ECONOMETRICA

JOURNAL OF THE ECONOMETRIC SOCIETY

An International Society for the Advancement of Economic Theory in its Relation to Statistics and Mathematics

http://www.econometricsociety.org/

Econometrica, Vol. 83, No. 3 (May, 2015), 1185–1197

IDENTIFICATION OF NONSEPARABLE MODELS USING INSTRUMENTS WITH SMALL SUPPORT

ALEXANDER TORGOVITSKY
Northwestern University, Evanston, IL 60208, U.S.A.

Econometrica, Vol. 83, No. 3 (May, 2015), 1185–1197

IDENTIFICATION OF NONSEPARABLE MODELS USING INSTRUMENTS WITH SMALL SUPPORT

BY ALEXANDER TORGOVITSKY[1]

I consider nonparametric identification of nonseparable instrumental variables models with continuous endogenous variables. If both the outcome and first stage equations are strictly increasing in a scalar unobservable, then many kinds of continuous, discrete, and even binary instruments can be used to point-identify the levels of the outcome equation. This contrasts sharply with related work by Imbens and Newey (2009) that requires continuous instruments with large support. One implication is that assumptions about the dimension of heterogeneity can provide nonparametric point-identification of the distribution of treatment response for a continuous treatment in a randomized controlled experiment with partial compliance.

KEYWORDS: Nonseparable models, endogeneity, unobserved heterogeneity, quantile treatment effects, nonparametric identification, instrumental variables.

1. INTRODUCTION AND MODEL

SUPPOSE THAT A SCALAR RESPONSE VARIABLE Y is determined as

$$Y = g^*(X, W, \varepsilon), \tag{1}$$

where X is a d_x vector of continuous explanatory variables (treatments), W are observed covariates, ε is a scalar unobservable, and g^* is an unknown function. This specification is nonseparable (not additively separable) in the latent term, ε, which allows it to capture unobserved heterogeneity in the effect of X on Y. Economic theory and empirical evidence strongly suggest that such heterogeneous treatment effects are a pervasive feature of economic data (Heckman (2001); Imbens (2007)).

A common concern is that X may be endogenous, that is, statistically dependent with ε, even conditional on W. For example, if Y is a schooling outcome for a school or an individual student, X is a measure of class size, and W are observable socioeconomic variables, then family sorting on latent preferences

[1]This paper is based on a portion of my job market paper, which was circulated under the title "Identification and Estimation of Nonparametric Quantile Regressions With Endogeneity" and dated November 12, 2010. I thank D. Andrews, X. Chen, and E. Vytlacil, who were gracious with their advice, support, and feedback. I have benefited from discussions with J. Altonji, L. Benkard, S. Berry, K. Evdokimov, J. Hahn, P. Haile, J. Heckman, K. Hirano, B. Honoré, G. Imbens, Y. Kitamura, S. Lee, O. Linton, G. Ridder, J.-M. Robin, and E. Tamer. This work was presented at the 2010 Cowles Foundation Conference, the 2010 World Congress of the Econometric Society, the 2010 New York Area Econometrics Colloquium, the 2011 Review of Economic Studies May Meetings at LBS, IIES, and CEMFI, and at seminars at Yale, Chicago, Northwestern, Columbia, Michigan, UCLA, USC, UT Austin, Penn State, Brown, Stanford, UC Berkeley, Princeton, UCSD, Harvard/MIT, NYU, IUPUI, Chicago Booth, and Ohio State. My thanks to those audiences.

 DOI: 10.3982/ECTA9984

may lead to $\varepsilon \not\perp\!\!\!\perp X|W$. I consider nonparametric identification of g^* in the presence of such endogeneity. Under a common normalization discussed subsequently, g^* determines the distribution of the counterfactual random variable $Y_x \equiv g^*(x, w, \varepsilon)$, conditional on $W = w$. This counterfactual distribution is of interest because it describes the impact on Y of exogenously setting X to x for the population subgroup determined by $W = w$. (For ease of notation, I suppress W throughout the identification analysis. All assumptions and results can be understood as conditional on W.)

A popular strategy for addressing endogeneity is to utilize the variation of an observable instrument Z that is excluded from (1). The existing nonparametric point-identification results for nonseparable models like (1) with X continuously distributed require Z to also be continuously distributed.[2] Yet discrete instruments are widely used in practice. Many instruments employed in empirical work are only binary, such as the intent-to-treat in a randomized controlled experiment with partial compliance.

This paper shows that instruments with small support, that is, instruments that are binary, discrete, or continuous but without large support, can in fact be used to point-identify g^* under two commonly used assumptions about the dimension of unobserved heterogeneity. The first assumption (called G.S below) is that ε is scalar and that g^* is strictly monotone in ε. The second assumption (called FS below) is the existence of a first stage equation that is strictly monotone in a scalar unobservable that (jointly with ε) is independent of Z. Formally, the assumptions are as follows.

ASSUMPTION C: *The random variables $Y|X = x$, $Z = z$ and $X|Z = z$ are (absolutely) continuously distributed for all x and z.*

ASSUMPTION G: *Let $\mathcal{G}$ denote the collection of admissible outcome functions. For simplicity, every $g \in \mathcal{G}$ is assumed to be defined everywhere on $\mathbb{R}^{d_x+1}$. Assume that $g^* \in \mathcal{G}$ and that the following statements hold*:

G.C (Continuity). *Each $g \in \mathcal{G}$ is everywhere continuous.*

G.S (Scalar outcome heterogeneity). *The function $g(x, \cdot)$ is strictly increasing for every x and every $g \in \mathcal{G}$.*

G.N (Normalization). *If $g, g' \in \mathcal{G}$ are distinct on $\operatorname{supp}(X, \varepsilon)$, then there does not exist a strictly increasing function ψ such that $g(x, e) = g'(x, \psi(e))$ for all $(x, e) \in \operatorname{supp}(X, \varepsilon)$.*[3]

[2]These results include those found in Chesher (2003), Chernozhukov and Hansen (2005), Florens, Heckman, Meghir, and Vytlacil (2008), and Imbens and Newey (2009). A partial exception to this statement is Chesher (2007), who allows for Z to be discrete but point-identifies $g^*(x', e) - g^*(x'', e)$ at values x', x'', and e that depend on the distribution of (X, ε, Z).

