\partial Z_1} \\
&= \frac{\partial}{\partial Z_1}\int\int_{-\infty}^{\vartheta_1(Z_1)-\vartheta_2(Z_2)}\int_{-\infty}^{\vartheta_1(Z_1)-\vartheta_3(Z_3)} y_1 f_{Y_1,V_1-V_2,V_1-V_3}\left(y_1, v_1-v_2, v_1-v_3\right) d\left(v_1-v_3\right) d\left(v_1-v_2\right) dy_1\Big|_{Z=z} \\
&= \left.\frac{\partial \vartheta_1\left(Z_1\right)}{\partial Z_1}\right|_{Z_1=z_1}\left[\begin{array}{l} \int y_1 \int_{-\infty}^{\vartheta_1(z_1)-\vartheta_3(z_3)} f_{Y_1,V_1-V_2,V_1-V_3}\left(y_1, \vartheta_1\left(z_1\right)-\vartheta_2\left(z_2\right), v_1-v_3\right) d\left(v_1-v_3\right) dy_1 \\ +\int y_1 \int_{-\infty}^{\vartheta_1(z_1)-\vartheta_2(z_2)} f_{Y_1,V_1-V_2,V_1-V_3}\left(y_1, v_1-v_2, \vartheta_1\left(z_1\right)-\vartheta_3\left(z_3\right)\right) d\left(v_1-v_2\right) dy_1 \end{array}\right].
\end{aligned} \tag{23}$$

By similar reasoning, we can recover the following combination of parameters from the data on $Y_2 D_2$:

$$\begin{aligned}
&\frac{\partial E\left(Y_2 D_2 \mid Z=z\right)}{\partial Z_1} \\
&= \frac{\partial}{\partial Z_1}\int y_2 \int_{-\infty}^{\vartheta_2(Z_2)-\vartheta_1(Z_1)}\int_{-\infty}^{\vartheta_2(Z_2)-\vartheta_3(Z_3)} f_{Y_2,V_2-V_1,V_2-V_3}(y_2, v_2-v_1, v_2-v_3)\, d\left(v_2-v_3\right)\, d\left(v_2-v_1\right) dy_2\Big|_{Z=z} \\
&= \left.\frac{-\partial \vartheta_1\left(Z_1\right)}{\partial Z_1}\right|_{Z_1=z_1}\left[\int y_2 \int_{-\infty}^{\vartheta_2(z_2)-\vartheta_3(z_3)} f_{Y_2,V_2-V_1,V_2-V_3}(y_2, \vartheta_2(z_2)-\vartheta_3(z_1), v_2-v_3) d\left(v_2-v_3\right) dy_2\right].
\end{aligned} \tag{24}$$

[37] Recall that Z_1 only affects the utility associated with choice 1.

From data on $Y_3 D_3$, we obtain the following combination of parameters:

$$\frac{\partial E\left(Y_3 D_3 \mid Z_1 = z\right)}{\partial Z_1} = \left.\frac{-\partial \vartheta_1\left(Z_1\right)}{\partial Z_1}\right|_{Z_1 = z_1} \int y_3 \int_{-\infty}^{\vartheta_3(z_3)-\vartheta_2(z_2)} f_{Y_3, V_3 - V_1, V_3 - V_2}\left(y_3, \vartheta_3\left(z_3\right) - \vartheta_1\left(z_1\right), v_3 - v_2\right) d\left(v_3 - v_2\right) dy_3 \,. \tag{25}$$

Agents induced into 1 come from 2 and 3. There are two margins:

$$\left(R_1 = R_2\right) \text{ and } \left(R_1 \geq R_3\right) \qquad \text{(margin of indifference between 1 and 2),}$$

and

$$\left(R_1 = R_3\right) \text{ and } \left(R_1 \geq R_2\right) \qquad \text{(margin of indifference between 1 and 3).}$$

Unaided, IV does not enable analysts to identify the returns at each of the different margins. Instead, it identifies a weighted average of returns. It does not identify the density of persons at the various margins, i.e., the proportion of people induced into (or out of) 1 from each possible alternative state by a change in the instrument.

Collecting terms and rewriting in more easily interpretable components, which generalize the MTE developed for a two choice model to a multiple choice unordered model:[38]

$$\left.\frac{\left(\frac{\partial E(Y|Z)}{\partial Z_1}\right)}{\left(\frac{\partial \vartheta_1}{\partial Z_1}\right)}\right|_{Z=z} =$$

$$\left[\begin{array}{l} \overbrace{\left[E\left(Y_1 - Y_2 \mid R_1\left(z_1\right) = R_2\left(z_2\right), R_1\left(z_1\right) \geq R_3\left(z_3\right)\right)\right]}^{\substack{\text{Generalization of MTE for persons indifferent} \\ \text{between 1 and 2, where choice 3 is dominated}}} \Pr\left(R_1\left(z_1\right) = R_2\left(z_2\right), R_1\left(z_1\right) \geq R_3\left(z_3\right)\right) \\ + \underbrace{\left[E\left(Y_1 - Y_3 \mid R_1\left(z_1\right) = R_3\left(z_3\right), R_1\left(z_1\right) \geq R_2\left(z_2\right)\right)\right]}_{\substack{\text{Generalization of MTE for persons indifferent} \\ \text{between 1 and 3, where choice 2 is dominated}}} \Pr\left(R_1\left(z_1\right) = R_3\left(z_3\right), R_1\left(z_1\right) \geq R_2\left(z_2\right)\right) \end{array}\right] .$$

This is a weighted return to alternative 1 for persons coming from two separate margins: alternative 1 *versus* alternative 2, and alternative 1 *versus* alternative 3, i.e., the return to people induced into 1 from their next best choice. The weights are the proportion of people induced into 1 from each

[38] Heckman, Urzua, and Vytlacil (2006) generalize the MTE to an ordered choice model. See also Heckman and Vytlacil (2007b).

margin. This *combination* of parameters can be identified from IV. The components of the sum cannot be identified by IV without further assumptions. Note that it is possible that a group at one margin gains while a group at another margin loses. IV only estimates a net effect, which might be zero.

Notice that from representation (2.1) and the assumption that the Z_j ($j \epsilon \mathcal{J}$) are distinct, *pairwise monotonicity*, an extension of the monotonicity assumption invoked by Imbens and Angrist (1994) for the binary choice case, is satisfied.[39] In the context of a model with multiple choices, pairwise monotonicity means the same pattern of flow between any two states is experienced by everyone. Thus, as Z_j increases, there is a flow from i to j but not from j to i (or vice versa). From (20), changing Z_1 induces all persons to move in the same direction (*i.e.* from 1 to 2 or 2 to 1 but not both, and from 1 to 3 or 3 to 1 but not both). Pairwise monotonicity does not rule out the possibility that a change in an instrument causes people to move in the direction from j to i but to move away from the direction from k to i for $j \neq k$, and $j, k \neq i$.

By the chain rule, the derivative of $\Pr(D_1 = 1 \mid Z)$ is:

$$\frac{\partial \Pr(D_1 = 1 \mid Z = z)}{\partial Z_1} = \frac{\partial \vartheta_1}{\partial Z_1}\Big|_{Z_1 = z_1} \left[\begin{array}{c} \Pr(R_1(z_1) = R_2(z_2), R_1(z_1) \geq R_3(z_3)) \\ + \Pr(R_1(z_1) = R_3(z_3), R_1(z_1) \geq R_2(z_2)) \end{array} \right].$$

We can define LIV in terms of the preceding ingredients as

$$\mathrm{LIV}(z) = \frac{\left(\frac{\partial E(Y|Z)}{\partial Z_1}\right)}{\left(\frac{\partial \Pr(D_1=1|Z)}{\partial Z_1}\right)}\Bigg|_{Z=z} = \left[\begin{array}{c} E(Y_1 - Y_2 \mid R_1(z_1) = R_2(z_2), R_1(z_1) \geq R_3(z_3))\, \omega_{12} \\ + E(Y_1 - Y_3 \mid R_1(z_1) = R_3(z_3), R_1(z_1) \geq R_2(z_2))\, \omega_{13} \end{array} \right]. \tag{26}$$

The *combination* of terms can be identified by LIV from the data on (Y, D, Z).

