**UCD GEARY INSTITUTE
DISCUSSION PAPER
SERIES**

Instrumental Variables in Models with Multiple Outcomes: The General Unordered Case

15th December 2008

The views expressed here do not necessarily reflect those of the Geary Institute.
All errors and omissions remain those of the author.

Geary/WP/30/2008

INSTRUMENTAL VARIABLES IN MODELS WITH MULTIPLE OUTCOMES: THE GENERAL UNORDERED CASE[1]

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

This paper develops the method of local instrumental variables for models with multiple, unordered treatments when treatment choice is determined by a nonparametric version of the multinomial choice model. Responses to interventions are permitted to be heterogeneous in a general way and agents are allowed to select a treatment (e.g. participate in a program) with at least partial knowledge of the idiosyncratic response to the treatments. We define treatment effects in a general model with multiple treatments as differences in counterfactual outcomes that would have been observed if the agent faced different choice sets. We show how versions of local instrumental variables can identify the corresponding treatment parameters. Direct application of local instrumental variables identifies the marginal treatment effect of one option versus the next best alternative without requiring knowledge of any structural parameters from the choice equation or any large support assumptions. Using local instrumental variables to identify other treatment parameters requires either large support assumptions or knowledge of the latent index function of the multinomial choice model.

JEL: C31

This paper extends the choice-theoretic analysis of local instrumental variables (*LIV*) and local average treatment effect (*LATE*) by Heckman and Vytlacil (2001, 2005) developed for a two treatment model to the case of multiple treatments with choices generated by a general multinomial choice model. Heckman and Vytlacil (2001) use *LIV* to identify the marginal treatment effect (*MTE*) when the treatment choice is characterized by a binary choice threshold crossing model and interpret this version of *LIV* using choice theory. Vytlacil (2002) shows that the assumptions of Imbens and Angrist (1994) used to define *LATE* both imply and are implied by a nonparametric choice model generated by an index crossing a threshold.

Heckman, Urzua, and Vytlacil (2006) analyze multiple treatment effect models. This paper extends that paper by considering multiple treatments generated by a general unordered choice model. We define treatment parameters for a general multiple treatment problem and present conditions for the application of instrumental variables for identifying a variety of new treatment parameters. Our identification conditions are weaker than the ones used in Heckman and Vytlacil (2007) who establish conditions under which it is possible to nonparametrically identify a full multinomial selection model. Additionally, we illustrate the empirical consequences of our analysis with two examples: GED certification and randomized trial with imperfect compliance.

Our approach relies on choice theory in an essential way. One particularly helpful result we draw on is the representation of the multinomial choices in terms of the choice between a particular choice and the best option among all other choices. This representation is crucial for understanding why *LIV* allows one to identify the *MTE* for the effect of one choice versus the best alternative option. The representation was introduced in Domencich and McFadden (1975), and has been used in the analysis of parametric multinomial selection models by Lee (1983) and Dahl (2002). Unlike those authors, we systematically explore treatment effect heterogeneity, consider nonparametric identification, and examine the application of the LIV methodology to such models.

Our analysis proceeds as follows. We first introduce our nonparametric, multinomial selection model and state our assumptions in Section 1. In Section 2, we define treatment effects in a general unordered model as the differences in the counterfactual outcomes that would have been observed if the agent faced different choice sets, i.e., the effects observed if individuals are forced to choose from one choice set instead of another. We also define the corresponding treatment parameters. Treatment effects in this context exhibit a form of treatment effect heterogeneity not present in the binary treatment case. The new form of heterogeneity arises from agents facing different choice sets.

Section 3 establishes that LIV and the nonparametric Wald-IV estimand produce identification of the $MTE/LATE$ versions of the effect of one choice versus the best alternative option without requiring knowledge of the latent index functions generating choices or large support assumptions. Mean treatment effects comparing one option versus the best alternative are the easiest treatment effects to study using instrumental variable methods because we effectively collapse a multiple outcome model to a series of two outcome models, picking one outcome relative to the rest. In Section 4, we consider a more general case and state conditions for identifying the mean effect of the outcome associated with the best option in one choice set to the mean effect of the best option not in that choice set. We show that identification of the corresponding $MTE/LATE$ parameters requires knowledge of the latent index functions of the multinomial choice model. Thus, to identify the parameters by using IV or LIV requires an explicit choice model. In Section 5, we analyze the identification of treatment parameters corresponding to the mean effect of one specified choice versus another specified choice. Identification of marginal treatment parameters in this case requires the use of "identification at infinity" arguments relying on large support assumptions, but does not require knowledge of the latent index functions of the multinomial choice problem. This use of large support assumptions is closely related to the need for large support assumptions to identify the full model developed in Heckman and Vytlacil (2007). Section 6 concludes.

1 Model and Assumptions

We analyze the following model with multiple choices and multiple outcome states. Let $\mathcal{J}$ denote the agent's choice set, where $\mathcal{J}$ contains a finite number of elements. The value to the agent of choosing option $j \in \mathcal{J}$ is

$$R_j(Z_j) = \vartheta_j(Z_j) - V_j, \tag{1.1}$$

where Z_j are the agent's observed characteristics that affect the utility from choosing choice j, and V_j is the unobserved shock to the agent's utility from choice j. To simplify notation, we will sometimes suppress the argument and write R_j for $R_j(Z_j)$. Let Z denote the random vector containing all unique elements of $\{Z_j\}_{j\in\mathcal{J}}$, i.e., $Z =$ union of $\{Z_j\}_{j\in\mathcal{J}}$. We also sometimes write $R_j(Z)$ for $R_j(Z_j)$, leaving implicit that $R_j(\cdot)$ only depends on those elements of Z that are contained in Z_j. Let $D_{\mathcal{J},j}$ be an indicator variable for whether the agent would choose option j if confronted with choice set $\mathcal{J}$:[2]

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

Let $I_{\mathcal{J}}$ denote the choice that would be made if the agent is confronted with choice set $\mathcal{J}$:

$$I_{\mathcal{J}} = j \iff D_{\mathcal{J},j} = 1.$$

Let $Y_{\mathcal{J}}$ be the outcome variable that would be observed if the agent faced choice set $\mathcal{J}$, determined by

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

where Y_j is the 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 variable.[3] Let X denote the random vector containing all unique elements of $\{X_j\}_{j \in \mathcal{J}}$, i.e., $X =$ union of $\{X_j\}_{j \in \mathcal{J}}$. $(Z, X, I_{\mathcal{J}}, Y_{\mathcal{J}})$ is assumed to be observed.[4] Define $R_{\mathcal{J}}$ as the maximum obtainable value given choice set $\mathcal{J}$:

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

We thus obtain the traditional representation of the decision process that choice j is optimal implies that choice j is better than the "next best" option:

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

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

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

As we will show, this simple, well-known representation of the choice problem is the key intuition for understanding how nonparametric instrumental variables identify the effect of a given choice versus the "next best" alternative.

Analogous to our 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 $Z = z$,

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

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

We invoke the following assumptions, which generalize the assumptions invoked in Heckman and Vytlacil (2001) and later used in Heckman and Vytlacil (2005) and Heckman, Urzua, and Vytlacil (2006) to the general unordered case.

(A-1) The distribution of $(\{V_j\}_{j \in \mathcal{J}})$ is continuous,[5] with support equal to $\Re^{\#\mathcal{J}}$ where $\#\mathcal{J}$ denotes the cardinality of the set $\mathcal{J}$.

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

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

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

Assumption (A-1) and (A-2) imply that $R_j \neq R_k$ w.p.1 for $j \neq k$, so that $\arg\max\{R_j\}$ is unique w.p.1. Assumption (A-3) is required for the mean treatment parameters to be well defined. It allows us to integrate to the limit, which is a crucial step in our identification analysis. Assumption (A-4) requires that at least some individuals participate in each program for all X.

Our definition and analysis of the treatment parameters only uses assumptions (A-1) to (A-4). However, we will also impose an exclusion restriction for our identification analysis. Let $Z^{[l]}$ denote the lth component of Z. Let $Z^{[-l]}$ denote all elements of Z except for the lth component. We work with two alternative assumptions for the exclusion restriction.[6] Consider

(A-5a) 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 nondegenerate,

or

(A-5b) 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.[7]

Assumption (A-5a) 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. Assumption (A-5b) strengthens (A-5a) by adding a smoothness assumption. A necessary condition for (A-5b) is for the excluded variable to have a density with respect to Lebesgue measure conditional on all other regressors and for $\vartheta_j(\cdot)$ to be a continuous and nontrivial function of the excluded variable.[8] Assumption (A-5a) will be used to identify a generalization of the *LATE* parameter. Assumption (A-5b) will be used to identify a generalization of the *MTE* parameter. For certain portions of our analysis we strengthen (A-5b) to a large support condition, though the large support assumption will not be required for most of our results. Note that the required exclusion restriction is for an exogenous covariate that changes the value of one option but (1) does not affect the value of the other options, and (2) does not affect the outcome. We discuss two potential examples of such exclusion restrictions in the next section.

