cemmap
centre for microdata methods and practice

Identifying effects of multivalued treatments

Sokbae Lee
Bernard Salanié

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP34/18

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Identifying Effects of Multivalued Treatments*

Sokbae Lee† **Bernard Salanié**‡

May 30, 2018

Abstract

Multivalued treatment models have typically been studied under restrictive assumptions: ordered choice, and more recently unordered monotonicity. We show how treatment effects can be identified in a more general class of models that allows for multidimensional unobserved heterogeneity. Our results rely on two main assumptions: treatment assignment must be a measurable function of threshold-crossing rules, and enough continuous instruments must be available. We illustrate our approach for several classes of models.

KEYWORDS: Identification, selection, multivalued treatments, instruments, monotonicity, multidimensional unobserved heterogeneity.

*We are grateful to the co-editor and to four anonymous referees for their comments. We also thank Stéphane Bonhomme, Eric Gautier, Joe Hotz, Thierry Magnac, Lars Nesheim, Rodrigo Pinto, Adam Rosen, Christoph Rothe, Azeem Shaikh, Alex Torgovitsky, Ed Vytlacil, and especially Jim Heckman for their very useful suggestions. We also benefited from comments of seminar audiences in Cambridge, Chicago, Duke, Georgetown, Harvard, MIT, NYU, UCL, and Yale. We would like to thank Junlong Feng and Cameron LaPoint for proofreading the paper. This research has received financial support from the European Research Council under the European Community's Seventh Framework Program FP7/2007-2013 grant agreement No. 295298-DYSMOIA and under the Horizon 2020 Framework Program grant agreement No. 646917-ROMIA.

†Columbia University and Institute for Fiscal Studies, sl3841@columbia.edu.

‡Columbia University, bsalanie@columbia.edu.

1 Introduction

Since the seminal work of Heckman (1979), selection problems have been one of the main themes in both empirical economics and econometrics. One popular approach in the literature is to rely on instruments to uncover the patterns of the self-selection into different levels of treatments, and thereby to identify treatment effects. The main branches of this literature are the local average treatment effect (LATE) framework of Imbens and Angrist (1994) and the local instrumental variables (LIV) framework of Heckman and Vytlacil (2005).

The LATE and LIV frameworks emphasize different parameters of interest and suggest different estimation methods. However, they both focus on binary treatments, and restrict selection mechanisms to be "monotonic". Vytlacil (2002) establishes that the LATE and LIV approaches rely on the same monotonicity assumption. For binary treatment models, these approaches require that selection into treatment be governed by a single index crossing a threshold.

Many real-world selection problems are not adequately described by single-crossing models. The literature has developed ways of dealing with less restrictive models of assignment to treatment. Angrist and Imbens (1995) analyze ordered choice models. Heckman, Urzua, and Vytlacil (2006, 2008) show how (depending on restrictions and instruments) a variety of treatment effects can be identified in discrete choice models that are additively separable in instruments and errors. More recently, Heckman and Pinto (2018) define an "unordered monotonicity" condition that is weaker than monotonicity when treatment is multivalued. They show that given unordered monotonicity, several treatment effects can be identified.

Even the most generally applicable of these approaches can still only deal with models of treatment that are formally analogous to an additively separable discrete choice model, as proved in Section 6 of Heckman and Pinto (2018). The key condition is that the data contain changes in instruments that create only *one-way flows* in or out of the treatment cells the analyst is interested in. In binary treatment models, this is exactly the meaning of monotonicity: there cannot be both compliers and defiers, so that LATE estimates the average treatment effect on compliers[1]. Things are somewhat more complex in multivalued treatment models. Unless selection only

[1]de Chaisemartin (2017) shows that under a weaker condition, LATE estimates the average treatment effect on a specific subset of the compliers.

depends on one function of the instruments, there exist changes in instruments that generate two-way flows in and out of any treatment cell. Unordered monotonicity requires that we observe *some* changes in instruments that only induce one way-flows.

This is still too restrictive for important applications. For instance, many transfer programs (or many educational tests) rely on several criteria and combine them in complex ways to assign agents to treatments; and agents add their own objectives and criteria to the list. An additively separable discrete choice model may not describe such a selection mechanism. To see this, start from a very simple and useful application: the double hurdle model, which treats agents only if each of *two* indices passes a threshold[2]. While this is a binary treatment model, the existence of two thresholds makes it non-monotonic: if a change in instruments increases a threshold but reduces the other, some agents move into the treatment group and some move out of it.

The double hurdle model is still unordered monotonic, as any change in instruments that moves the two thresholds in the same direction only creates one-way flows. Now let us change the structure of the model slightly: there are still two thresholds, but we only treat agents who are above one threshold and below the other. As we will see in Section 2, *any* change in instruments that moves both thresholds generates two-way flows, and standard approaches to identification fail. This model of *selection with two-way flows* cannot be represented by a discrete choice model; it is formally equivalent to a discrete choice model with three alternatives in which the analyst only observes partitioned choices (e.g. the analyst only observes whether alternative 2 is chosen or not). Our identification results apply to this variant of the double hurdle model, and to all treatment models generated by a finite family of threshold-crossing rules. In fact, one way to describe our contribution is that it encompasses all additively separable discrete choice models in which the analyst only observes a partition of the set of alternatives.

To illustrate the applicability of our framework, assume that assignment to treatment can be described by a random utility model of choice. Now imagine that, as is common in practice, the analyst only observes choices between *sets* of treatments: e.g., various vocational programs have been aggregated into a "training" category in her dataset. Our methods allow identification of the effect of these different training programs on outcomes, provided that continuous instruments shift their mean

[2]See e.g. Poirier (1980) for a parametric version of this model.

utilities. Variables such as distance to the locations of the training centers or other components of the "full cost" of treatment could serve as instruments in this application. For another example, consider a dynamic sequence of treatments such as the curriculum of a college student or the career of a worker. This could be represented as a "decision tree" in which various threshold-crossing rules govern the path of the individual through time. Again, this type of model can be analyzed using the techniques in this paper. Here we could use measures of performances of the worker, or the grades of the student, as (quasi) continuous instruments in order to infer the effect of each of the possible paths on outcomes. We study related examples more formally in Section 4.

Our analysis allows selection to be determined by a vector of threshold-crossing rules. Each of these rules compares a scalar unobservable to a threshold; these unobservables can be correlated with each other and with potential outcomes. We proceed in two steps. First assume that the thresholds are known to the analyst. We use their values as control variables to deal with multidimensional unobserved heterogeneity. One important difference with the unidimensional case is that in our setting LATE-type estimators can only recover a mixture of causal parameters on groups that cross different thresholds, and are therefore harder to interpret. We establish conditions under which one can identify a generalized version of the marginal treatment effects (MTE) of Heckman and Vytlacil (2005), as well as the probability distribution of unobservables governing the selection mechanism, and more aggregated treatment effects such as the average treatment effect (ATE), quantile treatment effects, the average treatment effect on the treated (ATT), and the policy-relevant treatment effect (PRTE).

Since thresholds often are not known a priori, the second step requires identifying them from the data. This is highly model-specific and the family of models encompassed in this paper is too large and diverse to allow for a general result. We limit our discussion to a few applications; in particular, we provide what we believe are new identification theorems for the double-hurdle model.

We give a detailed comparison of our paper to the existing literature in Section 5. Let us here mention a few points in which our paper differs from the literature. Unlike Imbens (2000), Hirano and Imbens (2004), Cattaneo (2010), and Yang, Imbens, Cui, Faries, and Kadziola (2016), we allow for selection on unobservables. Gautier and Hoderlein (2015) study binary treatment when selection is driven by a rule that is

linear in a vector of unobservable heterogeneity. Lewbel and Yang (2016) consider a different non-monotonic rule for binary treatment to identify the average treatment effect. These two papers break monotonicity in different ways than ours. We focus on the point identification of marginal treatment effects, unlike the research on partial identification (see e.g. Manski (1990), Manski (1997) and Manski and Pepper (2000)). Chesher (2003), Hoderlein and Mammen (2007), Florens, Heckman, Meghir, and Vytlacil (2008), Imbens and Newey (2009), D'Haultfœuille and Février (2015), and Torgovitsky (2015) study models with continuous endogenous regressors. Each of these papers develops identification results for various parameters of interest. Our paper complements this literature by considering multivalued (but not continuous) treatments with more general types of selection mechanisms.

Heckman and Vytlacil (2007, Appendix B) and Heckman, Urzua, and Vytlacil (2008) and more recently Heckman and Pinto (2018) and Pinto (2015) are more closely related to our paper. But they focus on the selection induced by multinomial discrete choice models, whereas our paper allows for more general selection problems.

The paper is organized as follows. Section 2 sets up our framework; it motivates our central assumptions by way of examples. We present and prove our identification results in Section 3. Section 4 applies our results to three important classes of applications, including the models mentioned in this introduction. We relate our contributions to the literature in Section 5. Finally, Section 6 gives the proof of the main theorem. Some further results and details of the omitted proofs are collected in Online Appendices.

2 The Model and our Assumptions

We assume throughout that treatments take values in a finite set of treatments $\mathcal{K}$. This set may be naturally ordered, as with different tax rates. But it may not be, as when welfare recipients enroll in different training schemes for instance; this makes no difference to our results. We assume that treatments are exclusive. This involves no loss of generality as treatment values could easily be redefined otherwise. We denote $K = |\mathcal{K}|$ the number of treatments, and we map the set $\mathcal{K}$ into $\{0, \ldots, K-1\}$ for notational convenience.

We denote $\{Y_k : k \in \mathcal{K}\}$ the potential outcomes. Let D_k be 1 if the k treatment is realized and 0 otherwise. The observed outcome and treatment are $Y := \sum_{k \in \mathcal{K}} Y_k D_k$

and $D := \sum_{k \in \mathcal{K}} kD_k$, respectively.

In addition to the covariates $\boldsymbol{X}$, observed treatment D and outcomes Y, the data contain a random vector $\boldsymbol{Z}$ that will serve as instruments. We always condition on the value of $\boldsymbol{X}$ in our analysis of identification, and thus suppress it from the notation. Observed data consist of a sample $\{(Y_i, D_i, \boldsymbol{Z}_i) : i = 1, \ldots, N\}$ of $(Y, D, \boldsymbol{Z})$, where N is the sample size. We denote the generalized propensity scores by $P_k(\boldsymbol{Z}) := \Pr(D = k|\boldsymbol{Z})$; they are directly identified from the data. Our models of treatment assignment rely on functions of the instruments $Q_j(\boldsymbol{Z})$ that are a priori unknown to the econometrician and will need to be identified. We also introduce random vectors $\boldsymbol{V}$ to represent unobserved heterogeneity.

Let G denote a function defined on the support $\mathcal{Y}$ of Y, which can be discrete, continuous, or multidimensional. We focus on identification of the conditional counterfactual expectations $E\left(G(Y_k)|\boldsymbol{V} = \boldsymbol{v}\right)$ and on measures of treatment effects that can be derived from them. For example, a possible object of interest is the marginal treatment effect (MTE), defined as $E\left(Y_k - Y_l|\boldsymbol{V} = \boldsymbol{v}\right)$. This is similar to the MTE in the binary treatment model, in that it conditions on the value of unobserved heterogeneity in treatment. One important difference is that the link between the unobserved heterogeneity vector $\boldsymbol{V}$ and the generalized propensity scores $\Pr(D = k|\boldsymbol{Z})$ is now more indirect.

Aggregating up would give the mean of the counterfactual outcome $G(Y_k)$ (conditional on the omitted covariates $\boldsymbol{X}$). Once we identify $EG(Y_k)$ for each k, we also identify the average treatment effect $E(G(Y_k) - G(Y_j))$ between any two treatments k and j. Alternatively, if we let $G(Y_k) = \mathbb{1}(Y_k \leq y)$ for some y, where $\mathbb{1}(\cdot)$ is the usual indicator function, then the object of interest is the marginal distribution of Y_k. This leads to the identification of quantile treatment effects.

One of our aims is to relax the usual monotonicity assumption that underlies the LATE and LIV estimators. Consider the following, simple example where $K = 3$, and treatment assignment is driven by a pair of random variables V_1 and V_2 whose marginal distributions are normalized to be $U[0, 1]$.

Example 1 (Selection with Two-Way Flows)**.** Assume that there are two thresholds $Q_1(\boldsymbol{Z})$ and $Q_2(\boldsymbol{Z})$ such that

- $D = 0$ iff $V_1 < Q_1(\boldsymbol{Z})$ and $V_2 < Q_2(\boldsymbol{Z})$,
- $D = 1$ iff $V_1 > Q_1(\boldsymbol{Z})$ and $V_2 > Q_2(\boldsymbol{Z})$,

- $D = 2$ iff $(V_1 - Q_1(\boldsymbol{Z}))$ and $(V_2 - Q_2(\boldsymbol{Z}))$ have opposite signs.

We could interpret Q_1 and Q_2 as minimum grades or scores in a two-part exam or an eligibility test based on two criteria: failing both parts/criteria assigns you to $D = 0$, passing both to $D = 1$, and failing only one to $D = 2$.

If F is the joint cdf of (V_1, V_2), it follows that the generalized propensity scores are

$$\begin{aligned} P_0(\boldsymbol{Z}) &= F\left(Q_1(\boldsymbol{Z}), Q_2(\boldsymbol{Z})\right), \\ P_1(\boldsymbol{Z}) &= 1 - Q_1(\boldsymbol{Z}) - Q_2(\boldsymbol{Z}) + F\left(Q_1(\boldsymbol{Z}), Q_2(\boldsymbol{Z})\right), \\ P_2(\boldsymbol{Z}) &= Q_1(\boldsymbol{Z}) + Q_2(\boldsymbol{Z}) - 2F\left(Q_1(\boldsymbol{Z}), Q_2(\boldsymbol{Z})\right). \end{aligned} \tag{2.1}$$

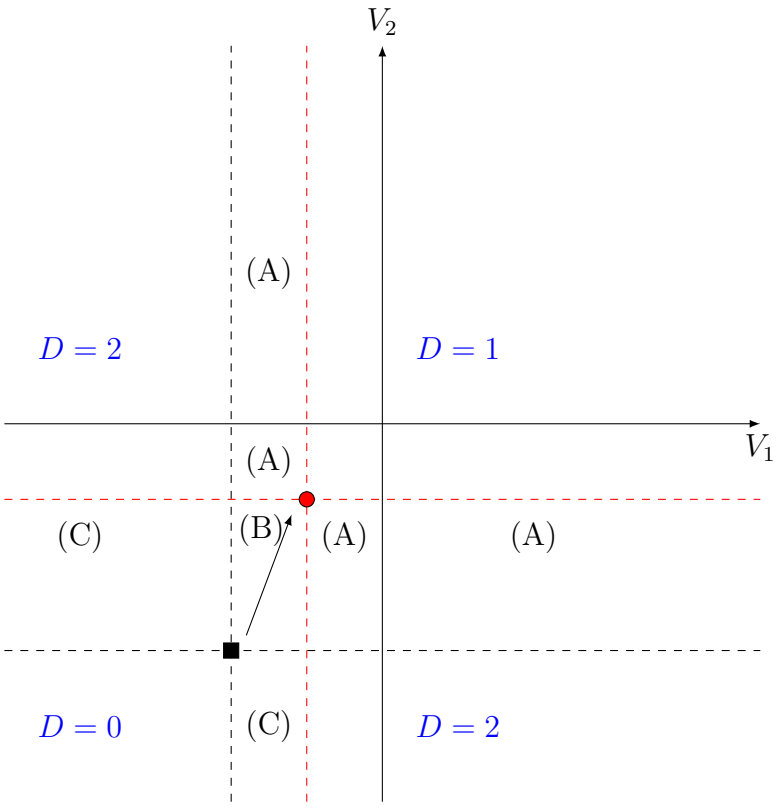
Take a change in the values of the instruments that increases both $Q_1(\boldsymbol{Z})$ and $Q_2(\boldsymbol{Z})$: both criteria, or both parts of the exam, become more demanding. Figure 1 plots this change in (V_1, V_2) space. The black square represents the initial marginal observation, with $V_1 = Q_1(\boldsymbol{Z})$ and $V_2 = Q_2(\boldsymbol{Z})$; and the red circle at the other end of the arrow is the new marginal observation. In both cases, the quadrants delimited by the axes that intersect at the marginal observation define treatment cells. Observations in region (A) move from $D = 1$ to $D = 2$, those in region (B) move from $D = 1$ to $D = 0$, and those in regions (C) move from $D = 2$ to $D = 0$. This violates monotonicity, and even the weaker assumption that generalized propensity scores are monotonic in the instruments. Note also that observations in region (C) leave $D = 2$, while those in region (A) move into $D = 2$: there are *two-way flows* in and out of $D = 2$. Moreover, it is easy to see that *any* change in the thresholds creates such two-way flows; Figure 2 illustrates it for changes in opposite directions, with observations in region (E) moving from $D = 0$ to $D = 2$, observations (F) moving from $D = 2$ to $D = 1$, observations (G) moving from $D = 1$ to $D = 2$, and observations (H) moving from $D = 2$ to $D = 0$.