The IV weights are:

$$\omega_{12} = \frac{\Pr(R_1(z_1) = R_2(z_2), R_1(z_1) \geq R_3(z_3))}{\left[\begin{array}{c} \Pr(R_1(z_1) = R_2(z_2), R_1(z_1) \geq R_3(z_3)) \\ + \Pr(R_1(z_1) = R_3(z_3), R_1(z_1) \geq R_2(z_2)) \end{array} \right]} \tag{27}$$

(28)

[39] This is defined as "uniformity" in Heckman, Urzua, and Vytlacil (2006).

and

$$\omega_{13} = \frac{\Pr\left(R_1\left(z_1\right)=R_3\left(z_3\right), R_1\left(z_1\right)\geq R_2\left(z_2\right)\right)}{\left[\begin{array}{c}\Pr\left(R_1\left(z_1\right)=R_2\left(z_2\right), R_1\left(z_1\right)\geq R_3\left(z_3\right)\right)\\ +\Pr\left(R_1\left(z_1\right)=R_3\left(z_3\right), R_1\left(z_1\right)\geq R_2\left(z_2\right)\right)\end{array}\right]}. \tag{29}$$

The weights can be identified from a structural discrete choice analysis.[40] They cannot be identified by an unaided instrumental variable analysis. Thus it is not possible to identify the component parts of (24) by LIV alone, i.e., one cannot separately identify the generalized MTEs:

$$E\left(Y_1-Y_2 \mid R_1\left(z_1\right)=R_2\left(z_2\right), R_1\left(z_1\right)\geq R_3\left(z_3\right)\right)$$

and

$$E\left(Y_1-Y_3 \mid R_1\left(z_1\right)=R_3\left(z_3\right), R_1\left(z_1\right)\geq R_2\left(z_2\right)\right),$$

unless one invokes "identification at infinity" arguments or alternative arguments using local variation in regressors developed by Fox and Gandhi (2008).[41] **?** and Heckman and Vytlacil (2007b) develop classes of economically interpretable parameters that do not require "identification at infinity" and that can be identified using estimated marginal treatment effects.

Using a structural model, one can estimate the components of (26) and determine the flow into (or out of) state 1 from all sources. We illustrate this point in Section 2.3. First we consider what standard IV estimates.

[40] Conditions for nonparametric identification of the multinomial discrete choice model are presented in Matzkin (1993, 1994). Conditions for nonparametric identification of the full choice model with outcomes are given in Heckman and Vytlacil (2007a, Appendix B). Conditions for identification of general dynamic discrete choice models are presented in Abbring and Heckman (2007). Conditions for identification in multinomial models that do not require "identification at infinity" are given in Fox and Gandhi (2008).

[41] See Heckman and Vytlacil (2007b) who show how to vary Z_3 or Z_2 to effectively shut down one margin of choice. Specifically, for any fixed $Z_1 = z_1$, if $\lim_{Z_2 \to \tilde{Z}_2} R_2(Z_2) \to -\infty$ and $\lim_{Z_3 \to \tilde{Z}_3} R_3(Z_3) \to -\infty$ where $\tilde{Z}_2$ and $\tilde{Z}_3$ represent limit sets, then we can identify, respectively, the gains at the $3 \to 1$ margin in the limit set, and the gains in the $2 \to 1$ margin in the limit set. These assumptions require that one can vary Z_2 and Z_3 to shut down one or the other margin of choice. Under these assumptions and some additional mild regularity assumptions, the structural approach can identify distributions of $(Y_1 - Y_2)$ and $(Y_2 - Y_3)$ as we demonstrate in the example in Section 2.3 of this paper. "Identification at infinity" is a model-specific misnomer. It is an assumption that there are different sets each with non-negligible probability such that the probabilities of attaining various outcome states are arbitrarily close to one. "Identification at infinity" assumptions are justified naturally in truncated regression models. See Heckman (1987).

2.2 What does standard IV estimate?

To see what standard IV estimates, consider the following linear-in-schooling model of earnings that receives much attention in the literature in labor economics.[42] Let Y denote log earnings and write S as years of schooling. The model writes

$$Y = \alpha + \beta S + U \tag{30}$$

where

$$S = \sum_{j=1}^{3} j D_j, \tag{31}$$

and Y is defined as in Section 2. It is interpreted in this section as an approximation to the general model presented in Section 2. S is assumed to be correlated with U, and β is a random variable that may be statistically dependent on S. The model of Section 2 does *not*, in general, imply (30). Indeed, there is much empirical evidence against model (30).[43] An analysis of what IV estimates when linearity in S is imposed as an approximation, even though it may be inappropriate, is an interesting exercise because linearity is so often invoked.

Suppose Z_1 is a valid instrument. We now interpret what

$$\Delta^{\text{IV}}_{Z_1} = \frac{\text{Cov}(Z_1, Y)}{\text{Cov}(Z_1, S)} \tag{32}$$

estimates. We do this by decomposing $\Delta^{\text{IV}}_{Z_1}$ into components analogous to the decomposition produced by Heckman et al. (2006, 2009) and Heckman and Vytlacil (2007b). The Appendix presents the derivation of the following decomposition of IV into our pairwise generalization of

[42] We keep conditioning on X implicit.

[43] See Heckman, Lochner, and Todd (2006) for discussions of this model and various justifications for it. Heckman, Layne-Farrar, and Todd (1996) present evidence against linearity of the earnings function in terms of years of schooling.

MTE for the unordered case:

$$\Delta_{Z_1}^{\text{IV}} = \frac{\text{Cov}(Z_1, Y)}{\text{Cov}(Z_1, S)} = \tag{33}$$

$$\left(\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \overbrace{E\left(Y_1 - Y_2 \mid V_2 - V_1 = v_2 - v_1, \vartheta_2(z_2) - \vartheta_3(z_3) \geq V_2 - V_3\right)}^{\text{Generalized MTE } (2 \to 1) \text{ not identified from LIV}} \right.$$
$$\left. \times \underbrace{\eta_{\vartheta_2(Z_2)-\vartheta_3(Z_3), V_2 - V_1}\left(\vartheta_2(z_2) - \vartheta_3(z_3), v_2 - v_1\right)}_{\substack{\text{weight identified from discrete} \\ \text{choice analysis}}} d(v_2 - v_1)\, d\left(\vartheta_2(z_2) - \vartheta_3(z_3)\right) \right)$$
$$+ \left(\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \overbrace{E\left(Y_1 - Y_3 \mid V_3 - V_1 = v_3 - v_1, \vartheta_3(z_3) - \vartheta_2(z_2) \geq V_3 - V_2\right)}^{\text{Generalized MTE}(3 \to 1) \text{ not identified from LIV}} \right.$$
$$\left. \times \underbrace{\eta_{\vartheta_3(Z_3)-\vartheta_2(Z_2), V_3 - V_1}\left(\vartheta_3(z_3) - \vartheta_2(z_2), v_3 - v_1\right)}_{\substack{\text{weight identified from discrete} \\ \text{choice analysis}}} d(v_3 - v_1)\, d\left(\vartheta_3(z_3) - \vartheta_2(z_2)\right) \right)$$

$$\Big/ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \underbrace{\left[-\eta_{\vartheta_2(Z_2)-\vartheta_3(Z_3), V_2 - V_1}\left(\vartheta_2(z_2) - \vartheta_3(z_3), v_2 - v_1\right)\right]}_{\substack{\text{weight identified from discrete} \\ \text{choice analysis}}} d\left(\vartheta_2(z_2) - \vartheta_3(z_3)\right)\, d(v_2 - v_1)$$
$$+2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \underbrace{\left[-\eta_{\vartheta_3(Z_3)-\vartheta_2(Z_2), V_3 - V_1}\left(\vartheta_3(z_3) - \vartheta_2(z_2), v_3 - v_1\right)\right]}_{\substack{\text{weight identified from discrete} \\ \text{choice analysis}}} d(v_3 - v_1)\, d\left(\vartheta_3(z_3) - \vartheta_2(z_2)\right).$$

