Online Appendix:
Judging Judge Fixed Effects

Brigham R. Frandsen[*] Lars J. Lefgren[*] Emily C. Leslie[*]

November 15, 2022

Proofs

Proof of Theorem 1. Condition 1 satisfies the conditions of Theorem 1 in Imbens and Angrist (1994), which implies

$$\begin{aligned} & E\left[Y_i | J_i = j, X_i = x\right] && (1) \\ = \; & \left(p_{j,x} - p_{1,x}\right) E\left[Y_i\left(1\right) - Y_i\left(0\right) | D_i\left(j\right) > D_i\left(0\right), X_i = x\right] + E\left[Y_i | J_i = 1, X_i = x\right]. \end{aligned}$$

By Condition 1c, monotonicity, for each individual i with $X_i = x$, one can define a marginal propensity, $\bar{p}_{i,x} := \inf\left\{p : p \in \left\{p_{1,x}, \ldots, p_{J,x}\right\}, D_i\left(j\right) = 1, X_i = x\right\}$, such that when assigned a judge with $p_{j,x} \geq \bar{p}_{i,x}$, the individual is treated, and otherwise is untreated. For never-takers, we define $\bar{p}_{i,x} = \infty$. For always-takers, $\bar{p}_{i,x} = p_{1,x}$. Note that $\bar{p}_{i,x}$ depends only on $D_i\left(j\right)$, and by Condition 1a is therefore independent of J_i conditional on $X_i = x$ and that potential treatment status for individual i with $X_i = x$ can be written $D_i\left(J_i\right) = \bar{p}_{i,x} \leq p_{J_i,x}$. The right hand side of equation (1) then can be written

$$\phi_x\left(p_{j,x}\right) = \left(p_{j,x} - p_{1,x}\right) E\left[Y_i\left(1\right) - Y_i\left(0\right) | p_{1,x} < \bar{p}_{i,x} \leq p_{j,x}, X_i = x\right] + E\left[Y_i | J_i = 1, X_i = x\right],$$

which depends on j only through $p_{j,x}$. By monotonicity the average slope of ϕ through two points p and p' (where $p' \geq p$) can be written:

$$\phi_x\left(p'\right) - \phi_x\left(p\right) = \left(p' - p\right) E\left[Y_i\left(1\right) - Y_i\left(0\right) | p \leq \bar{p}_{i,x} \leq p', X_i = x\right].$$

Let $\mathcal{Y}$ be the compact support of Y_i. Noting that $K := \sup\mathcal{Y} - \inf\mathcal{Y}$ is finite and that $\left|E\left[Y_i\left(1\right) - Y_i\left(0\right) | p \leq \bar{p}_{i,x} \leq p', X_i = x\right]\right| \leq K$ yields the result. ■

[*] Department of Economics, Brigham Young University

Proof of Theorem 2. Define

$$\begin{aligned} T_{1,j} &= \frac{\hat{y}_j - y_j - K\left(\hat{p}_j - p_j\right)}{\sqrt{\left(\hat{\sigma}^2_{Yj} + K^2\hat{\sigma}^2_{Dj} - 2K\hat{\sigma}_{YDj}\right)/n_j}}, \\ T_{2,j} &= \frac{\hat{y}_j - y_j + K\left(\hat{p}_j - p_j\right)}{\sqrt{\left(\hat{\sigma}^2_{Yj} + K^2\hat{\sigma}^2_{Dj} + 2K\hat{\sigma}_{YDj}\right)/n_j}}, \\ T_n &= \max\left\{\left|T_{1,j}\right|, \left|T_{2,j}\right|\right\}_{j=1}^J. \end{aligned}$$

Note that by Condition 1 $\{y_j, p_j\}$ satisfy constraint (2). By the central limit theorem

$$\sqrt{n_j}\begin{pmatrix} T_{1,j} \\ T_{2,j} \end{pmatrix} \to_d N\left(0, \rho_j\right), \qquad j = 1, \ldots, J,$$

independently across j. Therefore T_n is asymptotically distributed as the maximum element of a $2J$ Gaussian variate with block-diagonal correlation matrix with correlations ρ_j. Standard results on order statistics imply that the cdf of T_n is $F(t)$. Since by construction $\hat{T} \leq T_n$, we have

$$\lim_{n\to\infty} \Pr\left(\hat{T} > C_{1-\alpha}\right) \leq \lim_{n\to\infty} \Pr\left(T_n > C_{1-\alpha}\right) = \alpha.$$

■

Proof of Theorem 3. Define the following:

$$\begin{aligned} Y_{ij} &= Y_i\left(D_i\left(j\right), j\right) \\ \bar{D}_i &= \sum_{j=1}^J \lambda_j D_i\left(j\right) \\ p &= \sum_{j=1}^J \lambda_j p_j \\ \bar{Y}_i &: = \sum_{j=1}^J \lambda_j Y_{ij} \end{aligned}$$