2 Definition of Treatment Effects and Treatment Parameters

Treatment effects are defined as the difference in the counterfactual outcomes that would have been observed if the agent faced different choice sets. For any two choice sets, $\mathcal{K}, \mathcal{L} \subset \mathcal{J}$, define

$$\Delta_{\mathcal{K},\mathcal{L}} = Y_{\mathcal{K}} - Y_{\mathcal{L}}.$$

This is the effect of the individual being forced to choose from choice set $\mathcal{K}$ versus choice set $\mathcal{L}$. The conventional treatment effect is defined as the difference in potential outcomes between two specified states,

$$\Delta_{k,l} = Y_k - Y_l,$$

which is nested within this framework by taking $\mathcal{K} = \{k\}$, $\mathcal{L} = \{l\}$.

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

$$\begin{aligned} \Delta_{\mathcal{K},\mathcal{L}} &= 1(I_{\mathcal{K}} \neq I_{\mathcal{L}})\Delta_{\mathcal{K}\setminus I_{\mathcal{L}},\mathcal{L}} \\ &= 1(I_{\mathcal{K}} \neq I_{\mathcal{L}}) \left(\textstyle\sum_{j\in\mathcal{K}\setminus I_{\mathcal{L}}} D_{\mathcal{K},j}\Delta_{j,\mathcal{L}} \right). \end{aligned} \tag{2.1}$$

In the special case where $\mathcal{L} \subset \mathcal{K}$, $I_{\mathcal{K}} \neq I_{\mathcal{L}}$ implies $I_{\mathcal{K}} \in \mathcal{K} \setminus \mathcal{L}$, and equation (2.1) becomes

$$\begin{aligned} \Delta_{\mathcal{K},\mathcal{L}} &= 1(I_{\mathcal{K}} \in \mathcal{K} \setminus \mathcal{L})\Delta_{\mathcal{K}\setminus\mathcal{L},\mathcal{L}} \\ &= 1(I_{\mathcal{K}} \in \mathcal{K} \setminus \mathcal{L}) \left(\textstyle\sum_{j\in\mathcal{K}\setminus\mathcal{L}} D_{\mathcal{K},j}\Delta_{j,\mathcal{L}} \right). \end{aligned} \tag{2.2}$$

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

To fix ideas regarding these alternative definitions of treatment effects, we consider two examples. The first example concerns GED certification. The GED is an exam that certifies high school dropouts who pass a test as the equivalents of high school graduates.[9]

Example: GED Certification. Consider studying the effect of GED certification on later wages. Consider the case where $\mathcal{J} = \{\ \{\text{GED}\}, \{\text{HS Degree}\}, \{\text{Permanent Dropout}\}\}$. Let $j = \{\text{GED}\}$, $k =\{\text{HS Degree}\}$, and $l =\{\text{Permanent Dropout}\}$. Suppose one wishes to study the effect of the GED on later earnings. Then possible definitions of the effect of the GED include:

- $\Delta_{j,k}$ is the individual's outcome if he or she received the GED versus if he or she had graduated from High School;
- $\Delta_{j,l}$ is the individual's outcome if he or she received the GED versus if he or she had been a permanent dropout;
- $\Delta_{j,\mathcal{J}\setminus j}$ is the individual's outcome if he or she had received the GED versus what the outcome would have been if he or she had not had the option of receiving the GED;
- $\Delta_{\mathcal{J},\mathcal{J}\setminus j}$ is the individual's outcome if he or she had the option of receiving the GED versus the outcome if he or she did not have the option of receiving the GED. Notice that $\Delta_{\mathcal{J},\mathcal{J}\setminus j}$ is a version of an option value treatment effect.

In this example, we assume access to a variable that influences the value function for GED but not the value function for the other choices and not earnings directly. Examples of variables that might satisfy this condition include state level variation in the age at which one can obtain the GED and state level variation in the minimum test score for GED certification.[10] The exclusion restriction in this case is that a lower minimum age or lower minimum test score to obtain the GED makes it easier to obtain the GED but does not directly affect the value of being a permanent dropout, does not directly affect the value of a high school degree, and does not directly affect the wages associated with these counterfactual states. This exclusion restriction rules out, e.g., the possibility that a lower test score threshold for GED certification causes some individuals who otherwise would have been permanent dropouts to not become GED recipients but instead to become high school graduates.

Example: Randomized Trial with Imperfect Compliance. Another example is a randomized trial with multiple treatments and imperfect compliance. For example, the randomized trial might provide funding for different types of treatment, but some individuals who are provided funding might not take up the form of training for which they are funded, and others who are not funded might still receive the training.[11] For this example one possible exclusion might be that funding for a particular treatment increases the value function of that type of treatment but does not directly affect the value of other forms of treatment or the value of no treatment, and does not directly affect earnings. For example, we might have $\mathcal{J} = \{$ {No Training}, {Classroom Training}, {Job Search Assistance}}. Let $j =$ {No Training}, $k =${Classroom Training}, and $l =${Job Search Assistance}. In this example, $\Delta_{\mathcal{J},\mathcal{J}\setminus k}$ is the individual's outcome if he or she had the option of receiving the classroom training versus the outcome if he or she did not have the option of receiving the classroom training. People may be randomly assigned to either receive funding for classroom assistance, to receive funding for job search assistance, or not to receive funding for any form of training. A possible exclusion restriction is that funding for classroom training increases the value to the agent of receiving classroom training but does not directly affect the value of no training or the value of job search assistance. This exclusion restriction rules out the possibility that funding for classroom training causes some individuals who otherwise would not have received any training to receive no classroom training but instead to receive job search assistance. Any value-function argument exclusion will work.

2.1 Treatment Parameters

The conventional definition of the average treatment effect (ATE) is

$$\Delta_{k,l}^{ATE}(x,z) = E(\Delta_{k,l}|X=x, Z=z),$$

which immediately generalizes to the class of parameters discussed in this section as:

$$\Delta_{\mathcal{K},\mathcal{L}}^{ATE}(x,z) = E(\Delta_{\mathcal{K},\mathcal{L}}|X=x, Z=z).$$

The conventional definition of the treatment on the treated (TT) parameter is

$$\Delta_{k,l}^{TT}(x,z) = E(\Delta_{k,l}|X=x, Z=z, I_{\mathcal{J}}=k),$$

which we generalize to

$$\Delta_{\mathcal{K},\mathcal{L}}^{TT}(x,z) = E(\Delta_{\mathcal{K},\mathcal{L}}|X=x, Z=z, I_{\mathcal{J}} \in \mathcal{K}).$$

There are connections across parameters for different choice sets. For example, from equation (2.2), we have

$$\Delta_{\mathcal{J},\mathcal{J}\backslash k} = D_{\mathcal{J},k}\Delta_{k,\mathcal{J}\backslash k}.$$

Thus, there is a trivial connection between the ATE parameter for $\Delta_{\mathcal{J},\mathcal{J}\backslash k}$ and the TT parameter for $\Delta_{k,\mathcal{J}\backslash k}$:

$$\Delta^{ATE}_{\mathcal{J},\mathcal{J}\backslash k}(x,z) = \Pr[D_{\mathcal{J},k} = 1|X = x, Z = z]\Delta^{TT}_{k,\mathcal{J}\backslash k}(x,z).$$

More generally, using equation (2.2), we have for $\mathcal{K} \subset \mathcal{J}$,

$$\Delta_{\mathcal{J},\mathcal{J}\backslash\mathcal{K}} = \mathbf{1}[I_{\mathcal{J}} \in \mathcal{K}]\Delta_{\mathcal{K},\mathcal{J}\backslash\mathcal{K}}$$

so that

$$\Delta^{ATE}_{\mathcal{J},\mathcal{J}\backslash\mathcal{K}}(x,z) = \Pr[I_{\mathcal{J}} \in \mathcal{K}|X = x, Z = z]\Delta^{TT}_{\mathcal{K},\mathcal{J}\backslash\mathcal{K}}(x,z).$$

We will focus on $\Delta_{k,\mathcal{J}\backslash k}$, the effect of being forced to choose option k versus being denied option k. However, from the above relationships, our analysis of identification for $\Delta^{TT}_{k,\mathcal{J}\backslash k}(x,z)$ in Section 3 has implications for the identification of $\Delta^{ATE}_{\mathcal{J},\mathcal{J}\backslash k}(x,z)$. Likewise, our results for $\Delta^{TT}_{\mathcal{K},\mathcal{J}\backslash\mathcal{K}}(x,z)$ in Section 4 have implications for the identification of $\Delta^{ATE}_{\mathcal{K},\mathcal{J}\backslash\mathcal{K}}(x,z)$.