IV identifies a weighted average of gains to state 1 compared to the next best alternative which may be 2 or 3. The two terms of the decomposition are defined as generalized MTEs and are weighted averages of the gain of moving from state 2 to state 1 for persons on the margin of indifference between 1 and 2 and for whom 2 is a better choice than 3 (the first term) and the gain of moving from 3 to 1 for persons on the margin of indifference between 1 and 3 and for whom 3 is a better choice than 2 (the second term).[44]

In the Appendix, we derive the weights on the generalized MTEs and show that they do not sum to 1 even when normalized by the denominator. The mathematical reason for this result is simple. The weights in the numerator do not sum to the weights in the denominator. The second term in the denominator receives twice as much weight as the corresponding term in the numerator. This is a consequence of the definition of S (31), which plays no role in the numerator term. Thus, IV applied to the general model produces an arbitrarily weighted sum of generalized MTEs with weights that do not sum to 1, and which, in general, places more weight on the first generalized

[44] Since Z_1 only affects $R(Z_1)$, it has no direct effect on the margin $2 \to 3$.

MTE term than on the second term, compared to the weights placed on the corresponding terms in the denominator.[45] Using IV alone, we cannot decompose (33) into its component parts, even though the weights can be identified from discrete choice analysis.[46,47]

2.3 An Example

It is instructive to summarize our analysis with an example. Consider a 3 choice model with associated outcomes. This corresponds to the GED, high school dropout and high school graduate example that we have used throughout the paper. Under conditions presented in Heckman and Vytlacil (2007a, Appendix B), the structural model is nonparametrically identified. A key assumption in their proof is the "identification at infinity" assumption previously discussed.[48] This assumes the ability to vary (Z_1, Z_2, Z_3) freely and the existence of limit sets such that fixing any two of (Z_1, Z_2, Z_3), one makes the R_j associated with Z_j arbitrarily small.[49]

Heckman and Vytlacil (2007b) show that if one augments the IV assumptions with the same identification at infinity assumptions used in structural models, one can use IV in the limit to identify the components of (3.5). In the limit sets, one can identify

$$E(Y_1 - Y_2 | R_1(z_1) = R_2(z_2)) \tag{35}$$

[45] Thus "2" appears only in the denominator and not in the numerator.

[46] The structural model is nonparametrically identified under the conditions in Appendix B of Heckman and Vytlacil (2007a).

[47] Decomposition (33) is not unique. It arises from decomposing Y into

$$Y = D_1 Y_1 + D_2 Y_2 + D_3 Y_3$$

where we solve out $D_1 = 1 - D_2 - D_3$, to obtain

$$Y = Y_1 + D_2(Y_2 - Y_1) + D_3(Y_3 - Y_1).$$

We could also solve out $D_2 = 1 - D_1 - D_3$ to obtain

$$Y = Y_2 + D_1(Y_1 - Y_2) + D_3(Y_3 - Y_2)$$

or $D_3 = 1 - D_1 - D_2$ to obtain

$$Y = Y_3 + D_1(Y_1 - Y_3) + D_2(Y_2 - Y_3).$$

Each decomposition can be used to represent $\Delta^{IV}_{Z_1}$. For each decomposition, the leading terms on the right-hand side, (Y_1, Y_2, Y_3), respectively, are uncorrelated with Z_1 by virtue of (A-2). Corresponding generalized MTEs can be defined for each decomposition. Z_1 affects the lower boundary of the opportunity set in

$$E(Y_2 - Y_3 | R(z_2) \geq R(z_1), R(z_3) \geq R(z_1)).$$

We choose the decomposition reported in the text for its greater interpretability.

[48] Alternatively, one can make functional form assumptions about the distribution of the error terms.

[49] See the conditions in footnote 41.

and

$$E(Y_1 - Y_3 | R_1(z_1) = R_3(z_3)) \tag{36}$$

by setting Z_3 and Z_2 respectively to limit set values. Essentially one can use the limit sets to make a three choice model into a two choice model, and the standard results for the two choice model apply.[50] Under these assumptions, and additional mild regularity assumptions, using structural methods, one can identify the distributions of (Y_1, Y_2) and (Y_1, Y_3) so that one can identify *distributions* of treatment effects, $Y_2 - Y_1$ and $Y_3 - Y_1$, in addition to the mean parameters identified by IV.[51] One can also identify the proportion of people induced into 1 from each alternative state using variation in the instrument.

Consider the model with the parameters presented in Table 5. This is a discrete choice model with associated outcome variables. The $Z_j, j = 1, \ldots, 3$, are assumed to be scalar and mutually independent. They are normally distributed so they satisfy large support ("identification at infinity") conditions. Table 6 shows how a change in Z_1, which increases it by .75 standard deviations, shifts people across categories. This corresponds to making GED attainment easier.[52] The estimates reported in Table 6 can be obtained from a structural discrete choice model. The percentage initially in 1 (GED) increases from 33.17% to 38.8%. The percentage in 2 (dropout) decreases from 29.11% to 25.91%. The percentage in 3 (graduating high school) declines from 37.72% to 35.29%.

The IV estimate is -.032. (See the base of Table 7) This is the only number produced by an IV analysis using Z_1 as an instrument that changes within the specified range. The structural analysis in Table 7 shows that the net effect produced by the change in Z_1 is composed of 2 terms. It arises from a gain of .199 for the switchers $2 \to 1$ (dropout to GED) and a loss of .336 ($3 \to 1$) (graduate to GED).

Figure 1 shows what can be identified from the structural model. It plots the distributions of gains for persons going from 2 to 1 and from 3 to 1 as well as the overall distribution of gains to the switchers. Persons switching from 3 to 1 are harmed in gross terms by the policy that changes Z_1, while those who switch from 2 to 1 gain in gross terms. In utility terms, (R_j), people are

[50] See Heckman and Vytlacil (2007b)

[51] The literature on "quantile treatment effects" uses IV to identify the quantiles of Y_1 and Y_2 separately but not the quantiles of $Y_1 - Y_2$. See Abbring and Heckman (2007).

[52] Heckman, LaFontaine, and Rodríguez (2008) show that easing GED requirements promotes dropping out of school and causes some dropouts to become GEDs.

better off.[53] In terms of gross gains, about 56.8% of the people who switch from 2 to 1 are better off while 39.3% of the people who switch from 3 to 1 are better off. Overall, 49.2% are better off in gross terms even though the IV estimate is slightly negative. If one seeks to understand the distributional effects of the policy associated with a change Z_1, the structural analysis is clearly much more revealing. The IV estimate, which is a mean gross gain aggregating over origin states, does not capture the rich information about choices afforded by a structural analysis. However, it does identify the average gain to the program compared to the next-best alternatives. If that is the object of interest, linear IV is the right tool to use.