[3]The support of a random vector X, denoted $\operatorname{supp}(X)$, is defined as the smallest closed set $\mathcal{S}$ such that $\mathbb{P}[X \in \mathcal{S}] = 1$.

ASSUMPTION FS: *There exists an unobserved* d_x *vector* $\eta \equiv (\eta_1, \dots, \eta_{d_x})$ *and functions* h_k *such that* $X_k = h_k(Z, \eta_k)$ *for each* $k = 1, \dots, d_x$ *and such that the following statements hold*:

FS.E (Exogenous instrument). *The instrument is independent of the latent variables*: $Z \perp\!\!\!\perp (\eta, \varepsilon)$.

FS.S (Scalar first stage heterogeneity). *The function* $h_k(z, \cdot)$ *is strictly increasing for every* z *and* k.

Assumption C requires both Y and X to be continuously distributed.

The scalar heterogeneity assumptions (G.S and FS) are the key restrictions in this model.[4] Assumption G.S imposes *rank invariance* on Y with respect to X, meaning that $F_{Y_x}(Y_x) = F_\varepsilon(\varepsilon) = F_{Y_{x'}}(Y_{x'})$ for any x, x'.[5] Rank invariance can be interpreted as positing an underlying proneness or ranking of units for Y that is not affected by counterfactual manipulations of X. For example, if rank invariance holds, then conditional on observed covariates, a relatively high performing school (Y_x) with a small class size ($X = x$) would also be relatively high performing ($Y_{x'}$) if it had a large class size ($X = x'$). The scalar first stage heterogeneity assumption, FS, similarly imposes rank invariance in the effect of Z on X.[6]

Matzkin (2003) showed that G.N is a necessary condition for point-identification of g^* under G.S and provided sufficient conditions for G.N. These conditions can be viewed as normalizations with the effect of fixing the units of g^* and ε, thereby endowing g^* with a concrete interpretation. One easily interpretable normalization is that $\varepsilon^g \equiv g^{-1}(X, Y) \sim \text{Unif}[0, 1]$ for every $g \in \mathcal{G}$, where $g^{-1}(x, \cdot)$ denotes the inverse of g with respect to its last component. In this case, $Q_{Y_x}(e) = g^*(x, e)$ is the eth quantile of the counterfactual distribution of Y_x. The eth quantile treatment effect of an exogenous (or causal) shift from x to x' is then given by $Q_{Y_{x'}}(e) - Q_{Y_x}(e) = g^*(x', e) - g^*(x, e)$.[7] See Matzkin (2003) for other normalizations. The results in this paper hold for any normalization that implies G.N.

[4]Note that FS.S by itself is without loss of generality given Assumption C. (Take $h_k(z, \cdot)$ to be the conditional-on-[$Z = z$] quantile function of X_k.) The restrictiveness of FS.S comes from simultaneously maintaining FS.E. Hence, I refer to the entirety of FS as the scalar heterogeneity assumption for the first stage.

[5]Throughout the paper, I use the notation F_A and $F_{A|B}(\cdot|b)$ for the unconditional and conditional-on-[$B = b$] distribution functions of a random variable A. Similarly, Q_A and $Q_{A|B}(\cdot|b)$ denote unconditional and conditional quantile functions.

[6]Rank invariance was introduced by Doksum (1974) and was revisited more recently by Heckman, Smith, and Clements (1997) and Chernozhukov and Hansen (2005). Chernozhukov and Hansen (2005) introduced a slightly weaker alternative to rank invariance that they call rank similarity. Rank similarity allows the ranks to deviate from a common ranking as long as the deviations are exogenous. Assumption G.S can be replaced by a rank similarity condition without affecting the results of this paper.

[7]Quantile treatment effects have attracted considerable attention among both theoretical and applied researchers interested in distributional effects. See, for example, Abadie, Angrist, and Imbens (2002), Bitler, Gelbach, and Hoynes (2006), Firpo (2007), and Djebbari and Smith (2008).

Imbens and Newey (2009) considered identification of the same model (with $d_x = 1$), except they did not impose G.S, instead allowing ε to be a vector of arbitrary dimension.[8] Both G.S and FS are arguably strong assumptions. As emphasized by Imbens (2007) and Hoderlein and Mammen (2007, 2009), many ideal structural relationships in economics cannot be characterized as depending monotonically on a single latent term. The main point of this paper is that G.S and FS together also have tremendous identifying content, and, in particular, allow one to dispense with the undesirable large support assumption in Imbens and Newey (2009).[9] The extreme and counterintuitive manifestation of this point is that G.S and FS combined enable nonparametric point-identification of the infinite-dimensional function g^* under relatively weak conditions even when $Z \in \{0, 1\}$ is binary.

2. THE IDENTIFIED SET

Assumption FS implies that $(\varepsilon, V) \perp\!\!\!\perp Z$, where $V_k \equiv F_{X_k|Z}(X_k|Z)$ is the *conditional rank* of X_k and $V \equiv (V_1, \ldots, V_{d_x})$ is the vector of these ranks.[10] This implication characterizes the identified set. A misspecification test can be based on the nonemptyness of this set.