Therefore this model violates the weaker requirement of unordered monotonicity of Heckman and Pinto (2018), which we describe in Section 5.3—unless we are only interested in treatment values 0 and 1. □

To take a slightly more complicated example, consider the following entry game.

Example 2 (Entry Game)**.** Two firms $j = 1, 2$ are considering entry into a new market. Firm j has profit π_j^m if it becomes a monopoly, and $\pi_j^d < \pi_j^m$ if both firms

Figure 1: Example 1



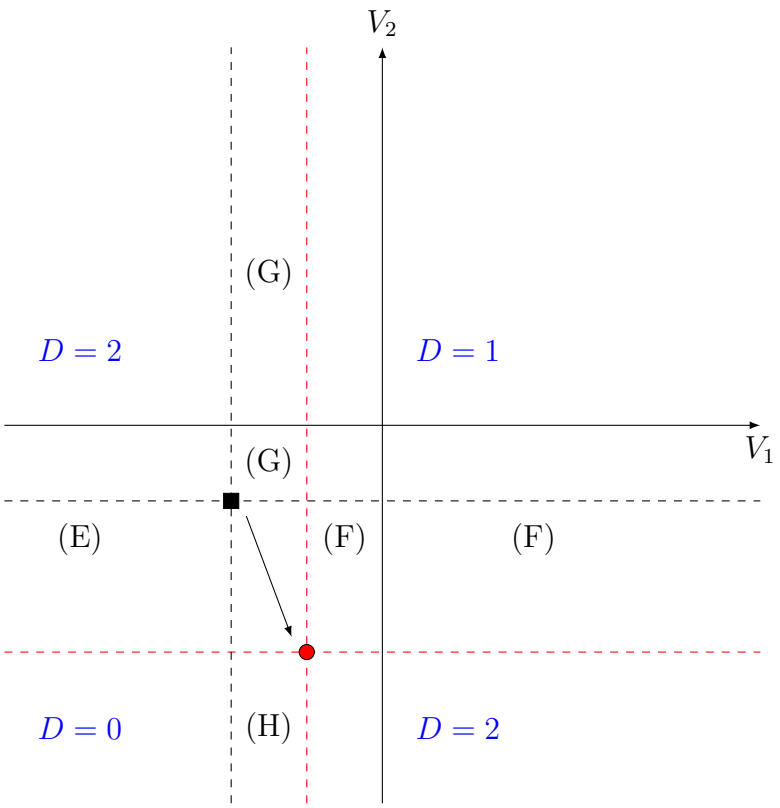
enter. The static Nash equilibria are simple:

- if for both firms $\pi_j^m < 0$, then no firm enters;
- if $\pi_j^m > 0$ and $\pi_k^m < 0$, then only firm j enters;
- if for both firms $\pi_j^d > 0$, then both firms enter;
- if $\pi_j^d > 0$ and $\pi_k^d < 0$, then only firm j enters;
- if $\pi_j^m > 0 > \pi_j^d$ for both firms, then there are two symmetric equilibria, with only one firm operating.

Now let $\pi_j^m = V_j - Q_j(\boldsymbol{Z})$ and $\pi_j^d = \bar{V}_j - \bar{Q}_j(\boldsymbol{Z})$, and suppose we only observe the number $D = 0, 1, 2$ of entrants. Then

- $D = 0$ iff $V_1 < Q_1(\boldsymbol{Z})$ and $V_2 < Q_2(\boldsymbol{Z})$
- $D = 2$ iff $\bar{V}_1 > \bar{Q}_1(\boldsymbol{Z})$ and $\bar{V}_2 > \bar{Q}_2(\boldsymbol{Z})$

Figure 2: Example 1 (continued)



- $D = 1$ otherwise.

This is very similar to the structure of Example 1; in fact it coincides with it in the degenerate case when for each firm, π_m^j and π_d^j have the same sign with probability one[3]. □

2.1 The Selection Mechanism

These two examples motivate the weak assumption we impose on the underlying selection mechanism. In the following we use $\boldsymbol{J}$ to denote the set $\{1, \ldots, J\}$.

Assumption 2.1 (Selection Mechanism)**.** *There exist a finite number J, a vector of unobserved random variables $\boldsymbol{V} := \{V_j : j \in \boldsymbol{J}\}$, and a vector of known functions*

[3]If the econometrician observes the identity of the entrants and not only their numbers, we face the usual partial identification problem generated by the existence of multiple equilibria (see e.g. Tamer, 2003). If equilibrium selection is modeled as an additional threshold-crossing rule, then our approach actually encompasses this case. We refer the reader to Online Appendix C, where we explain this in more detail.

$\{\boldsymbol{Q}_j(\boldsymbol{Z}) : j \in \boldsymbol{J}\}$ *such that any of the following three equivalent statements holds:*

(i) the treatment variable D is measurable with respect to the σ-field generated by the events

$$E_j\left(\boldsymbol{V}, \boldsymbol{Q}(\boldsymbol{Z})\right) := \{V_j < Q_j(\boldsymbol{Z})\} \text{ for } j \in \boldsymbol{J};$$

(ii) each event $\{D = k\} = \{D_k = 1\}$ is a member of this σ-field;

(iii) for each k, there exists a function d_k that is measurable with respect to this σ-field such that $D_k = d_k(\boldsymbol{V}, \boldsymbol{Q}(\boldsymbol{Z}))$.

Moreover, every treatment value k has positive probability.

The threshold conditions in Assumption 2.1 have the "rectangular" form $V_j < Q_j(\boldsymbol{Z})$. Appendix E discusses a more general form of linear inequalities $\boldsymbol{\beta}_j \cdot \boldsymbol{V} < Q_j(\boldsymbol{Z})$. Note that the fact that every observation belongs to one and only one treatment group imposes further constraints. We defer discussion of these constraints to section 4, where we show how they can be used for overidentification tests.

In this notation, the validity of the instruments translates into:

Assumption 2.2 (Conditional Independence of Instruments)**.** *Y_k and $\boldsymbol{V}$ are jointly independent of $\boldsymbol{Z}$ for each $k = 0, \ldots, K-1$.*

To describe the class of selection mechanisms defined in Assumption 2.1 more concretely, we focus on a treatment value k. We define $S_j(\boldsymbol{V}, \boldsymbol{Q}(\boldsymbol{Z})) := \mathbb{1}(V_j < Q_j(\boldsymbol{Z}))$ for $j = 1, \ldots, J$. The σ-field generated by $\{E_j(\boldsymbol{V}, \boldsymbol{Q}(\boldsymbol{Z})) : j = 1, \ldots, J\}$ is obtained by taking unions, intersections, and complements of these E_j sets. These three operations correspond to taking sums, products, and differences of their indicator functions S_j. Therefore the function d_k referred to in Assumption 2.1.(iii) can be written as an algebraic sum of products of the S_j indicator functions. Let $\mathcal{L}$ denote the set of all subsets $l = \left\{l_1, \ldots, l_{|l|}\right\}$ of $\boldsymbol{J}$. Then

$$d_k(\boldsymbol{V}, \boldsymbol{Q}(\boldsymbol{Z})) = \sum_{l \in \mathcal{L}} c_l^k \prod_{j \in l} S_j(\boldsymbol{V}, \boldsymbol{Q}(\boldsymbol{Z})) = \sum_{l \in \mathcal{L}} c_l^k \prod_{m=1}^{|l|} S_{l_m}(\boldsymbol{V}, \boldsymbol{Q}(\boldsymbol{Z})) \tag{2.2}$$

where the c_l^k are algebraic integers. Moreover, this decomposition is unique.

Since $d_k(\boldsymbol{V}, \boldsymbol{Q}(\boldsymbol{Z}))$ depends on $\boldsymbol{V}$ and $\boldsymbol{Q}(\boldsymbol{Z})$ only through $\boldsymbol{S} := \{S_j(\boldsymbol{V}, \boldsymbol{Q}(\boldsymbol{Z})) : j \in \boldsymbol{J}\}$, it will sometimes be convenient to express d_k as a function of $\boldsymbol{S}$, which we

denote $\mathcal{D}_k(\boldsymbol{S})$. For example, if $J = 2$, we have $\mathcal{D}_k(\boldsymbol{S}) = c^k_\emptyset + c^k_{\{1\}} S_1 + c^k_{\{2\}} S_2 + c^k_{\{1,2\}} S_1 S_2$ for some algebraic integers $c^k_\emptyset, c^k_{\{1\}}, c^k_{\{2\}}$, and $c^k_{\{1,2\}}$.

To illustrate this, let us return to Example 1, with $J = 2$ and $K = 3$. For $k = 0$, the selection mechanism is described by the intersection $E_1 \cap E_2$, whose indicator function is $\mathcal{D}_0(\boldsymbol{S}) = S_1 S_2$. Similarly, for $k = 1$ we find $\mathcal{D}_1(\boldsymbol{S}) = (1 - S_1)(1 - S_2)$. Finally, for $k = 2$ we have

$$\mathcal{D}_2(\boldsymbol{S}) = S_1(1 - S_2) + (1 - S_1)S_2 = S_1 + S_2 - 2S_1 S_2.$$

It is useful to think of the products in (2.2) as alternatives in a discrete choice model. For instance, $(1 - S_1)S_2$ could be interpreted as "item 1" having negative value and "item 2" having positive value. In Example 1, $D_2 = 1$ informs us that the values of item 1 and of item 2 have opposite signs. In essence, we are dealing with discrete choice models with only partially observed choices. This analogy will prove useful.

2.2 Indices and Degrees

The term $l = \{1, \ldots, J\} = \boldsymbol{J}$, which corresponds to the product of all J indicator functions S_j in (2.2), plays an important role in our analysis. We will call its c_l the *index of the treatment*.

Definition 2.1. Take a treatment value k in a treatment model with J thresholds. We call the coefficient $c^k_{\boldsymbol{J}}$ in (2.2) the index of treatment k.

In Example 1, the highest order term has a coefficient $c^k_{\{1,2\}} = -1$. With $J = 2$ as in Example 1, the only treatments with a zero index are those which depend on only one threshold: e.g. $\mathbb{1}(V_1 < Q_1)$. But with three or more thresholds ($J > 2$), it is not hard to generate cases in which a treatment value k depends on all J thresholds and still has zero index, as shown in Example 3.

Example 3 (Zero Index)**.** Assume that $J = K = 3$ and take treatment 0 such that

$$\begin{aligned} D_0 = {} & \mathbb{1}(V_1 < Q_1(\boldsymbol{Z}), V_2 < Q_2(\boldsymbol{Z}), V_3 < Q_3(\boldsymbol{Z})) \\ & + \mathbb{1}(V_1 > Q_1(\boldsymbol{Z}), V_2 > Q_2(\boldsymbol{Z}), V_3 > Q_3(\boldsymbol{Z})). \end{aligned}$$

Then the indicator function for $\{D_0 = 1\}$ is

$$d_0 = S_1S_2S_3 + (1 - S_1)(1 - S_2)(1 - S_3) = 1 - S_1 - S_2 - S_3 + S_1S_2 + S_1S_3 + S_2S_3,$$

which has no degree three term. □

When the index is zero as in Example 3, the indicator function of the corresponding treatment k has degree strictly smaller than J. Since Assumption 2.1 rules out the uninteresting cases when treatment k occurs with probability zero or one, its indicator function cannot be constant; and its leading terms have degree $m \geq 1$. We call m the *degree* of treatment k. In Example 3, treatment value 0 has index 0 and degree 2.

The following lemma summarizes the discussion in Sections 2.1 and 2.2:

Lemma 2.1. *Under Assumption 2.1, for each $k \in \mathcal{K}$ there exists a unique family of algebraic integers (c_l^k) such that*

$$d_k(\boldsymbol{V}, \boldsymbol{Q}(\boldsymbol{Z})) = \sum_{l \in \mathcal{L}} c_l^k \prod_{j \in l} S_j(\boldsymbol{V}, \boldsymbol{Q}(\boldsymbol{Z}))$$

where $\mathcal{L}$ is the set of all subsets $l = \{l_1, \ldots, l_{|l|}\}$ of $\boldsymbol{J}$.

The leading terms of the multivariate polynomial $\mathcal{D}_k(\boldsymbol{S})$ have degree $1 \leq m \leq J$, which we also call the degree *of treatment k.*

- *If $m = J$, then the leading term in $\mathcal{D}_k(\boldsymbol{S})$ is*

$$c_{\boldsymbol{J}}^k \prod_{j=1}^{J} S_j.$$

- *if $m < J$, then $c_{\boldsymbol{J}}^k = 0$.*

We call $c_{\boldsymbol{J}}^k$ the index *of treatment k.*

3 Identification Results

In this section we fix $\boldsymbol{x}$ in the support of $\boldsymbol{X}$ and we suppress it from the notation. All the results obtained below are local to this choice of $\boldsymbol{x}$. Global (unconditional)

identification results follow immediately if our assumptions hold for almost every $\boldsymbol{x}$ in the support of $\boldsymbol{X}$.

We only treat the non-zero index in the text. We make this explicit in the following assumption.

Assumption 3.1 (Non-zero index)**.** *The index $c^k_{\boldsymbol{J}}$ defined in Lemma 2.1 is nonzero.*

We analyze zero-index treatments in Appendix A.1.

We require that $\boldsymbol{V}$ have full support:

Assumption 3.2 (Continuously Distributed Unobserved Heterogeneity in the Selection Mechanism)**.** *The joint distribution of $\boldsymbol{V}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^J$ and its support is $[0,1]^J$.*

Note that when $J = 1$, Assumptions 2.1 and 3.2 define the usual threshold-crossing model that underlies the LATE and LIV approaches. However, our assumptions allow for a much richer class of selection mechanisms when $J > 1$. Our Example 1 illustrates that our "multiple thresholds model" does not impose any multidimensional extension of the monotonicity condition that is implicit with a single threshold model. Even when $K = 2$, so that treatment is binary, J could be larger than one. This would allow for flexible treatment assignment: just modify Example 1 to obtain the double hurdle model

$$D = \mathbb{1}\left(V_1 < Q_1(\boldsymbol{Z}) \text{ and } V_2 < Q_2(\boldsymbol{Z})\right).$$

Let $f_{\boldsymbol{V}}(\boldsymbol{v})$ denote the joint density function of $\boldsymbol{V}$ at $\boldsymbol{v} \in [0,1]^J$. Our identification argument relies on continuous instruments that generate enough variation in the thresholds. This motivates the following three assumptions.

For any function ψ of $\boldsymbol{q}$, define "local equicontinuity at $\overline{\boldsymbol{q}}$" by the following property: for any subset $I \subset \boldsymbol{J}$, the family of functions $\boldsymbol{q}_I \mapsto \psi(\boldsymbol{q}_I, \boldsymbol{q}_{-I})$ indexed by $\boldsymbol{q}_{-I} \in [0,1]^{|\boldsymbol{J}-I|}$ is equicontinuous in a neighborhood of $\overline{\boldsymbol{q}}_I$.

Assumption 3.3 (Local equicontinuity at $\boldsymbol{q}$)**.** *The functions $\boldsymbol{v} \mapsto f_{\boldsymbol{V}}(\boldsymbol{v})$ and $\boldsymbol{v} \mapsto E\left(G(Y_k)|\boldsymbol{V} = \boldsymbol{v}\right)$ are locally equicontinuous at $\boldsymbol{v} = \boldsymbol{q}$.*

Assumption 3.3 allows us to differentiate the relevant expectation terms. It is fairly weak: Lipschitz-continuity for instance implies local equicontinuity[4].

[4]It would be easy to adapt our results to cases where, for instance, $\boldsymbol{Q}$ has discontinuities. We do not pursue it in this paper.

Definition 3.1. Let $\mathcal{Z}$ denote the support of $\boldsymbol{Z}$; and $\mathcal{Q} = \boldsymbol{Q}(\mathcal{Z})$ the range of variation of $\boldsymbol{Q}(\boldsymbol{Z})$.

The next two assumptions apply to the functions $\boldsymbol{Q}(\boldsymbol{Z})$ and in particular to their range of variation over the support $\mathcal{Z}$ of $\boldsymbol{Z}$. The functions $\boldsymbol{Q}$ are unknown in most cases, and need to be identified; in this part of the paper we assume that they are known. We will return to identification of the $\boldsymbol{Q}$ functions in Section 3.2.

Assumption 3.4 (Open Range at $\boldsymbol{q}$)**.** *The point $\boldsymbol{q}$ belongs to the interior of the range of variation of the thresholds $\mathcal{Q}$.*

Assumption 3.4 ensures that we can generate any small variation in $\boldsymbol{Q}(\boldsymbol{Z})$ around $\boldsymbol{q}$ by varying the instruments around $\boldsymbol{z}$. This makes the instruments strong enough to deal with multidimensional unobserved heterogeneity $\boldsymbol{V}$.