3 Derivation of the Standard IV Estimator

We first study the numerator of $\Delta^{\text{IV}}_{Z_1}$ in the text. Recall that we keep the conditioning on X implicit. Using $\tilde{Z}_1 = Z_1 - \bar{Z}_1$,

$$\text{Cov}\left(Y, Z_1\right) = E\left(\tilde{Z}_1 \left(Y_1 D_1 + Y_2 D_2 + Y_3 D_3\right)\right).$$

Using $D_1 = 1 - D_2 - D_3$, we obtain

$$\begin{aligned}
\text{Cov}\left(Y, Z_1\right) &= E\left(\tilde{Z}_1 \left(Y_1 + \left(Y_2 - Y_1\right) D_2 + \left(Y_3 - Y_1\right) D_3\right)\right) \\
&= E\left(\tilde{Z}_1 Y_1\right) + E\left(\tilde{Z}_1 \left(Y_2 - Y_1\right) D_2\right) + E\left(\tilde{Z}_1 \left(Y_3 - Y_1\right) D_3\right),
\end{aligned}$$

[53] This is imposed in a discrete choice model.

where $E\left(\tilde{Z}_1 Y_1\right) = 0$. It is natural to decompose this expression using choice "1" as the base, because Z_1 only shifts $R_1(Z_1)$. The final two terms can be written as

$$
\begin{aligned}
&\operatorname{Cov}(Y, Z_1) \\
&= E\left(\tilde{Z}_1 (Y_2 - Y_1) \mathbf{1}\left(R_2(Z_2) \geq R_1(Z_1), R_2(Z_2) \geq R_3(Z_3)\right)\right) + E\left(\tilde{Z}_1 (Y_3 - Y_1) \mathbf{1}\left(R_3(Z_3) \geq R_1(Z_1), R_3(Z_3) \geq R_2(Z_2)\right)\right) \\
&= E\left[\tilde{Z}_1 (Y_2 - Y_1) \mathbf{1}\left(\left(\vartheta_2(Z_2) - \vartheta_1(Z_1) \geq V_2 - V_1\right), \left(\vartheta_2(Z_2) - \vartheta_3(Z_3) \geq V_2 - V_3\right) \right)\right] \\
&\quad + E\left[\tilde{Z}_1 (Y_3 - Y_1) \mathbf{1}\left(\left(\vartheta_3(Z_3) - \vartheta_1(Z_1) \geq V_3 - V_1\right), \left(\vartheta_3(Z_3) - \vartheta_2(Z_2) \geq V_3 - V_2\right) \right)\right] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tilde{z}_1 (y_2 - y_1) \\
&\quad \times \left(\int_{-\infty}^{\vartheta_2(z_2) - \vartheta_1(z_1)} \int_{-\infty}^{\vartheta_2(z_2) - \vartheta_3(z_3)} f_{Y_2 - Y_1, V_2 - V_1, V_2 - V_3}(y_2 - y_1, v_2 - v_1, v_2 - v_3)\, d(v_2 - v_3)\, d(v_2 - v_1)\, d(y_2 - y_1)\right) \\
&\quad \times f_{\tilde{Z}_1, \vartheta_2(Z_2) - \vartheta_1(Z_1), \vartheta_2(Z_2) - \vartheta_3(Z_3)}\left(\tilde{z}_1, \vartheta_2(z_2) - \vartheta_1(z_1), \vartheta_2(z_2) - \vartheta_3(z_3)\right) d\left(\vartheta_2(z_2) - \vartheta_3(z_3)\right) d\left(\vartheta_2(z_2) - \vartheta_1(z_1)\right) d\tilde{z}_1 \\
&\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tilde{z}_1 (y_3 - y_1) \\
&\quad \times \left(\int_{-\infty}^{\vartheta_3(z_3) - \vartheta_1(z_1)} \int_{-\infty}^{\vartheta_3(z_3) - \vartheta_2(z_2)} f_{Y_3 - Y_1, V_3 - V_1, V_3 - V_2}(y_3 - y_1, v_3 - v_1, v_3 - v_2)\, d(v_3 - v_2)\, d(v_3 - v_1)\, d(y_3 - y_1)\right) \\
&\quad \times f_{\tilde{Z}_1, \vartheta_3(Z_3) - \vartheta_1(Z_1), \vartheta_3(Z_3) - \vartheta_2(Z_2)}\left(\tilde{z}_1, \vartheta_3(z_3) - \vartheta_1(z_1), \vartheta_3(z_3) - \vartheta_2(z_2)\right) d\left(\vartheta_3(z_3) - \vartheta_2(z_2)\right) d\left(\vartheta_3(z_3) - \vartheta_1(z_1)\right) d\tilde{z}_1.
\end{aligned}
$$

By Fubini's Theorem, we can simplify the expressions and obtain for the first term:

$$
\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E\left(Y_2 - Y_1 \mid V_2 - V_1 = v_2 - v_1, \vartheta_2(z_2) - \vartheta_3(z_3) \geq V_2 - V_3\right) \\
&\quad \times \left\{\int_{-\infty}^{\infty} \tilde{z}_1 \left[\left(\int_{-\infty}^{\vartheta_2(z_2) - \vartheta_3(z_3)} h_{V_2 - V_1, V_2 - V_3}(v_2 - v_1, v_2 - v_3)\, d(v_2 - v_3)\right)\right.\right. \\
&\quad \left.\left. \times \left(\int_{v_2 - v_1}^{\infty} f_{\tilde{Z}_1, \vartheta_2(Z_2) - \vartheta_1(Z_1), \vartheta_2(Z_2) - \vartheta_3(Z_3)}\left(\tilde{z}_1, \vartheta_2(z_2) - \vartheta_1(z_1), \vartheta_2(z_2) - \vartheta_3(z_3)\right) d\left(\vartheta_2(z_2) - \vartheta_1(z_1)\right)\right)\right] d\tilde{z}_1\right\} \\
&\qquad\qquad \times d\left(\vartheta_2(z_2) - \vartheta_3(z_3)\right) d(v_2 - v_1). \qquad (37)
\end{aligned}
$$

$h_{V_2 - V_1, V_2 - V_3}(.)$ is the joint density of $V_2 - V_1$, $V_2 - V_3$. Define the weighting term in braces in (37) as $\eta_{\vartheta_2(Z_2) - \vartheta_3(Z_3), V_2 - V_1}(\vartheta_2(z_2) - \vartheta_3(z_3), v_2 - v_1)$. It is necessary to fix both $\vartheta_2(z_2) - \vartheta_3(z_3)$ and $v_2 - v_1$ in forming the weight. This weight can be estimated from a structural discrete choice analysis and the joint distribution of (Z, D_1, D_2, D_3). The terms multiplying the weight are marginal treatment effects generalized to the unordered case. (A.1) cannot be decomposed using IV. An alternative

representation of the term in braces, $\eta_{\vartheta_2(Z_2)-\vartheta_3(Z_3),V_2-V_1}(\vartheta_2(z_2)-\vartheta_3(z_3),v_2-v_1)$ is

$$\begin{aligned}
&\eta_{\vartheta_2(Z_2)-\vartheta_3(Z_3),V_2-V_1}\left(\vartheta_2(z_2)-\vartheta_3(z_3),v_2-v_1\right)= \\
&\quad E\left(Z_1-E\left(Z_1\right)\mid \vartheta_2(Z_2)-\vartheta_3(Z_3)=\vartheta_2(z_2)-\vartheta_3(z_3),\vartheta_2(Z_2)-\vartheta_1(Z_1)\geq v_2-v_1\right) \\
&\qquad \times \Pr\left(\vartheta_2(Z_2)-\vartheta_3(Z_3)=\vartheta_2(z_2)-\vartheta_3(z_3),\vartheta_2(Z_2)-\vartheta_1(Z_1)\geq v_2-v_1\right).
\end{aligned}$$