We also generalize the Marginal Treatment Effect (MTE) and Local Average Treatment Effect ($LATE$) parameters considered in Heckman and Vytlacil (2001). We generalize the MTE parameter to be the average effect conditional on being indifferent between the best option among choice set $\mathcal{K}$ versus the best option among choice set $\mathcal{L}$ at some fixed value of the instruments, $Z = z$:

$$\Delta^{MTE}_{\mathcal{K},\mathcal{L}}(x,z) = E\big(\Delta_{\mathcal{K},\mathcal{L}}|X = x, Z = z, R_{\mathcal{K}}(z) = R_{\mathcal{L}}(z)\big). \quad (2.3)$$

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

$$\Delta^{LATE}_{\mathcal{K},\mathcal{L}}(x, z, \tilde{z}) = E\big(\Delta_{\mathcal{K},\mathcal{L}} | X = x, Z = z, R_{\mathcal{K}}(\tilde{z}) \geq R_{\mathcal{L}}(\tilde{z}), R_{\mathcal{L}}(z) \geq R_{\mathcal{K}}(z)\big). \tag{2.4}$$

An important special case of this parameter arises when $z = \tilde{z}$ except for elements that enter the index functions only for choices in $\mathcal{K}$ and not for any choice in $\mathcal{L}$. In that special case, equation (2.4) simplifies to

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

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

As a concrete example, return to the case of a randomized trial with imperfect compliance. Suppose Z is a discrete variable denoting whether funding is provided for classroom training, or for job search assistance, or no funding is provided. Let z denote the value that funding is provided for classroom training, and $\tilde{z}$ denotes the value that no funding is provided for any form of training. Let k denote classroom training. Then $\Delta^{LATE}_{k,\mathcal{J}\backslash k}(x, z, \tilde{z})$ denotes the effect of choosing classroom training compared to the option that would have been chosen if classroom training was not available, among those who would have received classroom training if they received funding for it but not otherwise.

We have defined each of these parameters as conditional not only on X but also on the "instruments" Z. In general, the parameters depend on the Z evaluation point. For example, $\Delta^{ATE}_{\mathcal{K},\mathcal{L}}(x, z)$ generally depends on the z evaluation point. To see this, note that $Y_{\mathcal{K}} = \sum_{k \in \mathcal{K}} D_{\mathcal{K},k} Y_k$, and $Y_{\mathcal{L}} = \sum_{l \in \mathcal{L}} D_{\mathcal{L},l} Y_l$. Even if we assume that $Z \perp\!\!\!\perp \{Y_j\}_{j \in \mathcal{J}} \mid X$, but $D_{\mathcal{K},k}$ and $D_{\mathcal{L},l}$ depend on Z conditional on X and thus $Y_{\mathcal{K}} - Y_{\mathcal{L}}$

in general is dependent on Z conditional on X.[12] In other words, even though Z is conditionally independent of each individual potential outcome, it is correlated with which choice is optimal within the sets $\mathcal{K}$ and $\mathcal{L}$ and thus is related to $Y_{\mathcal{K}} - Y_{\mathcal{L}}$. This dependence of the ATE parameters on Z is one of the differences between our analysis for multinomial treatment and the Heckman and Vytlacil (2001) analysis for binary treatment.

2.2 Heterogeneity in Treatment Effects

Consider heterogeneity in the pairwise treatment effect $\Delta_{j,k}$ (with $(j,k) \in \mathcal{J}$) defined as

$$\Delta_{j,k} = Y_j - Y_k = \mu_j(X_j, U_j) - \mu_k(X_k, U_k),$$

which in general will vary with both observables (X_j, X_k) and unobservables (U_j, U_k). Since we have not assumed that the error terms are additively separable, the treatment effect will in general vary with unobservables even if $U_j = U_k$.

The mean treatment parameters for $\Delta_{j,k}$ will differ if the effect of treatment is heterogeneous and agents base participation decisions, in part, on their idiosyncratic treatment effect. In general, the *ATE*, *TT*, and the marginal treatment parameters for $\Delta_{j,k}$ will differ as long as there is dependence between (U_j, U_k) and the decision rule, i.e., if there is dependence between (U_j, U_k) and $\{V_l\}_{l \in \mathcal{J}}$. If we impose that $\{V_l\}_{l \in \mathcal{J}}$ is independent of (U_j, U_k), then the treatment effect is still heterogeneous, but the average treatment effect, average effect of treatment on the treated, and the marginal average treatment effects all coincide.

The literature often imposes additive separability in outcomes between observables and unobservables. In particular, it is commonly assumed that U_j and U_k are scalar random variables and that $Y_j = \mu_j(X_j) + U_j$, $Y_k = \mu_k(X_k) + U_k$. In that case, a common treatment effect model is equivalent to a model with an additive error term that does not vary with the treatment state: $U_j = U_k$.[13] In the special case of additive separability, the treatment parameters for $\Delta_{j,k}$ will be the same even if there is dependence between $\{V_l\}_{l\in\mathcal{J}}$ and (U_j, U_k) as long as $U_j = U_k$.[14]

There is an additional source of treatment heterogeneity in the more general case of $\Delta_{\mathcal{K},\mathcal{L}}$ arising from heterogeneity in which states are being compared. Consider, for example, $\Delta_{j,\mathcal{J}\backslash j}$. We have that

$$\Delta_{j,\mathcal{J}\backslash j} = \sum_{k\in\mathcal{J}\backslash j} D_{\mathcal{J}\backslash j,k}\Delta_{j,k},$$

which will vary over individuals even if each individual has the same $\Delta_{j,k}$ treatment effect. Consider the corresponding ATE and TT parameters:

$$\begin{aligned}\Delta^{ATE}_{j,\mathcal{J}\backslash j}(x,z) &= E(\Delta_{j,\mathcal{J}\backslash j}|X=x,Z=z)\\ &= \sum_{k\in\mathcal{J}\backslash j} Pr(I_{\mathcal{J}\backslash j}=k \mid X=x,Z=z)E(\Delta_{j,k}\mid X=x,Z=z,I_{\mathcal{J}\backslash j}=k),\end{aligned}$$

and

$$\begin{aligned}&\Delta^{TT}_{j,\mathcal{J}\backslash j}(x,z)\\ &= E(\Delta_{j,\mathcal{J}\backslash j}|X=x,Z=z,I_{\mathcal{J}}=j)\\ &= \sum_{k\in\mathcal{J}\backslash j} Pr(I_{\mathcal{J}\backslash j}=k|X=x,Z=z,I_{\mathcal{J}}=j)E(\Delta_{j,k}|X=x,Z=z,I_{\mathcal{J}}=j,I_{\mathcal{J}\backslash j}=k).\end{aligned}$$

Even in the case where $\{U_j\}_{j\in\mathcal{J}}$ is independent of $\{V_j\}_{j\in\mathcal{J}}$, so that $E(\Delta_{j,k}|X=x,Z=z,I_{\mathcal{J}\backslash j}=k) = E(\Delta_{j,k}|X=x,Z=z,I_{\mathcal{J}}=j,I_{\mathcal{J}\backslash j}=k)$, in general $\Delta^{ATE}_{j,\mathcal{J}\backslash j}(x,z) \neq \Delta^{TT}_{j,\mathcal{J}\backslash j}(x,z)$ since in general $\Pr(I_{\mathcal{J}\backslash j}=k \mid X=x,Z=z) \neq \Pr(I_{\mathcal{J}\backslash j}=k|X=x,Z=z,I_{\mathcal{J}}=j)$. The ATE and TT parameters differ in part because they place different weights on the alternative pairwise treatment effects, and differ even in the case where the pairwise (j versus k) treatment effects are common across all individuals. That ATE might not equal TT even when all pairwise treatment effects are common across individuals is another one of the distinctions between our analysis for multinomial treatments and the Heckman and Vytlacil (2001) analysis for binary treatments.