With J thresholds, Assumption 3.4 requires that $\mathcal{Q}$ contains a J-dimensional neighborhood of $\boldsymbol{q}$. This in turn can only happen (given Assumption 3.3) if the range of variation of the instruments $\mathcal{Z}$ contains an open subset of $\mathbb{R}^J$. Having J-dimensional continuous variation in the instruments is crucial to our approach.

For some corollaries, we use a global version of Assumptions 3.3 and 3.4. To state it formally, we need one last definition.

Definition 3.2. Let $\tilde{\mathcal{Q}} \subset \mathcal{Q}$ denote the set of values $\boldsymbol{q}$ where Assumptions 3.3 and 3.4 both hold.

Assumption 3.5 (Global Condition)**.** $\tilde{\mathcal{Q}}$ *contains* $(0,1)^J$.

Assumption 3.5 requires both that the variation in the instruments generate all possible values of the J thresholds and that Assumptions 3.3 and 3.4 hold everywhere. We do not need this rather stringent assumption to identify the marginal treatment effects; but it is useful to derive various parameters of interest that aggregate the marginal treatment effects.

3.1 Identification with a Non-Zero Index

We are now ready to prove identification of $E\left(G(Y_k)|\boldsymbol{V} = \boldsymbol{q}\right)$ when treatment k has a non-zero index. In the following theorem, for any real-valued function $\boldsymbol{q} \mapsto h(\boldsymbol{q})$, the notation

$$Th(\boldsymbol{q}) \equiv \frac{\partial^J h}{\prod_{j=1}^J \partial q_j}(\boldsymbol{q})$$

refers to the J-order derivative that obtains by taking derivatives of the function h at $\boldsymbol{q}$ in each direction of $\boldsymbol{J}$, when this derivative exists.

Theorem 3.1 (Identification with a non-zero index)**.** *Let Assumptions 2.1, 2.2, 3.1, and 3.2 hold. Fix a value $\boldsymbol{q}$ where Assumptions 3.3 and 3.4 hold; that is, $\boldsymbol{q} \in \tilde{\mathcal{Q}}$. Then the density of $\boldsymbol{V}$ and the conditional expectation of $G(Y_k)$ are given by*[5]

$$f_{\boldsymbol{V}}(\boldsymbol{q}) = \frac{1}{c_{\boldsymbol{J}}^k} T \Pr(D = k | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q})$$

$$E[G(Y_k)|\boldsymbol{V} = \boldsymbol{q}] = \frac{TE\left(G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}\right)}{T \Pr(D = k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q})}.$$

Proof of Theorem 3.1. See section 6. □

For two treatment values k and ℓ, define the marginal treatment effect as

$$\Delta_{\text{MTE}}^{(k,\ell)}(\boldsymbol{v}) := E[G(Y_k)|\boldsymbol{V} = \boldsymbol{v}] - E[G(Y_\ell)|\boldsymbol{V} = \boldsymbol{v}]. \tag{3.1}$$

The MTE function $\boldsymbol{v} \mapsto \Delta_{\text{MTE}}^{(k,\ell)}(\boldsymbol{v})$ is the average treatment effect conditional on $\boldsymbol{V} = \boldsymbol{v}$. Since $\boldsymbol{V}$ is the vector of unobservables that determine the selection mechanism, the MTE function reveals how treatment effects vary with the unobservables governing selection. As such, it captures the effect of selection and it allows the analyst to simulate counterfactual policies. It follows from Theorem 3.1 that if k and ℓ are two treatments to which all of our assumptions apply, then we can identify the marginal treatment effect of moving between these two treatments, as well as the quantile version of this MTE. We also identify the joint density function $\boldsymbol{v} \mapsto f_{\boldsymbol{V}}(\boldsymbol{v})$, which is an object of interest since it describes the dependence among elements of $\boldsymbol{V}$. Appendix A.1 extends Theorem 3.1 to zero-index treatment values; it shows that similar formulæ identify marginal treatment effects averaged over the missing threshold rules.

As in Heckman and Vytlacil (2005), we can identify various treatment effect parameters using Theorem 3.1. The following corollary shows that one can identify the average treatment effect (ATE), the average treatment effect on the treated (ATT), and the policy relevant treatment effect (PRTE) of Heckman and Vytlacil (2001).

[5]The proof of the theorem shows that these derivatives are well-defined.

The PRTE measures the average effect of moving from a baseline policy to an alternative policy. To define the PRTE, consider a class of policies that change $\boldsymbol{Q}$ but that do not affect $E[G(Y_k)|\boldsymbol{V}=\boldsymbol{v}]$. Let D_k^* and Y^*, respectively, denote the treatment choice indicator and the outcome under a new policy $\boldsymbol{Q}^*$. Define $D^* \equiv \sum_{k\in\mathcal{K}} kD_k^*$.

Corollary 3.2. *If Assumption 3.5 holds in addition to the conditions assumed in Theorem 3.1, then the average treatment effect (ATE) and the average treatment effect on the treated (ATT) are identified by*

$$E[G(Y_k) - G(Y_\ell)] = \int \Delta_{MTE}^{(k,\ell)}(\boldsymbol{v})\omega_{ATE}(\boldsymbol{v})d\boldsymbol{v}, \tag{3.2}$$

$$E[G(Y_k) - G(Y_\ell)|D=k] = \int \Delta_{MTE}^{(k,\ell)}(\boldsymbol{v})\omega_{ATT}^{k}(\boldsymbol{v})d\boldsymbol{v}, \tag{3.3}$$

where

$$\begin{aligned}
\omega_{ATE}(\boldsymbol{v}) &:= f_{\boldsymbol{V}}(\boldsymbol{v}), \\
\omega_{ATT}^{k}(\boldsymbol{v}) &:= \frac{\Pr\left[d_k(\boldsymbol{v},\boldsymbol{Q}(\boldsymbol{Z}))=1|\boldsymbol{V}=\boldsymbol{v}\right] f_{\boldsymbol{V}}(\boldsymbol{v})}{\Pr(D=k)}.
\end{aligned}$$

Furthermore, policy relevant treatment effects (PRTEs) are identified by

$$\begin{aligned}
E[G(Y^*)] - E[G(Y)] &= \sum_{k\in\mathcal{K}} \int \Upsilon_k(\boldsymbol{v},\boldsymbol{Q}^*,\boldsymbol{Q}) E[G(Y_k)|\boldsymbol{V}=\boldsymbol{v}] f_{\boldsymbol{V}}(\boldsymbol{v})d\boldsymbol{v}, \\
E[D^*] - E[D] &= \sum_{k\in\mathcal{K}} k \int \Upsilon_k(\boldsymbol{v},\boldsymbol{Q}^*,\boldsymbol{Q}) f_{\boldsymbol{V}}(\boldsymbol{v})d\boldsymbol{v}, \\
E[D_k^*=1] - E[D_k=1] &= \int \Upsilon_k(\boldsymbol{v},\boldsymbol{Q}^*,\boldsymbol{Q}) f_{\boldsymbol{V}}(\boldsymbol{v})d\boldsymbol{v},
\end{aligned}$$

where

$$\Upsilon_k(\boldsymbol{v},\boldsymbol{Q}^*,\boldsymbol{Q}) := \Pr[d_k(\boldsymbol{v},\boldsymbol{Q}^*(\boldsymbol{Z}))=1|\boldsymbol{V}=\boldsymbol{v}] - \Pr[d_k(\boldsymbol{v},\boldsymbol{Q}(\boldsymbol{Z}))=1|\boldsymbol{V}=\boldsymbol{v}].$$

Proof of Corollary 3.2. See Appendix B.1. □

In many applications, the range of variation of the thresholds may be limited so that Assumption 3.5 will not hold. However, it is still possible to construct bounds for the ATE, ATT and PRTE if $G(Y_k)$ is bounded. For example, consider the ATE with $G(Y_k) = \mathbb{1}(Y_k \leq y)$. As shown in the proof of Theorem 3.2, we can point-

identify $E[G(Y_k)|\boldsymbol{V}=\boldsymbol{q}]f_{\boldsymbol{V}}(\boldsymbol{q})$ by $\left(c_{\boldsymbol{J}}^k\right)^{-1} TE\left(G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z})=\boldsymbol{q}\right)$ for each $\boldsymbol{q}\in\tilde{\mathcal{Q}}$. In addition, we know that $G(Y_k)$ lies between 0 and 1. As in Manski (1990) and Heckman and Vytlacil (2000), using this fact we can bound $EG(Y_k)$ within the interval defined by

$$\frac{1}{c_{\boldsymbol{J}}^k}\int_{\boldsymbol{q}\in\mathcal{S}_{\boldsymbol{Q}(\boldsymbol{Z})}} TE\left(G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z})=\boldsymbol{q}\right)d\boldsymbol{q}$$

and

$$\frac{1}{c_{\boldsymbol{J}}^k}\int_{\boldsymbol{q}\in\mathcal{S}_{\boldsymbol{Q}(\boldsymbol{Z})}} TE\left(G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z})=\boldsymbol{q}\right)d\boldsymbol{q}+1-\Pr\left(\boldsymbol{Q}(\boldsymbol{Z})\in\mathcal{Q}\right);$$

and without further information, these bounds are sharp.

Finally, the analyst may only have discrete-valued instruments. A recent literature on the MTE focuses on this case (with binary treatment); it relies on assumed restrictions on the shape of the MTE function (see for example Brinch, Mogstad, and Wiswall (2017), Kowalski (2016) and Mogstad, Santos, and Torgovitsky (2017)). In future work it would be interesting to consider relaxing these restrictions within our framework.

3.2 Identification of Q

So far we assumed that the functions $\{\boldsymbol{Q}_j(\boldsymbol{Z}) : j=1,\ldots,J\}$ were known (see Assumption 2.1). In practice we often need to identify them from the data before applying Theorems 3.1 or A.1. The most natural way to do so starts from the generalized propensity scores $\{P_k(\boldsymbol{Z}) : k=0,\ldots,K-1\}$, which are identified as the conditional probabilities of treatment[6].

First note that by definition (and by Assumption 2.2),

$$\begin{aligned} P_k(\boldsymbol{z}) &= \Pr(D=k|\boldsymbol{Z}=\boldsymbol{z}) \\ &= \int \mathbb{1}\left(d_k\left(\boldsymbol{v},\boldsymbol{Q}(\boldsymbol{z})\right)=1\right) f_{\boldsymbol{V}}(\boldsymbol{v})d\boldsymbol{v}. \end{aligned}$$

Note that this is a J-index model. Ichimura and Lee (1991) consider identification of multiple index models when the indices are specified parametrically. Matzkin (1993,

[6]It would also be possible to seek identification jointly from the generalized propensity scores and from the cross-derivatives that appear in Theorems 3.1 or A.1, especially when they are over-identified. We do not pursue this here.

2007) obtains nonparametric identification results for discrete choice models[7]; but her results only apply to a subset of the types of selection mechanisms we consider (discrete choice models when all choices are observed). Section 4 discusses identification of the $\boldsymbol{Q}$'s through the lens of several models.

4 Applications

Our framework covers a wide variety of commonly used models. For simplicity, we only illustrate its usefulness on two-threshold selection models in this section. These models generate different selection patterns. Not surprisingly, the identification conditions require somewhat stronger instruments as the number of treatment values—the information available to the analyst—decreases.

4.1 Selection with Two-Way Flows

Let us return to Example 1, in which

- $D = 0$ iff $V_1 < Q_1(\boldsymbol{Z})$ and $V_2 < Q_2(\boldsymbol{Z})$,
- $D = 1$ iff $V_1 > Q_1(\boldsymbol{Z})$ and $V_2 > Q_2(\boldsymbol{Z})$,
- $D = 2$ iff $(V_1 - Q_1(\boldsymbol{Z}))$ and $(V_2 - Q_2(\boldsymbol{Z}))$ have opposite signs.

It is useful to start with some exclusion restrictions that help us identify $Q_1(\mathbf{Z})$ and $Q_2(\mathbf{Z})$ separately from the generalized propensity scores given in (2.1). Assume that

Assumption 4.1 (Two Continuous Instruments with Exclusion Restrictions)**.**

1. *The density of* (V_1, V_2) *is continuous on* $[0,1]^2$*, with marginal uniform distributions.*
2. *The instruments* $\mathbf{Z} \equiv (Z_1, Z_2)$ *consist of two scalar random variables whose joint distribution is absolutely continuous with respect to the Lebesgue measure on its support* $\mathcal{Z}$*.*

[7]See Heckman and Vytlacil (2007, Appendix B) for an application to treatment models.

3. *$Q_1(\mathbf{Z})$ does not depend on Z_2, and it is continuously differentiable with respect to Z_1.*

4. *$Q_2(\mathbf{Z})$ does not depend on Z_1, and it is continuously differentiable with respect to Z_2.*

The first condition in Assumption 4.1 is just a normalization of the marginal distribution of each $V_j \in \boldsymbol{V}$. The crucial part of Assumption 4.1 is in the exclusion restrictions: Z_1 affects Q_1 but not Q_2, and Z_2 affects Q_2 but not Q_1. For example, if Q_1 and Q_2 represent minimum required grades on two parts of an exam, then Z_1 should affect only the requirement on the first part, and Z_2 should only affect the second part.

Theorem 4.1 (Identification of Q_1 and Q_2)**.** *Under Assumption 4.1,*

(i) the function

$$P(\boldsymbol{Z}) \equiv 2P_0(\boldsymbol{Z}) + P_1(\boldsymbol{Z})$$

is additively separable in Z_1 and Z_2 on $\mathcal{Z}$.

(ii) Q_1 and Q_2 are identified up to an additive constant. More precisely, take any $(z_1^0, z_2^0) \in \mathcal{Z}$. Then

$$\begin{aligned} Q_1(z_1) &= P(z_1, z_2^0) - P(z_1^0, z_2^0) + C_1^0 \\ Q_2(z_2) &= P(z_1^0, z_2) - C_1^0 \end{aligned}$$

where the constant C_1^0 must satisfy the restrictions $\Pr(D = k) > 0$ for each $k = 0, 1, 2$.[8]

Proof of Theorem 4.1. See Appendix B.2. □

Suppose that the analyst has picked a point in the partially identified (Q_1, Q_2) set. Using these Q_1 and Q_2, since the indices are all nonzero ($c_{\boldsymbol{J}}^0 = c_{\boldsymbol{J}}^1 = 1$ and $c_{\boldsymbol{J}}^2 = -2$) we apply Theorem 3.1 to identify the joint density by

$$f_{V_1,V_2}(q_1, q_2) = \frac{1}{c_{\boldsymbol{J}}^k} \frac{\partial^2 \Pr[D = k | Q_1(\boldsymbol{Z}) = q_1, Q_2(\boldsymbol{Z}) = q_2]}{\partial q_1 \partial q_2}, \tag{4.1}$$

[8]The precise form of these restrictions in terms of C_1^0 and $P(Z_1, Z_2)$ is given in the proof of Theorem 4.1.

where $k = 0, 1, 2$.

Note that $f_{V_1,V_2}(q_1, q_2)$ is overidentified; checking equality between the right-hand sides of (4.1) for $k = 0, 1, 2$ provides a specification test[9]. Similar remarks apply to the conditional expectations $E(Y_k|V_1 = q_1, V_2 = q_2)$; and as

$$E(Y_k|V_1 = q_1, V_2 = q_2) = \frac{\partial^2 E[Y D_k|Q_1(\boldsymbol{Z}) = q_1, Q_2(\boldsymbol{Z}) = q_2]/\partial q_1 \partial q_2}{\partial^2 \Pr[D = k|Q_1(\boldsymbol{Z}) = q_1, Q_2(\boldsymbol{Z}) = q_2]/\partial q_1 \partial q_2}$$

for each $k = 0, 1, 2$, the identification of the marginal and average treatment effects follows immediately.

In practice, Q_1 and Q_2 are only identified up to the (restricted) additive constant C_1^0 in Theorem 4.1(ii). As a consequence, $f_{V_1,V_2}(q_1, q_2)$ and $E(Y_k|V_1 = q_1, V_2 = q_2)$ are only identified up to the corresponding location shift in (q_1, q_2). However, it is easy to check that (4.1) still yields a usable specification test.

4.2 The Double Hurdle Model

Let us now return to the double hurdle model of the introduction, where treatment is binary and the selection mechanism is governed by

$$D = 1 \text{ iff } V_1 < Q_1(\boldsymbol{Z}) \text{ and } V_2 < Q_2(\boldsymbol{Z}), \tag{4.2}$$

and $D = 0$ otherwise.

