Supplementary materials for "Testing for causal effects in a generalized regression model with endogenous regressors" (by Jason Abrevaya, Jerry A. Hausman, and Shakeeb Khan)

A Proofs of asymptotic results

In this supplemental document, we outline the asymptotic theory for the three-stage testing procedure. We repeat the statement the main asymptotic-normality results in the paper and also restate the sufficient regularity conditions for these results. The proofs, which are somewhat standard given previous results in this literature, are provided in this Supplement.

The following linear representation of the first-step estimator is assumed:

$$\hat{\delta} - \delta_0 = \frac{1}{n}\sum_{i=1}^{n} \psi_{\delta i} + o_p(n^{-1/2}), \tag{A.1}$$

where $\psi_{\delta i}$ is an influence-function term with zero mean and finite variance. This representation exists for the available $\sqrt{n}$-consistent semiparametric estimators. We do not specify a particular form for the influence-function term $\psi_{\delta i}$ since it will depend upon the particular estimator chosen.

The first result concerns the asymptotic distribution for the second-stage estimator of β_0. Since β_0 is only identified up to scale, we normalize its last component to 1 and denote its other components by θ_0 and the corresponding estimator by $\hat{\theta}$, where

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \frac{1}{n(n-1)} \sum_{i\neq j} 1[y_{2i} = y_{2j}] k_h(z_i'\hat{\delta} - z_j'\hat{\delta}) 1[y_{1i} > y_{1j}] 1[z_{1i}'\beta(\theta) > z_{1j}'\beta(\theta)] \tag{A.2}$$

We impose the following regularity conditions:

Assumption CPS (Parameter space) θ_0 lies in the interior of Θ, a compact subset of $\mathbf{R}^{k-1}$.

Assumption FS The first stage estimator used to estimate δ_0 will be the maximum rank correlation estimator of Han (1987). Consequently, the same regularity conditions in that paper and Sherman (1993) will be assumed so we will have a linear representation as discussed above. We normalize one of the coefficients of δ_0 to 1 and assume the corresponding regressor is continuously distributed on its support.

Assumption K (Matching stages kernel function) The kernel function $k(\cdot)$ used in the second stage and the third stage is assumed to have the following properties:

K.1 $k(\cdot)$ is twice continuously differentiable, has compact support and integrates to 1.

K.2 $k(\cdot)$ is symmetric about 0.

K.3 $k(\cdot)$ is a p^{th} order kernel, where p is an even integer:

$$\begin{aligned}\int u^l k(u)du &= 0 \quad \text{for } l = 1, 2, ...p-1 \\ \int u^p k(u)du &\neq 0\end{aligned}$$

Assumption H (Matching stages bandwidth sequence) The bandwidth sequence h_n used in the second stage and the third stage satisfies $\sqrt{n}h_n^p \to 0$ and $\sqrt{n}h_n^3 \to \infty$.

Assumption RD (Last regressor and index properties) $z_{1i}^{(k)}$ is continuously distributed with positive density on the real line conditional on $z_i'\delta_0$ and all other elements of z_{1i}. Moreover, $z_i'\delta_0$ is nondegenerate conditional on $z_{1i}'\beta_0$.

Assumption ED (Error distribution) (ϵ_i, η_i) is distributed independently of z_i and is continuously distributed with positive density on $\mathbf{R}^2$.

Assumption FR (Full rank condition) Conditional on $(z_i'\delta_0, y_{2i})$, the support of z_{1i} does not lie in a proper linear subspace of R^k.

The following lemma establishes the asymptotic properties of the second stage estimator of θ_0. Some additional notation is used in the statement of the lemma. The reduced-form linear index is denoted $\zeta_{\delta i} = z_i'\delta_0$ and $f_{\zeta\delta}()$ denotes its density function. F_{Z_1} denotes the distribution function of z_{1i}. Also, $\nabla_{\theta\theta}$ denotes the second-derivative operator.

Lemma 1 *If Assumptions CPS, FS, K, H, RD, ED, and FR hold, then*

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow N(0, V^{-1}\Omega V^{-1}) \tag{A.3}$$

or, alternatively, $\hat{\theta} - \theta_0$ *has the linear representation*

$$\hat{\theta} - \theta_0 = \frac{1}{n}\sum_{i=1}^{n} \psi_{\beta i} + o_p(n^{-1/2}) \tag{A.4}$$

with $V = \nabla_{\theta\theta}\mathcal{N}(\theta) \mid_{\theta=\theta_0}$ *and* $\Omega = E[\delta_{1i}\delta_{1i}']$, *and* $\psi_{\beta i} = V^{-1}\delta_{1i}$, *where*

$$\mathcal{N}(\theta) = \int 1[z_{1i}'\beta(\theta) > z_{1j}'\beta(\theta)]\mathcal{H}(\zeta_j, \zeta_j)\mathcal{F}(z_{1i}'\beta_0, z_{1j}'\beta_0, \zeta_j, \zeta_j)dF_{Z_1,\zeta}(z_{1i}, \zeta_j)dF_{Z_1,\zeta}(z_{1j}, \zeta_j) \tag{A.5}$$

with $\zeta_i = z_i'\delta_0$, *whose density function is denoted by* f_ζ, *and where*

$$\mathcal{F}(z_{1i}'\beta_0, z_{1j}'\beta_0, \zeta_i, \zeta_j) = P(y_{1i} > y_{1j} | y_{2i} = y_{2j}, z_{1i}, z_{1j}, \zeta_i, \zeta_j) \tag{A.6}$$

$$\mathcal{H}(\zeta_i, \zeta_j) = P(y_{2i} = y_{2j} | \zeta_i, \zeta_j) \tag{A.7}$$

and the mean-zero vector δ_{1i} *is given by*

$$\delta_{1i} = \left(\int f_\zeta(\zeta_i)\mu_{31}(\zeta_i, \zeta_i, \beta_0)d\zeta_i\right)\psi_{\delta i} \tag{A.8}$$

where

$$\mu(t,\zeta,\beta) = \mathcal{H}(t,\zeta)\mathcal{M}(t,\zeta,\beta)f_\zeta(t) \tag{A.9}$$

with

$$\mathcal{M}(t,\zeta,\beta) = E[\mathcal{F}(z_{1i}'\beta_0, z_{1j}'\beta_0, \zeta_i, \zeta_j)1[z_{1i}'\beta > z_{1j}'\beta]z_i' \mid \zeta_i = t, \zeta_j = \zeta] \tag{A.10}$$

and $\mu_1(\cdot,\cdot,\cdot)$ *denotes the partial derivative of* $\mu(\cdot,\cdot,\cdot)$ *with respect to its first argument and* $\mu_{31}(\cdot,\cdot,\cdot)$ *denotes the partial derivative of* $\mu_1(\cdot,\cdot,\cdot)$ *with respect to its third argument.*

Although the particular expressions for V and Ω are quite involved, note that V represents the second derivative of the limit of the expectation of the maximand and Ω represents the variance of the limit of its projection.