In summary, $\Delta_{j,k}$ will be heterogeneous depending on the functional form of the $\mu_j(\cdot)$ and $\mu_k(\cdot)$ equations and on the pairwise dependence between the U_j and U_k terms. The $\Delta_{j,k}$ mean treatment parameters will also vary depending on the dependence between $\{V_l\}_{l\in\mathcal{J}}$ and (U_j, U_k). For $\Delta_{j,\mathcal{J}\setminus j}$, there is an additional source of heterogeneity—which option is optimal in the set $\mathcal{J} \setminus j$. Even if there is no heterogeneity in the pairwise $\Delta_{j,k}$ terms, there will still be heterogeneity in $\Delta_{j,\mathcal{J}\setminus j}$, and heterogeneity in the corresponding mean treatment parameters.

3 *LIV* and Nonparametric Wald Estimands for One Choice *vs* the Best Alternative

We first consider identification of treatment parameters corresponding to averages of $\Delta_{j,\mathcal{J}\setminus j}$ using either a discrete change (Wald form for the instrumental variables estimand) or using the local instrumental variables (LIV) estimand.[15] The discrete change instrumental variables estimand will allow us to recover a version of the local average treatment effect ($LATE$) parameter.[16] Impose assumption (A-5a), and let $Z^{[l]}$ denote the excluded variable for option j with properties assumed in (A-5a). Define

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

where $z = (z^{[-l]}, z^{[l]})$, $\tilde{z} = (z^{[-l]}, \tilde{z}^{[l]})$, and where for notational convenience we are assuming that $Z^{[l]}$ is the last element of Z. Note that all components of z and $\tilde{z}$ are the same except for the lth component. Without loss of generality, we assume that $\vartheta_j(\tilde{z}) > \vartheta_j(z)$.

If there were no X regressors, and if Z was a scalar, binary random variable, then $\Delta_j^{Wald}(x, z^{[-l]}, z^{[l]}, \tilde{z}^{[l]})$ would be the probability limit of the Wald form of two-stage least squares regression ($2SLS$). With X regressors, and with Z a vector possibly including continuous components, it no longer corresponds to a Wald/*2SLS*, but rather to a nonparametric version of the Wald estimator where the analyst nonparametrically conditions on X and on Z taking one of two specified values.

The local instrumental variables estimator (LIV) estimand introduced in Heckman (1997), and developed further in Heckman and Vytlacil (1999, 2000) and Florens, Heckman, Meghir, and Vytlacil (2002), allows us to recover a version of the Marginal Treatment Effect (MTE) parameter. Impose (A-5b), and let $Z^{[l]}$ denote the excluded variable for option j with properties assumed in (A-5b). The results will be invariant to which particular variable satisfying (A-5b) is used if there is more than one variable with the property assumed in (A-5b). Define

$$\Delta_j^{LIV}(x,z) \equiv \tfrac{\partial}{\partial z^{[l]}} E(Y|X=x, Z=z) \Big/ \tfrac{\partial}{\partial z^{[l]}} Pr(D_{\mathcal{J},j}=1|X=x,Z=z).$$

$\Delta_j^{LIV}(x,z)$ is thus the limit form of $\Delta_j^{Wald}(x, z^{[-l]}, z^{[l]}, \tilde{z}^{[l]})$ as $\tilde{z}^{[l]}$ approaches $z^{[l]}$. Given our previous assumptions, one can easily show that this limit exists w.p.1. LIV corresponds to a nonparametric, local version of indirect least squares. It is a function of the distribution of the observable data, and it can be consistently estimated using any nonparametric estimator of the derivative of a conditional expectation.

Given these definitions, we have the following identification Theorem which is also included in Heckman, Urzua, and Vytlacil (2006).

Theorem 1. *1. Assume (A-1)-(A-4) and (A-5a). Then*

$$\Delta_j^{Wald}(x, z^{[-l]}, z^{[l]}, \tilde{z}^{[l]}) = \Delta_{j,\mathcal{J}\backslash j}^{\mathrm{LATE}}(x, z, \tilde{z})$$

where $\tilde{z} = (z^{[-l]}, \tilde{z}^{[l]})$ *and* $z = (z^{[-l]}, z^{[l]})$.

2. Assume (A1)-(A-4) and (A-5b). Then

$\Delta_j^{LIV}(x, z) = \Delta_{j,\mathcal{J}\backslash j}^{MTE}(x, z).$

Proof. See the Appendix. □

The basic idea is that in this case we can bring the J outcome model into a two outcome model using outcome j versus the next best outcome for all $j = 1, \ldots, J$.

$\Delta_{j,\mathcal{J}\backslash j}^{LATE}(x, z, \tilde{z})$ is the average effect of switching to state j from state $I_{\mathcal{J}\backslash j}$ for individuals who would choose $I_{\mathcal{J}\backslash j}$ at $Z = z$ but would choose j at $Z = \tilde{z}$. $\Delta_{j,\mathcal{J}\backslash j}^{MTE}(x, z)$ is the average effect of switching to state j from state $I_{\mathcal{J}\backslash j}$ (the best option besides state j) for individuals who are indifferent between state j and $I_{\mathcal{J}\backslash j}$ at the given values of the selection indices (i.e., at $Z = z$, $\{\vartheta_k(Z_k) = \vartheta_k(z_k)\}_{k\in\mathcal{J}}$).

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

$$\begin{aligned}\Delta_{j,\mathcal{J}\backslash j}^{LATE}(x, z, \tilde{z}) &= E\big(\Delta_{j,\mathcal{J}\backslash j} | X = x, Z = z, R_j(\tilde{z}) \geq R_{\mathcal{J}\backslash j}(z) \geq R_j(z)\big)\\ &= \sum_{k\in\mathcal{J}\backslash j}\Big[Pr\big(I_{\mathcal{J}\backslash j} = k | Z = z, R_j(\tilde{z}) \geq R_{\mathcal{J}\backslash j}(z) \geq R_j(z)\big)\\ &\qquad \times E\big(\Delta_{j,k} | X = x, Z = z, R_j(\tilde{z}) \geq R_{\mathcal{J}\backslash j}(z) \geq R_j(z), I_{\mathcal{J}\backslash j} = k\big)\Big].\end{aligned}$$

where we are using the result that $R_{\mathcal{J}\backslash j}(z) = R_{\mathcal{J}\backslash j}(\tilde{z})$ since $z = \tilde{z}$ except for one component that only enters the index for the jth option. How heavily each option is weighted in this average depends on the probability $\Pr\big(I_{\mathcal{J}\backslash j} = k | Z = z, R_j(\tilde{z}_j) \geq R_k(z_k) \geq R_j(z_j)\big)$, which in turn depends on $\{\vartheta_k(z_k)\}_{k\in\mathcal{J}\backslash j}$. The higher $\vartheta_k(z_k)$, holding the other indices constant, the larger the weight given to state k as the base state.

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

$$\begin{aligned}
\int \Delta_j^{Wald}(x, z^{[-l]}, z^{[l]}, \tilde{z}^{[l]}) dF_{Z^{[-l]}}(z^{[-l]}) & \\
&= \int \Delta_{j,\mathcal{J}\setminus j}^{LATE}(x, z, \tilde{z}) dF_{Z^{[-l]}}(z^{[-l]}) \\
&= E\big(\Delta_{j,\mathcal{J}\setminus j} | X = x, R_j(Z^{[-l]}, \tilde{z}^{[l]}) \geq R_{\mathcal{J}\setminus j}(Z^{[-l]}) \geq R_j(Z^{[-l]}, z^{[l]})\big),
\end{aligned}$$

and

$$\begin{aligned}
\int \Delta_j^{LIV}(x, z) dF_Z(z) &= \int \Delta_{j,\mathcal{J}\setminus j}^{MTE}(x, z) dF_Z(z) \\
&= E\big(\Delta_{j,\mathcal{J}\setminus j} | X = x, R_j(Z) = R_{\mathcal{J}\setminus j}(Z)\big).
\end{aligned}$$

An examination of the proof of Theorem 1 shows the role of the exclusion restriction, that $Z^{[-l]}$ be excluded from the outcome equation and the value function of other options besides option j. The role of the first aspect of the exclusion restriction, that $Z^{[-l]}$ be excluded from the outcome equation, is completely standard. If this exclusion did not hold, then shifting $Z^{[-l]}$ would not only change the fraction of people entering treatment j but would also shift Y directly, and it would not be possible to disentangle the indirect effect of $Z^{[-l]}$ through treatment choice from the direct effect on Y. The second aspect of the exclusion restriction, that $Z^{[-l]}$ be excluded from the value function of other options besides option j, is perhaps less familiar but is equally important in this context. Given this exclusion restriction, shifting $Z^{[-l]}$ only shifts the value of option j relative to the other options, and does not shift the value of the other options relative to each other. If this exclusion did not hold, shifting $Z^{[-l]}$ would not only cause some people to switch into/out of option j, but also cause some people to switch between the other options, and it would not be possible to disentangle the effect of $Z^{[-l]}$ shifting people into/out of option j versus shifting people between the other options.