Both treatment values have non-zero indices: $c_{\boldsymbol{J}}^1 = 1$ and $c_{\boldsymbol{J}}^0 = -1$. But identification of Q_1 and Q_2, which is a premise of Theorem 3.1, is far from straightforward. In fact, this case is more demanding than the selection model with two-way flows in Section 4.1 since we only have two treatment values. We observe the propensity score

$$\Pr(D = 1|\boldsymbol{Z}) = F_{V_1,V_2}\left(Q_1(\boldsymbol{Z}), Q_2(\boldsymbol{Z})\right), \tag{4.3}$$

which we denote $H(\boldsymbol{Z})$. This is a nonparametric double index model in which both the link function F_{V_1,V_2} and the indices Q_1 and Q_2 are unknown; it is clearly underidentified without stronger restrictions. Matzkin (1993, 2007) considers nonparametric identification and estimation of polychotomous choice models. Our multiple hurdle model has a similar but not identical structure.

[9]Since probabilities add up to one, only one of these equalities generates a specification test.

In order to identify $\boldsymbol{Q}$, we assume that there exist two instruments that are excluded from one of the thresholds. More precisely, let the vector of instruments be $\boldsymbol{Z} = (Z_1, Z_2, \boldsymbol{Z}_{-12})$, with Z_1 and Z_2 scalar; we require that

- $Q_1(\boldsymbol{Z})$ does not depend on Z_2, and
- $Q_2(\boldsymbol{Z})$ does not depend on Z_1.

To simplify notation, we fix the value of $\boldsymbol{Z}_{-12}$ and we denote $Q_1(\boldsymbol{Z}) = G_1(Z_1)$ and $Q_2(\boldsymbol{Z}) = G_2(Z_2)$, where G_1 and G_2 are two unknown functions. Note that the propensity score becomes $H(Z_1, Z_2) = F_{V_1,V_2}(G_1(Z_1), G_2(Z_2))$.

We give two identification results under these exclusion restrictions. We first build on Lewbel (2000) and on Matzkin (1993, 2007)'s results to identify $\boldsymbol{Q}$ and rely on full support restrictions (conditional on the value of $\boldsymbol{Z}_{-12}$):

Assumption 4.2. *$Q_1 = G_1(Z_1)$ and $Q_2 = G_2(Z_2)$. Moreover,*

1. *The density of (V_1, V_2) is continuous on $[0,1]^2$, with marginal uniform distributions.*
2. *G_1 and G_2 are strictly increasing C^1 functions from possibly unbounded intervals (a_1, b_1) and (a_2, b_2) to $(0,1)$; that is, for every $t \in (0,1)$ there exist $z_1 \in (a_1, b_1)$ and $z_2 \in (a_2, b_2)$ such that $G_1(z_1) = G_2(z_2) = t$.*
3. *$\mathcal{Z}$ is the rectangle $(a_1, b_1) \times (a_2, b_2)$.*

Theorem 4.2. *Under Assumption 4.2, the functions $F_{\boldsymbol{V}}, G_1$ and G_2 are identified from the propensity score* $\Pr(D = 1|\boldsymbol{Z})$.

Proof. See Appendix B.3. □

While Theorem 4.2 requires two continuous instruments that generate all possible values of the thresholds, various additional restrictions would relax this requirement. If for instance G_1 and G_2 were linear, we would be back to the linear multiple index model of Ichimura and Lee (1991).

Theorem 4.3 provides a complementary result and is useful when the instruments have limited support. It relies on a semiparametric restriction. Remember that we normalized the marginal distributions of V_1 and V_2 to be uniform over $[0,1]$; we now

assume that the codependence between V_1 and V_2 is described by a strict symmetric Archimedean copula[10]:

$$F_{V_1,V_2}(v_1, v_2) = \phi^{-1}\left(\phi(v_1) + \phi(v_2)\right), \tag{4.4}$$

where ϕ belongs to the set Ψ of C^2, strictly decreasing, and convex functions from $[0,1]$ unto $[0,+\infty]$.

Assumption 4.3. *$Q_1 = G_1(Z_1)$ and $Q_2 = G_2(Z_2)$. Moreover,*

(a) the propensity score and the distribution of (V_1, V_2) are described by (4.3) and (4.4) for unknown functions ϕ, G_1, and G_2,

(b) the interior of the support $\mathcal{Z}$ of (Z_1, Z_2) contains a connected set $\mathcal{N}$, and

(c) G_1 and G_2 are C^1 functions over the projections of $\mathcal{N}$, with derivatives bounded away from zero.

Let $H_k(z_1, z_2)$ denote the derivative of the propensity score $H(z_1, z_2)$ with respect to its kth argument ($k = 1, 2$) and $H_{12}(z_1, z_2)$ the second-order cross derivative of $H(z_1, z_2)$. Note that the scale of ϕ is not identifiable in view of (4.4). Furthermore, if ϕ is specified nonparametrically as an element of Ψ, the location of ϕ is only identifiable when the argument of ϕ takes values close to 1:

Theorem 4.3. *Let Assumption 4.3 hold. Then*

1. *Over $\mathcal{N}$, the ratio*
$$\frac{H_{12}}{H_1 H_2}(z_1, z_2)$$
is non-negative and only depends on the value $h = H(z_1, z_2)$.

2. *The function ϕ is identified up to scale and location in Ψ on the image $H(\mathcal{N}) = (\underline{h}, \bar{h}) \subset (0,1)$ of $\mathcal{N}$ under H, where $\mathcal{N}$ is given in Assumption 4.3(b).*

3. *The scale parameter for ϕ is an arbitrary negative number, which we normalize by imposing $\phi'(\bar{h}) = -1$; given this normalization, the location parameter for ϕ is bounded by*
$$0 \leq \phi(\bar{h}) \leq 1 - \bar{h}.$$

[10]The class of Archimedean copulas include the Clayton, Frank, and Gumbel families among others (see Nelsen, 2006, ch. 4).

If moreover $\sup_{\boldsymbol{z}\in\mathcal{N}} \Pr(D=1|\boldsymbol{Z}=\boldsymbol{z})=1$*, then* $\bar{h}=1$ *and* ϕ *is point-identified.*

4. *For any admissible value of the location parameter of* ϕ*, the functions* G_1 *and* G_2 *are identified on* $\mathcal{N}$ *up to a common constant* k*: any other admissible* $(\tilde{G}_1, \tilde{G}_2)$ *must satisfy*

$$\phi(\tilde{G}_1(z_1)) = \phi(G_1(z_1)) - k$$
$$\phi(\tilde{G}_2(z_2)) = \phi(G_2(z_2)) + k$$

over the projections of $\mathcal{N}$*. The number* k *is bounded above and below. If moreover* $\sup_{\boldsymbol{z}\in\mathcal{N}} \Pr(D=1|\boldsymbol{Z}=\boldsymbol{z})=1$*, then* G_1 *and* G_2 *are point-identified on the projections of* $\mathcal{N}$*.*

Proof. See Appendix B.4. □

Our constructive identification starts by writing

$$\frac{\phi''}{\phi'}(h) = -\frac{H_{12}}{H_1 H_2}(z_1, z_2) \tag{4.5}$$

for all (z_1, z_2) such that $H(z_1, z_2) = h$. Once ϕ is identified from (4.5), then we proceed to identify G_1 and G_2. The function G_1, for instance, would be identified by

$$\phi(G_1(z_1)) = \phi(H(z_1, z_2^0)) - \phi(g_2^0)$$

for a fixed z_2^0 and a value g_2^0 of $G_2(z_2^0)$.

Given more a priori restrictions on the function ϕ, identification results can be sharper. The following example illustrates this point by taking a parametric family of ϕ.

Example 4. Take the strict Clayton copula, which is generated by $\phi(u) = (u^{-\theta}-1)/\theta$ for $\theta > 0$. This yields

$$H(z_1, z_2) = \left(G_1(z_1)^{-\theta} + G_2(z_2)^{-\theta} - 1\right)^{-1/\theta}.$$

In this example, $\frac{\phi''}{\phi'}(h)$ is simply $-(1+\theta)/h$. Therefore, it follows from (4.5) that θ

can be identified in closed form as

$$\theta = h\frac{H_{12}}{H_1 H_2}(z_1, z_2) - 1 \tag{4.6}$$

for all (h, z_1, z_2) such that $H(z_1, z_2) = h$. Note that the scale and location of ϕ are point-identified, given the parametric restriction.

Conversely, the constancy of the right-hand side of (4.6) characterizes a Clayton copula. To identify G_1 and G_2, note that

$$G_1(z_1)^{-\theta} + G_2(z_2)^{-\theta} = H(z_1, z_2)^{-\theta} + 1.$$

Thus it is easy to see that G_1 and G_2 are identified up to a location constant[11]. □

Once $Q_1(\boldsymbol{Z})$ and $Q_2(\boldsymbol{Z})$ are identified, then under our assumptions we identify the joint density by

$$f_{V_1,V_2}(q_1, q_2) = \frac{\partial^2 \Pr[D = 1 | Q_1(\boldsymbol{Z}) = q_1, Q_2(\boldsymbol{Z}) = q_2]}{\partial q_1 \partial q_2}, \tag{4.7}$$

and the marginal treatment effect is given by

$$E(Y_1 - Y_0 | V_1 = q_1, V_2 = q_2) f_{V_1,V_2}(q_1, q_2) = \frac{\partial^2 E[Y | Q_1(\boldsymbol{Z}) = q_1, Q_2(\boldsymbol{Z}) = q_2]}{\partial q_1 \partial q_2}. \tag{4.8}$$

Furthermore, it follows from Corollary 3.2 that under the additional Assumption 3.5, the ATE, ATT and PRTE parameters are identified as well.

4.3 Dynamic Treatment

To conclude our examples, let us consider a two-period model of dynamic treatment where treatment assignment D^2 in the second period depends on the first-period treatment D^1 and outcome Y^1:

$$D^1 = \mathbb{1}(V_1 < Q_1(\boldsymbol{Z}^1)) \text{ and } D^2 = \mathbb{1}\left(V_2 < Q_2(\boldsymbol{Z}^2, D^1, Y^1)\right).$$

The analyst observes $(D^1, D^2, Y^1, Y^2, \boldsymbol{Z}^1, \boldsymbol{Z}^2)$. Theorem 3.1 applies to this model, provided only that the functions Q_1 and Q_2 are identified. The identification of Q_1

[11] This location constant plays the role of k in part 4 of Theorem 4.3.

is straightforward. To identify Q_2, we use the results of Shaikh and Vytlacil (2011), which considers a model similar to our second-period treatment assignment. While they stress partial identification, their Remark 2.2 (p. 954) gives a sufficient condition for point identification. Translated in our notation, this requires that

1. the support of $(\boldsymbol{Z}^2, Q_1(\boldsymbol{Z}^1))$ is the product of the support of $\boldsymbol{Z}^2$ and the support of $Q_1(\boldsymbol{Z}^1)$, and that

2. for every value $(\boldsymbol{z}^2, y^1)$ of $(\boldsymbol{Z}^2, Y^1)$, there is a value $\bar{\boldsymbol{z}}^2$ such that $Q_2(\bar{\boldsymbol{z}}^2, 1, y^1) = Q_2(\boldsymbol{z}^2, 0, y^1)$; and there is a value $\underline{\boldsymbol{z}}^2$ such that $Q_2(\underline{\boldsymbol{z}}^2, 0, y^1) = Q_2(\boldsymbol{z}^2, 1, y^1)$.

Assumption 1 above requires that the set of instruments in the second period has a component that does not affect treatment in the first period, and whose range of variation does not depend on the propensity score of the first period. Assumption 2 adds the requirement that the ranges of the second-period propensity scores are independent of the first-period treatment, for all values of the first-period outcome.

These assumptions require overlap between treatment branches. They would not hold, for instance, in a medical trial when patients are oriented towards completely different treatments depending on how they fare early on.

5 Relation to the Existing Literature

The existing literature is very large; we only discuss here the most directly relevant papers.

5.1 Ordered Treatments with Discrete Instruments

Angrist and Imbens (1995) consider two-stage least-squares (TSLS) estimation of a model in which the ordered treatment takes a finite number of values, and a discrete-valued instrument is available. They show that the TSLS estimator obtained by regressing outcome Y on a preestimated $E(D|Z)$ converges to a weighted sum of *average causal responses* under some monotonicity assumption. Heckman, Urzua, and Vytlacil (2006, 2008) go beyond Angrist and Imbens (1995) by showing how the TSLS estimate can be reinterpreted in more transparent ways in the MTE framework. They also analyze a family of discrete choice models, to which we now turn.

5.2 Discrete Choice Models

Heckman, Urzua, and Vytlacil (2008, see also Heckman and Vytlacil (2007)) consider a multinomial discrete choice model of treatment. They posit

$$(5.1)\quad D = k \iff R_k(\boldsymbol{Z}) - U_k > R_l(\boldsymbol{Z}) - U_l \text{ for } l = 0, \ldots, K-1 \text{ such that } l \neq k,$$

where the U's are continuously distributed and independent of $\boldsymbol{Z}$. Then they study the identification of marginal and local average treatment effects under assumptions that are similar to ours: continuous instruments that generate enough dimensions of variation in the thresholds.

As they note, the discrete choice model with an additive structure implicitly imposes monotonicity, in the following form: if the instruments $\boldsymbol{Z}$ change in a way that increases $R_k(\boldsymbol{Z})$ relative to all other $R_l(\boldsymbol{Z})$, then no observation with treatment value k is assigned to a different treatment. We make no such assumption, as Example 1 and Figure 1 illustrate. Our results extend those of Heckman, Urzua, and Vytlacil (2008) to any model with identified thresholds.[12] We consider a discrete choice model with three alternatives as an example.

Example 5 (Discrete Choice Model with Three Alternatives)**.** Suppose that $\mathcal{K} = \{0, 1, 2\}$ with $K = 3$. Let $\tilde{R}_{0,1}(\boldsymbol{Z}) = R_0(\boldsymbol{Z}) - R_1(\boldsymbol{Z})$, $\tilde{R}_{0,2}(\boldsymbol{Z}) = R_0(\boldsymbol{Z}) - R_2(\boldsymbol{Z})$ and $\tilde{R}_{1,2}(\boldsymbol{Z}) = R_1(\boldsymbol{Z}) - R_2(\boldsymbol{Z})$. Similarly, let $\tilde{U}_{0,1} = U_0 - U_1$, $\tilde{U}_{0,2} = U_0 - U_2$ and $\tilde{U}_{1,2} = U_1 - U_2$. Let $V_{0,1} = F_{\tilde{U}_{0,1}}(\tilde{U}_{0,1})$ and $Q_{0,1}(\boldsymbol{Z}) = F_{\tilde{U}_{0,1}}(\tilde{R}_{0,1}(\boldsymbol{Z}))$. Define $V_{0,2}$, $V_{1,2}$, $Q_{0,2}(\boldsymbol{Z})$ and $Q_{1,2}(\boldsymbol{Z})$ similarly. Then the selection mechanism in (5.1) can be rewritten as

- $D = 0$ iff $V_{0,1} < Q_{0,1}(\boldsymbol{Z})$ and $V_{0,2} < Q_{0,2}(\boldsymbol{Z})$
- $D = 1$ iff $V_{0,1} > Q_{0,1}(\boldsymbol{Z})$ and $V_{1,2} < Q_{1,2}(\boldsymbol{Z})$
- $D = 2$ iff $V_{0,2} > Q_{0,2}(\boldsymbol{Z})$ and $V_{1,2} > Q_{1,2}(\boldsymbol{Z})$.

Our general result in Section 3 applies immediately once the $Q_{j,k}$'s are identified. This can be done, for example, by applying the results of Matzkin (1993, 2007). □

There is a growing empirical literature on multivalued unordered treatments. Dahl (2002) develops a semiparametric Roy model for migration across U.S. states. In his

[12] Appendix D compares our results with those of Heckman, Urzua, and Vytlacil (2008) in more detail.

empirical work, the number of unordered treatment is 51 (50 states plus the District of Columbia) and he controls for selection bias by conditioning on migration probabilities. Kirkeboen, Leuven, and Mogstad (2016) use discrete instruments to obtain TSLS estimates of returns to different fields of study in postsecondary education in Norway. In their setup, the unordered treatments are different fields of study. Kline and Walters (2016) use data from the Head Start Impact Study to estimate a semiparametric selection model. Their model has three treatment cells: Head Start, competing preschool programs, and no preschool (that is, home care).

Broadly speaking, these papers are in the same vein as Roy models and discrete choice models. Our approach complements this literature by focusing on the role of unobserved heterogeneity and the selection mechanism.

5.3 Unordered Monotonicity

In an important recent paper, Heckman and Pinto (2018) introduce a new concept of monotonicity. Their "unordered monotonicity" assumption can be rephrased in our notation in the following way. Take two values $\boldsymbol{z}$ and $\boldsymbol{z}'$ of the instruments $\boldsymbol{Z}$ and any treatment value k.