The asymptotic theory for the third-stage statistic is based on the above conditions and the following additional smoothness condition:

Assumption S (Order of smoothness of density and conditional expectation functions)

- **S.1** Letting $\zeta_{\beta i}$ denote $z_{1i}'\beta_0$, and let $f_{\zeta\beta}(\cdot)$ denote its density function, we assume $f_{\zeta\beta}(\cdot)$ is p times continuously differentiable with derivatives that are bounded on the support of $\zeta_{\beta i}$.
- **S.2** The functions $G_{11}(\cdot)$ and $G_x(\cdot)$, defined as follows:

$$G_{11}(\cdot) = E[sgn(y_{1i} - y_{1j})f_{Z_k|Z_{-k}}(\Delta z_{-kij}'\delta_0^{(-k)})\Delta z_{-kij}'|\zeta_{\beta i} = \cdot, \zeta_{\beta j} = \cdot] \tag{A.11}$$

$$G_x(\cdot) = E[(sgn(y_{1i} - y_{1j})sgn(z_i'\delta_0 - z_j'\delta_0) - \tau_0)(z_{1i} - z_{1j})'|z_{1i} - z_{1j} = \cdot] \tag{A.12}$$

  where $f_{Z_k|Z_{-k}}()$ in (A.11) denotes the density function of the last component of $z_i - z_j$, conditional on its other components, and Δz_{-kij} denotes the difference for all the components of z_i except the last one, are all assumed to be all p times continuously differentiable with derivatives that are bounded on the support of $\zeta_{\beta i}$.

The main theorem establishes the asymptotic distribution of the statistic $\hat{\tau}$:

Theorem 1 *If Assumptions CPS, FS, K, H, RD, ED, FR, and S hold, then*

$$\sqrt{n}(\hat{\tau} - \tau_0) \Rightarrow N(0, V_2^{-2}\Omega_2) \tag{A.13}$$

with $V_2 = E[f_{\zeta\beta}(\zeta_{\beta i})]$ *and* $\Omega_2 = E[\delta_{2i}^2]$. *The mean-zero random variable* δ_{2i} *is*

$$\delta_{2i} = 2f_{\zeta\beta}(\zeta_{\beta i})G(y_{1i}, z_i, \zeta_{\beta i}) + E[G_x'(\zeta_{\beta i})f_{\zeta\beta}(\zeta_{\beta i})]\psi_{\beta i} + E[G_{11}(\zeta_{\beta i})f_{\zeta\beta}(\zeta_\beta i)]\psi_{\delta i}, \tag{A.14}$$

where $G_x'()$ *denotes the derivative of* G_x *and* $G(\cdot,\cdot,\cdot)$ *is given by*

$$G(y_1, z, \zeta) = E[sgn(y_{1i} - y_1)sgn(z_i'\delta_0 - z'\delta_0)|\zeta_{\beta i} = \zeta]. \tag{A.15}$$

A.1 Proof of Lemma 1

The proof strategy will be along the lines of Sherman (1994), and we deliberately aim to keep notation as similar as possible to that used in that paper. Specifically, let $G_n(\theta)$ and $\hat{G}_n(\theta)$ be defined as[1]

$$G_n(\theta) = \frac{1}{n(n-1)}\sum_{i\neq j} 1[y_{2i}=y_{2j}]k_h(z_i'\delta_0 - z_j'\delta_0)1[y_{1i}>y_{1j}]1[z_{1i}'\beta(\theta)>z_{1j}'\beta(\theta)] \quad \text{(A.16)}$$

$$\hat{G}_n(\theta) = \frac{1}{n(n-1)}\sum_{i\neq j} 1[y_{2i}=y_{2j}]k_h(z_i'\hat{\delta} - z_j'\hat{\delta})1[y_{1i}>y_{2j}]1[z_{1i}'\beta(\theta)>z_{1j}'\beta(\theta)] \quad \text{(A.17)}$$

Similar to Sherman (1994), the proof strategy involves the following stages:

1. Establish consistency of the estimator.
2. Show the estimator converges at the parametric ($\sqrt{n}$) rate.
3. Establish asymptotic normality of the estimator.

For consistency, we apply Theorem 2.1 in Newey and McFadden (1994). Compactness follows from Assumption CPS. To show uniform convergence, we note that the estimated first-stage index converges uniformly to the true index, by Assumption FS, so we can replace estimated indexes with true values inside the objective function, and work with $G_n(\theta)$. Next, we note by Theorem 2 in Sherman (1994),

$$\sup_{\theta\in\Theta}\left(G_n(\theta) - E[G_n(\theta)]\right) = o_p(1) \quad \text{(A.18)}$$

by the Euclidean property of the indicator function $1[z_{1i}'\beta > z_{1j}'\beta]$, and the uniform (in n) boundedness of $E[G_n(\theta)]$. By a change of variables and Assumptions K and H,

$$\sup_{\theta\in\Theta} E[G_n(\theta)] - \mathcal{G}(\theta) \xrightarrow{p} 0 \quad \text{(A.19)}$$

where we will define $\mathcal{G}(\theta)$ as follows. First, define the indicator $\tilde{d}_{ij}$ as $1[y_{2i}=y_{2j}]$, and define

$$\mathcal{H}(z_i'\delta_0, z_j'\delta_0) = E[\tilde{d}_{ij}|z_i'\delta_0, z_j'\delta_0].$$

Furthermore, define

$$\mathcal{F}(z_{1i}'\beta_0, z_{1j}'\beta_0, z_i'\delta_0, z_j'\delta_0) = P(y_{1i}>y_{1j}|\tilde{d}_{ij}=1, z_{1i}'\beta_0, z_{1j}'\beta_0, z_i'\delta_0, z_j'\delta_0).$$

[1] Implicit in the proofs that follow, we are subtracting the function $1[z_{1i}'\beta_0 > z_{1j}'\beta_0]$ from $1[z_{1i}'\beta(\theta) > z_{1j}'\beta(\theta)]$ in each of the two objective functions, analogous to Sherman (1993). The terms subtracted do not affect the value of the estimator, and are omitted for notational convenience.

So we can define

$$\mathcal{G}(\theta) = E[\mathcal{F}(z_{1i}'\beta_0, z_{1j}'\beta_0, z_i'\delta_0, z_i'\delta_0)\mathcal{H}(z_i'\delta_0, z_i'\delta_0)1[z_{1i}'\beta(\theta) > z_{1j}'\beta(\theta)]] \tag{A.20}$$

where the above expectation is taken with respect to (z_i, z_j). This establishes uniform convergence of $\hat{G}_n(\theta)$ to $\mathcal{G}(\theta)$. $\mathcal{G}(\theta)$ is continuous by the smoothness assumptions on the regressor vector z_{1i} and the index $z_i'\delta_0$ distribution. Finally, as a last condition to apply Theorem 2.1 in Newey and McFadden (1994), we need to show that $\mathcal{G}(\theta)$ is uniquely maximized at θ_0. This follows from the distributional assumption on ϵ_i (Assumption ED), the index distributional assumption (Assumption RD), and the full rank condition (Assumption FR). This establishes consistency.