In the GED example above, if the age at which one is allowed to take the GED changes only the value of a GED but not the value of being a permanent dropout or the value of a high school diploma, then a drop in the minimum age of GED certification only causes individuals to shift from permanent dropout to GED and from high school graduate into GED, but does not cause individuals to shift from the permanent dropout state to the high school graduate state. On the other hand, if the minimum age of GED receipt also changes the value of being a permanent dropout or the value of being a being a high school graduate, then changes in the minimum age of GED receipt would also cause individuals to shift from permanent dropout to high school graduate (or vice versa). In this case, it would be impossible to disentangle the effect of a change in the minimum age for GEDs due to the flow of people into or out of GED status from the effect of a change in the minimum age of GED through people switching from being permanent dropouts to becoming high school graduates.

As another example, again consider our job training example with imperfect compliance. If funding for classroom training only affects the value of classroom training but not the value of no training or the value of job search assistance (JSA), then provision of funding for classroom training will cause people to switch from no training and from JSA into classroom training. On the other hand, if provision of funding for classroom training also affects the value of JSA, then provision of classroom funding will not only cause individuals to shift from other categories to classroom training but also induce shifts from no training to JSA (or vice versa). In that case, it is not possible to disentangle the indirect effect of provision of funding for classroom training through increased receipt of classroom training from the effect of people switching from no training to JSA (or vice versa).

Thus far we have only considered identification of marginal treatment effect parameters, $LATE$ and MTE, and not of the more standard treatment parameters like ATE and TT. However, following Heckman and Vytlacil (1999, 2001), $LATE$ can approximate ATE or TT arbitrarily well given the appropriate support conditions.

Theorem 1 shows that we can use Wald estimands to identify $LATE$ for $\Delta_{j,\mathcal{J}\backslash j}$, and we can thus adapt the analysis of Heckman and Vytlacil to identify ATE or TT for $\Delta_{j,\mathcal{J}\backslash j}$. Suppose that $Z^{[l]}$ denotes the excluded variable for option j with properties assumed in (A-5a), and suppose that: (i) the support of the distribution of $Z^{[l]}$ conditional on all other elements of Z is the full real line; and (ii) $\vartheta_j(z_j) \to \infty$ as $z^{[l]} \to \infty$, and $\vartheta_j(z_j) \to -\infty$ as $z^{[l]} \to -\infty$. Then $\Delta^{ATE}_{j,\mathcal{J}\backslash j}(x,z)$ and $\Delta^{LATE}_{j}(x, z^{[-l]}, z^{[l]}, \tilde{z}^{[l]})$ are arbitrarily close when evaluated at a sufficiently large value of $\tilde{z}^{[l]}$ and a sufficiently small value of $z^{[l]}$. Following Heckman and Vytlacil (1999), $\Delta^{TT}_{j,\mathcal{J}\backslash j}(x,z)$ and $\Delta^{LATE}_{j}(x, z^{[-l]}, z^{[l]}, \tilde{z}^{[l]})$ are arbitrarily close for sufficiently small $z^{[l]}$. Using Theorem 1, we can use Wald estimands to identify the $LATE$ parameters, and thus can use the Wald estimand to identify the ATE and TT parameters provided that there is sufficient support for the Z. While this discussion has used the Wald estimands, alternatively we could also follow Heckman and Vytlacil (1999) in expressing ATE and TT as integrated versions of *MTE*. By Theorem 1, we can use LIV to identify MTE and can thus express ATE and TT as integrated versions of the LIV estimand. We next consider a more general class of treatment effects.

4 Identification: Effect of Best Option in $\mathcal{K}$ Versus Best Option not in $\mathcal{K}$

We just presented an analysis of identification for treatment parameters defined as averages of $\Delta_{j,\mathcal{J}\backslash j}$, the effect of choosing option j versus the preferred option in $\mathcal{J}$ if j was not available. We now consider $\Delta_{\mathcal{K},\mathcal{J}\backslash\mathcal{K}}$, the effect of choosing the preferred choice among set $\mathcal{K}$ versus the preferred choice among $\mathcal{J}$ if no option in $\mathcal{K}$ were available. Thus, in this section we compare sets of options, and not just a single option compared to the rest.

We first start with an analysis that varies the $\{\vartheta_k(\cdot)\}_{k\in\mathcal{J}}$ indices directly. This analysis would be useful if one first identifies the index function, e.g. through an identification at infinity argument.[17] We then perform an analysis shifting Z directly. We show that it is possible to identify MTE and $LATE$ averages of the $\Delta_{\mathcal{K},\mathcal{J}\setminus\mathcal{K}}$ effect if one has knowledge of the $\{\vartheta_k(\cdot)\}_{k\in\mathcal{J}}$ index functions but is not possible using shifts in Z without knowledge of the index functions. The one exception to this result is the special case already considered, when $\mathcal{K} = k$, i.e., the set only contains one element, in which case it is possible to identify the marginal parameters using shifts in Z directly without knowledge of the index functions.

Let $\vartheta_{\mathcal{J}}(Z)$ denote a random vector stacking the indices, $\vartheta_{\mathcal{J}}(Z) =$ union of $\{\vartheta_k(Z) : k \in \mathcal{J}\}$. Let $\vartheta_{\mathcal{J}}$ be a vector denoting a potential evaluation point of $\vartheta_{\mathcal{J}}(Z)$, $\vartheta_{\mathcal{J}} = \{\vartheta_k : k \in \mathcal{J}\}$, so that $\vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}$ denotes the event $\{\vartheta_k(Z) = \vartheta_k : k \in \mathcal{J}\}$.[18] Let $\vartheta_{\mathcal{J}} + h$ denote $\{\vartheta_k + h : k \in \mathcal{J}\}$. We now define a version of the Wald estimand that uses the indices directly as instruments instead of using Z as instruments:

$$\tilde{\Delta}^{Wald}_{\mathcal{K}}(x, \vartheta_{\mathcal{J}}, h) \equiv \frac{E(Y|X = x, \vartheta_{\mathcal{K}}(Z) = \vartheta_{\mathcal{K}} + h, \vartheta_{\mathcal{J}\setminus\mathcal{K}}(Z) = \vartheta_{\mathcal{J}\setminus\mathcal{K}}) - E(Y|X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}})}{Pr(I_{\mathcal{J}} \in \mathcal{K}|X = x, \vartheta_{\mathcal{K}}(Z) = \vartheta_{\mathcal{K}} + h, \vartheta_{\mathcal{J}\setminus\mathcal{K}}(Z) = \vartheta_{\mathcal{J}\setminus\mathcal{K}}) - Pr(I_{\mathcal{J}} \in \mathcal{K}|X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}})}.$$

$\tilde{\Delta}^{Wald}_{\mathcal{K}}(x, \vartheta_{\mathcal{J}}, h)$ corresponds to the effect of a shift in each index in $\mathcal{K}$ upward by h while holding each index in $\mathcal{J} \setminus \mathcal{K}$ constant. We define a version of the LIV estimand using indices directly. We define $\tilde{\Delta}^{LIV}_{\mathcal{K}}(x, \vartheta_{\mathcal{J}})$ through a limit expression:

$$\tilde{\Delta}^{LIV}_{\mathcal{K}}(x, \vartheta_{\mathcal{J}}) = \lim_{h\to 0} \tilde{\Delta}^{Wald}_{\mathcal{K}}(x, \vartheta_{\mathcal{J}}, h).$$

Likewise, we define versions of the $LATE$ and MTE parameters that are functions of the ϑ indices instead of functions of z evaluation points,

$$\tilde{\Delta}^{LATE}_{\mathcal{K},\mathcal{L}}(x, \vartheta_{\mathcal{J}}, h) = E\big(\Delta_{\mathcal{K},\mathcal{L}}|X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_{\mathcal{K}}(Z) + h \geq R_{\mathcal{L}}(Z) \geq R_{\mathcal{K}}(Z)\big)$$

$$\tilde{\Delta}^{MTE}_{\mathcal{K},\mathcal{L}}(x,\vartheta_{\mathcal{J}}) = E\big(\Delta_{\mathcal{K},\mathcal{L}}|X=x,\vartheta_{\mathcal{J}}(Z)=\vartheta_{\mathcal{J}},R_{\mathcal{K}}(Z)=R_{\mathcal{L}}(Z)\big)$$

We state the following identification Theorem:

Theorem 2.