THEOREM 1: *The identified set is* $\mathcal{G}^* \equiv \{g \in \mathcal{G} : (g^{-1}(X, Y), V) \perp\!\!\!\perp Z\}$.

Since X, Y, Z, and V are all features of the observed data, Theorem 1 enables one to determine whether a given $g \in \mathcal{G}$ is in the identified set. In Torgovitsky (2013), I use Theorem 1 to construct an estimator of g^* under the point-identifying assumptions discussed subsequently.

3. SUFFICIENT CONDITIONS FOR POINT IDENTIFICATION

In this section, I show that g^* is point-identified on $\text{supp}(X, \varepsilon)$ under natural assumptions about the strength of the dependence between X and Z. These assumptions take different forms, depending on whether Z has a continuous or discrete distribution. The results for the continuous and discrete cases use the same preliminary arguments, but differ at a crucial stage. Since the result when Z is continuous is less remarkable, it is discussed in the Supplemental Material (Torgovitsky (2015)). Here, I discuss the preliminary arguments common to both results and then discuss the sequencing argument used in the case where

[8]As a consequence, they do not require G.N and their model also covers cases where Y is discrete.

[9]A similar point is made by Florens et al. (2008), who show that large support can be weakened to "measurable separability" if g^* has a particular polynomial structure in unobservables. Their analysis still requires Z to be continuously distributed.

[10]As observed by Imbens and Newey (2009) (their Theorem 1), Assumptions C and FS imply that $V_k = F_{\eta_k}(\eta_k)$ and that $(\varepsilon, F_{\eta_1}(\eta_1), \ldots, F_{\eta_{d_x}}(\eta_{d_x})) \perp\!\!\!\perp Z$, that is, $(\varepsilon, V) \perp\!\!\!\perp Z$.

Z is discrete. This analysis is most straightforward when X is scalar ($d_x = 1$), so I maintain that assumption for the remainder of the main text. Results for $d_x > 1$ are discussed in the Supplemental Material.

The identification analysis is based on the implication of Assumption FS that $\varepsilon \perp\!\!\!\perp Z|V$ and the implication of Assumption C that for $x \in \mathcal{X}_z^\circ \equiv \operatorname{int}\operatorname{supp} X|Z = z$, the event $[X = x, Z = z]$ is equivalent to the event $[V = F_{X|Z}(x|z), Z = z]$.[11] Together, these imply that if

$$F_{X|Z}(x^a|z^a) = \overline{v} = F_{X|Z}(x^b|z^b) \quad \text{for} \quad x^a \in \mathcal{X}_{z^a}^\circ \text{ and } x^b \in \mathcal{X}_{z^b}^\circ,$$

then the distributions of $\varepsilon|X = x^a, Z = z^a$ and $\varepsilon|X = x^b, Z = z^b$ are the same and equal to that of $\varepsilon|V = \overline{v}$. As a result, any differences between the observed distributions of $Y|X = x^a, Z = z^a$ and $Y|X = x^b, Z = z^b$ should be solely due to the direct effect that g^* has on Y when X is shifted from x^a to x^b.

This direct effect can be isolated with the aid of G.S, which allows realizations of ε to be expressed uniquely as realizations of Y.[12] Specifically, if $\overline{y} \in \mathcal{Y}_{x^a,z^a}^\circ \equiv \operatorname{int}\operatorname{supp}(Y|X = x^a, Z = z^a)$, then by G.S there exists a unique $\overline{e}$ such that $\overline{y} = g^*(x^a, \overline{e})$. The value of $g^*(x^b, \overline{e})$ can be recovered by inverting the relationship

$$\begin{aligned} F_{Y|XZ}(\overline{y}|x^a, z^a) &= \mathbb{P}[\varepsilon \le \overline{e}|X = x^a, Z = z^a] \\ &= \mathbb{P}[\varepsilon \le \overline{e}|X = x^b, Z = z^b] = F_{Y|XZ}(g^*(x^b, \overline{e})|x^b, z^b). \end{aligned}$$

If $X \perp\!\!\!\perp Z$, then necessarily $x^a = x^b$ and the preceding expression has no identifying content. As in other instrumental variable models, dependence between X and Z is required to generate exogenous variation in X.

The preceding intuition is incomplete for two reasons. First, to establish point-identification of g^* as a function, it needs to be shown that $g^*(x^a, \overline{e})$ can be exogenously compared to $g^*(x, \overline{e})$ for any other supported x, not just x^b. This is demonstrated in the Supplemental Material using differential arguments when Z is continuously distributed. It will be shown in the main text using a sequencing argument when Z is binary or more generally discrete. Second, the value of $\overline{e}$, which was defined implicitly in terms of g^*, is still unknown. As a result, the intuition characterizes the effect of exogenously shifting X to x^b for an agent who endogenously attained $Y = \overline{y}$ and $X = x^a$. That is, it point-identifies $g^*(x^b, (g^*)^{-1}(x^a, \overline{y}))$. To point-identify $g^*(x^b, e)$ for a particular e of interest requires a more involved argument.

[11]Calligraphic capital letters $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{Z}$, and $\mathcal{E}$ are used to denote supports in the remainder of the paper. Subscripts denote conditional supports (where the conditioning will be obvious) and open circles denote interiors.

[12]This part of the argument is similar to arguments used in Altonji and Matzkin (2005) and Athey and Imbens (2006).