With consistency established, the next two stages can be established along the lines of Sherman (1994) and Khan (2001). For root-n consistency, we will apply Theorem 1 of Sherman (1994), whose sufficient conditions for rates of convergence are that

1. $\hat{\theta} - \theta_0 = O_p(\delta_n)$.
2. There exists a neighborhood of θ_0 and a constant $\kappa > 0$ such that $\mathcal{G}(\theta) - \mathcal{G}(\theta_0) \geq \kappa\|\theta - \theta_0\|^2$ for all θ in this neighborhood.
3. Uniformly over $O_p(\delta_n)$ neighborhoods of θ_0

$$\hat{G}_n(\theta) = \mathcal{G}(\theta) + O_p(\|\theta - \theta_0\|/\sqrt{n}) + o_p(\|\theta - \theta_0\|^2) + O_p(\epsilon_n). \tag{A.21}$$

which suffices for $\hat{\theta} - \theta_0 = O_p(\max(\epsilon_n^{1/2}, n^{-1/2}))$. Having already established consistency, we will first set $\delta_n = o(1)$, $\epsilon_n = O(n^{-1})$, and show the above three conditions are satisfied. This will imply that the estimator is root-n consistent.

To show these conditions, note we set $\delta_n = o(1)$, so the first holds by our consistency proof. Under Assumptions RD and ED, the function $\mathcal{G}(\cdot)$ is sufficiently smooth in θ such that we can apply a second order expansion (see, e.g., Sherman (1993) and Khan (2001)) to show the second condition. For the the third condition, first replace $\hat{G}_n(\theta)$ with $G_n(\theta)$. Under Assumptions K and H and equation (A.1), the remainder term of such a replacement is (uniformly in $o_p(1)$ neighborhoods of θ_0) $o_p(\|\theta - \theta_0\|/\sqrt{n})$ (see Khan (2001) for the complete arguments used). Therefore, it remains to show that

$$G_n(\theta) = \mathcal{G}(\theta) + O_p(\|\theta - \theta_0\|/\sqrt{n}) + o_p(\|\theta - \theta_0\|^2) + O_p(\epsilon_n) \tag{A.22}$$

The difference $G_n(\theta) - \mathcal{G}(\theta)$ can be viewed as a centered U-process to which locally uniform rate results, such as Theorem 3 in Sherman (1994), can be applied to given our smoothness, boundedness, and bandwidth conditions. (A.22) follows immediately, establishing root-n consistency of $\hat{\theta}$.

Turning to asymptotic normality of the estimator, note that we can apply Theorem 2 of Sherman (1994), of which a sufficient condition will be to show that uniformly over $O_p(1/\sqrt{n})$ neighborhoods of θ_0,

$$\hat{G}_n(\theta) = \frac{1}{2}\theta' V \theta + \frac{1}{\sqrt{n}}\theta' W_n + o_p(n^{-1}) \tag{A.23}$$

where V is a negative definite matrix whose form will given below, and W_n is asymptotically normal, with mean 0 and variance Ω (whose form is given below).

To show (A.23), we will work with the following expansion:

$$\hat{G}_n(\theta) = G_n(\theta) + G_n'(\theta) + R_n \tag{A.24}$$

where

$$G_n'(\theta) = \frac{1}{n(n-1)} \sum_{i \neq j} 1[y_{2i} = y_{2j}] h_n^{-1} k_h'(z_i'\hat{\delta} - z_j'\hat{\delta}) 1[y_{1i} > y_{1j}] 1[z_{1i}'\beta(\theta) > z_{1j}'\beta(\theta)](\Delta z_{ij}'\hat{\delta} - \Delta z_{ij}'\delta_0) \tag{A.25}$$

(recall that $\Delta z_{ij} = z_i - z_j$) with $k_h'(\cdot)$ denoting the derivative of the function $k_h(\cdot)$. R_n in (A.24) denotes the remainder term in the expansion, whose asymptotic properties will be dealt with after we derive the asymptotic properties of $G_n(\theta)$ and $G_n'(\theta)$.

The following Lemma establishes a representation for $G_n(\theta)$:

Lemma 2 *Under the conditions RD, ED, FR, uniformly over $O_p(n^{-1/2})$ neighborhoods of θ_0, we have*

$$G_n(\theta) = \frac{1}{2}\theta' V \theta + \frac{1}{\sqrt{n}}\theta' W_n + o_p(n^{-1}) \tag{A.26}$$

where V is negative definite and W_n is asymptotically normal with mean 0 and variance Ω.

Proof: We will first evaluate a representation for $E[G_n(\theta)]$. We do this because we will later work with the U-statistic representation theorems found in, e.g., Serfling (1978). Letting $\zeta_i = z_i'\delta_0$, we write $E[G_n(\theta)]$ as the integral:

$$\int k_h(\Delta\zeta_{ij}) 1[z_{1i}'\beta > z_{1j}'\beta] \mathcal{H}(\zeta_i, \zeta_j) \mathcal{F}(z_{1i}'\beta_0, z_{1j}'\beta_0, \zeta_i, \zeta_j) dF_{Z_1,\zeta}(z_{1i}, \zeta_i) dF_{Z_1,\zeta}(z_{1j}, \zeta_j) \tag{A.27}$$

Next, we do the change of variables $u = \frac{\Delta\zeta_{ij}}{h_n}$ and obtain the following integral

$$\int k(u) 1[z_{1i}'\beta > z_{1j}'\beta] \mathcal{H}(\zeta_j + uh_n, \zeta_j) \mathcal{F}(z_{1i}'\beta_0, z_{1j}'\beta_0, \zeta_j + uh_n, \zeta_j)$$

$$dF_{Z_1,\zeta}(z_{1i}, \zeta_j) dF_{Z_1,\zeta}(z_{1j}, \zeta_j + uh_n) du \tag{A.28}$$

Taking a second-order expansion inside the integral around $uh_n = 0$, the lead term is of the form:

$$\int k(u)1[z_{1i}'\beta > z_{1j}'\beta]\mathcal{H}(\zeta_j, \zeta_j)\mathcal{F}(z_{1i}'\beta_0, z_{1j}'\beta_0, \zeta_j, \zeta_j)dF_{Z_1,\zeta}(z_{1i}, \zeta_j)dF_{Z_1,\zeta}(z_{1j}, \zeta_j)du \tag{A.29}$$

Note the term $\mathcal{F}(z_{1i}'\beta_0, z_{1j}'\beta_0, \zeta_j, \zeta_j)$ controls for selection bias. Note also we can use the same arguments as in Sherman (1993) to conclude that the integral in (A.29) is of the form

$$\frac{1}{2}\theta' V\theta + o_p(n^{-1}) \tag{A.30}$$

for θ uniformly in $O_p(n^{-1/2})$ neighborhoods of θ_0. The first-order term in the second-order expansion is 0 since $\int uk(u)du = 0$. The second-order term can be bounded above by

$$\left(C\int 1[z_{1i}'\beta > z_{1j}'\beta]dF_{Z_1,\zeta}(z_{1i}, \zeta_i)dF_{Z_1,\zeta}(z_{1j}, \zeta_j)\right)h_n^2 \tag{A.31}$$

where C is a finite constant. This term is $O\left((\theta-\theta_0)'h_n^2\right)$, which is $o(n^{-1})$ for θ in a $O(n^{-1/2})$ neighborhood of θ_0 by the assumptions on h_n which imply $\sqrt{n}h_n^2 \to 0$.

Next, we establish a representation for $E[G_n(\theta)|z_{1i}, y_{1i}, z_i]$. Using the same arguments as in the unconditional expectation, we conclude that $E[G_n(\theta)|z_{1i}, y_{1i}, z_i]$ is of the above form, now no longer integrating over the variables z_{1i}, y_{1i}, z_i.