An analysis parallel to the preceding one shows that the second term can be written as

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E\left(Y_3-Y_1\mid V_3-V_1=v_3-v_1,\vartheta_3(z_3)-\vartheta_2(z_2)\geq V_3-V_2\right) \\
&\times\left\{\int_{-\infty}^{\infty}\tilde{z}_1\left[\left(\int_{-\infty}^{\vartheta_3(z_3)-\vartheta_2(z_2)} h_{V_3-V_1,V_3-V_2}\left(v_3-v_1,v_3-v_2\right)\,d\left(v_3-v_2\right)\right)\right.\right. \\
&\left.\left.\times\left(\int_{v_3-v_1}^{\infty} f_{\tilde{Z}_1,\vartheta_3(Z_3)-\vartheta_1(Z_1),\vartheta_3(Z_3)-\vartheta_2(Z_2)}\left(\tilde{z}_1,\vartheta_3(z_3)-\vartheta_1(z_1),\vartheta_3(z_3)-\vartheta_2(z_2)\right)\,d\left(\vartheta_3(z_3)-\vartheta_1(z_1)\right)\right)\right]d\tilde{z}_1\right\} \\
&\qquad\qquad \times d\left(\vartheta_3(z_3)-\vartheta_2(z_2)\right)\,d\left(v_3-v_1\right). \qquad (38)
\end{aligned}$$

Define the term in braces in (38) as the weight $\eta_{\vartheta_3(Z_3)-\vartheta_2(Z_2),V_3-V_1}(\vartheta_3(z_3)-\vartheta_2(z_2),v_3-v_1)$.

To obtain the denominator for the IV, recall that $S=\sum_{j=1}^{3} j\,D_j$. Substitute $D_1=1-D_2-D_3$,

$$\begin{aligned}
\sum_{j=1}^{3} j\,D_j &= (1-D_2-D_3)+2D_2+3D_3 \\
&= 1+D_2+2D_3.
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{Cov}(S,\tilde{Z}_1) &= E\left(\tilde{Z}_1D_2\right)+2E\left(\tilde{Z}_1D_3\right) \\
&= E\left(\tilde{Z}_1\left(\mathbf{1}\left(R_2\geq R_1,R_2\geq R_3\right)\right)\right) \qquad (39) \\
&\quad +2E\left(\tilde{Z}_1\left(\mathbf{1}\left(R_3\geq R_1,R_3\geq R_2\right)\right)\right).
\end{aligned}$$

Using reasoning similar to that invoked for the analysis of the numerator terms, we obtain expressions for the terms corresponding to the two terms of (37) and (38). We obtain for the first term

of (39)

$$\int_{-\infty}^{\infty}\tilde{z}_1\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\vartheta_2(z_2)-\vartheta_1(z_1)} f_{\tilde{Z}_1,\vartheta_2(Z_2)-\vartheta_1(Z_1),\vartheta_2(Z_2)-\vartheta_3(Z_3)}\left(\tilde{z}_1,\vartheta_2\left(z_2\right)-\vartheta_1\left(z_1\right),\vartheta_2\left(z_2\right)-\vartheta_3\left(z_3\right)\right)\right.$$
$$\times\left(\int_{-\infty}^{\vartheta_2(z_2)-\vartheta_3(z_3)} h_{V_2-V_1,V_2-V_3}\left(v_2-v_1,v_2-v_3\right)\,d\left(v_2-v_3\right)\right)d\left(v_2-v_1\right)$$
$$\left.\times d\left(\vartheta_2\left(z_2\right)-\vartheta_3\left(z_3\right)\right)d\left(\vartheta_2\left(z_2\right)-\vartheta_1\left(z_1\right)\right)\right]d\tilde{z}_1. \tag{40}$$

By Fubini's Theorem, we obtain:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\tilde{z}_1\left[\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\vartheta_2(z_2)-\vartheta_3(z_3)} h_{V_2-V_1,V_2-V_3}\left(v_2-v_1,v_2-v_3\right)\,d\left(v_2-v_3\right)\right)\right.$$
$$\times\left(\int_{v_2-v_1}^{\infty} f_{\tilde{Z}_1,\vartheta_2(Z_2)-\vartheta_1(Z_1),\vartheta_2(Z_2)-\vartheta_3(Z_3)}\left(\tilde{z}_1,\vartheta_2\left(z_2\right)-\vartheta_1\left(z_1\right),\vartheta_2\left(z_2\right)-\vartheta_3\left(z_3\right)\right)\right.$$
$$\left.\left.\times d\left(\vartheta_2\left(z_2\right)-\vartheta_1\left(z_1\right)\right)\right)d\left(\vartheta_2\left(z_2\right)-\vartheta_3\left(z_3\right)\right)\right]d\left(v_2-v_1\right)\,d\tilde{z}_1 \tag{41}$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\eta_{\vartheta_2(Z_2)-\vartheta_3(Z_3),V_2-V_1}\left(\vartheta_2\left(z_2\right)-\vartheta_3\left(z_3\right),v_2-v_1\right)d\left(v_2-v_1\right)\,d\left(\vartheta_2\left(z_2\right)-\vartheta_3\left(z_3\right)\right).$$

By parallel logic, we obtain for the second term in 39:

$$2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\tilde{z}_1\left[\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\vartheta_3(z_3)-\vartheta_2(z_2)} h_{V_3-V_1,V_3-V_2}\left(v_3-v_1,v_3-v_2\right)\,d\left(v_3-v_2\right)\right)\right.$$
$$\times\left(\int_{v_3-v_1}^{\infty} f_{\tilde{Z}_1,\vartheta_3(Z_3)-\vartheta_1(Z_1),\vartheta_3(Z_3)-\vartheta_2(Z_2)}\left(\tilde{z}_1,\vartheta_3\left(z_3\right)-\vartheta_1\left(z_1\right),\vartheta_3\left(z_3\right)-\vartheta_2\left(z_2\right)\right)d\left(\vartheta_3\left(z_3\right)-\vartheta_1\left(z_1\right)\right)\right)$$
$$\left.\times\,d\left(\vartheta_3\left(z_3\right)-\vartheta_2\left(z_2\right)\right)\right]d\left(v_3-v_1\right)\,d\tilde{z}_1$$
$$= 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\eta_{\vartheta_3(Z_3)-\vartheta_2(Z_2),V_3-V_1}\left(\vartheta_3\left(z_3\right)-\vartheta_2\left(z_2\right),v_3-v_1\right)\,d\left(v_3-v_1\right)\,d\left(\vartheta_3\left(z_3\right)-\vartheta_2\left(z_2\right)\right).$$

These terms can be identified from a structural analysis using the joint distribution of (Z, D_1, D_2, D_3). Collecting results, we obtain decomposition (33) in the text if we multiply both the numerator and denominator by -1.