Iterating expectations in the numerator and denominator of the definition of β_{2SLS}, the expression becomes:

$$\begin{aligned} & \frac{\sum_{j=1}^k \lambda_j \left(E\left[\left(p_j - p\right)\left(Y_i - E\left[Y_i\right]\right)|J_i = j\right]\right)}{\sum_{j=1}^k \lambda_j \left(p_j - p\right)^2} \\ = & \frac{\sum_{j=1}^k \lambda_j \left(E\left[\left(p_j - p\right)\left(Y_{ij} - \bar{Y}_i\right)|J_i = j\right]\right)}{\sum_{j=1}^k \lambda_j \left(p_j - p\right) E\left[\left(D_i\left(j\right) - \bar{D}_i\right)\right]}, \end{aligned}$$

where the second line follows from random assignment which implies $E\left[\bar{Y}_i | J_i = j\right] = E\left[\bar{Y}_i\right] = E\left[Y_i\right]$. Noting that $\lambda_j\left(p_j - p\right)$ is deterministic and that random assignment implies $E\left[Y_{ij} | J_i = j\right] = E\left[Y_{ij}\right]$, the IV estimand can be written:

$$\begin{aligned} & \frac{\sum_{j=1}^{k} \lambda_j\left(p_j - p\right)\left(E\left[Y_{ij}\right] - E\left[\bar{Y}_i\right]\right)}{E\left[\sum_{j=1}^{k} \lambda_j\left(p_j - p\right)\left(D_i\left(j\right) - \bar{D}_i\right)\right]} \\ = & \frac{E\left[\sum_{j=1}^{k} \lambda_j\left(p_j - p\right)\left(Y_{ij} - \bar{Y}_i\right)\right]}{E\left[\sum_{j=1}^{k} \lambda_j\left(p_j - p\right)\left(D_i\left(j\right) - \bar{D}_i\right)\right]} \\ = & \frac{E\left[\omega_i\left(\bar{Y}_i\left(1\right) - \bar{Y}_i\left(0\right)\right)\right]}{E\left[\omega_i\right]} \\ & + \frac{E\left[\sum_{j=1}^{k} \lambda_j\left(p_j - p\right)\gamma_{ij}\right]}{\sum_{j=1}^{k} \lambda_j\left(p_j - p\right)^2} \end{aligned}$$

where the first equality follows from the interchangeability of integration and summation, and the final equality from the definitions of Y_{ij}, $\bar{Y}_i$ and ω_i. ■

Formal Motivation for Test Statistics

This section shows that the judge-level test statistics,

$$\begin{aligned} \hat{T}_{1,j} &= \frac{\hat{y}_j - \tilde{y}_j - K\left(\hat{p}_j - \tilde{p}_j\right)}{\sqrt{\left(\hat{\sigma}_{Yj}^2 + K^2\hat{\sigma}_{Dj}^2 - 2K\hat{\sigma}_{YDj}\right)/n_j}}, \\ \hat{T}_{2,j} &= \frac{\hat{y}_j - \tilde{y}_j + K\left(\hat{p}_j - \tilde{p}_j\right)}{\sqrt{\left(\hat{\sigma}_{Yj}^2 + K^2\hat{\sigma}_{Dj}^2 + 2K\hat{\sigma}_{YDj}\right)/n_j}}, \end{aligned}$$

are proportional to the distance between the sample judge quantities $(\hat{p}_j, \hat{y}_j)$ and candidate population judge quantities $(\tilde{p}_j, \tilde{y}_j)$ when the slope constraints bind. To see this, note that the boundaries of the constraints satisfied by judge j's candidate population quantities $(\tilde{p}_j, \tilde{y}_j)$ are lines with slope $\pm K$ of the form

$$y = c \pm Kp.$$

The distance from the sample quantities $(\hat{p}_j, \hat{y}_j)$ to such a line is

$$d_j = \frac{\left|\hat{y}_j \mp K\hat{p}_j - c\right|}{\sqrt{1 + K^2}}.$$

When the constraint binds the line will pass through $(\tilde{p}_j, \tilde{y}_j)$, implying that $c = \tilde{y}_j \mp K\tilde{p}_j$. Substituting this in, the distance is

$$d_j = \frac{\left|\hat{y} - \tilde{y}_j \mp K\left(\hat{p} - \tilde{p}_j\right)\right|}{\sqrt{1 + K^2}},$$

which is proportional to $\left|\hat{T}_{1,j}\right|$ and $\left|\hat{T}_{2,j}\right|$. Thus, the test statistics are proportional to the distance between the sample quantities $(\hat{p}_j, \hat{y}_j)$ and the candidate population quantities $(\tilde{p}_j, \tilde{y}_j)$ when the slope constraints bind.

Testing Procedure with Covariates

The main text describes how covariates may be incorporated into the testing procedure in one of two ways: (1) assuming covariates have linear and separable effects on the outcome and treatment; and (2) performing the test within discrete covariate cells, which we'll refer to as a saturated specification. This section provides formal results to supplement the discussion in the text.