Next, we derive a linear representation for (A.25). Regarding the term $\Delta z_{ij}'\hat{\delta}$, we will only derive the linear representation involving the component $\hat{\zeta}_i - \zeta_i$ as the term involving $\hat{\zeta}_j - \zeta_j$ can be dealt with similarly. The first step is to plug in a linear representation for the estimator $\hat{\delta}$ of δ_0:

$$\frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq k} h^{-1}k_h'((z_i - z_j)'\delta_0)1[y_{1i} > y_{1j}]z_i'\psi_{\delta k}1[z_{1i}'\beta > z_{1j}'\beta] \tag{A.32}$$

where $\psi_{\delta k}$ denotes the influence function in the linear representation of the first stage estimator $\hat{\delta}$, evaluated at the k-th observation. (Note the remainder term from the linear representation can effectively be ignored; when interacted with the other terms in (A.32), we get a neglible term (see, e.g., Khan (2001) for details).) We have a centered third-order U-process. Again, we note the its unconditional mean is 0, as is its mean conditional on each of its first two arguments. Consequently, we derive a linear representation for its mean conditional on its third argument:

$$\frac{1}{n}\sum_{k=1}^{n}\left(\int \mathcal{H}(\zeta_i, \zeta_j)\mathcal{F}(z_{1i}'\beta_0, z_{1j}'\beta_0, \zeta_i, \zeta_j)h^{-1}k_h'(\zeta_i - \zeta_j)1[z_{1i}'\beta > z_{1j}'\beta]z_i'\times\right.$$

$$\left. dF_Z(z_i)dF_Z(z_j)\right)\psi_{\delta_k} \tag{A.33}$$

While the above integral is expressed with respect to (z_i, z_j), it will prove convenient to express the integral in terms of (ζ_i, ζ_j) as follows:

$$\int \mathcal{H}(\zeta_i, \zeta_j)\mathcal{M}(\zeta_i, \zeta_j, \beta)h^{-1}k_h'(\zeta_i - \zeta_j)f_\zeta(\zeta_i)f_\zeta(\zeta_j)d\zeta_i d\zeta_j \tag{A.34}$$

where f_ζ denotes the density function of ζ, and

$$\mathcal{M}(\zeta_i,\zeta_j,\beta)=E[\mathcal{F}(z_{1i}'\beta_0,z_{1j}'\beta_0,\zeta_i,\zeta_j)1[z_{1i}'\beta>z_{1j}'\beta]z_i'|\zeta_i,\zeta_j] \tag{A.35}$$

Now we do a change of variables in (A.34) $v=(\zeta_i-\zeta_j)/h$, noting that under our assumptions we have $\int k'(v)dv=0$ and $\int k'(v)vdv=-1$ so that the lead term in the expansion (inside the integral) around $vh_n=0$ yields the integral

$$\left(\int \mu_1(\zeta_j,\zeta_j,\beta)f_\zeta(\zeta_j)d\zeta_j\right) \tag{A.36}$$

where

$$\mu(t,\zeta,\beta)=\mathcal{H}(t,\zeta)\mathcal{M}(t,\zeta,\beta)f_\zeta(t) \tag{A.37}$$

and $\mu_1(\cdot,\cdot,\cdot)$ denotes its partial derivative with respect to its first argument. The remaining terms in the expansion are negligible—i.e. $o_p(n^{-1})$ uniformly in $O_p(n^{-1/2})$ neighborhoods of β around β_0.

The next step is to expand the function $\mu_1(\cdot,\cdot,\cdot)$ in (A.36) around $\beta=\beta_0$. A second-order expansion of $\mu_1(\cdot,\cdot,\beta)$ around $\beta=\beta_0$, (inside the above integral) yields the term

$$\frac{1}{n}\sum_{k=1}^{n}\left(\int \mu_{31}(\zeta_j,\zeta_j,\beta_0)f_\zeta(\zeta_j)d\zeta_j\right)(\psi_{\delta k})'(\beta-\beta_0)+R_n \tag{A.38}$$

where $\mu_{31}(\cdot,\cdot,\cdot)$ denotes the partial derivative of $\mu_1(\cdot,\cdot,\cdot)$ with respect to its third argument, and the remainder term R_n is $o_p(n^{-1})$ uniformly in β in $O_p(n^{-1/2})$ neighborhoods of β_0. This concludes the linear representation of the term in the objective function involving $\hat{\zeta}_i-\zeta_i$. We note analogous arguments can be used to derive the linear representation for the term involving $\hat{\zeta}_j-\zeta_j$, which concludes the proof that $\hat{\beta}$ is a root-n consistent and asymptotically normal estimator of β_0. These results are used below in deriving the asymptotic properties of the rank-based statistic.

A.2 Proof of Theorem 1

Our strategy is to derive a linear representation for $\hat{\tau}-\tau_0$, where

$$\tau_0=E[sgn(y_{1i}-y_{1j})sgn(z_i'\delta_0-z_j'\delta_0)|z_{1i}'\beta_0-z_{1j}'\beta_0=0].$$

Note that

$$\hat{\tau}-\tau_0=\left(\frac{1}{n(n-1)}\sum_{i\neq j}\hat{\omega}_{ij}\right)^{-1}\left(\frac{1}{n(n-1)}\sum_{i\neq j}\hat{\omega}_{ij}(sgn(y_{1i}-y_{1j})sgn(z_i'\hat{\delta}-z_j'\hat{\delta})-\tau_0)\right) \tag{A.39}$$

from which we will derive the probability limit of the denominator term, a linear representation of the numerator term, and then apply Slutsky's theorem.

For the denominator term in (A.39), a mean-value expansion around the true index difference, the root-n consistency of $\hat{\beta}$, and a LLN for U-statistics implies:

$$\frac{1}{n(n-1)}\sum_{i\neq j}\hat{\omega}_{ij}\xrightarrow{p}E[f_{Z\beta}(\zeta_{\beta i})] \tag{A.40}$$

where $f_{Z\beta}(\cdot)$ denotes the density function of $\zeta_{\beta i}=z_{1i}'\beta_0$.

For the numerator term in (A.39), we first work with the term

$$\frac{1}{n(n-1)}\sum_{i\neq j}\omega_{ij}(sgn(y_{1i}-y_{1j})sgn(z_i'\delta_0-z_j'\delta_0)-\tau_0) \tag{A.41}$$

(where ω_{ij} is the expression obtained after replacing $\hat{\beta}$ with β_0 in $\hat{\omega}_{ij}$) to which we apply a U-statistic projection theorem. By a change of variables and the higher order properties of the kernel function inside ω_{ij}, the expectation of the term inside the double summation is $o_p(n^{-1/2})$; therefore, it remains to derive expressions for conditional expectations of the term inside the double summation conditional on its first and second arguments. Again, using a change of variables and the higher order properties of the kernel function inside the weighting function, we get the following expression for these conditional expectations:

$$\frac{1}{n}\sum_{i=1}^{n}2f_{Z\beta}(\zeta_{\beta i})(G(y_{1i},z_i,\zeta_{\beta i})-\tau_0)+o_p(n^{-1/2}) \tag{A.42}$$

where

$$G(y_1,z,\zeta_\beta)=E[sgn(y_{1i}-y_1)sgn(z_i'\delta_0-z'\delta_0)|\zeta_{\beta i}=\zeta_\beta]. \tag{A.43}$$

Therefore, by the projection theorem (Powell et al. (1989)) (which is applicable due to the properties of the kernel function and bandwidth, which imply that the variance of the statistic is $o(n)$), (A.41) can be represented as (A.42).