Table 1: Treatment Effects and Estimands as Weighted Averages of the Marginal Treatment Effect

$\text{ATE}(x) = E\left(Y_1 - Y_0 \mid X = x\right) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, du_D$
$\text{TT}(x) = E\left(Y_1 - Y_0 \mid X = x, D = 1\right) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{TT}}(x, u_D)\, du_D$
$\text{TUT}(x) = E\left(Y_1 - Y_0 \mid X = x, D = 0\right) = \int_0^1 \Delta^{\text{MTE}}\left(x, u_D\right)\, \omega_{\text{TUT}}\left(x, u_D\right)\, du_D$
Policy Relevant Treatment Effect $(x) = E\left(Y_{a'} \mid X = x\right) - E\left(Y_a \mid X = x\right) = \int_0^1 \Delta^{\text{MTE}}\left(x, u_D\right)\, \omega_{\text{PRTE}}\left(x, u_D\right)\, du_D$ for two policies a and a' that affect the Z but not the X
$\text{IV}_J(x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{IV}}^J(x, u_D)\, du_D$, given instrument J
$\text{OLS}(x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{OLS}}(x, u_D)\, du_D$

Table 2: Weights

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

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

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

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

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

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

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

$\omega_0(x, u_D) = \left[\int_{0}^{u_D} f(p \mid X = x)\, dp\right] \dfrac{1}{E((1 - P) \mid X = x)}$

Source: Heckman and Vytlacil (2005)

Table 3: Weights for Different Treatment Parameters

(a) Treatment Effects and Estimands as Weighted Averages of the Marginal Treatment Effect

$$ATE(x) = \int_0^1 MTE(x, u_S)du_S$$

$$TT(x) = \int_0^1 MTE(x, u_S)h_{\text{IV}}(x, u_S)du_S$$

$$TUT(x) = \int_0^1 MTE(x, u_S)h_{TUT}(x, u_S)du_S$$

$$PRTE\,(x) = \int_0^1 MTE\,(x, u_S)\,h_{PRTE}\,(x, u_S)\,du_S.$$

$$\text{IV}(x) = \int_0^1 MTE(x, u_S)h_{\text{IV}}(x, u_S)du_S$$

$$\text{OLS}(x) = \int_0^1 MTE(x, u_S)h_{\text{OLS}}(x, u_S)du_S$$

(b) Weights for ATE, TT, TUT, PRTE, IV and OLS

$$h_{ATE}(x, u_S) = 1$$

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

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

$$h_{PRTE}(x, u_S) = \left[\frac{F_{P^*,X}(u_S) - F_{P,X}(u_S)}{\Delta P}\right]$$

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

$$h_{\text{OLS}} = \frac{E(U_1 \mid X = x, U_S = u_S)h_1(x, u_D)\text{-}E(U_0 \mid X = x, U_S = u_S)h_0(x, u_S)}{MTE(x, u_S)}$$

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

$$h_0(x, u_S) = \left[\int_0^{u_S} f(p \mid X = x)dp\right] \frac{1}{E((1 - P) \mid X = x)}$$

Source: Heckman and Vytlacil (2005).
*For a general instrument, see Heckman, Urzua, and Vytlacil (2006) or Heckman and Vytlacil (2007b).

Table 4: Weights for $MPRTE$

Measure of Distance for People Near the Margin	Definition of Policy Change	Weight
$\lvert\mu_S(Z) - V\rvert < e$	$Z_\alpha^k = Z^k + \alpha$	$h_{MPRTE}(x, u_S) = \frac{f_{P\lvert X}(u_S) f_{V\lvert X}(F^{-1}_{V\lvert X}(u_S))}{E(f_{V\lvert X}(\mu_S(Z))\lvert X)}$
$\lvert P - U\rvert < e$	$P_\alpha = P + \alpha$	$h_{MPRTE}(x, u_S) = f_{P\lvert X}(u_S)$
$\left\lvert \frac{P}{U} - 1\right\rvert < e$	$P_\alpha = (1+\alpha) P$	$h_{MPRTE}(x, u_S) = \frac{u_S f_{P\lvert X}(u_S)}{E(P\lvert X)}$

Source: Carneiro, Heckman and Vytlacil (Forthcoming, 2009).

Table 5: Potential Outcomes, Choice Model and Parameterizations

Outcomes	Choice Model
$Y_j = \alpha_j + U_j \quad \text{with} \quad j \in \mathcal{J} = \{1, 2, 3\}$	$D_j = \begin{cases} 1 & \text{if } R_j \geq R_k \ \forall \ j \in \mathcal{J} \\ 0 & \text{otherwise} \end{cases}$
$Y = \sum_{j \in \mathcal{J}} Y_j D_j$	$R_j = \gamma_j Z_j - V_j \quad \text{with} \quad j \in \mathcal{J}$

Parameterization

$$(U_1, U_2, U_3, V_1, V_2, V_3) \sim N(\mathbf{0}, \mathbf{\Sigma}_{UV}) \quad , \quad (Z_1, Z_2, Z_3) \sim N(\mu_Z, \mathbf{\Sigma}_Z)$$

$$\mathbf{\Sigma}_{UV} = \begin{bmatrix} 0.64 & 0.16 & 0.16 & 0.024 & -0.32 & 0.016 \\ 0.16 & 1 & 0.20 & 0.020 & -0.30 & 0.010 \\ 0.16 & 0.20 & 1 & 0.020 & -0.40 & 0.040 \\ 0.024 & 0.020 & 0.020 & 1 & 0.6 & 0100 \\ -0.32 & -0.30 & -0.40 & 0.6 & 1 & 0.2 \\ 0.016 & 0.01 & 0.040 & 0100 & 0.2 & 1 \end{bmatrix}, \ \mu_Z = (1.0, 0.5, 1.5) \text{ and } \mathbf{\Sigma}_Z = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 \end{bmatrix} = [\ 0.3 \quad 0.1 \quad 0.7\ ], \ \begin{bmatrix} \gamma_1 & \gamma_2 & \gamma_3 \end{bmatrix} = [\ 0.2 \quad 0.3 \quad 0.1\ ]$$

Table 6: Transition Matrix Obtained from the Change in the Instrument Z_1

The Instrument Increases by 0.75 Standard Deviation

		New Value of Instrument ($\widetilde{Z} = Z_1 + 0.75$)			
		$D_1 = 1$	$D_2 = 1$	$D_3 = 1$	*Total*
Original Value of Instrument (Z_1)	$D_1 = 1$	33.17%	0%	0%	*33.17%*
	$D_2 = 1$	3.20%	25.91%	0%	*29.11%*
	$D_3 = 1$	2.43%	0%	35.29%	*37.72%*
	Total	*38.80%*	*25.91%*	*35.29%*	*100%*

Table 7: Marginal Gains Identified from the Change in the Instrument Z_1

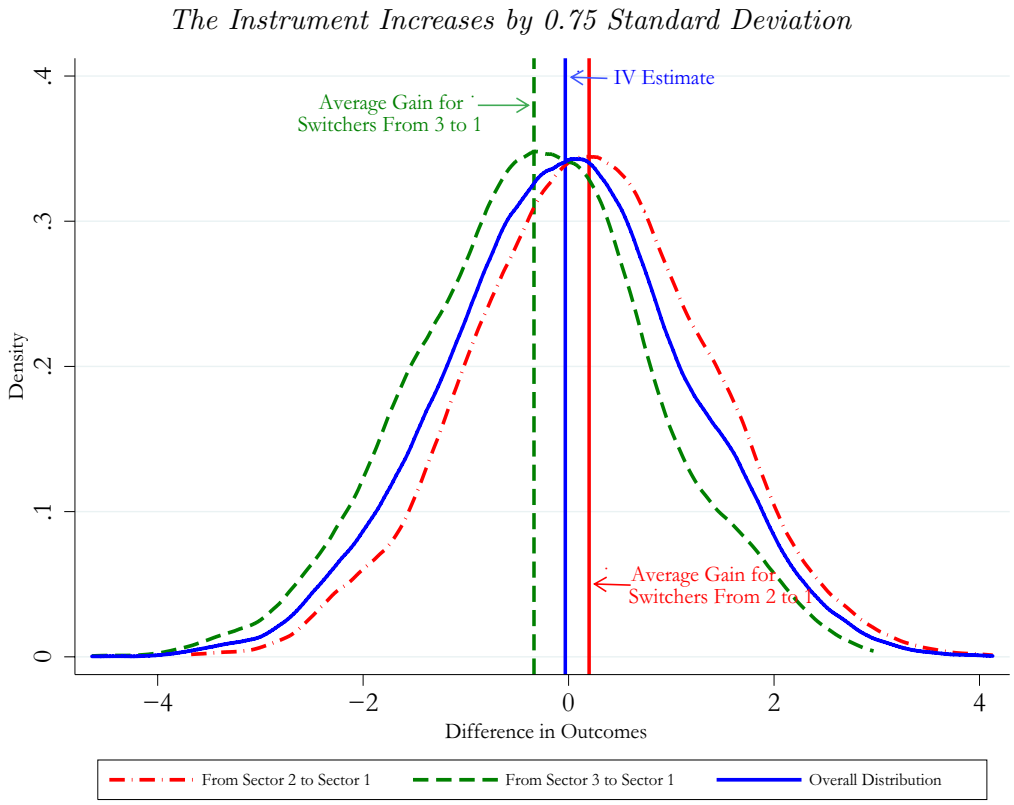
The Instrument Increases by 0.75 Standard Deviation