To fill in these gaps and formalize the intuition, consider $I^g(x, e) \equiv g^{-1}(x, g^*(x, e))$ as a measure of the difference between g^* and any $g \in \mathcal{G}$. By construction, $I^g(x, e) = e$ if and only if $g(x, e) = g^*(x, e)$. In addition, since $I^g(x, \cdot)$ is strictly increasing, the normalization G.N implies that if $I^g(x, e)$ is not a function of x for $(x, e) \in \operatorname{supp}(X, \varepsilon)$, say $I^g(x, e) = J^g(e)$, then $g(x, e) = g^*(x, e)$ on $\operatorname{supp}(X, \varepsilon)$. Hence, point-identification can be established by showing that if $g \in \mathcal{G}^*$, then $I^g(x, e) = J^g(e)$ for all such (x, e).[13]

With this goal in mind, suppose that $g \in \mathcal{G}^*$ and recall the definition $\varepsilon^g \equiv g^{-1}(X, Y)$. By G.S,

$$\tag{2} \begin{aligned} &Q_{\varepsilon^g|XZ}(t|x, z) = g^{-1}\bigl(x, Q_{Y|XZ}(t|x, z)\bigr) \quad \text{and} \\ &Q_{Y|XZ}(t|x, z) = g^*\bigl(x, Q_{\varepsilon|XZ}(t|x, z)\bigr). \end{aligned}$$

Combining the two expressions in (2) gives

$$\tag{3} I^g\bigl(x, Q_{\varepsilon|XZ}(t|x, z)\bigr) = Q_{\varepsilon^g|XZ}(t|x, z).$$

Substituting $t = F_{\varepsilon|XZ}(e|x, z)$ for any $e \in \mathcal{E}^\circ_{x,z} \equiv \operatorname{int}\operatorname{supp}(\varepsilon|X = x, Z = z)$ into (3) yields

$$\tag{4} I^g(x, e) = Q_{\varepsilon^g|XZ}\bigl(F_{\varepsilon|XZ}(t|x, z)|x, z\bigr).$$

Since $\varepsilon \perp\!\!\!\perp Z|V$ and $\varepsilon^g \perp\!\!\!\perp Z|V$ for $g \in \mathcal{G}^*$ by Theorem 1, and because the event $[X = x, Z = z]$ is equivalent to the event $[V = F_{X|Z}(x|z), Z = z]$ for $x \in \mathcal{X}^\circ_z$, (4) can be rewritten with $D^g(v, e) \equiv Q_{\varepsilon^g|V}(F_{\varepsilon|V}(e|v)|v)$ as

$$\tag{5} I^g(x, e) = D^g\bigl(F_{X|Z}(x|z), e\bigr) \quad \text{for} \quad x \in \mathcal{X}^\circ_z, e \in \mathcal{E}^\circ_{x,z}.$$

From (5), it follows that

$$F_{X|Z}\bigl(x^a|z^a\bigr) = F_{X|Z}\bigl(x^b|z^b\bigr) \quad \text{implies} \quad I^g\bigl(x^a, e\bigr) = I^g\bigl(x^b, e\bigr)$$

for two supported points (x^a, z^a) and (x^b, z^b), and all $e \in \mathcal{E}^\circ_{x^a,z^a}$. This observation is vacuous if $x^a = x^b$ for all choices of z^a and z^b, which would be the case, for example, if $X \perp\!\!\!\perp Z$. However, if such distinct pairs exist, then (5) establishes the constancy of I^g as a function of x on $\{x^a, x^b\}$. If Z is continuously distributed and has a nonzero effect on the conditional distribution of X at any x^a, then such an x^b can be found arbitrarily close to x^a, enabling a straightforward proof of the constancy of $I^g(x, e)$ as a function of x and, hence, of point-identification. This is shown in the Supplemental Material.

If Z is discretely distributed (with finite support), then any distinct points x^a and x^b for which $F_{X|Z}(x^a|z^a) = F_{X|Z}(x^b|z^b)$ will necessarily be far apart. Nevertheless, under certain conditions, it is still possible to show that I^g is constant

[13]In fact, since every $g \in \mathcal{G}^*$ is continuous, it suffices to show this for the interior of the support of (X, ε).

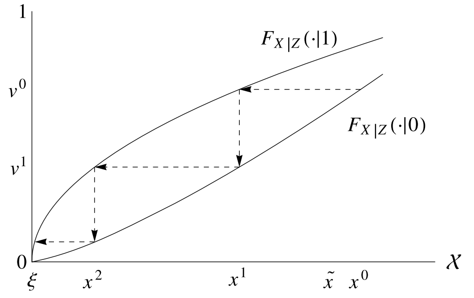

FIGURE 1.—The point-identification argument when $Z \in \{0, 1\}$ is binary with $\mathcal{X}_z = [\xi, \infty)$ and $F_{X|Z}(x|1) > F_{X|Z}(x|0)$ for all $x > \xi > -\infty$.

as a function of x through a sequencing argument. For example, suppose for the sake of exposition that $\mathcal{X} = \mathcal{X}_z = [\xi, \infty)$ for $z \in \mathcal{Z} = \{0, 1\}$ with $\xi > -\infty$ and that $F_{X|Z}(x|1) > F_{X|Z}(x|0)$ for all $x > \xi$. This configuration is depicted in Figure 1. Also assume that $\mathcal{E}_{x,z} = \mathcal{E}$ for all x and z. Consider the mapping

$$\pi : (\xi, \infty) \to (\xi, \infty) : \pi(x) \equiv Q_{X|Z}\big(F_{X|Z}(x|0)|1\big),$$

which satisfies $F_{X|Z}(\pi(x)|1) = F_{X|Z}(x|0)$ and, hence, $I^g(\pi(x), e) = I^g(x, e)$ for any $e \in \mathcal{E}^\circ$ by (5). Pick an initial point $x^0 > \xi$ and form a recursive sequence $x^{n+1} = \pi(x^n)$ for $n \geq 0$. The sequence is decreasing because

$$\pi(x) \equiv Q_{X|Z}\big(F_{X|Z}(x|0)|1\big) < Q_{X|Z}\big(F_{X|Z}(x|1)|1\big) = x$$

for $x > \xi$. Since $\xi > -\infty$, the sequence therefore converges to a limiting point. As X is continuously distributed, this limiting point must satisfy $F_{X|Z}(\lim x^n|1) = \lim F_{X|Z}(x^{n+1}|1) = \lim F_{X|Z}(x^n|0) = F_{X|Z}(\lim x^n|0)$. The only point that satisfies this property (see Figure 1) is ξ, so it must be the case that $x^n \to \xi$. Since $I^g(x, e)$ is a continuous function by Assumption G and because it remains invariant when switching x to $\pi(x)$, it follows that

$$I^g\big(x^0, e\big) = I^g\big(x^1, e\big) = \cdots = I^g\big(x^n, e\big) = I^g(\xi, e).$$