Turning attention to the linear term in the expansion of $\hat{\omega}_{ij}$ around ω_{ij}

$$\frac{1}{n(n-1)}\sum_{i\neq j}h^{-1}k_h'(\zeta_i-\zeta_j)(z_{1i}-z_{1j})'(\hat{\beta}-\beta_0)(sgn(y_{1i}-y_{1j})sgn(z_i'\delta_0-z_j'\delta_0)-\tau_0), \tag{A.44}$$

we can plug in the derived linear representation for $\hat{\beta}-\beta_0$, yielding a third order U-statistic plus a negligible remainder term. The U-statistic is of the form:

$$\frac{h_n^{-1}}{n(n-1)(n-2)}\sum_{i\neq j\neq k}k_h'(\zeta_{\beta i}-\zeta_{\beta j})(z_{1i}-z_{1j})'\psi_{\beta i}(sgn(y_{1i}-y_{1j})sgn(z_i'\delta_0-z_j'\delta_0)-\tau_0). \tag{A.45}$$

We note the unconditional expectation of the above term is 0, as is the expectation conditional on each of its first two arguments. Using similar arguments as before, it follows that the expectation conditional on its third argument can be expressed as:

$$\frac{1}{n}\sum_{i=1}^{n}E[G_x'(\zeta_{\beta i})f_{Z\beta}(\zeta_{\beta i})]\psi_{\beta i}+o_p(n^{-1/2}) \tag{A.46}$$

where

$$G_x(\cdot) = E[(sgn(y_{1i} - y_{1j})sgn(z_i'\delta_0 - z_j'\delta_0) - \tau_0)(z_{1i} - z_{1j})'|z_{1i} - z_{1j} = \cdot]$$

and $G_x'(\cdot)$ denotes the derivative of $G_x(\cdot)$ with respect to its argument.

A final term to deal with in the linear representation of the statistic is

$$\frac{1}{n(n-1)} \sum_{i \neq j} \omega_{ij} sgn(y_{1i} - y_{1j})(sgn((z_i - z_j)'\hat{\delta}) - sgn((z_i - z_j)'\delta_0)) \tag{A.47}$$

Here we can effectively expand the above term with $\hat{\delta}$ around δ_0. To do so, since the sign function is not differentiable, we take the expectation of $sgn((z_i - z_j)'\delta)$ for any δ. Recall, as a normalization, we set the last component of $\delta = 1$ and assume its associated regressor was continuously distributed. Here we let $F_{Z_k|Z_{-k}}(\cdot)$ denote the cdf of $z_{ki} - z_{kj}$ conditional on (z_{-ki}', z_{-kj}') (where z_{ki} denotes the last component of z_i and z_{-ki}' denotes the other components of z_i). Consequently, we have:

$$E[sgn((z_i - z_j)'\delta)|z_{-ki}, z_{-kj}] = F_{Z_k|Z_{-k}}(\Delta z_{-kij}'\delta^{(-k)})$$

where Δz_{-kij} denotes the difference in the corresponding components of z_i and z_j, and $\delta^{(-k)}$ denotes the subvector of δ corresponding to z_{-ki}. We can expand this conditional expectation evaluated at δ around the conditional expectation evaluated at δ_0:

$$F_{Z_k|Z_{-k}}(\Delta z_{-kij}'\hat{\delta}^{(-k)}) = F_{Z_k|Z_{-k}}(\Delta z_{-kij}'\delta_0^{(-k)}) + f_{Z_k|Z_{-k}}(\Delta z_{-kij}'\delta_0^{(-k)})\Delta z_{-kij}'(\hat{\delta} - \delta_0) + O_p(\|\hat{\delta} - \delta_0\|^2)$$

where $f_{Z_k|Z_{-k}}(\cdot)$ denotes the density function of $z_{ki} - z_{kj}$ conditional on z_{-ki}', z_{-kj}'. Note that since the sgn function is Euclidean and $\hat{\delta} - \delta_0$ is $O_p(n^{-1/2})$ (by using, e.g., Theorem 1 in Sherman (1994)), the remainder term that arises from replacing the difference in sgn functions with their expectations in (A.47) is $o_p(n^{-1/2})$.

Next, by applying the same arguments as before, involving plugging in a linear representation (this time for $(\hat{\delta} - \delta_0)$), and decomposing the resulting third order U-statistic, we get a linear representation for (A.47). Specifically, let

$$G_{11}(\cdot) = E[sgn(y_{1i} - y_{1j})f_{Z_k|Z_{-k}}(\Delta z_{-kij}'\delta_0^{(-k)})\Delta z_{-kij}'|\zeta_{\beta i} = \cdot, \zeta_{\beta j} = \cdot].$$

Then, we may conclude that (A.47) has the following linear representation

$$\frac{1}{n}\sum_{i=1}^{n} E[G_{11}(\zeta_{\beta i})f_{Z\beta}(\zeta_{\beta i})]\psi_{\delta i} + o_p(n^{-1/2}). \tag{A.48}$$

Finally, we note that is easy to show that the remainder term, which involves the product of $\hat{\beta} - \beta_0$ and $\hat{\delta} - \delta_0$, is $o_p(n^{-1/2})$.

Therefore, collecting all our results, we may conclude that

$$\hat{\tau} - \tau_0 = E[f_{Z\beta}(\zeta_{\beta i})]^{-1}\frac{1}{n}\sum_{i=1}^{n}\Big(2f_{Z\beta}(\zeta_{\beta i})(G(y_{1i}, z_i, \zeta_{\beta i}) - \tau_0) + E[G'_x(\zeta_{\beta i})f_{Z\beta}(\zeta_{\beta i})]\psi_{\beta i} + E[G_{11}(\zeta_{\beta i})f_{Z\beta}(\zeta_{\beta i})]\psi_{\delta i}\Big) + o_p(n^{-1/2}), \tag{A.49}$$

which establishes the theorem.

B Monte Carlo simulations

In this section, we report results of Monte Carlo simulations that examine the performance of the proposed statistic ($\hat{\tau}$) as well as other alternative approaches for testing for presence of causal effects (most notably the two-stage least-squares estimator (2SLS)). First, we consider a design (Design 1 below) in which both $\hat{\tau}$ and 2SLS (as well as a correctly specified MLE estimator) are theoretically justified testing approaches. This design has a binary outcome and a binary endogenous regressor (with continuous instrument) but no additional covariates in the outcome equation; for this design, the results of Shaikh and Vylacil (2005) and Bhattacharya, Shaikh, and Vytlacil (2005) imply that the probability limit of 2SLS identifies the correct sign of the treatment effect. The goal of these simulations is to compare size and power of the alternative approaches to make sure that the flexibility of the semiparametric ($\hat{\tau}$) approach does not adversely affect practical performance in a simple setting. Second, we provide a simple design (Design 2 below) that serves as a cautionary tale for using 2SLS (or ordinary least squares (OLS)) in order to test for causal effects in non-linear models. Specifically, we consider a probit model (binary outcome) with no treatment effect and no endogeneity. In general, when one moves beyond the binary-outcome/binary-treatment model with no additional covariates, the non-linearity of the outcome equation will generally lead the OLS or 2SLS estimators for the endogenous coefficient to have non-zero probability limits even when the true coefficient is zero. The implication is that, under the null of no treatment effect, rejection rates can approach 100% as the sample size grows. Finally, to demonstrate the robustness of the semiparametric approach, we consider two designs (Designs 4 and 5 below) having additional non-linearity and interactions in the covariate specification; for these designs, as in Design 3, we compare the performance of the rank-based test statistic to both OLS and 2SLS.