	Gains to Switchers	Fraction of Population Switching
From 2 to 1	0.199	3.20%
From 3 to 1	-0.336	2.43%
Overall (IV estimate)	-0.032	5.63%

IV Estimate:

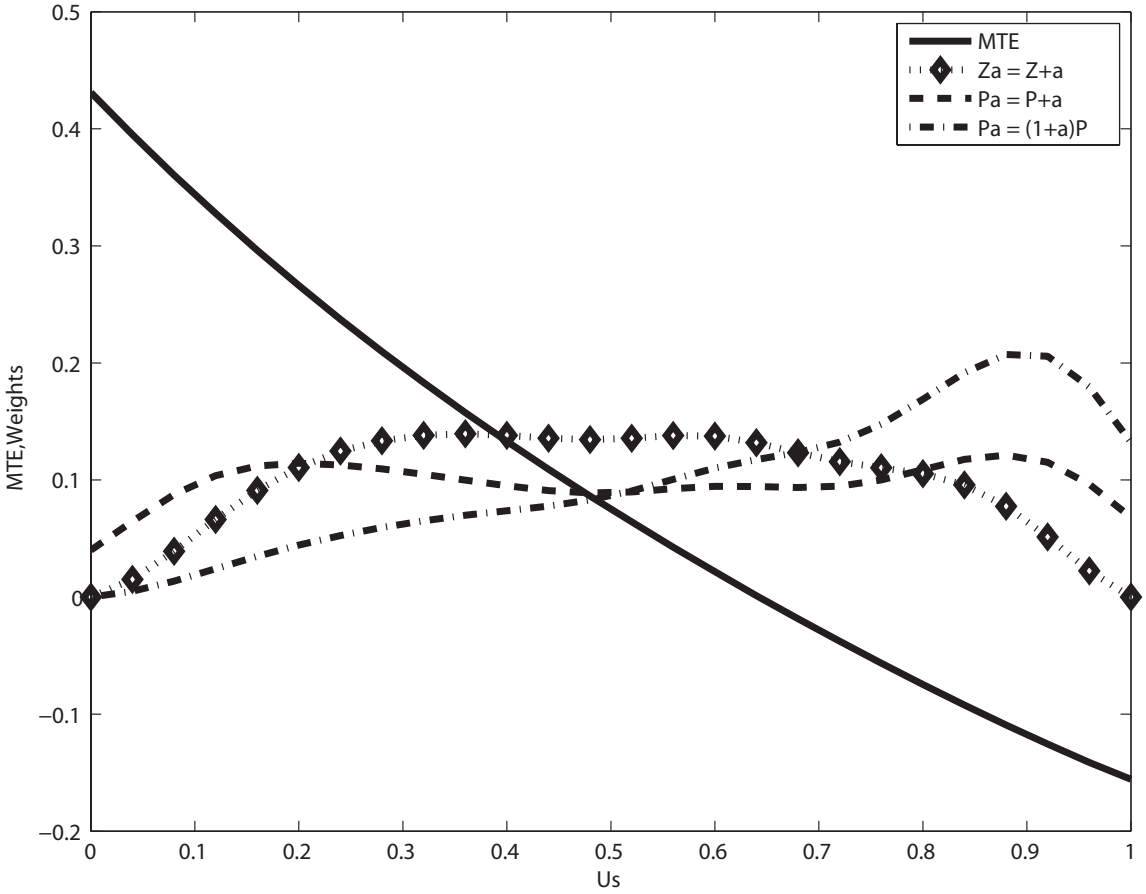
$$E\left[Y|\widetilde{Z}_1\right] - E\left[Y|Z_1\right] = \frac{3.20}{3.20+2.43} \times 0.199 - \frac{2.43}{3.20+2.43} \times 0.336 = -0.032$$

Figure 1: Distribution of Gains in Outcomes Induced by the Change in the Instrument Z_1



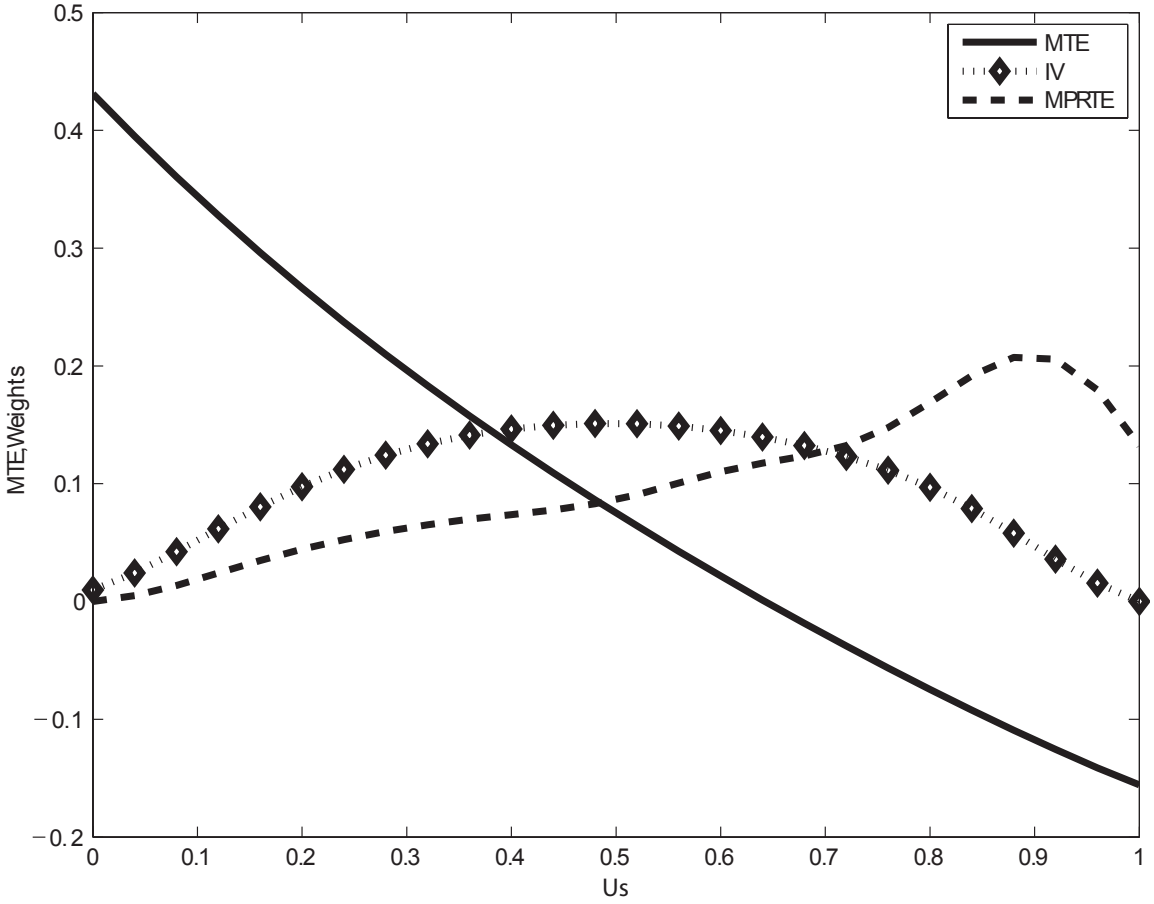
Fraction of Gross Gainers by Source	
% Gross Gainers from 2 to 1	56.8%
% Gross Gainers from 3 to 1	39.3%
% Gross Gainers from all Sources	49.2%