Noting that x^0 was an arbitrary point, this argument shows that $I^g(x, e) = I^g(\xi, e) = J^g(e)$ for any $x > \xi$ and $e \in \mathcal{E}^\circ$. As discussed, this implies that g^* is point-identified.[14]

Figure 1 illustrates the main features of this argument. The horizontal arrow between (x^0, v^0) and (x^1, v^0) represents the equality $I^g(x, e) = I^g(\pi(x), e)$

[14]The preceding argument establishes point-identification implicitly (vs. constructively) because it shows that under the given conditions, there can be only one $g \in \mathcal{G}$ that satisfies $(g^{-1}(X, Y), V) \perp\!\!\!\perp Z$. There is also a constructive proof based on iterating (2) along the sequence defined by π. This constructive proof is harder to generalize and does not appear to facilitate estimation, so it is not included here. The details are available from the author on request.

that is implied by (5). The equality represents the implication of FS that observations with $(X, Z) = (x^0, 0)$ and $(X, Z) = (x^1, 1)$ have the same unobservable distribution $\varepsilon|V = v^0$, even though they have different values of X. The vertical arrow between (x^1, v^0) and (x^1, v^1) represents the shift in V induced by shifting Z from 1 back to 0 while holding $X = x^1$. This has no effect on $I^g(x^1, e)$ because Z has no effect on g^*, due to the exclusion restriction in (1). Horizontal and vertical shifts are repeated until the limiting point ξ is reached. Intuitively, this procedure establishes an exogenous comparison between treatment levels x^0 and ξ, because each horizontal shift moves X while keeping the conditional distribution of ε fixed. An exogenous comparison between treatment level x^0 and any other treatment level $\widetilde{x}$ results from their mutual comparisons with treatment level ξ.

For an intuitive explanation of the source of point-identification, consider an experiment that randomly assigns students across various schools to a large or small class ($Z = 0$ or 1, respectively).[15] The definition of a large or small class may vary across schools and/or grades so that the distribution of X is roughly continuous. Partial compliance can arise if some students randomly assigned to large classes end up attending small classes, for example, because of parental interference. If the degree of parental interference is correlated with the outcome Y (e.g., test scores), then X will be dependent with ε. However, the randomly assigned intent-to-treat $Z \in \{0, 1\}$ will still serve as a valid and relevant instrument for X.

Suppose that the distribution of class size X conditional on intent-to-treat status Z is given by Figure 1. The model assumptions imply that students with class size $X = x^0$ in the group assigned to large classes ($Z = 0$) are unobservably identical to students with class size $X = x^1$ in the group assigned to small classes ($Z = 1$). The connection between these two types of students is that they have the same rank in their respective X distributions, that is, $F_{X|Z}(x^0|0) = v^0 = F_{X|Z}(x^1|1)$. This rank can be interpreted as the latent propensity (due to parental interference, etc.) of the students for being in a large-sized class. Since the unobservable characteristics of the two groups are identical and because Z has no direct effect on Y, the differences in test scores between these two groups of students must be due solely (that is, causally) to the change in X from x^0 to x^1. The class-size production functions are traced out at all levels by continuing this sequence of binary comparisons from x^1 to x^2 and so on, and by varying the initial point of reference, x^0.

Essentially the same point-identification argument can be used under much more general conditions than those discussed in the preceding text and displayed visually in Figure 1. In particular, $\mathcal{X}$ need not be bounded, $\mathcal{Z}$ need not be binary, and the conditional distribution functions $\{F_{X|Z}(\cdot|z)\}_{z\in\mathcal{Z}}$ can cross up to a certain extent. The following result proves that g^* is point-identified

[15] A well known example of such an experiment is Project STAR; see, for example, Krueger (1999).

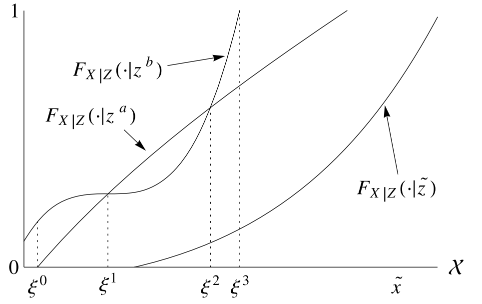

FIGURE 2.—An example of dividing $\mathcal{X}$ into smaller intervals in Theorem 2. Here $\mathcal{Z}$ has three points of support and $\mathcal{X}^{\dagger}(z^a, z^b) = \{\xi^1, \xi^2\}$.

under these more general conditions by effectively dividing the support of X into smaller pieces and then repeating the preceding argument on each piece. An example of the construction used in the formal proof is illustrated in Figure 2.