Design 1: binary outcome, binary treatment, no additional covariates

The following simple design, with a continuous instrumental variable for the binary endogenous regressor, is considered:

$$y_{1i} = 1[\alpha_0 y_{2i} + \epsilon_i > 0] \tag{A.50}$$

$$y_{2i} = 1[z_i + \eta_i > 0] \tag{A.51}$$

where z_i is a standardized χ^2_3 random variable (normalized to have mean zero and standard deviation one) and $\begin{pmatrix} \epsilon_i \\ \eta_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_0 \\ \rho_0 & 1 \end{pmatrix}\right)$. Two parameters, α_0 (the coefficient on the binary endogenous variable) and ρ_0 (the correlation between η_i and ϵ_i), are chosen to vary over the simulation designs. In particular, the values $\rho_0 = 0, 0.25, 0.50, 0.75$ and $\alpha_0 = 0, 0.1, 0.2, 0.3$ are considered, yielding 16 different designs. For each of the 16 designs, 1,000 simulations were conducted with a sample size of $n = 500$. Three different approaches to testing the significance of $\hat{\alpha}$ (i.e., testing $H_0 : \alpha_0 = 0$) were considered: (i) a full MLE estimation strategy (with the test based upon the z-statistic of $\hat{\alpha}_{mle}$), (ii) a linear IV estimation strategy (with the test based upon the z-statistic of $\hat{\alpha}_{iv}$, and (iii) the third-stage statistic $\hat{\tau}$ proposed above (which, in this case, is the Kendall's tau correlation between y_{1i} and z_i). Table S1 summarizes the results, with rejection rates reported for the 5% and 10% levels for the three approaches (labeled MLE, IV, and $\hat{\tau}$, respectively). The first four rows of the table correspond to $\alpha_0 = 0$ and, therefore, provide evidence on the size of the test. The rejection rates for the three tests are in line with the 5% and 10% levels, although at the higher levels of correlation ($\rho_0 = 0.50$ and $\rho_0 = 0.75$) between the error disturbances, the MLE approach exhibits some over-rejection. The remaining rows of the table provide evidence on the power of the test (for $\alpha_0 = 0.1, 0.2, 0.3$). Overall, the power of the alternative approaches is remarkably similar across these designs. For $\rho_0 = 0.75$, the MLE approach does have higher rejection rates, but these are likely the result of the over-rejection phenomenon seen in the $\alpha_0 = 0/\rho_0 = 0.75$ design; for the other ρ_0 values, the MLE approach has rejection rates which are basically indistinguishable from the other two approaches. For this simple Monte Carlo design, the semiparametric approach to testing for significance of the binary endogenous variable compares favorably with both an MLE approach and a linear IV approach.

Design 2: probit model with no treatment effect and no endogeneity

We consider a probit model having no treatment effect (a coefficient of zero on the y_2 variable) and no endogeneity. The y_1 outcome equation has a covariate w, and y_2 is related to an additional "instrument" z:

$$y_{1i} = 1[-3 + w_i + \eta_i > 0] \tag{A.52}$$

$$y_{2i} = 1[-5 + w_i + z_i + \epsilon_i > 0] \tag{A.53}$$

with $w_i \sim \chi^2_3$, $z_i \sim U[0, 4]$, $\epsilon_i \sim N(0, 1)$, and $\eta_i \sim N(0, 1)$ independent of each other. For the Monte Carlo simulations, we compared the performance of three approaches (OLS, 2SLS, and the $\hat{\tau}$-based test) for testing significance of the causal effect of y_2 upon y_1. For OLS, we ran a regression of y_1 on w and y_2 (and an intercept) and computed the heteroskedasticity-robust z-statistic on the y_2

coefficient. For 2SLS, the first-stage regression was OLS of y_2 on w and z, and the second-stage regression was OLS of y_1 on w and the fitted y_2 values from the first stage; we then computed the heteroskedasticity-robust z-statistic on the y_2 coefficient. For the $\hat{\tau}$ statistic, we used the bootstrap (50 replications) in order to compute standard errors and, in turn, z-statistics; the normal-density kernel was used, and results for a bandwidth of $h = 0.1$ are reported (though results were similar across a range of different bandwidths). Table S2 reports the results from 1,000 simulations for sample sizes $n \in \{50, 100, 200, 400, 1000\}$. Rejection rates for the 5%-level and 10%-level tests are reported for the three different methods. Both the OLS and 2SLS methods are clearly incorrectly sized. The OLS rejection rates for $n = 50$ are 16.0% and 25.5% and quickly diverge at higher sample sizes (nearly 100% rejection rates at $n = 1000$). The 2SLS performs even worse for the smallest sample size (28.9% and 38.5% rejection rates for $n = 50$), and the rejection rates remain fairly stable (and incorrect) across the sample sizes considered. In contrast, the rank-based test statistic appears to be properly sized. There is some evidence of over-rejection at $n = 50$ (9.1% and 15.2$ rejection rates) but, at higher samples, the rejection rates are very close to the correct size.

Design 3: non-linear relationship between the instrument and endogenous variable

We consider a design similar to Design 1, except that the endogenous variable is continuous and is a non-linear function of the instrumental variable. As in Design 1, we consider the simplest case where there are no additional covariates. The goal is to examine what happens in a case where we use 2SLS assuming that the instrument is z_i but the reduced-form model for the endogenous variable is instead a non-linear function of z_i (specifically a square-root form in this design):

$$y_{1i} = 1[5\alpha_0 y_{2i} + \epsilon_i > 0] \tag{A.54}$$

$$y_{2i} = -1 + \sqrt{z_i/5} + \eta_i \tag{A.55}$$

with $\sqrt{6}z_i \sim \chi^2_3$ and (ϵ_i, η_i) has a bivariate normal distribution with zero means, unit variances, and correlation parameter ρ_0. Like Design 1, we consider a set of 16 designs (with $\rho_0 \in \{0.00, 0.25, 0.50, 0.75\}$ and $\alpha_0 \in \{0.0, 0.1, 0.2, 0.3\}$). To give a sense of how the non-zero α_0 specifications differ from the $\alpha_0 = 0$ (no treatment) specification, we computed the average marginal effects for these values — approximately 0.17, 0.26, and 0.30 for $\alpha = 0.1$, $\alpha = 0.2$, and $\alpha = 0.3$, respectively.

Using 1,000 simulations for $n = 200$, Table S3 reports rejection rates for OLS, 2SLS, and the $\hat{\tau}$-based test at the 5% level. The OLS method clearly performs very badly, with the rejection rate at $\alpha_0 = \rho_0 = 0$ being fine (5.6%) but increasing to 81.6% for $\alpha_0 = 0/\rho_0 = 0.25$ and 100% for all other specifications. While the 2SLS appears to have greater power than the $\hat{\tau}$ method, there is evidence that the 2SLS method is incorrectly sized. In particular, there is significant under-rejection for the $\alpha_0 = 0$ case when $\rho_0 \in \{0, 0.25, 0.5\}$.