1. *Assume (A-1) to (A-4) and (A-5a). Then:*
   $\tilde{\Delta}^{Wald}_{\mathcal{K}}(x,\vartheta_{\mathcal{J}},h) = \tilde{\Delta}^{\text{LATE}}_{\mathcal{K},\mathcal{J}\setminus\mathcal{K}}(x,\vartheta_{\mathcal{J}},h)$,
2. *Assume (A-1) to (A-4) and (A-5b). Then:*
   $\tilde{\Delta}^{LIV}_{\mathcal{K}}(x,\vartheta_{\mathcal{J}}) = \tilde{\Delta}^{MTE}_{\mathcal{K},\mathcal{J}\setminus\mathcal{K}}(x,\vartheta_{\mathcal{J}})$

Proof. Follows with trivial modifications from the proof of Theorem 1. □

Now consider the same analysis shifting Z directly instead of shifting the indices directly. First consider *LATE*. If one knew what shifts in Z corresponded to shifting each index in $\mathcal{K}$ upward by the same amount while holding each index in $\mathcal{J}\setminus\mathcal{K}$ constant, then one could apply the previous analysis to recover $E\big(\Delta_{\mathcal{K},\mathcal{J}\setminus\mathcal{K}}|X = x,\vartheta_{\mathcal{J}}(Z)=\vartheta_{\mathcal{J}},R_{\mathcal{K}}(Z)+h \geq R_{\mathcal{J}\setminus\mathcal{K}}(Z) \geq R_{\mathcal{K}}(Z)\big)$. However, unless $\mathcal{K}$ is a singleton, without knowledge of the index functions one does not know what shifts in Z will have this property. One possible approach would be to only shift elements of Z that are elements of Z_k for $k \in \mathcal{K}$ but are excluded from Z_j for $j \in \mathcal{J}\setminus\mathcal{K}$. However, unless the shifts move the indices for choices in $\mathcal{K}$ all by the same amount, the shift in Z will result in movement not only from the set $\mathcal{J}\setminus\mathcal{K}$ to the set $\mathcal{K}$ but also cause movement between choices within $\mathcal{K}$. Thus, one can use shifts in Z to recover a *LATE*-type parameter for $\Delta_{\mathcal{K},\mathcal{J}\setminus\mathcal{K}}$ only if either (i) the index functions are known, or (ii) $\mathcal{K}=k$, i.e., the set $\mathcal{K}$ contains only one element.

Thus far, we have only considered identification of marginal treatment effect parameters for $\Delta_{\mathcal{K},\mathcal{J}\setminus\mathcal{K}}$ and not of the more standard treatment parameters *ATE* and *TT* for $\Delta_{\mathcal{K},\mathcal{J}\setminus\mathcal{K}}$. As in the previous section, we can follow Heckman and Vytlacil

(1999) in expressing ATE and TT as integrated versions of MTE or show that ATE and TT can be approximated arbitrarily well by $LATE$ parameters. Given appropriate support conditions, we can again identify MTE over the appropriate range or identify the appropriate $LATE$ parameters and thus identify ATE and TT given the required support conditions.

5 Identification: Effect of One Fixed Choice Versus Another

Consider evaluating the effect of fixed option j versus fixed option k, $\Delta_{j,k}$, i.e., the effect for the individual of having no choice except to choose state j versus no choice except to choose state k. We show that it is possible to identify averages of $\Delta_{j,k}$ if one has sufficient support conditions. These conditions supplement the standard IV conditions developed for the binary case (Heckman, Urzua, and Vytlacil, 2006) with the conditions more commonly used in semiparametric estimation. We start by considering the analysis assuming knowledge of the ϑ index functions, and then show that knowledge of the ϑ index functions is not necessary.

For notational purposes, for any $j, k \in \mathcal{J}$, define $U_{j,k} = U_j - U_k$, and let $\vartheta_{j,k}(Z) = \vartheta_j(Z_j) - \vartheta_k(Z_k)$. One could follow our previous strategy to identify treatment parameters for $\Delta_{j,k}$ if one could shift $\vartheta_j - \vartheta_k = \vartheta_{j,k}$ while holding constant $\{\vartheta_{l,m}\}_{(l,m)\in\mathcal{J}\times\mathcal{J}\setminus(j,k)}$, i.e., while holding all other utility contrasts fixed.[19] However, given the structure of the latent variable model determining choices, these are incompatible conditions. To see this, note that $\vartheta_{j,k} = \vartheta_{l,k} - \vartheta_{l,j}$ for any l, and thus $\vartheta_{j,k}$ cannot be shifted while holding $\vartheta_{l,j}$ and $\vartheta_{l,k}$ constant.[20]

To bypass this problem we develop a limit strategy to make the consequences of shifting indices negligible. This strategy relies on an identification at infinity argument. For example, consider the case where $\mathcal{J} = \{1, 2, 3\}$, and consider identification of the *MTE* parameter for option 3 versus option 1. Recall that $D_{\mathcal{J}\backslash 3,l}$ is an indicator variable for whether option l would be chosen if option 3 were not available, so that $D_{\mathcal{J}\backslash 3,l}\Delta_{3,\mathcal{J}\backslash 3} = \sum_{l \in \mathcal{J}\backslash 3} D_{\mathcal{J}\backslash 3,l}\Delta_{3,l}$. Since 1 and 2 are the only options if 3 is not available, it follows that $\Delta_{3,\mathcal{J}\backslash 3} = D_{\mathcal{J}\backslash 3,1}\Delta_{3,1} + D_{\mathcal{J}\backslash 3,2}\Delta_{3,2}$, and we have

$$\begin{aligned} &E\left(\Delta_{3,\mathcal{J}\backslash 3} \mid X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_3(Z) = R_{\mathcal{J}\backslash 3}(Z)\right) \\ &\qquad = E\left(D_{\mathcal{J}\backslash 3,1}\Delta_{3,1} \mid X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_3(Z) = R_{\mathcal{J}\backslash 3}(Z)\right) \\ &\qquad\qquad + E\left(D_{\mathcal{J}\backslash 3,2}\Delta_{3,2} \mid X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_3(Z) = R_{\mathcal{J}\backslash 3}(Z)\right). \end{aligned}$$

The smaller ϑ_2 (holding ϑ_1 and ϑ_3 fixed), the larger the probability that the "next best option" is 1 and not 2. Note that $E\left(\Delta_{3,1} \mid X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_3(Z) = R_1(Z)\right)$ does not depend on the ϑ_2 evaluation point given our independence assumption (A-2) so that

$$\begin{aligned} &E\left(\Delta_{3,1} \mid X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_3(Z) = R_1(Z)\right) \\ &\qquad\qquad = E\left(\Delta_{3,1} \mid X = x, \vartheta_{\mathcal{J}\backslash 2}(Z) = \vartheta_{\mathcal{J}\backslash 2}, R_3(Z) = R_1(Z)\right). \end{aligned}$$

Thus, by Assumptions (A-2)–(A-3) and the Dominated Convergence Theorem, we have that

$$\begin{aligned} &\lim_{\vartheta_2 \to -\infty} E\left(D_{\mathcal{J}\backslash 3,1}\Delta_{3,1} \mid X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_3(Z) = R_{\mathcal{J}\backslash 3}(Z)\right) \\ &\qquad\qquad = E\left(\Delta_{3,1} \mid X = x, \vartheta_{\mathcal{J}\backslash 2}(Z) = \vartheta_{\mathcal{J}\backslash 2}, R_3(Z) = R_1(Z)\right) \end{aligned}$$

while

$$\lim_{\vartheta_2 \to -\infty} E\left(D_{\mathcal{J}\backslash 3,2}\Delta_{3,2} \mid X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_3(Z) = R_{\mathcal{J}\backslash 3}(Z)\right) = 0,$$

so that

$$\lim_{\vartheta_2 \to -\infty} E\left(\Delta_{3,\mathcal{J}\setminus 3} \mid X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_3(Z) = R_{\mathcal{J}\setminus 3}(Z)\right)$$
$$= E\left(\Delta_{3,1} \mid X = x, \vartheta_{\mathcal{J}\setminus 2}(Z) = \vartheta_{\mathcal{J}\setminus 2}, R_3(Z) = R_1(Z)\right).$$

In other words, as the value of option 2 becomes arbitrarily small, the probability of the "next best option" being 1 becomes arbitrarily close to one, and thus the *MTE* parameter for option 3 versus the next best option becomes arbitrarily close to the *MTE* parameter for option 3 versus option 1.