THEOREM 2: *Suppose that $\mathcal{Z}$ is finite, that $d_x = 1$, and that $\mathcal{X}_z \equiv \operatorname{supp}(X|Z = z)$ is an interval for every $z \in \mathcal{Z}$. Also assume that $\mathcal{E}_{x,z} \equiv \operatorname{supp}(\varepsilon|X = x, Z = z)$ does not depend on x or z. Define $\mathcal{X}^{\dagger}(z^a, z^b) \equiv \{x \in \mathcal{X}_{z^a} \cap \mathcal{X}_{z^b} : F_{X|Z}(x|z^a) = F_{X|Z}(x|z^b)\}$. Then g^* is point-identified on $\operatorname{supp}(X, \varepsilon)$ if there exist z^a, z^b such that $\mathcal{X}^{\dagger}(z^a, z^b)$ is nonempty and finite.*

Theorem 2 requires the existence of instrument values z^a and z^b for which the distribution functions $F_{X|Z}(\cdot|z^a)$ and $F_{X|Z}(\cdot|z^b)$ cross at least once—but do not overlap—on the intersection of their supports.[16] This assumption is important and potentially restrictive. It rules out a linear relationship such as $X = \pi Z + V$ with $\mathcal{Z} = \{0, 1\}$, $\pi \neq 0$, and $\operatorname{supp} V = \mathbb{R}$, because in this case, $F_{X|Z}(x|1) = F_{X|Z}(x - \pi|0)$, so $F_{X|Z}(x|1)$ and $F_{X|Z}(x|0)$ are never equal and, hence, $\mathcal{X}^{\dagger}(0, 1)$ is empty. On the other hand, $\mathcal{X}^{\dagger}(z^a, z^b)$ will be nonempty if $\mathcal{X}_{z^a}$ and $\mathcal{X}_{z^b}$ share a lower bound (e.g., 0), which is common in economic applications.

Imbens and Newey (2009) considered the same model as in this paper, except they did not maintain G.S, instead allowing ε to be of arbitrary dimension. To obtain point-identification of the counterfactual quantity $Q_{Y_x}(t)$ (which is equal to $g^*(x, Q_\varepsilon(t))$ under G.S), they require $F_{X|Z}(x|Z)|X = x$ (that is,

[16]As shown in previous versions of this paper, these conditions can be generalized somewhat further to allow for the possibility that $F_{X|Z}(\cdot|z^a)$ and $F_{X|Z}(\cdot|z^b)$ overlap on certain nonsingleton subsets of $\mathcal{X}_{z^a} \cap \mathcal{X}_{z^b}$, as long as they do not overlap on any given nonsingleton subset of $\mathcal{X}$ for all $z^a, z^b \in \mathcal{Z}$. These generalizations complicate the proof of Theorem 1 considerably and seem to add little empirical relevance.

$V|X=x$) to have support $[0,1]$.[17] Intuitively, this type of "large support" condition requires that for a given x of interest, there exist observed instrument values z^a and z^b such that any agent with z^a would never attain anything smaller than x, while any agent with z^b would never attain anything larger than x. In general, this requires Z to be continuous with support $\mathbb{R}$. This is a strong condition that is unlikely to be satisfied by the types of instruments used in practice. In contrast, the conditions given in Theorem 2 (and in the related results in the Supplemental Material) only require the instrument to have a nonzero effect on the probability that an agent attains x, for almost every x. This is a requirement that can be met by instruments that are binary, discrete, or continuous but without large support. Since the important difference between the current model and that of Imbens and Newey (2009) is G.S and the dimension of ε, the implication is that this assumption about heterogeneity contains a great deal of identifying content.

D'Haultfœuille and Février (2015) prove a result similar to Theorem 2 using group theory.[18] Their approach requires several additional assumptions, including the empirically relevant (and typically false) restriction that the support of (X, Z) be rectangular. The benefit of their approach is the ability to apply theorems due to Hölder and Denjoy that can be used to establish point-identification in situations where $\mathcal{X}^\dagger(z^a, z^b)$ is empty for all z^a and z^b under an additional high-level "nonperiodicity" condition. Their results are weaker than Theorem 2 for cases in which $\mathcal{X}^\dagger(z^a, z^b)$ is nonempty and finite for some z^a and z^b.

The Supplemental Material contains an extension of Theorem 2 for $d_x > 1$ under stronger conditions. It also contains two examples that discuss the restrictiveness of the identifying assumptions for the problem of determining the causal effect of class size on schooling outcomes.

4. CONCLUSION

This paper has shown that g^* in the nonseparable specification (1) can be nonparametrically point-identified using the exogenous variation generated by an instrument with small support. The instrument can be continuous, but, more surprisingly, it can also be discrete or even binary. The key to the result is the imposition of rank invariance on both the outcome and the first stage equations. Since other point-identification results for nonseparable models with continuous endogenous variables require continuous instruments (sometime with large support), the implication is that these assumptions about the dimension of heterogeneity have a tremendous amount of identifying content.

[17]The authors also provide sharp partial identification results when Z does not satisfy this condition.

[18]Their $s_{ij}(x)$ is essentially π from the special case of Theorem 2. A more complete discussion of D'Haultfœuille and Février (2015) can be found on my website.

A practical consequence is that assumptions about the dimension of heterogeneity can provide nonparametric point-identification of the distribution of treatment response for a continuous treatment in a randomized controlled experiment with partial compliance.