Design 4: non-linearity and interaction in the outcome equation

The following design again exhibits non-linearity. In addition, however, we allow for an interaction between the endogenous variable y_2 and the exogenous-covariate index (in this case, just a single exogenous variable w). Specifically, y_1 has the following exponential specification:

$$y_{1i} = \exp(-w_i - \alpha_0 w_i y_{2i} + \epsilon_i) \tag{A.56}$$

$$y_{2i} = -1 + z_i + \eta_i \tag{A.57}$$

with $w_i \sim \text{Bernoulli}(0.5)$, $\sqrt{6}z_i \sim \chi^2_3$, and (ϵ_i, η_i) has a bivariate normal distribution with zero. This outcome equation falls within the generalized regression model since the endogenous variable is interacting with the full covariate index (which in this simple case is just the w_i variable itself). Note that w_i is a binary covariate here, with $y_{1i} = \exp(\epsilon_i)$ when $w_i = 0$ and $y_{1i} = \exp(-1-\alpha_0 y_{2i}+\epsilon_i)$ when $w_i = 1$. For this design, we consider what happens when we use 2SLS of y_1 on w and y_2, with z instrumenting for y_2. Thus, the underlying model for this 2SLS regression is misspecified in two ways: (i) the non-linearity is ignored (using y_1 rather than $\ln(y_1)$) and (ii) the interaction is ignored. For the OLS results, we regress y_1 on w and y_2. Again, we consider a set of 16 designs (with $\rho_0 \in \{0.00, 0.25, 0.50, 0.75\}$ and $\alpha_0 \in \{0.0, 0.1, 0.2, 0.3\}$). The last three columns of Table S3 report rejection rates for OLS, 2SLS, and $\hat{\tau}$ at the 5% level, based upon 1,000 simulations for $n = 200$. In contrast to Design 3, the 2SLS test appears to be appropriately sized in this design (even in the presence of the non-linearity). For the $\alpha_0 = 0.1$ alternatives (the alternatives closest to the null), the 2SLS and $\hat{\tau}$ tests have comparable power. For the other alternatives ($\alpha_0 = 0.2$ and $\alpha_0 = 0.3$), however, the power of the $\hat{\tau}$ test becomes much stronger than that of the 2SLS test. Finally, as in the previous designs, we find that OLS vastly over-rejects in the presence of endogeneity when there is no treatment effect present (50.5% rejection for $\alpha_0 = 0/\rho_0 = 0.25$ and increasing at higher ρ_0 values).

C Empirical application

In this section, we apply our estimation and testing methodology to an empirical application concerning the effects of fertility on female labor supply. In particular, we adopt the approach of Angrist and Evans (1998), who use the gender mix of a woman's first two children to instrument for the decision to have a third child. This instrumental-variable strategy allows one to identify the effect of having a third child upon the woman's labor-supply decision. The rationale for this strategy is that child gender is arguably randomly assigned and that, in the United States, families whose first two children are the same gender are significantly more likely to have a third child.

Using 1980 and 1990 Census data, Angrist and Evans (1998) find that married women whose first two children are the same gender are 5–8% more likely to have a third child; using the same-sex indicator as an instrument for having a third child, they find that having a third child lowers the probability of a married women working for pay by about 10–12%. Rather than using the 1980 and 1990 Census data, the sample for the current study is drawn from the 2000 Census data (5-percent public-use microdata sample (PUMS)) in order to see if any interesting changes have occurred in the relationship between fertility and labor supply. Starting from the household PUMS data, a mother was retained in the sample if all of the following criteria were satisfied: (i) mother has two or more children, (ii) mother is white, (iii) mother is a United States citizen, (iv) mother is married with spouse present in household, and (v) oldest child is 12 years of age or younger.[2] In addition, to eliminate any families that might have twin births (or higher-order multiple births), any family with same-aged first and second children or same-aged second and third children were dropped from the sample. The resulting sample consists of 293,771 observations.

Summary statistics for the variables to be used in the analysis are provided in Table S4. The table shows that 69.7% of mothers in the sample worked for pay during 1999 and 25.7% of mothers had a third child. The percentage of women working for pay represents a very slight increase over the comparable percentage from the 1990 Census data, and the percentage having a third child represents a decline from 1990. In the analysis, the outcome of interest (y_1) is whether the mother worked in 1999, the binary endogenous explanatory variable (y_2) is the presence of a third child, and the instrument is whether the mother's first two children were of the same gender.

Table S5 reports the first-stage regression results, i.e. regressing the have-third-child indicator upon the same-sex indicator variable and the other z variables. The linear probability estimates and the probit estimates indicate that mothers whose first two children are the same gender are 5.6–5.8 percentage points more likely to have a third child than mothers whose first two children are of different gender. These estimates are very similar, although slightly lower in magnitude, to those found by Angrist and Evans (1998, Table 5) for the earlier 1980 and 1990 samples. Table S6 reports the second-stage estimates for the (linear) two-stage least squares estimator, along with the OLS estimates for comparison. Again, the results are very similar to those found by Angrist and Evans (1998, Tables 7 and 8), with the OLS estimates of the had-third-child effect on labor supply larger in magnitude than the 2SLS estimates.

Table S7 considers the alternative tests for significance of the binary endogenous regressor, comparing the semiparametric $\hat{\tau}$ test proposed in this paper with the z-test based upon the 2SLS estimates. In order to examine the effect of additional covariates, testing results are reported starting from a model with no exogenous covariates and then adding covariates onb-by-one until

[2] The PUMS data contains information on children under the age of 18 that are living in the household. Unlike earlier editions of the PUMS data, the 2000 edition does not contain a data item for the "total number of children ever born." Therefore, the last criterion is used in order to make it more likely that the oldest child in the household is actually the mother's first child. The cutoff could be lowered further to increase this certainty but at the expense of decreasing the sample size.

the full set of three exogenous covariates are included. (Note that the 2SLS estimates and standard errors for the no-covariate and three-covariate models correspond to those presented in Table S6.) In the model with no exogenous covariates, the z-statistics associated with $\hat{\tau}$ and the 2SLS coefficient are extremely similar. This finding is very much in line with the Monte Carlo simulation evidence (Design 1) of the previous section. The 2SLS z-statistic for the larger models is basically unchanged from the no-covariate model, which is not too surprising given that the same-sex instrument is uncorrelated with the other exogenous covariates in the model. In contrast, the magnitude of the z-statistic for the semiparametric $\hat{\tau}$ method does decline. The addition of covariates to the model forces the semiparametric method to make comparisons based upon observation-pairs with similar first-stage (estimated) index values associated with these exogenous covariates. It is encouraging, however, that the z-statistic magnitude does not decline by much as the second and third covariates are added to the model. Table S7 highlights the inherent robustness-power tradeoff between the semiparametric and parametric methodologies. Although one might have worried that the tradeoff would be so drastic to render the semiparametric method useless in practice, the results indicate that this is not the case. Even in the model with three covariates, the $\hat{\tau}$ estimate provides strong statistical evidence ($z = -2.69$) that the endogenous third-child indicator variable has a causal effect upon mothers' labor supply. Importantly, this finding is not subject to the inherent misspecification of the linear probability model or any type of parametric assumption on the error disturbances.