Assumption 5.1 (Unordered Monotonicity)**.** *Denote $d_k(\boldsymbol{v}, \boldsymbol{z})$ and $d_k(\boldsymbol{v}, \boldsymbol{z'})$ the counterfactual values of the variable $d_k = \mathbf{1}(D = k)$ for an observation with unobserved heterogeneity $\boldsymbol{v}$. Then*

$$d_k(\boldsymbol{v}, \boldsymbol{z}) \geq d_k(\boldsymbol{v}, \boldsymbol{z'}) \; \forall \boldsymbol{v};$$
$$\textit{or: } d_k(\boldsymbol{v}, \boldsymbol{z}) \leq d_k(\boldsymbol{v}, \boldsymbol{z'}) \; \forall \boldsymbol{v}.$$

Unordered monotonicity for treatment value k requires that if some observations move out of (resp. into) treatment value k when instruments change value from $\boldsymbol{z}$ to $\boldsymbol{z}'$, then no observation can move into (resp. out of) treatment value k. For binary treatments, unordered monotonicity is equivalent to the usual monotonicity assumption: there cannot be both compliers and defiers. When $K > 2$, it is weaker than ordered choice. For example, suppose that there are three options $\{0, 1, 2\}$ and that a change of instruments makes option 1 less appealing. Under ordered choice, all agents who give up option 1 must fall back on option 0, or all must fall back on option 2. Unordered monotonicity allows different agents to fall back on different

options. It still rules out two-way flows, that is agents moving from option 0 or 2 into option 1.

Heckman and Pinto (2018) show that unordered monotonicity (for well-chosen changes in instruments) is essentially equivalent to a treatment model based on rules that are additively separable in the unobserved variables—that is, the model of section 5.2. In this interpretation, changes in instruments that increase the mean utility of an alternative relative to all others are unordered monotonic for that alternative, for instance. We refer the reader to Section 6 of Heckman and Pinto (2018) for a more rigorous discussion, and to Pinto (2015) for an application to the Moving to Opportunity program.

Unlike us, Heckman and Pinto (2018) do not require continuous instruments; all of their analysis is framed in terms of discrete-valued instruments and treatments. Beyond this (important) difference, unordered monotonicity clearly obeys our assumptions. On the other hand, we allow for much more general models of treatment. It would be impossible, for instance, to rewrite our Examples 1, 2 and 3 so that they obey unordered monotonicity. We illustrate this point using Example 1 below.

Example 1 (continued)**.** In Example 1, $D = 2$ iff $(V_1 - Q_1(\boldsymbol{Z}))$ and $(V_2 - Q_2(\boldsymbol{Z}))$ have opposite signs. Note that there are two unobserved categories within $D = 2$:

$$\begin{aligned} D = 2a \text{ iff } V_1 < Q_1 \text{ and } V_2 > Q_2, \\ D = 2b \text{ iff } V_1 > Q_1 \text{ and } V_2 < Q_2. \end{aligned}$$

Each one is unordered monotonic; but because we only observe their union, $D = 2$ is not unordered monotonic—increasing Q_1 brings more people into $2a$ but moves some out of $2b$, so that in the end we have two-way flows, contradicting unordered monotonicity. To put it differently, the selection mechanism in Example 1 becomes a discrete choice model when each of four alternatives $d = 0, 1, 2a, 2b$ is observed; however, we only observe whether alternative $d = 0$, $d = 1$ or $d = 2$ is chosen in Example 1. This amounts to an unordered monotonic treatment that is observed through a coarser information partition; this coarsening destroys unordered monotonicity. □

Appendix A.2 provides a characterization of Heckman and Pinto's unordered monotonicity property as a subcase of our more general framework.

5.4 Other Nonmonotonic Models

It is also worth commenting on other papers that break monotonicity. Gautier and Hoderlein (2015) consider a triangular random coefficients model for the binary treatment case. Their model is motivated by a single agent Roy model with random coefficients. Its selection mechanism is governed by

$$D = 1\{V_1 - Z_1 - g(Z_1, \ldots, Z_L) - \sum_{j=2}^{J} V_j f_j(Z_j) > 0\},$$

where $\boldsymbol{V} = (V_1, \ldots, V_J)$ is a vector of unobserved random variables, $\boldsymbol{Z} = (Z_1, \ldots, Z_J)$ is a vector of instruments that are independent of $(Y_0, Y_1, \boldsymbol{V})$, and the functions $f_2, \ldots, f_J$ and g are unknown. If we limit our attention to the case of two unobservables as in the double hurdle model, then the selection equation in Gautier and Hoderlein (2015) reduces to

$$D = 1\{V_1 - Z_1 - g(Z_1, Z_2) - V_2 f_2(Z_2) > 0\}.$$

Here changes in Z_1 conform to monotonicity; but changes in Z_2 need not.

Lewbel and Yang (2016) consider a different non-monotonic selection mechanism for estimating the average treatment effect. They show that the average treatment effect is identified when a binary treatment is assigned by

$$D = \mathbb{1}\left(\alpha_0 \leq Z + V \leq \alpha_1\right),$$

where V is an unobserved random variable; Z is a continuous variable that satisfies $E(Y_j|V, Z) = E(Y_j|V)$ for $j = 0, 1$ and $V \perp\!\!\!\perp Z$; and α_0, α_1 are unknown parameters.

5.5 Models with Continuous Treatment

Chesher (2003) develops conditions to identify derivatives of structural functions in nonseparable models by functionals of quantile regression functions. In addition, Florens, Heckman, Meghir, and Vytlacil (2008) consider a potential outcome model with a continuous treatment. They assume a stochastic polynomial restriction and show that the average treatment effect can be identified if a suitable control function can be constructed using instruments.

Imbens and Newey (2009) also consider selection on unobservables with a continuous treatment. They assume that the treatment (more generally in their paper, an endogenous variable) is given by $D = g(Z, V)$, with g increasing in a scalar unobserved V. They identify the average structural function as well as quantile, average, and policy effects. Other more recent identification results along this line can be found in Torgovitsky (2015) and D'Haultfœuille and Février (2015) among others. One key restriction in this group of papers is the monotonicity in the scalar V in the selection equation. We do not rely on this type of restriction, but we only focus on the case of multivalued treatments. Hence, our approach and those of the papers cited in this subsection are complementary.

Finally, our approach shares some features with Hoderlein and Mammen (2007). They consider the identification of marginal effects in nonseparable models without monotonicity. They show how local average structural derivatives can be identified. Like ours, their approach relies on differentiation of observed functionals. The parameters of interest they study are quite different, however, and their selection mechanism is not as explicit as ours.

6 Proof of Theorem 3.1

Our proof has three steps. We first write conditional moments as integrals with respect to indicator functions. Then we show that these integrals are differentiable and we compute their multidimensional derivatives. Finally, we impose Assumption 3.1 and we derive the equalities in the theorem.

Step 1:

Under the assumptions imposed in the theorem, for any $\boldsymbol{q}$ in the range of $\boldsymbol{Q}$,

$$\begin{aligned}
&E[G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}] \\
&= E[G(Y_k)|D = k, \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}] \Pr(D = k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}) \\
&= E[G(Y_k)|d_k(\boldsymbol{V}, \boldsymbol{Q}(\boldsymbol{Z})) = 1, \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}] \Pr(d_k(\boldsymbol{V}, \boldsymbol{Q}(\boldsymbol{Z})) = 1|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}) \\
&= E[G(Y_k)|d_k(\boldsymbol{V}, \boldsymbol{q}) = 1] \Pr\left(d_k(\boldsymbol{V}, \boldsymbol{q}) = 1\right) \\
&= E[G(Y_k)\mathbb{1}\left(d_k(\boldsymbol{V}, \boldsymbol{q}) = 1\right)] \\
&= E\left(E[G(Y_k)\mathbb{1}\left(d_k(\boldsymbol{V}, \boldsymbol{q}) = 1\right)|\boldsymbol{V}]\right) \\
&= E\left(E[G(Y_k)|\boldsymbol{V}]\mathbb{1}\left(d_k(\boldsymbol{V}, \boldsymbol{q}) = 1\right)\right),
\end{aligned}$$

where the third equality follows from Assumption 2.2 and the others are obvious. As a consequence,

$$\begin{aligned}
&E[G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}] \\
&= \int \mathbb{1}\left(d_k(\boldsymbol{v}, \boldsymbol{q}) = 1\right) E[G(Y_k)|\boldsymbol{V} = \boldsymbol{v}] f_{\boldsymbol{V}}(\boldsymbol{v}) d\boldsymbol{v}. \qquad (6.1)
\end{aligned}$$

Let $b_k(\boldsymbol{v}) \equiv E[G(Y_k)|\boldsymbol{V} = \boldsymbol{v}] f_{\boldsymbol{V}}(\boldsymbol{v})$ and $B_k(\boldsymbol{q}) = E[G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}]$. Then (6.1) takes the form

$$B_k(\boldsymbol{q}) = \int \mathbb{1}(d_k(\boldsymbol{v}, \boldsymbol{q}) = 1) b_k(\boldsymbol{v}) d\boldsymbol{v}.$$

Now recall from Lemma 2.1 that the indicator function of $D = k$ is a multivariate polynomial of the indicator functions S_j for $j \in \boldsymbol{J}$. Moreover,

$$S_j(\boldsymbol{V}, \boldsymbol{Q}(\boldsymbol{Z})) = \mathbb{1}(V_j < Q_j(\boldsymbol{Z})) = H(Q_j(\boldsymbol{Z}) - V_j),$$

where $H(t) = \mathbb{1}(t > 0)$ is the one-dimensional Heaviside function. Therefore we can rewrite the selection of treatment k as

$$\mathbb{1}(d_k(\boldsymbol{v}, \boldsymbol{q}) = 1) = \sum_{l \in \mathcal{L}} c_l^k \prod_{j \in l} H(q_j - v_j) \qquad (6.2)$$

and it follows that

$$B_k(\boldsymbol{q}) = \sum_{l \in \mathcal{L}} c_l^k \int \left(\prod_{j \in l} H(q_j - v_j) \right) b_k(\boldsymbol{v}) d\boldsymbol{v}. \qquad (6.3)$$

Step 2:

By Assumption 3.3, the function $\boldsymbol{b}$ is locally equicontinuous; and by Assumption 3.4, it is defined over an open neighborhood of $\boldsymbol{q}$. This implies that all terms in (6.3) are differentiable along all dimensions of $\boldsymbol{q}$. To see this, start with dimension $j = 1$. Any term l in (6.3) that does not contain 1 is constant in q_1 and obviously differentiable. Take any other term and rewrite it as

$$A_l(q_1) \equiv c_l^k \int_0^{q_1} \int \left(\prod_{j \in l} H(q_j - v_j) \right) b_k(v_1, \boldsymbol{v}_{-1}) d\boldsymbol{v}_{-1} dv_1,$$

where $\boldsymbol{v}_{-1}$ collects all directions of $\boldsymbol{v}$ in $l - \{1\}$.

Then for any $\varepsilon \neq 0$,

$$\begin{aligned} \frac{A_l(q_1 + \varepsilon) - A_l(q_1)}{\varepsilon} - c_l^k \int \left(\prod_{j \in l - \{1\}} H(q_j - v_j) \right) b_k(q_1, \boldsymbol{v}_{-1}) d\boldsymbol{v}_{-1} \\ = \frac{c_l^k}{\varepsilon} \int_{q_1}^{q_1 + \varepsilon} \int \left(\prod_{j \in l - \{1\}} H(q_j - v_j) \right) \left(b_k(v_1, \boldsymbol{v}_{-1}) - b_k(q_1, \boldsymbol{v}_{-1}) \right) d\boldsymbol{v}_{-1} dv_1. \end{aligned}$$

Since the functions $(b_k(\cdot, \boldsymbol{v}_{-1}))$ are locally equicontinuous at q_1, for any $\eta > 0$ we can choose ε such that if $|q_1 - v_1| < \varepsilon$,

$$|b_k(q_1, \boldsymbol{v}_{-1}) - b_k(v_1, \boldsymbol{v}_{-1})| < \eta;$$

and since the Heaviside functions are bounded above by one, we have

$$\left| \frac{A_l(q_1 + \varepsilon) - A_l(q_1)}{\varepsilon} - c_l^k \int \left(\prod_{j \in l - \{1\}} H(q_j - v_j) \right) b_k(q_1, \boldsymbol{v}_{-1}) d\boldsymbol{v}_{-1} \right| < |c_l| \, \eta.$$

This proves that A_l is differentiable in q_1 and that its derivative with respect to q_1, which we denote A_l^1, is

$$A_l^1 = c_l \int \prod_{j \in l - \{1\}} H(q_j - v_j) \; b_k(q_1, \boldsymbol{v}_{-1}) d\boldsymbol{v}_{-1}.$$

But this derivative itself has the same form as A_l. Letting $\boldsymbol{v}_{-1,2}$ collect all com-

ponents of $\boldsymbol{v}$ except (q_1, q_2), the same argument would prove that since the functions $(b_k(\cdot, \boldsymbol{v}_{-1,2}))$ are locally equicontinuous at (q_1, q_2), the function A_l^1 is differentiable with respect to q_2 and its derivative is

$$c_l^k \int \left(\prod_{j \in l - \{1,2\}} H(q_j - v_j) \right) b_k(q_1, q_2, \boldsymbol{v}_{-1,2}) d\boldsymbol{v}_{-1,2}.$$

Continuing this argument finally gives us the cross-derivative with respect to $(\boldsymbol{q}^l)$ as

$$c_l^k \int b_k(\boldsymbol{q}^l, \boldsymbol{v}_{-l}) d\boldsymbol{v}_{-l},$$

where $\boldsymbol{v}_{-l}$ collects all components of $\boldsymbol{v}$ whose indices are not in l.

Step 3:

Lemma 2.1 and Assumption 3.1 also imply that the leading term in the sum $\sum_l c_l^l \prod_{j \in l} H(q_j - v_j)$ is

$$c_{\boldsymbol{J}}^k \prod_{j=1}^{J} H(q_j - v_j).$$

Now take the J-order derivative of $B(\boldsymbol{q})$ with respect to all q_j in turn. By Lemma 2.1, the highest-degree term of B in $\boldsymbol{q}$ is

$$c_{\boldsymbol{J}}^k \int \left(\prod_{j=1}^{J} H(q_j - v_j) \right) b_k(\boldsymbol{v}) d\boldsymbol{v}$$

as $c_{\boldsymbol{J}}^k \neq 0$ under Assumption 3.1; all other terms have a smaller number of indices j.

This term contributes a cross-derivative

$$c_{\boldsymbol{J}}^k b_k(\boldsymbol{q}),$$

and all other terms generate zero-value contributions since each of them is constant in at least one of the directions j.

More formally,

$$TB_k(\boldsymbol{q}) = \frac{\partial^J B_k(\boldsymbol{q})}{\prod_{j \in \boldsymbol{J}} \partial q_j} = c_{\boldsymbol{J}}^k b_k(\boldsymbol{q}). \tag{6.4}$$

Given Assumptions 3.3 and 3.4, we can apply (6.4) successively to the pair of functions

$$B_k(\boldsymbol{q}) = E[G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}] \text{ and } b_k(\boldsymbol{v}) = E[G(Y_k)|\boldsymbol{V} = \boldsymbol{v}]f_{\boldsymbol{V}}(\boldsymbol{v}),$$

as in (6.3), and to the pair of functions

$$B_k(\boldsymbol{q}) = \Pr[D = k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}] \text{ with } b_k(\boldsymbol{v}) = f_{\boldsymbol{V}}(\boldsymbol{v}).$$

The first pair gives us the second equality in the Theorem, and the second pair gives us the first equality. □

References

Angrist, J. D., and G. W. Imbens (1995): "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.

Brinch, C. N., M. Mogstad, and M. Wiswall (2017): "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, 125(4), 985–1039.

Cattaneo, M. D. (2010): "Efficient semiparametric estimation of multi-valued treatment effects under ignorability," *Journal of Econometrics*, 155(2), 138–154.

Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71(5), 1405–1441.

Dahl, G. B. (2002): "Mobility and the return to education: Testing a Roy model with multiple markets," *Econometrica*, 70(6), 2367–2420.

de Chaisemartin, C. (2017): "Tolerating Defiance? Local Average Treatment Effects without Monotonicity," *Quantitative Economics*, 8(2), 367–396.

D'Haultfœuille, X., and P. Février (2015): "Identification of Nonseparable Triangular Models With Discrete Instruments," *Econometrica*, 83(3), 1199–1210.

Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76(5), 1191–1206.

Gautier, E., and S. Hoderlein (2015): "A Triangular Treatment Effect Model With Random Coefficients in the Selection Equation," http://arxiv.org/abs/1109.0362v4.

Heckman, J., and R. Pinto (2018): "Unordered Monotonicity," *Econometrica*, 86(1), 1–35.

Heckman, J. J. (1979): "Sample selection bias as a specification error," *Econometrica*, 47(1), 153–161.