We can identify the *MTE* parameter for option 3 versus the next best option using the *LIV* estimand as in Theorem 1, and thus conditioning on ϑ_2 arbitrarily small we have that the *LIV* estimand is arbitrarily close to the *MTE* parameter for option 3 versus option 1. This analysis requires the appropriate support conditions in order for the limit operations to be well defined. The following Theorem formalizes this idea for the more general case where $\mathcal{J}$ is a general finite set.

Theorem 3. *Assume (A-1) to (A-4) and (A-5b). Assume that, for any $t \in \Re$,*

$$Pr\big(\vartheta_l(Z_l) \leq t \big| \vartheta_j(Z_j), \vartheta_k(Z_k)\big) \geq 0 \quad \forall \;\; l \in \mathcal{J} \setminus \{j,k\}.$$

Then

$$\lim_{\max\limits_{l \in \mathcal{J}\setminus\{j,k\}}\{\vartheta_l\} \to -\infty} \tilde{\Delta}_j^{LIV}(x, \vartheta_{\mathcal{J}}) = E\big(\Delta_{j,k} \big| X = x, \vartheta_{j,k}(Z) = \vartheta_{j,k}, R_j(Z) = R_k(Z)\big)$$

for any

$$x \in \lim_{t \to -\infty} Supp(X | \vartheta_j(Z_j) = \vartheta_j, \vartheta_k(Z_k) = \vartheta_k, \max_{l \in \mathcal{J}\setminus\{j,k\}}\{\vartheta_l(Z)\} \leq t).$$

Proof. By a trivial modification to the proof of Theorem 1, we have that $\tilde{\Delta}_j^{LIV}(x, \vartheta_{\mathcal{J}}) =$

$E(\Delta_{j,\mathcal{J}\setminus j}|X = x, \vartheta_{\mathcal{J}}(Z) = \vartheta_{\mathcal{J}}, R_j(Z) = R_{\mathcal{J}\setminus j}(Z)\big)$. The remainder of the proof follows from an immediate extension of the 3-option case analyzed in the text. $\square$

Thus, for x values in the appropriate limit support, we can approximate $E\big(\Delta_{j,k}\big|X = x, \vartheta_{j,k}(Z) = \vartheta_{j,k}, R_j(z) = R_k(z)\big)$ arbitrarily well by $\Delta_j^{LIV}(x, \vartheta_{\mathcal{J}})$ for an arbitrarily small $\max_{l\in\mathcal{J}\setminus\{j,k\}}\{\vartheta_l\}$.

This analysis uses the ϑ index functions directly, but the results can be restated without using the ϑ functions directly. Again consider the three-choice example. The central aspect of the identification strategy is to "zero-out" the second choice by making ϑ_2 arbitrarily small, allowing one to then use the LIV estimand to identify the MTE parameter for the first option versus the third as if the second choice was not an option. If we do not know the ϑ_2 function, we cannot condition on it. However, if we know that ϑ_2 is decreasing in a particular element of Z, say $Z^{[l]}$, where $Z^{[l]}$ does not enter the index function for choices 1 and 3 and where $\vartheta_2(z_2) \to 0$ as $z^{[l]} \to -\infty$, then we can follow the same strategy as if we knew the ϑ_2 index except conditioning on $Z^{[l]}$ being small instead of conditioning on ϑ_2 being small. The idea then naturally extends to the case of more than three options.

It is useful to compare and contrast the support condition here with those used by Heckman and Vytlacil (2007) to identify the full nonparametric selection model. In the case of three options, in order to identify the marginal treatment effect for choice 1 versus 3 in this paper we need to have a large support assumption on one index – the index for option 2 while holding constant the Z variables that enter the indices for options 1 and 3. In contrast, the required support assumption to identify the full nonparametric selection model is stronger. The condition in that case requires a large support assumption on all three indices.

We can follow Heckman and Vytlacil (1999) in following a two step identification strategy for ATE and TT parameters of $\Delta_{j,k}$, first identifying the appropriate MTE or $LATE$ parameters and then using them to identify ATE and TT given the appropriate support conditions. Notice that the support conditions are now stronger than what are required to identify the ATE and TT parameters of $\Delta_{j,\mathcal{J}\setminus j}$. For identification of the ATE and TT parameters of $\Delta_{j,\mathcal{J}\setminus j}$, we require a large support assumption only on the jth index. In particular, we require that it is possible to condition on Z values that make ϑ_j arbitrarily small or arbitrarily large while holding the remaining indices fixed. In contrast, for identification of the ATE and TT parameters of $\Delta_{j,k}$, we require a large support assumption on each index. We require that for each index we can condition on Z values that make ϑ arbitrarily small or arbitrarily large while holding the remaining indices fixed. The reason for this stronger condition is that for $\Delta_{j,k}$ we need to use an identification at infinity strategy on all but the j and k indices to even obtain the marginal parameters, and then need an additional identification at infinity step to use the marginal parameters to recover the ATE and TT parameters.

6 Conclusion

This paper extends local instrumental variables analysis to a model with multiple treatments in which treatment choices are determined by a general multinomial choice model. Our analysis extends the analysis developed by Heckman, Urzua, and Vytlacil (2006) to a general unordered case. Local instrumental variables identify the marginal treatment effect corresponding to the effect of one option versus the best alternative option without requiring large support assumptions or knowledge of the parameters of the choice model. This preserves the spirit of the $LATE$ analysis of Imbens and Angrist (1994) and the analysis of Heckman and Vytlacil (2001, 2005). More generally, LIV identifies the marginal treatment effect corresponding to the effect of choosing between one choice set versus not having that choice set available. However, in the general case, identification of the more general parameters requires knowledge (identification) of the structural, latent index functions of the multinomial choice model. *LIV* can also provide identification of the effect of one specified choice versus another, requiring large support assumptions but not knowledge of the latent index functions. In order to identify some treatment parameters we require identification of the latent index functions generating the multinomial choice model or else having large support assumptions. This connects the *LIV* analysis in this paper to the more ambitious but demanding identification conditions for the full multinomial selection model developed in Heckman and Vytlacil (2007) .

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

Dept. of Economics, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, U.S.A.; s-urzua@northwestern.edu,

Dept. of Economics, Yale University, 30 Hillhouse Avenue, New Haven, CT 06520, U.S.A.; edward.vytlacil@yale.edu

Appendix

Proof of Theorem 1

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

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

Using independence assumption (A-2), $R_{\mathcal{J}\setminus j}(z) - V_j$ is independent of Z conditional on X, so that

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

$\vartheta_k(\cdot)$ does not depend on $z^{[l]}$ for $k \neq j$ by assumption (A-5b), and thus $R_{\mathcal{J}\setminus j}(z)$ does not depend on $z^{[l]}$, and we therefore with an abuse of notation write $R_{\mathcal{J}\setminus j}(z^{[-l]})$ for $R_{\mathcal{J}\setminus j}(z)$. Write $F(\cdot; x, z^{[-l]})$ for the distribution function of $R_{\mathcal{J}\setminus j}(z^{[-l]}) - V_j$ conditional on $X = x$. Then

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

and

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

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

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

As a consequence of (A-1)-(A-3) and (A-5b) we have that $E\left(Y_{I_{\mathcal{J}\backslash j}} \mid X = x, Z = z\right)$ does not depend on $z^{[l]}$. Using the assumptions and the law of iterated expectations, we may write

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

Thus,

$$\begin{aligned} &\frac{\partial}{\partial z^{[l]}} E\left(Y \mid X = x, Z = z\right) \\ &= E\left(Y_j - Y_{I_{\mathcal{J}\backslash j}} \mid X = x, Z^{[-l]} = z^{[-l]}, R_j(z) = R_{\mathcal{J}\backslash j}(z)\right) \left[\frac{\partial}{\partial z_j^{[l]}} \vartheta_j(z_j)\right] f(\vartheta_j(z_j)). \end{aligned}$$