Figure 2: Weights for three different versions of the MPRTE



Notes: The scale of the y-axis is the scale of the MTE, not the scale of the weights, which are scaled to fit the picture.

Figure 3: Weights for IV and MPRTE



Notes: The scale of the y-axis is the scale of the MTE, not the scale of the weights, which are scaled to fit the picture.

References

Abbring, J. H. and J. J. Heckman (2007). Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, pp. 5145–5303. Amsterdam: Elsevier.

Ahn, H. and J. Powell (1993, July). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics 58*(1-2), 3–29.

Angrist, J. D. and G. W. Imbens (1995, June). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association 90*(430), 431–442.

Björklund, A. and R. Moffitt (1987, February). The estimation of wage gains and welfare gains in self-selection. *Review of Economics and Statistics 69*(1), 42–49.

Cameron, S. V. and J. J. Heckman (1998, April). Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of American males. *Journal of Political Economy 106*(2), 262–333.

Cunha, F. and J. J. Heckman (2007, October). The evolution of inequality, heterogeneity and uncertainty in labor earnings in the U.S. economy. *National Bureau of Economic Research Working Paper No. 13526*.

Cunha, F., J. J. Heckman, and S. Navarro (2005, April). Separating uncertainty from heterogeneity in life cycle earnings, The 2004 Hicks Lecture. *Oxford Economic Papers 57*(2), 191–261.

Fox, J. and A. Gandhi (2008, November). Identifying heterogeneity in economic choice and selection models using mixtures. Unpublished manuscript, University of Chicago.

Heckman, J. J. (1980). Addendum to sample selection bias as a specification error. In E. Stromsdorfer and G. Farkas (Eds.), *Evaluation Studies Review Annual*, Volume 5. Beverly Hills: Sage Publications.

Heckman, J. J. (1987). Selection bias and self-selection. In J. Eatwell, M. Milgate, and P. Newman (Eds.), *The New Palgrave: A Dictionary of Economics*, pp. 287–297. London: Palgrave Macmillan Press.

Heckman, J. J. (1990, May). Varieties of selection bias. *American Economic Review 80*(2), 313–318.

Heckman, J. J., H. Ichimura, J. Smith, and P. E. Todd (1998, September). Characterizing selection bias using experimental data. *Econometrica 66*(5), 1017–1098.

Heckman, J. J., P. A. LaFontaine, and P. L. Rodríguez (2008). Taking the easy way out: How the GED testing program induces students to drop out. Unpublished manuscript, University of Chicago, Department of Economics.

Heckman, J. J., A. Layne-Farrar, and P. E. Todd (1996, November). Human capital pricing equations with an application to estimating the effect of schooling quality on earnings. *Review of Economics and Statistics 78*(4), 562–610.

Heckman, J. J., L. J. Lochner, and P. E. Todd (2006). 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*, Chapter 7, pp. 307–458. Amsterdam: Elsevier.

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

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

Heckman, J. J. and R. Robb (1985). Alternative methods for evaluating the impact of interventions. In J. Heckman and B. Singer (Eds.), *Longitudinal Analysis of Labor Market Data*, Volume 10, pp. 156–245. New York: Cambridge University Press.

Heckman, J. J. and R. Robb (1986). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In H. Wainer (Ed.), *Drawing Inferences from Self-Selected Samples*, pp. 63–107. New York: Springer-Verlag. Reprinted in 2000, Mahwah, NJ: Lawrence Erlbaum Associates.

Heckman, J. J., S. Urzua, and E. J. Vytlacil (2006). Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics 88*(3), 389–432.

Heckman, J. J., S. Urzua, and E. J. Vytlacil (2009). Instrumental variables in models with multiple outcomes: The general unordered case. *Les Annales d'Economie et de Statistique*. Forthcoming.

Heckman, J. J. and E. J. Vytlacil (1999, April). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences 96*(8), 4730–4734.

Heckman, J. J. and E. J. Vytlacil (2001a). Instrumental variables, selection models, and tight bounds on the average treatment effect. In M. Lechner and F. Pfeiffer (Eds.), *Econometric Evaluation of Labour Market Policies*, pp. 1–15. New York: Center for European Economic Research.

Heckman, J. J. and E. J. Vytlacil (2001b). Local instrumental variables. 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*, pp. 1–46. New York: Cambridge University Press.

Heckman, J. J. and E. J. Vytlacil (2001c, May). Policy-relevant treatment effects. *American Economic Review 91*(2), 107–111.

Heckman, J. J. and E. J. Vytlacil (2005, May). Structural equations, treatment effects and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Heckman, J. J. and E. J. Vytlacil (2007a). 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 6B, pp. 4779–4874. Amsterdam: Elsevier.

Heckman, J. J. and E. J. Vytlacil (2007b). 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 6B, pp. 4875–5144. Amsterdam: Elsevier.

Imbens, G. W. (2004, February). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics 86*(1), 4 – 29.

Imbens, G. W. and J. D. Angrist (1994, March). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Matzkin, R. L. (1993, July). Nonparametric identification and estimation of polychotomous choice models. *Journal of Econometrics 58*(1-2), 137–168.

Matzkin, R. L. (1994). Restrictions of economic theory in nonparametric methods. In R. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, pp. 2523–58. New York: North-Holland.

Powell, J. L. (1994). Estimation of semiparametric models. In R. Engle and D. McFadden (Eds.), *Handbook of Econometrics, Volume 4*, pp. 2443–2521. Amsterdam: Elsevier.

Rosenbaum, P. R. and D. B. Rubin (1983, April). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Urzua, S. (2008, Fall). Racial labor market gaps: The role of abilities and schooling choices. *Journal of Human Resources 43*(4), 919–971.

Vytlacil, E. J. (2002, January). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica 70*(1), 331–341.

Vytlacil, E. J. (2006, August). A note on additive separability and latent index models of binary choice: Representation results. *Oxford Bulletin of Economics and Statistics 68*(4), 515–518.

Willis, R. J. and S. Rosen (1979, October). Education and self-selection. *Journal of Political Economy 87*(5, Part 2), S7–S36.

Yitzhaki, S. (1989). On using linear regression in welfare economics. Working Paper 217, Department of Economics, Hebrew University.

Yitzhaki, S. (1996, October). On using linear regressions in welfare economics. *Journal of Business and Economic Statistics 14*(4), 478–486.

Yitzhaki, S. and E. Schechtman (2004). The Gini Instrumental Variable, or the "double instrumental variable" estimator. *Metron 62*(3), 287–313.