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

——— (2008): "Instrumental variables in models with multiple outcomes: The general unordered case," *Annales d'économie et de statistique*, 91/92, 151–174.

Heckman, J. J., and E. Vytlacil (2000): "Local Instrumental Variables," Technical Working Paper No. 252, National Bureau of Economic Research.

——— (2001): "Policy-relevant treatment effects," *American Economic Review*, 91(2), 107–111.

——— (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 econometrics*, ed. by J. J. Heckman, and E. Leamer, vol. 6B, chap. 70, pp. 4779–4874. Elsevier, Amsterdam.

Hirano, K., and G. W. Imbens (2004): "The propensity score with continuous treatments," in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, ed. by A. Gelman, and X. Meng, pp. 73–84. Wiley-Blackwell.

Hoderlein, S., and E. Mammen (2007): "Identification of Marginal Effects in Nonseparable Models Without Monotonicity," *Econometrica*, 75(5), 1513–1518.

Ichimura, H., and L.-f. Lee (1991): "Semiparametric least squares estimation of multiple index models: single equation estimation," in *International Symposia in Economic Theory and Econometrics*, ed. by W. A. Barnett, J. Powell, and G. Tauchen, pp. 3–49. Cambridge University Press.

Imbens, G. W. (2000): "The role of the propensity score in estimating dose-response functions," *Biometrika*, 87(3), 706–710.

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

Imbens, G. W., and W. K. Newey (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77(5), 1481–1512.

Kirkeboen, L. J., E. Leuven, and M. Mogstad (2016): "Field of study, earnings, and self-selection," *Quarterly Journal of Economics*, 131(3), 1057–1111.

Kline, P., and C. R. Walters (2016): "Evaluating public programs with close substitutes: The case of Head Start," *Quarterly Journal of Economics*, 131(4), 1795–1848.

Kowalski, A. E. (2016): "Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments," Working Paper no. 22363, National Bureau of Economic Research.

Lewbel, A. (2000): "Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables," *Journal of Econometrics*, 97(1), 145–177.

Lewbel, A., and T. T. Yang (2016): "Identifying the average treatment effect in ordered treatment models without unconfoundedness," *Journal of Econometrics*, 195(1), 1–22.

Manski, C. F. (1990): "Nonparametric bounds on treatment effects," *American Economic Review*, 80, 319–323.

——— (1997): "Monotone treatment response," *Econometrica*, 65, 1311–1334.

Manski, C. F., and J. V. Pepper (2000): "Monotone instrumental variables: with an application to the returns to schooling," *Econometrica*, 68(4), 997–1010.

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

——— (2007): "Heterogeneous choice," in *Advances in economics and econometrics: theory and applications*, ed. by R. Blundell, W. Newey, and T. Persson, vol. 2, chap. 4, pp. 75–110. Cambridge University Press.

Mogstad, M., A. Santos, and A. Torgovitsky (2017): "Using Instrumental Variables for Inference about Policy Relevant Treatment Effects," NBER Working Paper No. 23568.

Nelsen, R. B. (2006): *An Introduction to Copulas*. Springer, New York, NY, second edn.

Pinto, R. (2015): "Selection bias in a controlled experiment: the case of Moving to Opportunity," University of Chicago, mimeo.

Poirier, D. J. (1980): "Partial observability in bivariate probit models," *Journal of Econometrics*, 12(2), 209–217.

Shaikh, A., and E. Vytlacil (2011): "Partial Identification in Triangular Systems of Equations With Binary Dependent Variables," *Econometrica*, 79, 949–955.

Tamer, E. (2003): "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria," *Review of Economic Studies*, 70(1), 147–165.

Torgovitsky, A. (2015): "Identification of Nonseparable Models Using Instruments With Small Support," *Econometrica*, 83(3), 1185–1197.

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

Yang, S., G. W. Imbens, Z. Cui, D. E. Faries, and Z. Kadziola (2016): "Propensity score matching and subclassification in observational studies with multi-level treatments," *Biometrics*, 72(4), 1055–1065.

Online Appendices to "Identifying Effects of Multivalued Treatments"

Sokbae Lee* **Bernard Salanié†**

May 30, 2018

Appendix A gives an identification result for the zero-index case, which was not dealt with in the text. It also provides a characterization of Heckman and Pinto's unordered monotonicity property as a subcase of our more general framework. Appendix B collects proofs of some of the results in the main text. Finally, Appendix C fills in the details of the entry game introduced in Section 2, and Appendix D compares our results with those of Heckman, Urzua, and Vytlacil (2008) in more detail. Appendix E discusses a more general form of threshold conditions than the "rectangular" threshold conditions in Assumption 2.1.

A Additional Results

A.1 Identification with a Zero Index

Theorem 3.1 required that the index of treatment k be non-zero (Assumption 3.1). It therefore does not apply to, for instance, Example 3. Recall that in that example,

$$D_0 = \mathcal{D}_0(\boldsymbol{S}) = 1 - S_1 - S_2 - S_3 + S_1S_2 + S_1S_3 + S_2S_3$$

and treatment 0 has degree $m^0 = 2 < J^0 = 3$.

Note, however, that steps 1 and 2 of the proof of Theorem 3.1 apply to zero-index treatments as well; the relevant polynomial of Heaviside functions has leading term

$$H(q_1 - v_1)H(q_2 - v_2) + H(q_1 - v_1)H(q_3 - v_3) + H(q_2 - v_2)H(q_3 - v_3),$$

*Columbia University and Institute for Fiscal Studies, sl3841@columbia.edu.
†Columbia University, bsalanie@columbia.edu.

and we can take the derivative in (q_1, q_2) for instance to obtain an equation that replaces (6.4):

$$\frac{\partial^2}{\partial q_1 \partial q_2} B_0(\boldsymbol{q}) = \int b_0(q_1, q_2, v_3) dv_3.$$

Applying this to $B_0(\boldsymbol{q}) = \Pr[D = 0 | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}]$ and $b_0(\boldsymbol{v}) = f_{\boldsymbol{V}}(\boldsymbol{v})$, and then to $B_0(\boldsymbol{q}) = E[Y D_0 | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}]$ and $b_0(\boldsymbol{v}) = E[G(Y_0) | \boldsymbol{V} = \boldsymbol{v}] f_{\boldsymbol{V}}(\boldsymbol{v})$, identifies

$$\int f_{V_1, V_2, V_3}(q_1, q_2, v_3) dv_3 = f_{V_1, V_2}(v_1, v_2)$$

and

$$\int E[G(Y_0) | V_1 = q_1, V_2 = q_2, V_3 = v_3] f_{V_1, V_2, V_3}(q_1, q_2, v_3) dv_3 \\ = E[G(Y_0) | V_1 = q_1, V_2 = q_2] f_{V_1, V_2}(v_1, v_2).$$

Dividing through identifies a local counterfactual outcome:

$$E[G(Y_0) | V_1 = q_1, V_2 = q_2].$$

Under Assumption 3.5, this also identifies $EG(Y_0)$. Moreover, we can apply the same logic to the pairs (q_1, q_3) and (q_2, q_3) to get further information on the treatment effects.

This argument applies more generally. It allows us to state the following theorem:

Theorem A.1 (Identification with a zero index)**.** *Let Assumptions 2.1, 2.2 and 3.2 hold. Fix a value $\boldsymbol{q}$ in $\tilde{\mathcal{Q}}$, so that Assumptions 3.3 and 3.4 also hold at $\boldsymbol{q}$. Let m be the degree of treatment k. Take l to be any subset of $\boldsymbol{J}$ that corresponds to a leading term in the expansion of the indicator function of $\{D = k\}$. Denote $\widetilde{T}$ the differential operator*

$$\widetilde{T} = \frac{\partial^m}{\prod_{i=1,\ldots,m} \partial_{l_i}}.$$

Then for $\boldsymbol{q} = (\boldsymbol{q}^l, \boldsymbol{q}^{\boldsymbol{J}-l})$,

$$f_{\boldsymbol{V}^l}(\boldsymbol{q}^l) = \frac{1}{c_l^k} \widetilde{T} \Pr[D = k | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}]$$

$$E[G(Y_k) | \boldsymbol{V}^l = \boldsymbol{q}^l] = \frac{\widetilde{T}\, E[G(Y) D_k | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}]}{\widetilde{T} \Pr[D = k | \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}]}.$$

Proof of Theorem A.1. The proof of Theorem A.1 is basically the same as that of Theorem 3.1. Steps 1 and 2 of the proof of Theorem 3.1 do not rely on any assumption about indices. They show that if we define

$$W_l(\boldsymbol{q}) = \int \prod_{j \in l} H(q_j - v_j) b_k(\boldsymbol{v}) d\boldsymbol{v}$$

where the set $l \subset \boldsymbol{J}$, then its cross-derivative with respect to $(\boldsymbol{p}^l)$ is

$$\int b_k(\boldsymbol{q}^l, \boldsymbol{v}_{-l}) d\boldsymbol{v}_{-l},$$

where $\boldsymbol{v}_{-l}$ collects all components of $\boldsymbol{v}$ whose indices are not in l.

Now let m be the degree of treatment k. In the sum (6.3), take any term l such that $|l| = m$. Recall that $\widetilde{T}$ denotes the differential operator

$$\widetilde{T} = \frac{\partial^m}{\prod_{i=1,\ldots,m} \partial_{j_i}}.$$

By the formula above, applying $\widetilde{T}$ to term l gives

$$c_l \int b_k(\boldsymbol{q}^l, \boldsymbol{v}_{-l}) d\boldsymbol{v}_{-l}.$$

Moreover, applying $\widetilde{T}$ to any other term l' obviously gives zero if term l' has degree less than m. Now take any other term l' of degree m. As $\widetilde{T}$ takes at least one derivative along a direction that is not in l', that term must also contribute zero.

This proves that

$$\widetilde{T} B_k(\boldsymbol{q}) = c_l^k \int b_k(\boldsymbol{q}^l, \boldsymbol{v}_{-l}) d\boldsymbol{v}_{-l};$$

note that it also implies that $\widetilde{T} B_k(\boldsymbol{q})$ only depends on $\boldsymbol{q}^l$.

Applying this first to $b_k(\boldsymbol{v}) = f_{\boldsymbol{V}}(\boldsymbol{v})$ and $B_k(\boldsymbol{q}) = \Pr(D = k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q})$, then to $b_k(\boldsymbol{v}) = E[G(Y_k)|\boldsymbol{V} = \boldsymbol{v}]f_{\boldsymbol{V}}(\boldsymbol{v})$ and $B_k(\boldsymbol{q}) = E[G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}]$ exactly as in the proof of Theorem 3.1, we get

$$\int f_{\boldsymbol{V}}(\boldsymbol{q}^l, \boldsymbol{v}_{-l})d\boldsymbol{v}_{-l} = \frac{1}{c_l^k}\widetilde{T}\Pr(D = k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q})$$
$$\int E[G(Y_k)|\boldsymbol{V} = (\boldsymbol{q}^l, \boldsymbol{v}_{-l})]f_{\boldsymbol{V}}(\boldsymbol{q}^l, \boldsymbol{v}_{-l})d\boldsymbol{v}_{-l} = \frac{1}{c_l^k}\widetilde{T}E(G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}).$$

Since the left-hand sides are simply $f_{\boldsymbol{V}^l}(\boldsymbol{v}^l)$ and $E[G(Y_k)|\boldsymbol{V}^l = \boldsymbol{q}^l]f_{\boldsymbol{V}^l}(\boldsymbol{v}^l)$, the conclusion of the Theorem follows immediately. □ □

Theorem A.1 is a generalization of Theorem 3.1 (just take $m = J$). It calls for three remarks. First, we could weaken its hypotheses somewhat. We could for instance replace $(0,1)^J$ with $(0,1)^m$ in the statement of Assumption 3.5.

Second, when $m < J$ the treatment effects are overidentified. This is obvious from the equalities in Theorem A.1, in which the right-hand side depends on $\boldsymbol{q}$ but the left-hand side only depends on $\boldsymbol{q}^I$.

Finally, considering several treatment values can identify even more, since $\boldsymbol{V}$ is assumed to be the same across k. Theorem 3.1 would then imply that if there is any treatment value k with a nonzero index, then the joint density $f_{\boldsymbol{V}}$ is identified from that treatment value.

A.2 Further Analysis of Unordered Monotonicity

Our formalism allows us to derive a new characterization of the unordered monotonicity property defined by Heckman and Pinto (2018). Take any treatment value k. In our model, a change in instruments $\boldsymbol{Z}$ acts on the treatment assigned to an observation with unobserved characteristics $\boldsymbol{V}$ through the indicator functions $S_j = \mathbf{1}(V_j < Q_j(\boldsymbol{Z}))$, which depend on the thresholds $\boldsymbol{Q}(\boldsymbol{Z})$.

Unordered monotonicity requires that there exist changes in thresholds $\Delta\boldsymbol{Q}$ such that for $\boldsymbol{Q}' = \boldsymbol{Q} + \boldsymbol{\Delta Q}$,

$$\Pr\{d_k(\boldsymbol{V}, \boldsymbol{Q}) = 0 \text{ and } d_k(\boldsymbol{V}, \boldsymbol{Q}') = 1\} \times \Pr\{d_k(\boldsymbol{V}, \boldsymbol{Q}) = 1 \text{ and } d_k(\boldsymbol{V}, \boldsymbol{Q}') = 0\} = 0,$$

where the probabilities are computed over the joint distribution of $\boldsymbol{V}$.

In our framework, several thresholds are typically relevant for each treatment value. This makes the analysis of unordered monotonicity complex in general. To understand why, we start from the expression (2.2) of D_k as a polynomial of $\boldsymbol{S} = (S_1, \ldots, S_J)$ for $S_j(\boldsymbol{V}, \boldsymbol{Q}) = \mathbb{1}(V_j < Q_j)$. For any change in thresholds $\Delta\boldsymbol{Q}$ that induces changes in the indicators $\Delta\boldsymbol{S}$, Taylor's theorem yields

$$\Delta D_k = \sum_{m=1}^{J} \sum_{\alpha_1+\ldots+\alpha_J=m} \frac{1}{\alpha_1!\alpha_2!\cdots\alpha_J!} \frac{\partial^m \mathcal{D}_k(\boldsymbol{S})}{\partial S_1^{\alpha_1} \partial S_2^{\alpha_2} \ldots \partial S_J^{\alpha_J}} \prod_{l=1}^{J} \Delta S_l^{\alpha_l}, \tag{A.1}$$

where α_j is a nonnegative integer for $j = 1, \ldots, J$. Note that this is an exact expansion since $\mathcal{D}_k$ is a polynomial. Moreover, note that given a change in one threshold ΔQ_j, only S_j changes and

$$\Delta S_j = \mathbb{1}(0 < V_j - Q_j < \Delta Q_j) - \mathbb{1}(\Delta Q_j < V_j - Q_j < 0). \tag{A.2}$$

(We do not need to distinguish between the weak and strict inequalities since the distribution of V_j is absolutely continuous with respect to the Lebesgue measure.)

The changes ΔS_j can only take the values 0 or ± 1. In general higher-order terms in expansion A.1 may be nonzero. However, if the changes in thresholds $\Delta\boldsymbol{Q}$ are small then we can neglect the higher order terms since the values of $\boldsymbol{V}$ for which several ΔS_j are nonzero occur with very small probability. To make this more precise, we use the following definition:

Definition A.1 (Two-Way Flows)**.** A change in thresholds $\Delta\boldsymbol{Q}$ generates two-way flows for treatment value k if and only if

$$\lim_{\varepsilon \to 0} \left(\frac{\Pr\left(D_k(0) = 0 \text{ and } D_k(\varepsilon) = 1\right)}{\varepsilon} \times \frac{\Pr\left(D_k(0) = 1 \text{ and } D_k(\varepsilon) = 0\right)}{\varepsilon} \right) > 0$$

for $D_k(\varepsilon) \equiv d_k(\boldsymbol{V}, \boldsymbol{Q} + \varepsilon\Delta\boldsymbol{Q})$.

We now provide new characterizations of unordered monotonicity.

Theorem A.2 (Characterizing Unordered Monotonicity in the Small)**.** *Fix a value $\boldsymbol{Q}$ of the thresholds. Denote*

$$\nabla\mathcal{D}_k(\boldsymbol{S}) = \frac{\partial \mathcal{D}_k}{\partial \boldsymbol{S}}(\boldsymbol{S}).$$

Assume that $J \geq 2$ and that there exist two values $j_1 \neq j_2$ such that $\nabla_{j_1}\mathcal{D}_k$ and $\nabla_{j_2}\mathcal{D}_k$ are not identically zero. Then:

1. *If each component of $\nabla\mathcal{D}_k(\boldsymbol{S})$ has a constant sign when $\boldsymbol{S}$ varies over $\{0,1\}^J$, then some changes in thresholds do not generate two-way flows, and some others do.*

2. *If the sign of any component $\nabla_j\mathcal{D}_k(\boldsymbol{S})$ changes when S_j switches between 0 and 1, then any change in thresholds generates two-way flows.*

(In these two statements, we take 0 to have the same sign as both -1 and $+1$.)