APPENDIX: PROOFS

PROOF OF THEOREM 1: If $g \in \mathcal{G}$ is in the identified set, then $Y = g(X, \varepsilon^g)$ for some ε^g with $(\varepsilon^g, \eta) \perp\!\!\!\perp Z$, and, hence, $(g^{-1}(X, Y), V) \perp\!\!\!\perp Z$ since $\varepsilon^g = g^{-1}(X, Y)$ and $V_k = F_{\eta_k}(\eta_k)$ for all k (see footnote 10). Conversely, if $g \in \mathcal{G}^*$, then $(\varepsilon^g, \eta) \perp\!\!\!\perp Z$ and $Y = g(X, \varepsilon^g)$ for $\varepsilon^g \equiv g^{-1}(X, Y)$, so that g is in the identified set. *Q.E.D.*

PROOF OF THEOREM 2: Suppose that z^a and z^b are such that $\mathcal{X}^\dagger(z^a, z^b)$ is nonempty and finite. Then $\mathcal{X}_{z^a} \cap \mathcal{X}_{z^b}$ is a closed interval with a nonempty interior, since by definition it contains $\mathcal{X}^\dagger(z^a, z^b)$, and by assumption both $\mathcal{X}_{z^a}$ and $\mathcal{X}_{z^b}$ are closed intervals. Let $\{\xi^l\}_{l=0}^L$ denote the unique elements of

$$\mathcal{X}^\dagger\big(z^a, z^b\big) \cup \{\inf \mathcal{X}_{z^a} \cap \mathcal{X}_{z^b}, \sup \mathcal{X}_{z^a} \cap \mathcal{X}_{z^b}\}$$

arranged in increasing order, where the infimum may be $-\infty$ and the supremum may be $+\infty$. Then $\xi^0 = \inf \mathcal{X}_{z^a} \cap \mathcal{X}_{z^b}$, $\xi^L = \sup \mathcal{X}_{z^a} \cap \mathcal{X}_{z^b}$, and $\mathrm{cl} \bigcup_{l=1}^L (\xi^{l-1}, \xi^l) = \mathcal{X}_{z^a} \cap \mathcal{X}_{z^b}$, where cl denotes the closure of a set in the real numbers. See Figure 2 for an example of these definitions.

Consider the set (ξ^0, ξ^1), which must have either $\xi^0 \in \mathcal{X}^\dagger(z^a, z^b)$ or $\xi^1 \in \mathcal{X}^\dagger(z^a, z^b)$, or both, since $\{\xi^l\}_{l=0}^L$ has at least one element from $\mathcal{X}^\dagger(z^a, z^b)$ and $L \geq 1$. The first goal is to show that $I^g(x, e)$ is constant as a function of x over (ξ^0, ξ^1). There are four cases to consider, depending on whether ξ^0 or ξ^1 is in $\mathcal{X}^\dagger(z^a, z^b)$ and whether $F_{X|Z}(\cdot|z^b) - F_{X|Z}(\cdot|z^a)$ is strictly positive or strictly negative over (ξ^0, ξ^1).[19] The following proof is for the case $\xi^1 \in \mathcal{X}^\dagger(z^a, z^b)$ and $F_{X|Z}(x|z^b) > F_{X|Z}(x|z^a)$ for all $x \in (\xi^0, \xi^1)$, as depicted in Figure 2. The other cases follow symmetrically.

Define $\pi(x) \equiv Q_{X|Z}(F_{X|Z}(x|z^b)|z^a)$ and for any $x^0 \in (\xi^0, \xi^1)$, define the recursive sequence $x^{n+1} = \pi(x^n)$ for $n \geq 0$. Then $x^n < \xi^1$ for all n. For otherwise, there would exist an N such that $x^N < \xi^1$ and $x^{N+1} \geq \xi^1$. However, since $\xi^1 \in \mathcal{X}^\dagger(z^a, z^b)$, this would imply

$$F_{X|Z}\big(\xi^1|z^b\big) = F_{X|Z}\big(\xi^1|z^a\big) \leq F_{X|Z}\big(x^{N+1}|z^a\big) = F_{X|Z}\big(x^N|z^b\big),$$

which contradicts the strict monotonicity of $F_{X|Z}(\cdot|z^b)$ on $(\xi^0, \xi^1] \subseteq \mathcal{X}_{z^b}$ implied by Assumption C. In addition, x^n is increasing in n because $\pi(x) \equiv$

[19] The quantity $F_{X|Z}(\cdot|z^b) - F_{X|Z}(\cdot|z^a)$ is continuous by Assumption C and nonzero on (ξ^0, ξ^1) by the definition of $\{\xi^l\}_{l=0}^L$.

$Q_{X|Z}(F_{X|Z}(x|z^b)|z^a) \geq Q_{X|Z}(F_{X|Z}(x|z^a)|z^a) = x$ for $x \in (\xi^0, \xi^1) \subseteq \mathcal{X}^\circ_{z^a}$. Hence, $\lim x^n$ exists and, by Assumption C, it satisfies

$$F_{X|Z}\big(\lim x^{n+1}|z^a\big) = \lim F_{X|Z}\big(x^{n+1}|z^a\big)$$
$$= \lim F_{X|Z}\big(x^n|z^b\big) = F_{X|Z}\big(\lim x^n|z^b\big).$$

This can only be true if $\lim x^n = \xi^1$, since $(\xi^0, \xi^1] \cap \mathcal{X}^\dagger(z^a, z^b) = \{\xi^1\}$. Also, note that $x^n \in \mathcal{X}^\circ_{z^a} \cap \mathcal{X}^\circ_{z^b}$ for all n since $x^n \in (\xi^0, \xi^1) \subseteq \mathcal{X}^\circ_{z^a} \cap \mathcal{X}^\circ_{z^b}$.