Table S1: Monte Carlo simulation results for Design 1 (binary outcome, binary treatment, no additional covariates) with $n = 500$. Rejection rates (over 1,000 simulations) for tests at the 5% and 10% levels are reported. The three different z-tests are based upon MLE estimation, IV estimation, and rank correlation.

			5%-level rejections			10%-level rejections		
	ρ_0	α_0	MLE	IV	$\hat{\tau}$	MLE	IV	$\hat{\tau}$
Size of the test	0.00	0.0	0.054	0.051	0.040	0.091	0.099	0.087
	0.25	0.0	0.062	0.050	0.056	0.109	0.094	0.109
	0.50	0.0	0.065	0.054	0.047	0.120	0.107	0.085
	0.75	0.0	0.069	0.048	0.057	0.131	0.102	0.105
Power of the test	0.00	0.1	0.071	0.075	0.071	0.134	0.144	0.126
	0.25	0.1	0.099	0.106	0.091	0.153	0.163	0.160
	0.50	0.1	0.087	0.092	0.083	0.155	0.157	0.144
	0.75	0.1	0.099	0.090	0.086	0.160	0.153	0.152
	0.00	0.2	0.161	0.168	0.162	0.265	0.269	0.251
	0.25	0.2	0.174	0.180	0.176	0.272	0.271	0.280
	0.50	0.2	0.180	0.176	0.167	0.279	0.256	0.261
	0.75	0.2	0.241	0.222	0.225	0.361	0.328	0.332
	0.00	0.3	0.310	0.303	0.298	0.419	0.399	0.414
	0.25	0.3	0.315	0.302	0.324	0.440	0.428	0.437
	0.50	0.3	0.321	0.338	0.327	0.447	0.442	0.430
	0.75	0.3	0.439	0.420	0.418	0.587	0.543	0.528

Table S2: Monte Carlo simulation results for Design 2 (probit model with no treatment effect and no endogeneity). Rejection rates (over 1,000 simulations) for z-tests at the 5% and 10% levels are reported.

	OLS		2SLS		$\hat{\tau}$	
n	5% level	10% level	5% level	10% level	5% level	10% level
50	0.169	0.255	0.289	0.385	0.091	0.152
100	0.309	0.427	0.273	0.356	0.055	0.104
200	0.524	0.648	0.290	0.386	0.045	0.091
400	0.831	0.901	0.256	0.350	0.055	0.106
1000	0.993	0.998	0.237	0.337	0.056	0.101

Table S3: Monte Carlo simulation results for Design 3 (non-linear relationship between the instrument and endogenous variable) and Design 4 (non-linearity and interaction in the outcome equation) with $n = 200$. Rejection rates (over 1,000 simulations) for tests at the 5% level are reported. The z-tests are based upon OLS estimation, IV estimation, and the weighted-rank-correlation statistic.

			Design 3 rejection rates			Design 4 rejection rates		
			OLS	IV	$\hat{\tau}$	OLS	IV	$\hat{\tau}$
Size of the test	0.00	0.0	0.056	0.014	0.047	0.048	0.047	0.055
	0.25	0.0	0.816	0.033	0.061	0.505	0.060	0.054
	0.5	0.0	1.000	0.033	0.051	0.971	0.044	0.062
	0.75	0.0	1.000	0.055	0.060	0.998	0.042	0.064
Power of the test	0.00	0.1	1.000	0.149	0.151	0.094	0.088	0.089
	0.25	0.1	1.000	0.188	0.150	0.317	0.080	0.106
	0.50	0.1	1.000	0.220	0.139	0.943	0.070	0.107
	0.75	0.1	1.000	0.237	0.122	0.991	0.087	0.108
	0.00	0.2	1.000	0.400	0.326	0.162	0.131	0.215
	0.25	0.2	1.000	0.367	0.254	0.151	0.127	0.241
	0.50	0.2	1.000	0.395	0.229	0.884	0.097	0.250
	0.75	0.2	1.000	0.365	0.177	0.992	0.095	0.263
	0.00	0.3	1.000	0.554	0.376	0.287	0.218	0.395
	0.25	0.3	1.000	0.503	0.331	0.086	0.169	0.419
	0.50	0.3	1.000	0.504	0.301	0.744	0.163	0.503
	0.75	0.3	1.000	0.453	0.250	0.986	0.155	0.540

Table S4: Summary statistics

Variable	Description	Mean (Stdev)
Worked in 1999	1 if worked for pay in 1999, 0 otherwise	0.697
Same-sex indicator	1 if first two children are the same gender, 0 otherwise	0.502
Had third child	1 if had third child, 0 otherwise	0.257
Age at first birth	Mother's age when first child was born	26.36 (5.03)
1st child's age	Age of first child in 2000	7.55 (3.03)
Education	Mother's education level (in years)	10.97 (2.19)

Table S5: First-stage regression results. The dependent variable is an indicator variable equal to one if the woman had a third child. Heteroskedasticity-robust standard errors are reported for the OLS estimates. Marginal effects (evaluated at the means of the explanatory variables) for the probit estimates are provided in brackets.

	OLS	OLS	Probit
Same-sex indicator	0.0564	0.0562	0.1865
	(0.0016)	(0.0015)	(0.0052)
			[0.0576]
Age at first birth		-0.0147	-0.0511
		(0.0002)	(0.0006)
			[-0.0158]
1st child's age		0.0318	0.1111
		(0.0002)	(0.0009)
			[0.0344]
Education		0.0104	0.0364
		(0.0004)	(0.0013)
			[0.0113]
# observations	293,771	293,771	293,771
R-squared	0.0042	0.0812	

Table S6: Second-stage regression results. The dependent variable is an indicator variable equal to one if the woman worked for pay in 1999. The 2SLS regressions use the same-sex indicator variable as an instrument for the had-third-child indicator variable. Heteroskedasticity-robust standard errors are reported.

	OLS	2SLS (Wald)	OLS	2SLS
Had third child	-0.1382	-0.1118	-0.1728	-0.1124
	(0.0020)	(0.0298)	(0.0020)	(0.0295)
Age at first birth			-0.0074	-0.0065
			(0.0002)	(0.0005)
1st child's age			0.0161	0.0142
			(0.0003)	(0.0010)
Education			0.0319	0.0313
			(0.0004)	(0.0005)
# observations	293,771	293,771	293,771	293,771
R-squared	0.0173	0.0167	0.0462	0.0432

Table S7: Testing significance of the binary endogenous regressor. The z-statistics for the semi-parametric and 2SLS estimation approaches are reported for several different model specifications. The 2SLS standard errors are heteroskedasticity-robust.

Exogenous covariates	Semiparametric			2SLS		
in the model	$\hat{\tau}$	s.e.	z-stat	$\hat{\alpha}$	s.e.	z-stat
None	-0.00316	0.00085	-3.72	-0.1118	0.0298	-3.75
Education	-0.00299	0.00102	-2.94	-0.1103	0.0296	-3.72
Education, Mother's age at first birth	-0.00655	0.00229	-2.86	-0.1111	0.0296	-3.75
Education, Mother's age at first birth, Age of 1st child	-0.00695	0.00258	-2.69	-0.1124	0.0295	-3.80