Proof of Theorem A.2. Take $\varepsilon > 0$ small. Remember that given a change in thresholds $\varepsilon\Delta Q_j$,

$$\Delta S_j = \mathbb{1}(0 < V_j - Q_j < \varepsilon\Delta Q_j) - \mathbb{1}(\varepsilon\Delta Q_j < V_j - Q_j < 0),$$

which is zero or has the sign of ΔQ_j.

Under our assumptions on the distribution of $\boldsymbol{V}$, the probability that $\Delta S_j \neq 0$ is of order ε; the probability that $\Delta S_j \Delta S_l \neq 0$ is of order ε^2, etc. Given Definition A.1, we only need to work on the first-order terms in expansion (A.1) since the other terms generate vanishingly small corrections. That is, we use

(A.3)

$$\begin{aligned}\Delta D_k &\simeq \sum_{j=1}^{J} \nabla_j\mathcal{D}_k(\boldsymbol{S}) \times \Delta S_j \\ &= \sum_{j=1}^{J} \nabla_j\mathcal{D}_k(\boldsymbol{S}) \times \left(\mathbb{1}(0 < V_j - Q_j < \varepsilon\Delta Q_j) - \mathbb{1}(\varepsilon\Delta Q_j < V_j - Q_j < 0)\right).\end{aligned}$$

- *Proof of part 1:*

  To prove part 1 of the theorem, assume that each derivative $\nabla_j\mathcal{D}_k$ has a constant sign, independent of $\boldsymbol{S} \in \{0,1\}^J$.

  Then it is easy to find changes $\Delta\boldsymbol{Q}$ that only generate one-way flows. First take each ΔQ_j to have the sign of $\nabla_j\mathcal{D}_k$.

  Since each ΔS_j has the sign of the corresponding ΔQ_j, each product term in the sum (A.3) is non-negative, and so is the change in D_k. Obviously, changing the sign of all ΔQ_j's would generate one-way flows in the opposite direction.

It is equally easy to find changes in instruments that generate two-way flows. Take the indices j_1 and j_2 referred to in the statement of the theorem. Take $\Delta Q_m = 0$ for $m \neq j_1, j_2$. Then expansion (A.3) becomes

$$\Delta D_k \simeq \nabla_{j_1} \mathcal{D}_k(\boldsymbol{S}) \times \Delta S_{j_1} + \nabla_{j_2} \mathcal{D}_k(\boldsymbol{S}) \times \Delta S_{j_2}.$$

Choose some $\Delta Q_{j_1}, \Delta Q_{j_2} \neq 0$ such that

$$\nabla_{j_1} \mathcal{D}_k(\boldsymbol{S}) \times \Delta Q_{j_1} \text{ and } \nabla_{j_2} \mathcal{D}_k(\boldsymbol{S}) \times \Delta Q_{j_2}$$

have opposite signs (which do not vary with $\boldsymbol{S}$ by assumption).

Take $|V_{j_1} - Q_{j_1}|$ small and $|V_{j_2} - Q_{j_2}|$ not small, so that ΔS_{j_1} has the sign of ΔQ_{j_1} and $\Delta S_{j_2} = 0$; then ΔD_k has the sign of $\nabla_{j_1} \mathcal{D}_k(\boldsymbol{S}) \times \Delta Q_{j_1}$. Permuting j_1 and j_2 generates the opposite sign; therefore such a change in thresholds generates two-way flows.

- *Proof of part 2:*

  To prove part 2 of the theorem, take j such that $\nabla_j \mathcal{D}_k$ changes sign when the sign of $V_j - Q_j$ changes (so that S_j switches between 0 and 1). Let $\Delta Q_m = 0$ for all $m \neq j$, so that

  $$\Delta D_k \simeq \nabla_j \mathcal{D}_k(\boldsymbol{S}) \times \Delta S_j.$$

  By the assumption in part 2, the sign of ΔD_k is the sign of ΔS_j for some values of $\boldsymbol{V}$ and the opposite sign for other values. Take any change in the threshold ΔQ_j. Since ΔS_j is zero or has the sign of ΔQ_j, ΔD_k must take opposite values as $\boldsymbol{V}$ varies. □

□

To illustrate the theorem, first consider the double hurdle model, for which $\nabla \mathcal{D}_1(\boldsymbol{S}) = (S_2, S_1) \geq 0$. This case is covered by part 1 of Theorem A.2. Changes such that ΔQ_1 and ΔQ_2 have the same sign do not generate two-way flows, but changes that generate $\Delta Q_1 \Delta Q_2 < 0$ do.

Now turn to the model of Example 1, where $\nabla \mathcal{D}_2(\boldsymbol{S}) = (1 - 2S_2, 1 - 2S_1)$. This corresponds to part 2 of the Theorem, since the sign of $(1 - 2s)$ depends on $s = 0, 1$.

Using the expansion (A.3) gives, with $j_1 = 1, j_2 = 2$:

$$\Delta D_2 \simeq (1 - 2S_2) \times \Delta S_1 + (1 - 2S_1) \times \Delta S_2.$$

Depending on the values of $\boldsymbol{V}$ and therefore of S_1 and S_2, this can be

$$\Delta S_1 + \Delta S_2, \Delta S_1 - \Delta S_2, \Delta S_2 - \Delta S_1, \text{ or } -\Delta S_1 - \Delta S_2.$$

To get one way flows only, we would need to chaneg thresholds to induce $\Delta S_1, \Delta S_2 = \pm 1$ such that the four numbers above have the same sign. But that is clearly impossible. Hence *any* change in instruments creates two-way flows.

B Additional Proofs

B.1 Proof of Corollary 3.2

First consider the average treatment effect. Under Assumption 3.5, we have that

$$EG(Y_k) = \int E\left(G(Y_k)|\boldsymbol{V} = \boldsymbol{v}\right) f_{\boldsymbol{V}}(\boldsymbol{v}) d\boldsymbol{v},$$

which implies (3.2) immediately.

Now consider $E[G(Y_k) - G(Y_\ell)|D = k]$. Note that

$$\begin{aligned}
&E[G(Y_k) - G(Y_\ell)|D = k, \boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}] \\
&= E[G(Y_k) - G(Y_\ell)|d_k(\boldsymbol{V}, \boldsymbol{q}) = 1] \\
&= \frac{\int \mathbb{1}\left(d_k(\boldsymbol{v}, \boldsymbol{q}) = 1\right) E[G(Y_k) - G(Y_\ell)|\boldsymbol{V} = \boldsymbol{v}] f_{\boldsymbol{V}}(\boldsymbol{v}) d\boldsymbol{v}}{\int \mathbb{1}\left(d_k(\boldsymbol{v}, \boldsymbol{q}) = 1\right) f_{\boldsymbol{V}}(\boldsymbol{v}) d\boldsymbol{v}}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&E[G(Y_k) - G(Y_\ell)|D = k] \\
&= EE[G(Y_k) - G(Y_\ell)|D = k, \boldsymbol{Q}(\boldsymbol{Z})]] \\
&= \int \frac{\int \mathbb{1}\left(d_k(\boldsymbol{v}, \boldsymbol{q}) = 1\right) E[G(Y_k) - G(Y_\ell)|\boldsymbol{V} = \boldsymbol{v}] f_{\boldsymbol{V}}(\boldsymbol{v}) d\boldsymbol{v}}{\int \mathbb{1}\left(d_k(\boldsymbol{v}, \boldsymbol{q}) = 1\right) f_{\boldsymbol{V}}(\boldsymbol{v}) d\boldsymbol{v}} dF_{\boldsymbol{Q}(\boldsymbol{Z})|D}(\boldsymbol{q}|k).
\end{aligned}$$

By Bayes' rule, we have that

$$dF_{\boldsymbol{Q}(\boldsymbol{Z})|D}(\boldsymbol{q}|k) = \frac{\Pr[D=k|\boldsymbol{Q}(\boldsymbol{Z})=\boldsymbol{q}]}{\Pr(D=k)} dF_{\boldsymbol{Q}(\boldsymbol{Z})}(\boldsymbol{q}).$$

Since

$$\Pr[D=k|\boldsymbol{Q}(\boldsymbol{Z})=\boldsymbol{q}] = \int \mathbb{1}\left(d_k(\boldsymbol{v},\boldsymbol{q})=1\right) f_{\boldsymbol{V}}(\boldsymbol{v})d\boldsymbol{v},$$

we have that

$$\begin{aligned}
&E[G(Y_k)-G(Y_\ell)|D=k] \\
&= \int \frac{\int \mathbb{1}\left(d_k(\boldsymbol{v},\boldsymbol{q})=1\right) E[G(Y_k)-G(Y_\ell)|\boldsymbol{V}=\boldsymbol{v}] f_{\boldsymbol{V}}(\boldsymbol{v})d\boldsymbol{v}}{\Pr(D=k)} dF_{\boldsymbol{Q}(\boldsymbol{Z})}(\boldsymbol{q}) \\
&= \frac{\int \Pr\left(d_k(\boldsymbol{v},\boldsymbol{Q}(\boldsymbol{Z}))=1\right) E[G(Y_k)-G(Y_\ell)|\boldsymbol{V}=\boldsymbol{v}] f_{\boldsymbol{V}}(\boldsymbol{v})d\boldsymbol{v}}{\Pr(D=k)} \\
&= \int \Delta^{(k,\ell)}_{\mathrm{MTE}}(\boldsymbol{v})\omega^k_{\mathrm{TT}}(\boldsymbol{v})d\boldsymbol{v}.
\end{aligned}$$

We now move to the identification of the policy relevant treatment effects. Recall that in the proof of Theorem 3.1 (see equation (6.1)), we have that

$$\begin{aligned}
&E[G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z})=\boldsymbol{q}] \\
&= \int \mathbb{1}\left(d_k(\boldsymbol{v},\boldsymbol{q})=1\right) E[G(Y_k)|\boldsymbol{V}=\boldsymbol{v}] f_{\boldsymbol{V}}(\boldsymbol{v})d\boldsymbol{v}.
\end{aligned}$$

Since $G(Y) = \sum_{k\in\mathcal{K}} G(Y)D_k$, we then have that

$$\begin{aligned}
E[G(Y)] &= \sum_{k\in\mathcal{K}} E[E[G(Y)D_k|\boldsymbol{Q}(\boldsymbol{Z})=\boldsymbol{q}]] \\
&= \sum_{k\in\mathcal{K}} \int \Pr[d_k(\boldsymbol{v},\boldsymbol{Q}(\boldsymbol{Z}))=1] E[G(Y_k)|\boldsymbol{V}=\boldsymbol{v}] f_{\boldsymbol{V}}(\boldsymbol{v})d\boldsymbol{v}.
\end{aligned}$$

Similarly, we have that

$$\begin{aligned}
E[D] &= \sum_{k\in\mathcal{K}} kE[E[D_k|\boldsymbol{Q}(\boldsymbol{Z})=\boldsymbol{q}]] \\
&= \sum_{k\in\mathcal{K}} k \int \Pr[d_k(\boldsymbol{v},\boldsymbol{Q}(\boldsymbol{Z}))=1] f_{\boldsymbol{V}}(\boldsymbol{v})d\boldsymbol{v}
\end{aligned}$$

and that

$$\begin{aligned} E[D_k = 1] &= E[E[D_k|\boldsymbol{Q}(\boldsymbol{Z}) = \boldsymbol{q}]] \\ &= \int \Pr[d_k(\boldsymbol{v}, \boldsymbol{Q}(\boldsymbol{Z})) = 1] f_{\boldsymbol{V}}(\boldsymbol{v}) d\boldsymbol{v}. \end{aligned}$$

Then the desired results follow immediately since the new policy only changes $\boldsymbol{Q}$ to $\boldsymbol{Q}^*$, while everything else remains the same. □

B.2 Proof of Theorem 4.1

It follows from (2.1) on page 7 that

$$Q_1(\boldsymbol{Z}) + Q_2(\boldsymbol{Z}) = 2P_0(\boldsymbol{Z}) + P_2(\boldsymbol{Z}). \tag{B.1}$$

The right hand side of (B.1) is identified directly from the data. Suppose that $\tilde{Q}_1(\boldsymbol{Z})$ and $\tilde{Q}_2(\boldsymbol{Z})$ also satisfy $\tilde{Q}_1(\boldsymbol{Z}) + \tilde{Q}_2(\boldsymbol{Z}) = 2P_0(\boldsymbol{Z}) + P_2(\boldsymbol{Z})$, as well as Assumption 4.1. Then writing $\Delta_j(\boldsymbol{Z}) = Q_j(\boldsymbol{Z}) - \tilde{Q}_j(\boldsymbol{Z})$ $(j = 1, 2)$ gives $\Delta_1(\boldsymbol{Z}) = -\Delta_2(\boldsymbol{Z})$. But by Assumption 4.1, Δ_1 does not depend on Z_2, and Δ_2 does not depend on Z_1. Therefore we must have $\tilde{Q}_1(Z_1) = Q_1(Z_1) + C$ and $\tilde{Q}_2(Z_2) = Q_2(Z_2) - C$, where C is a constant. This proves that Q_1 and Q_2 are identified up to an additive constant.

Further, take any $(z_1^0, z_2^0) \in \mathcal{Z}$. If we take $Q_2(z_2) = P(z_1^0, z_2) - C_1^0$ for some constant C_1^0, then by (B.1),

$$Q_1(z_1) = P(z_1, z_2) - P(z_1^0, z_2) + C_1^0. \tag{B.2}$$

Since the right-hand side of (B.2) should not depend on z_2, we set

$$\begin{aligned} Q_1(z_1) &= P(z_1, z_2^0) - P(z_1^0, z_2^0) + C_1^0 \\ Q_2(z_2) &= P(z_1^0, z_2) - C_1^0. \end{aligned}$$

To describe the possible range of C_1^0, note that we require that

$$\begin{aligned} &\Pr(D = 0) = \Pr[Q_1(Z_1) > 0 \text{ and } Q_2(Z_2) > 0] > 0, \\ &\Pr(D = 1) = \Pr[Q_1(Z_1) < 1 \text{ and } Q_2(Z_2) < 1] > 0, \\ &\Pr(D = 2) = \Pr[Q_1(Z_1) > 0 \text{ and } Q_2(Z_2) < 1] + \Pr[Q_1(Z_1) < 1 \text{ and } Q_2(Z_2) > 0] > 0. \end{aligned}$$

That is, C_1^0 must satisfy the following restrictions:

$$\begin{aligned}
&\Pr[P(z_1^0, z_2^0) - P(Z_1, z_2^0) < C_1^0 < P(z_1^0, Z_2)] > 0, \\
&\Pr[P(z_1^0, Z_2) - 1 < C_1^0 < 1 + P(z_1^0, z_2^0) - P(Z_1, z_2^0)] > 0, \\
&\Pr\left[\max\{P(z_1^0, z_2^0) - P(Z_1, z_2^0), P(z_1^0, Z_2) - 1\} < C_1^0\right] \\
&+ \Pr\left[C_1^0 < \min\{1 + P(z_1^0, z_2^0) - P(Z_1, z_2^0), P(z_1^0, Z_2)\}\right] > 0.
\end{aligned}$$

□

B.3 Proof of Theorem 4.2

Recall that we denote $H(z_1, z_2) = \Pr(D = 1 | Z_1 = z_1, Z_2 = z_2)$ the propensity score. Under our exclusion restrictions, $H(z_1, z_2) = F_{V_1,V_2}(G_1(z_1), G_2(z_2))$.