The definition and properties of x^n with (5) imply that

$$I^g\big(x^{n+1}, e\big) = D^g\big(F_{X|Z}\big(\pi\big(x^n\big)|z^a\big), e\big) = D^g\big(F_{X|Z}\big(x^n|z^b\big), e\big) = I^g\big(x^n, e\big)$$

for all n and any $e \in \mathcal{E}^\circ$, since $\mathcal{E}^\circ_{x,z} = \mathcal{E}^\circ$ for all x, z by assumption. Because I^g is continuous everywhere due to Assumption G,

$$I^g\big(x^0, e\big) = I^g\big(\lim x^n, e\big) = I^g\big(\xi^1, e\big) \equiv J^g(e). \tag{6}$$

Since $x^0 \in (\xi^0, \xi^1)$ was arbitrary, (6) shows that $I^g(x, e) = J^g(e)$ for every $x \in \mathrm{cl}(\xi^0, \xi^1)$ and $e \in \mathcal{E}$, by continuity. The three other cases reach the same conclusion through symmetric arguments.[20] If $L \geq 2$, then the same arguments can be used to show that $I^g(x, e)$ does not vary over $x \in \mathrm{cl}(\xi^1, \xi^2)$. Since $\mathrm{cl}(\xi^0, \xi^1) \cap \mathrm{cl}(\xi^1, \xi^2) = \{\xi^1\}$, it follows that $I^g(x, e) = J^g(e)$ for all $x \in \mathrm{cl}(\xi^0, \xi^2)$. Repeating the argument a finite number of times establishes that $I^g(x, e) = J^g(e)$ for all $x \in \mathrm{cl}(\xi^0, \xi^L)$ and all $e \in \mathcal{E}$.

Finally, consider an $\widetilde{x} \in \mathcal{X}^\circ \setminus \mathrm{cl}(\xi^0, \xi^L)$ (if any exist). Then there exists a $\widetilde{z} \in \mathcal{Z}$ such that $\widetilde{x} \in \mathcal{X}^\circ_{\widetilde{z}}$. By the definition of ξ^0 and ξ^L, there either exists an $x^a \in (\xi^0, \xi^L)$ such that $F_{X|Z}(x^a|z^a) = F_{X|Z}(\widetilde{x}|\widetilde{z})$ or an $x^b \in (\xi^0, \xi^L)$ such that $F_{X|Z}(x^b|z^b) = F_{X|Z}(\widetilde{x}|\widetilde{z})$. In either case, an application of (5) shows that $I^g(\widetilde{x}, e) = J^g(e)$ for all $e \in \mathcal{E}$. Thus, for any $e \in \mathcal{E}$, $I^g(x, e) = J^g(e)$ is not a function of x on $\mathrm{cl}(\xi^0, \xi^L) \cup (\mathcal{X}^\circ \setminus \mathrm{cl}(\xi^0, \xi^L)) = \mathcal{X}^\circ$, hence $\mathcal{X}$, by continuity. As discussed in the main text, this implies that g^* is point-identified on $\mathrm{supp}(X, \varepsilon)$. *Q.E.D.*

REFERENCES

ABADIE, A., J. ANGRIST, AND G. IMBENS (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117. [1187]

ALTONJI, J. G., AND R. L. MATZKIN (2005): "Cross Section and Panel Data Estimators for Nonseparable Models With Endogenous Regressors," *Econometrica*, 73, 1053–1102. [1189]

ATHEY, S., AND G. W. IMBENS (2006): "Identification and Inference in Nonlinear Difference-in-Differences Models," *Econometrica*, 74, 431–497. [1189]

[20]Specifically, switch z^a and z^b in defining π so that x^n tends toward whichever of ξ^0 or ξ^1 is in $\mathcal{X}^\dagger(z^a, z^b)$.

BITLER, M. P., J. B. GELBACH, AND H. W. HOYNES (2006): "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," *American Economic Review*, 96, 988–1012. [1187]

CHERNOZHUKOV, V., AND C. HANSEN (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261. [1186,1187]

CHESHER, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405–1441. [1186]

——— (2007): "Instrumental Values," *Journal of Econometrics*, 139, 15–34. [1186]

D'HAULTFŒUILLE, X., AND P. FÉVRIER (2015): "Identification of Nonseparable Triangular Models With Discrete Instruments," *Econometrica*, 83 (3), 1199–1210. [1194]

DJEBBARI, H., AND J. SMITH (2008): "Heterogeneous Impacts in PROGRESA," *Journal of Econometrics*, 145, 64–80. [1187]

DOKSUM, K. (1974): "Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case," *The Annals of Statistics*, 2, 267–277. [1187]

FIRPO, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75, 259–276. [1187]

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, 1191–1206. [1186,1188]

HECKMAN, J. J. (2001): "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *Journal of Political Economy*, 109, 673–748. [1185]

HECKMAN, J. J., J. SMITH, AND N. CLEMENTS (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64, 487–535. [1187]

HODERLEIN, S., AND E. MAMMEN (2007): "Identification of Marginal Effects in Nonseparable Models Without Monotonicity," *Econometrica*, 75, 1513–1518. [1188]

——— (2009): "Identification and Estimation of Local Average Derivatives in Non-Separable Models Without Monotonicity," *Econometrics Journal*, 12, 1–25. [1188]

IMBENS, G. (2007): "Nonadditive Models With Endogenous Regressors," in *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, Vol. 3, ed. by R. Blundell, W. Newey, and T. Persson. New York: Cambridge University Press. [1185, 1188]

IMBENS, G. W., AND W. K. NEWEY (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512. [1185,1186,1188, 1193,1194]

KRUEGER, A. B. (1999): "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, 114, 497–532. [1192]

MATZKIN, R. L. (2003): "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339–1375. [1187]

TORGOVITSKY, A. (2013): "Minimum Distance From Independence Estimation of Nonseparable Instrumental Variables Models," Working Paper. [1188]

——— (2015): "Supplement to 'Identification of Nonseparable Models Using Instruments With Small Support'," *Econometrica Supplemental Material*, 83, http://dx.doi.org/10.3982/ECTA9984. [1188]

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

Manuscript received April, 2011; final revision received November, 2014.