Combining results, we have

$$\begin{aligned} &\frac{\partial}{\partial z^{[l]}} E(Y|X = x, Z = z) \Big/ \frac{\partial}{\partial z^{[l]}} Pr(D_{\mathcal{J},j} = 1 | X = x, Z = z) \\ &\qquad = E\left(Y_j - Y_{I_{\mathcal{J}\backslash j}} \mid X = x, Z^{[-l]} = z^{[-l]}, R_j(z) = R_{\mathcal{J}\backslash j}(z)\right). \end{aligned}$$

Finally, noting that

$$E\left(Y_j - Y_{I_{\mathcal{J}\setminus j}} \mid X = x, Z^{[-l]} = z^{[-l]}, R_j(z) = R_{\mathcal{J}\setminus j}(z)\right)$$
$$= E\left(Y_j - Y_{I_{\mathcal{J}\setminus j}} \mid X = x, Z = z, R_j(z) = R_{\mathcal{J}\setminus j}(z)\right)$$

provides the stated result. The proof for the *LATE* result follows from the parallel argument. □

Notes

[1] This project was supported by NSF grants SES-0241858, SES-0099195, and SES-9709873, and NIH grant R01-HD043411, and a grant from the American Bar Foundation. An early version of this paper was presented at the World Congress of the Econometric Society, London, September 2005.

[2] We will impose conditions such that ties, $R_j = R_k$ for $j \neq k$, occur with probability zero.

[3] More generally, we can allow U_j to be an unobserved random vector.

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

[5] Absolutely continuous with respect to Lebesgue measure on $\Re^{\#\mathcal{J}}$.

[6] 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.

[7] Absolutely continuous with respect to Lebesgue measure.

[8] (A-5b) can be easily relaxed to the weaker assumption that the support of $\vartheta_j(Z_j)$ conditional on $(X, Z^{[-l]})$ contains an open interval, or further weakened to the assumption that the conditional support contains at least one limit point. In these cases, the analysis of this section goes through without change for analysis for points within the open interval or more generally for any limit point.

[9] For a detailed discussion of GED certification, see Cameron and Heckman (1993).

[10] See Heckman and LaFontaine (2008) for further examples.

[11] See Heckman, Hohmann, Smith, and Khoo (2000) for an analysis of noncompliance in the case of job training programs, along with a summary of evidence on the widespread problem of noncompliance.

[12] An exception is if $\mathcal{K} = \{k\}$, $\mathcal{L} = \{l\}$, i.e., both sets are singletons.

[13] More generally, if U_j, U_k are vector valued, then additive separability becomes $Y_j = \mu_{1j}(X_j) + \mu_{2j}(U_j)$, $Y_k = \mu_{1k}(X_k) + \mu_{2k}(U_k)$, and the standard result becomes that a common treatment effect is equivalent to $\mu_{2j}(U_j) = \mu_{2k}(U_k)$.

[14] Because the literature often assumes additive separability in outcome equations, questions of a common treatment effect becomes a question of whether the additively separable error terms differ by treatment state. If the errors terms differ by treatment state, there will be differences in the treatment parameters according to whether the differences in the error terms are stochastically dependent on the participation decision. Aakvik, Heckman, and Vytlacil (1999) examine the case where the outcome variable is binary so that an additive separability assumption is not appropriate and Heckman and Vytlacil (2001, 2005) consider cases without additive separability. Bhattacharya, Shaikh, and Vytlacil (2008), Vytlacil, Santos, and Shaikh (2007) and Vytlacil and Yildiz (2007) develop the case where $U_j = U_k$ but the model is not additively separable.

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

[16] We are using the Z directly in the following manipulations instead of manipulating the $\{\vartheta_l(Z_l)\}_{l\in\mathcal{J}}$ indices. One can modify the following analysis to use $\{\vartheta_l(Z_l)\}_{l\in\mathcal{J}}$, with the disadvantage of requiring identification of $\{\vartheta_l(Z_l)\}_{l\in\mathcal{J}}$ (e.g. by an identification at infinity argument) but with the advantage of being able to follow the analysis of Heckman and Navarro (2007) in not requiring an exclusion restriction if Z contains a sufficient number of continuous variables and there is sufficient variation in the ϑ_k function across k.

[17] See Heckman and Vytlacil (2007).

[18] Note that in our notation, $R_{\mathcal{J}} = \max\{R_j\}_{k\in\mathcal{J}}$ is a scalar, while $\vartheta_{\mathcal{J}}(Z) = \{\vartheta_k(Z) : k \in \mathcal{J}\}$ is a vector.

[19] Alternatively, one can allow $\vartheta_{l,m}(z) \neq \vartheta_{l,m}(z')$ if $\Pr(\varepsilon_{l,m} \in [\vartheta_{l,m}(z), \vartheta_{l,m}(z')]) = 0$. Such a possibility would be ruled out except "at the limit" by the standard assumption that the support of $\varepsilon_{l,m}$ is connected. Even without such an assumption, such a possibility occurring simultaneously for all $(l,m) \in \mathcal{J} \times \mathcal{J} \setminus \{j,k\}$ for a particular z, z' seems extremely implausible, and we will therefore not consider this possibility further.

[20] This restriction is specific to the multinomial choice model we consider, and is not a restriction of sequential models. In sequential models, unexpected innovations in agent information sets will act to shift the current decision without affecting previous decisions. Consider the following sequential model of GED certification. In the first period, the agent chooses to graduate from high school or to drop out of high school. If the agent drops out of high school in the first period, he or she has the option in the second period of attaining GED certification or remaining a dropout permanently. An unexpected shock in the second period to the relative value of GED certification versus permanent dropout status will shift the GED/permanent dropout choice without changing the probability of high school graduation.

References

Aakvik, A., J. J. Heckman, and E. J. Vytlacil (1999): "Training Effects on Employment When the Training Effects are Heterogeneous: An Application to Norwegian Vocational Rehabilitation Programs," University of Bergen Working Paper 0599; and University of Chicago.

Bhattacharya, J., A. Shaikh, and E. Vytlacil (2008): "Treatment Effect Bounds under Monotonicity Assumptions: An Application to Swan Ganz Catheterization," *American Economic Review, Papers and Proceedings*, forthcoming.

Cameron, S. V., and J. J. Heckman (1993): "The Nonequivalence of High School Equivalents," *Journal of Labor Economics*, 11(1, Part 1), 1–47.

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

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

Domencich, T., and D. L. McFadden (1975): *Urban Travel Demand: A Behavioral Analysis*. North-Holland, Amsterdam, Reprinted 1996.

Florens, J.-P., J. J. Heckman, C. Meghir, and E. J. Vytlacil (2002): "Instrumental Variables, Local Instrumental Variables and Control Functions," Discussion Paper CWP15/02, CEMMAP, Under revision, *Econometrica*.

Heckman, J. J. (1997): "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources*, 32(3), 441–462, Addendum published vol. 33 no. 1 (Winter 1998).

Heckman, J. J., N. Hohmann, J. Smith, and M. Khoo (2000): "Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment," *Quarterly Journal of Economics*, 115(2), 651–694.

Heckman, J. J., and P. A. LaFontaine (2008): *The GED and the Problem of Noncognitive Skills in America*. University of Chicago Press, Chicago, Forthcoming.

Heckman, J. J., and S. Navarro (2007): "Dynamic Discrete Choice and Dynamic Treatment Effects," *Journal of Econometrics*, 136(2), 341–396.

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

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

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

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

——— (2007): "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in *Handbook of Economet-*

rics, ed. by J. Heckman, and E. Leamer, vol. 6B, pp. 4779–4874. Elsevier, Amsterdam.

Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

Lee, L.-F. (1983): "Generalized Econometric Models with Selectivity," *Econometrica*, 51(2), 507–512.

Thompson, T. S. (1989): "Identification of Semiparametric Discrete Choice Models," Discussion Paper 249, University of Minnesota Center for Economic Research, Minneapolis, MN.

Vytlacil, E. J. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70(1), 331–341.

Vytlacil, E. J., A. Santos, and A. M. Shaikh (2007): "Limited Dependent Variable Models and Bounds on Treatment Effects: A Nonparametric Analysis," Unpublished manuscript, Columbia University, Department of Economics.

Vytlacil, E. J., and N. Yildiz (2007): "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, Forthcoming.