Let $f_{\boldsymbol{V}}(v_1, v_2)$ denote the density of $\boldsymbol{V} = (V_1, V_2)$. By construction,

$$H(z_1, z_2) = F_{\boldsymbol{V}}(G_1(z_1), G_2(z_2)) = \int_0^{G_1(z_1)} \int_0^{G_2(z_2)} f_{\boldsymbol{V}}(v_1, v_2) dv_1 dv_2. \tag{B.3}$$

Differentiating both sides of (B.3) with respect to z_1 gives

$$\frac{\partial H}{\partial z_1}(z_1, z_2) = G_1'(z_1) \int_0^{G_2(z_2)} f_{\boldsymbol{V}}(G_1(z_1), v_2) dv_2. \tag{B.4}$$

Now letting $z_2 \to b_2$ on both sides of (B.4) yields

$$\lim_{z_2 \to b_2} \frac{\partial H}{\partial z_1}(z_1, z_2) = G_1'(z_1) \left[\lim_{z_2 \to b_2} \int_0^{G_2(z_2)} f_{\boldsymbol{V}}(G_1(z_1), v_2) dv_2\right]. \tag{B.5}$$

The expression inside the brackets on the right side of (B.5) is 1 since $\lim_{z_2 \to b_2} G_2(z_2) = 1$ and the marginal distribution of V_2 is $U[0, 1]$. Therefore we identify G_1 by

$$G_1(z_1) = \int_{a_1}^{z_1} \lim_{t_2 \to b_2} \frac{\partial H}{\partial z_1}(t_1, t_2) dt_1. \tag{B.6}$$

Analogously, we identify G_2 by

$$G_2(z_2) = \int_{a_2}^{z_2} \lim_{t_1 \to b_1} \frac{\partial H}{\partial z_2}(t_1, t_2) dt_2. \tag{B.7}$$

Returning to (B.3), since G_1 and G_2 are strictly increasing we identify $F_{\boldsymbol{V}}$ by

$$F_{\boldsymbol{V}}(v_1, v_2) = H(G_1^{-1}(v_1), G_2^{-1}(v_2)).$$

□

B.4 Proof of Theorem 4.3

B.4.1 Proof of part 1

Given our differentiability assumptions, we can take derivatives of the formula

$$\phi\left(H(z_1, z_2)\right) = \phi(G_1(z_1)) + \phi(G_2(z_2)) \tag{B.8}$$

over $\mathcal{N}$. Using

$$\frac{\partial^2 \left(\phi \circ H\right)}{\partial z_1 \partial z_2}(z_1, z_2) = 0,$$

we obtain

$$\phi''(h)\frac{\partial H}{\partial z_1}(z_1, z_2)\frac{\partial H}{\partial z_2}(z_1, z_2) + \phi'(h)\frac{\partial^2 H}{\partial z_1 \partial z_2}(z_1, z_2) = 0$$

with $h = H(z_1, z_2)$.

Take any smooth curve contained in $\mathcal{N}$ and parameterize it as $h \to (z_1(h), z_2(h))$ with $h = H(z_1(h), z_2(h))$; then we have a differential equation

$$\phi''(h)\frac{\partial H}{\partial z_1}(z_1(h), z_2(h))\frac{\partial H}{\partial z_2}(z_1(h), z_2(h)) + \phi'(h)\frac{\partial^2 H}{\partial z_1 \partial z_2}(z_1(h), z_2(h)) = 0. \tag{B.9}$$

Using (B.8), the partial derivatives H_1 and H_2 cannot take the value zero on $\mathcal{N}$ since G_1' and G_2' are never zero. Therefore we can rewrite (B.9) as

$$\frac{\phi''}{\phi'}(h) = -\frac{H_{12}}{H_1 H_2}(z_1(h), z_2(h))$$

over $\mathcal{N}$.

We note that this equation incorporates a sign constraint and overidentifying restrictions. For ϕ to be strictly decreasing and convex, we require $H_{12}/(H_1 H_2) \geq 0$. Moreover, on any admissible curve the ratio $H_{12}/(H_1 H_2)$ must be the same function of h, which we denote $R(h)$. □

B.4.2 Proof of part 2

From now on we denote $(\underline{h}, \overline{h}) \subset (0,1)$ the image of $\mathcal{N}$ by H.

We use the fact that $\partial \log(-\phi'(h))/\partial h = \phi''(h)/\phi'(h)$ to obtain

$$\log\left(-\phi'(h)\right) = \int_h^{\bar{h}} R(t)dt + \log\left(-\phi'(\bar{h})\right),$$

so that

$$\phi'(h) = \phi'(\bar{h}) \exp\left(\int_h^{\bar{h}} R(t)dt\right).$$

Denoting

$$\mathbb{T}(h) := \int_h^{\bar{h}} dk \exp\left(\int_k^{\bar{h}} R(t)dt\right)$$

gives us $\phi(h) = \phi(\bar{h}) - \phi'(\bar{h})\mathbb{T}(h)$. Note that by construction $\mathbb{T}$ is a decreasing function and $\mathbb{T}(\bar{h}) = 0$. Moreover, $\phi'(\bar{h})$ cannot be zero since ϕ would be constant. $\square$

B.4.3 Proof of part 3

If ϕ solves (B.8) then clearly so does $\alpha\phi$ for any $\alpha > 0$; we normalize $\phi'(\bar{h}) = -1$. Hence, from now on, $\phi(h) = \phi(\bar{h}) - \mathbb{T}(h)$. The constant $\phi(\bar{h})$ must be non-negative since ϕ cannot take negative values. Moreover, since ϕ is convex, $\phi'(\bar{h}) = -1$, and $\phi(1) = 0$, we must have $\phi(\bar{h}) \leq 1 - \bar{h}$. If moreover $\bar{h} = \sup_{\boldsymbol{z} \in \mathcal{N}} \Pr(D = 1 | \boldsymbol{Z} = \boldsymbol{z}) = 1$, then $\phi(\bar{h}) = \phi(1) = 0$; this defines directly $\phi(h) = -\mathbb{T}(h)$ over $(\underline{h}, 1)$. $\square$

B.4.4 Proof of part 4

Since the model is well-specified, there is a solution G_1, G_2 (the thresholds of the true DGP). In addition, since any other admissible $(\tilde{G}_1, \tilde{G}_2)$ must satisfy

$$\phi(\tilde{G}_1(z_1)) + \phi(\tilde{G}_2(z_2)) = \phi(H(z_1, z_2)) = \phi(G_1(z_1)) + \phi(G_2(z_2))$$

on $\mathcal{N}$, it must be that

$$\begin{aligned}
\phi(\tilde{G}_1(z_1)) &= \phi(G_1(z_1)) - k \\
\phi(\tilde{G}_2(z_2)) &= \phi(G_2(z_2)) + k
\end{aligned}$$

for some constant k. Any such constant must be such that $\phi(G_1(z_1)) - k$ and $\phi(G_2(z_2)) + k$ are both nonnegative for all z_1 and z_2 in the projections of $\mathcal{N}$. That is,

$$-\inf \phi(G_2(z_2)) \leq k \leq \inf \phi(G_1(z_1)).$$

If moreover $\sup_{\boldsymbol{z} \in \mathcal{N}} \Pr(D = 1 | \boldsymbol{Z} = \boldsymbol{z}) = 1$, then $\bar{h} = 1$. Take a sequence $(\boldsymbol{z}_n)$ such that $H(\boldsymbol{z}_n)$ converges to $\bar{h} = 1$. Then $\phi(H(\boldsymbol{z}_n))$ converges to zero, so that both $\phi(G_1(z_{1n}))$ and $\phi(G_2(z_{2n}))$ must converge to zero. The double inequality above implies that $k = 0$, and G_1 and G_2 are point-identified on the projections of $\mathcal{N}$. $\square$

C The Entry Game

Let us return to Example 2, in which two firms $j = 1, 2$ are considering entry into a new market. Firm j has profit π_j^m if it becomes a monopoly, and $\pi_j^d < \pi_j^m$ if both firms enter. We saw that if $\pi_j^m > 0 > \pi_j^d$ for both firms, then there are two symmetric equilibria, with only one firm operating. Now assume that we observe not only the number of entrants as in Example 2, but also their identity. With profits given by $\pi_j^m = V_j - Q_j(\boldsymbol{Z})$ and $\pi_j^d = \bar{V}_j - \bar{Q}_j(\boldsymbol{Z})$, if only firm 1 entered then we know that $\pi_1^m > 0$ and $\pi_2^d < 0$, so that

$$V_1 > Q_1(\boldsymbol{Z}) \text{ and } \bar{V}_2 < \bar{Q}_2(\boldsymbol{Z}).$$

That still leaves two possible cases:

1. $\pi_2^m < 0$, and the unique equilibrium has only firm 1 entering the market;
2. and $\pi_2^m > 0$, and there is another, symmetric equilibrium with only firm 2 entering.

Now let us postulate an equilibrium selection rule that has a threshold structure: when both π_1^m and π_m^2 are positive, firm 1 is selected to be the unique entrant if and only if $U < q(\boldsymbol{Z})$. Then the necessary and sufficient set of conditions for the entry of firm 1 only is

$$V_1 > Q_1(\boldsymbol{Z}) \text{ and } \left(V_2 < Q_2(\boldsymbol{Z}) \text{ or } \left(\bar{V}_2 < \bar{Q}_2(\boldsymbol{Z}) \text{ and } U < q(\boldsymbol{Z})\right)\right).$$

This is again a special case of the general framework we analyze in this paper.

D Detailed Discussion of Heckman, Urzua, and Vytlacil (2008)

Heckman, Urzua, and Vytlacil (2008) consider a multinomial discrete choice model for treatment. They posit

$$D = k \iff R_k(\boldsymbol{Z}) - U_k > R_l(\boldsymbol{Z}) - U_l \text{ for } l = 0, \ldots, K-1 \text{ such that } l \neq k,$$

where the U's are continuously distributed and independent of $\boldsymbol{Z}$.

Define

$$\boldsymbol{R}(\boldsymbol{Z}) = (R_k(\boldsymbol{Z}) - R_l(\boldsymbol{Z}))_{l \neq k} \quad \text{and} \quad \boldsymbol{U} = (U_k - U_l)_{l \neq k}\,.$$

Then $D_k = \mathbb{1}(\boldsymbol{R}(\boldsymbol{Z}) > \boldsymbol{U})$; and defining $Q_l(\boldsymbol{Z}) = \Pr[\boldsymbol{U}_l < \boldsymbol{R}_l(\boldsymbol{Z})|\boldsymbol{Z}]$ allows us to write the treatment model as

$$D = k \text{ iff } \boldsymbol{V} < \boldsymbol{Q}(\boldsymbol{Z}), \tag{D.1}$$

where each V_l is distributed as $U[0,1]$.

The applications they consider are GED certification (with three treatments: permanent high school dropout, GED, high school degree) and randomized trials with imperfect compliance (for example, no training, classroom training, and job search assistance).

They then study the identification of marginal and local average treatment effects under assumptions that are similar to ours: continuous instruments that generate enough dimensions of variation in the thresholds. They assume that $\boldsymbol{V}$ is continuously distributed with full support; that $(\boldsymbol{U}, \boldsymbol{V}) \perp\!\!\!\perp Z$; and that all treatments have positive probabilities. More importantly, they make either

- assumption (a): for each treatment j, there is a component of $\boldsymbol{Z}$ that drives some variation in R_j conditional on the other components, and in R_j only;
- assumption (b): for each treatment j, there is a component of $\boldsymbol{Z}$ that drives continuous variation in R_j conditional on the other components, and no variation in the other components of R.

For any subset of treatments $\mathcal{J} \subset \mathcal{K}$, they define $Y_{\mathcal{J}}$ to be the outcome when the agent chooses the best treatment from $\mathcal{J}$. They also define $\Delta_{\mathcal{J},\mathcal{L}} = Y_{\mathcal{J}} - Y_{\mathcal{L}}$, and in

particular the MTE

$$E\left(\Delta_{\mathcal{J},\mathcal{L}}|\boldsymbol{Z}, R_{\mathcal{J}}(\boldsymbol{Z}) = R_{\mathcal{L}}(\boldsymbol{Z})\right).$$

They show that

- if we take $\mathcal{J} = \{j\}$ and $\mathcal{L} = \mathcal{K} - \{j\}$, then the LATE is identified under (a) and the MTE is identified under (b);
- if we take any $\mathcal{J}$ and $\mathcal{L} = \mathcal{K} - \mathcal{J}$, then the results are similar but the MTEs and LATEs are defined by conditioning on the values of the Q's rather than on the Z's.

They do not invoke any large support assumptions to obtain identification results mentioned just above.

However, if we take $\mathcal{J} = \{j\}$ and $\mathcal{L} = \{l\}$, then their corresponding identification results (see Theorem 3 of Heckman, Urzua, and Vytlacil (2008)) require a large support condition. To see their logic, suppose that $K = 3$ and that one of the R_j's is sufficiently negative that the probability of choosing one of the choices is arbitrarily small. This case effectively reduces to the binary treatment case; their LIV estimand, which is the limit of a sequence of Wald estimands, identifies the MTE.

We do not rely on this type of identification-at-infinity strategy since we identify the MTE via multidimensional cross derivatives. Note that our identification results are conditional on the assumption that $\boldsymbol{Q}$ is already identified. A more stringent assumption on the support of $\boldsymbol{Z}$ might be necessary to identify $\boldsymbol{Q}$, as demonstrated in Matzkin (1993, 2007). In this sense, our assumptions are not necessarily weaker than those of Heckman, Urzua, and Vytlacil (2008). We view our identification results and theirs as complementing each other.

E Non-rectangular Threshold Conditions

The threshold conditions we postulated in Assumption 2.1 have the "rectangular" form $V_j < Q_j(\boldsymbol{Z})$. Suppose that the threshold conditions $j = 1, \ldots, J$ have the more general form

$$\boldsymbol{\alpha}_j \cdot \boldsymbol{U} \leq R_j(\boldsymbol{Z})$$

where the $\boldsymbol{\alpha}_j$ are possibly unknown parameter vectors in $\mathbb{R}^L$ and $\boldsymbol{U} = (U_1, \ldots, U_L)$ is independent of $\boldsymbol{Z}$. For notational simplicity, assume that each (scalar) random

variable $u_j \equiv \boldsymbol{\alpha}_j \cdot \boldsymbol{U}$ has positive density everywhere; denote H_j its cdf. Then each threshold condition can be written equivalently as

$$V_j \equiv H_j(u_j) < H_j(R_j(\boldsymbol{Z})) \equiv Q_j(\boldsymbol{Z}).$$

By construction, each V_j is distributed uniformly over $[0, 1]$. Moreover, since each threshold Q_j is an increasing function of the corresponding R_j only, any exclusion restriction assumed on either form applies equally to the other, so that we can hope to identify the thresholds Q_j under suitable assumptions. If they are indeed identified, then we can apply Theorem 3.1 to recover the joint density of $\boldsymbol{V} = (V_1, \ldots, V_j)$ and the MTE conditional on $\boldsymbol{v}$.

The random variables $\boldsymbol{V}$ and the thresholds $\boldsymbol{Q}$ are only auxiliary objects, and the analyst is likely to be more interested in the $\boldsymbol{U}$ and $\boldsymbol{R}$. If the cdf H_j were known, then we could write $R_j = H_j^{-1}(Q_j)$ and by the change-of-variables formula,

$$f_{\boldsymbol{u}}(u_1, \ldots, u_J) = f_{\boldsymbol{V}}\left(H_1^{-1}(u_1), \ldots, H_J^{-1}(u_J)\right) \times \prod_{j=1}^{J} H_j'(u_j).$$

In turn, knowing the joint distribution of $\boldsymbol{u}$ directly gives the density of $\boldsymbol{U}$ if $L = J$ and the matrix $\boldsymbol{\alpha}$ whose rows are the vectors $\boldsymbol{\alpha}_j'$ is invertible:

$$f_{\boldsymbol{U}}(\boldsymbol{U}) = f_{\boldsymbol{u}}\left(\boldsymbol{\alpha}\boldsymbol{U}\right) \times \left|\boldsymbol{\alpha}\right|.$$

If more realistically the H_j and $\boldsymbol{\alpha}_j$ are unknown, we may still use other restrictions. As an illustration, take a recursive system, where the matrix $\boldsymbol{\alpha}$ is lower-triangular with diagonal terms equal to one. Then since $U_2 = u_2 - \alpha_{21}u_1 = H_2^{-1}(V_2) - \alpha_{21}H_1^{-1}(V_1)$, the independence of U_1 and U_2, for instance, would translate into the independence of V_1 and of the variable

$$W_2 \equiv H_2^{-1}(V_2) - \alpha_{21}H_1^{-1}(V_1).$$

Now $V_2 = H_2(W_2 + \alpha_{21}U_1)$, so this in turn implies that the (identified) distribution of V_2 conditional of V_1 must satisfy

$$F_{V_2|V_1}\left(H_2\left(w_2 + \alpha_{21}H_1^{-1}(v_1)\right)|v_1\right) = F_{W_2}(w_2) = H_2(w_2)$$

for all w_2 and v_1. But as the right-hand-side does not depend on v_1, this imposes restrictions that only hold for some choices of H_1, H_2 and α_{21}. If we only know H_2, then

$$w_2 + \alpha_{21} H_1^{-1}(v_1) = F_{V_2|V_1}^{-1}\left(H_2(w_2)|v_1\right)$$

overidentifies the product $\alpha_{21} H_1^{-1}(v_1)$; and if we also know H_1, then it overidentifies α_{21}. These results extend directly to higher-dimensional systems.

References

Heckman, J., and R. Pinto (2018): "Unordered Monotonicity," *Econometrica*, 86(1), 1–35.

Heckman, J. J., S. Urzua, and E. Vytlacil (2008): "Instrumental variables in models with multiple outcomes: The general unordered case," *Annales d'économie et de statistique*, 91/92, 151–174.

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

——— (2007): "Heterogeneous choice," in *Advances in economics and econometrics: theory and applications*, ed. by R. Blundell, W. Newey, and T. Persson, vol. 2, chap. 4, pp. 75–110. Cambridge University Press.