Separable, Linear Covariates

Let X_i be a $k \times 1$ vector of covariates such that Condition 1 in the main text holds. We assume in this section that the outcome and treatment depend on X_i linearly and separably:

$$E\left[Y_i | J_i = j, X_i\right] = \tilde{y}_j + X_i' \pi_Y, \tag{2}$$

$$E\left[D_i | J_i = j, X_i\right] = \tilde{p}_j + X_i' \pi_D. \tag{3}$$

Under this specification the Wald ratio between a pair of judges (j, j') is

$$\frac{E\left[Y_i | J_i = j, X_i\right] - E\left[Y_i | J_i = j', X_i\right]}{E\left[D_i | J_i = j, X_i\right] - E\left[D_i | J_i = j', X_i\right]} = \frac{\tilde{y}_j - \tilde{y}_{j'}}{\tilde{p}_j - \tilde{p}_{j'}}.$$

Theorem 1 implies that this Wald ratio will lie between $-K$ and K, where K is the width of the support of the outcome variable, which gives the following testable restrictions:

$$\left|\frac{\tilde{y}_j - \tilde{y}_{j'}}{\tilde{p}_j - \tilde{p}_{j'}}\right| \leq K$$

for any pair of judges (j, j'). This can be tested using a similar procedure to that described in the main text, but with the following adjustments. First, estimates of $\{\tilde{y}_j, \tilde{p}_j\}$ replace the estimated judge-level means $\{\hat{y}_j, \hat{p}_j\}$. The estimates of $\{\tilde{y}_j, \tilde{p}_j\}$ are obtained from regressions of Y_i and D_i on a set of judge dummies, Z_i, and covariates X_i. Let Z_i include dummies for each judge (no omitted category) and assume that the elements of X_i are mean zero and do not include a constant. Define $W_i = (Z_i, X_i')'$ and the least squares regression coefficients

$$\hat{\theta}_Y = \sum_{i=1}^{n} \left(W_i W_i'\right)^{-1} W_i Y_i,$$

$$\hat{\theta}_D = \sum_{i=1}^{n} \left(W_i W_i'\right)^{-1} W_i D_i.$$

Estimates of $\{\tilde{y}_j\}$ are the first J elements of $\hat{\theta}_Y$ and estimates of $\{\tilde{p}_j\}$ are the first J elements of $\hat{\theta}_D$. Under the assumptions of Theorem 2, the estimators $\hat{\theta} = \left(\hat{\theta}'_Y, \hat{\theta}'_D\right)'$ are jointly asymptotically normal:

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \rightarrow_d N\left(0, E\left[R_i R_i'\right]\right),$$

where

$$R_i = \begin{pmatrix} Q_W^{-1} W_i \left(Y_i - W_i'\theta_Y\right) \\ Q_W^{-1} W_i \left(D_i - W_i'\theta_D\right) \end{pmatrix}$$

and $Q_W = E\left[W_i W_i'\right]$. Let $\hat{\Sigma}$ be the sample version of $E\left[R_i R_i'\right]$. Note that the numerators of the statistics $T_{1,j}$ and $T_{2,j}$ are $G'_{1,j}\left(\hat{\theta} - \theta\right)$ and $G'_{2,j}\left(\hat{\theta} - \theta\right)$, respectively, where $G_{1,j}$ has a one in the jth component and a $-K$ in the $(J + k + j)$th component, and zeroes elsewhere and G_{2j} is the same, but with a K in the $(J + k + j)$th component. The full $2J \times 1$ vector of test statistic numerators is given by $G'\left(\hat{\theta} - \theta\right)$, where

$$G = \left[G_{1,1}, G_{2,1}, \ldots, G_{1,J}, G_{2,J}\right].$$

The vector of normalized test statistics is given by

$$\mathbf{T} = \begin{pmatrix} T_{1,1} \\ T_{2,1} \\ \vdots \\ T_{1,J} \\ T_{2,J} \end{pmatrix} = \sqrt{n \text{diag}\left(G'\hat{\Sigma}G\right)^{-1}} G'\left(\hat{\theta} - \theta\right),$$

where diag$(\cdot)$ indicates a diagonal matrix whose elements are the diagonals of the argument, and $\sqrt{\cdot}$ is elementwise. The vector of test statistics converges in distribution:

$$\begin{aligned} \mathbf{T} &\rightarrow_d N\left(0, \Omega\right), \\ \Omega &= \sqrt{\text{diag}\left(G'\hat{\Sigma}G\right)^{-1}} G' E\left[R_i R_i'\right] G \sqrt{\text{diag}\left(G'\hat{\Sigma}G\right)^{-1}}. \end{aligned}$$

Let $|\mathbf{T}|$ be the maximum absolute element of $\mathbf{T}$, and let $|Z|_{1-\alpha}$ be the $1-\alpha$ quantile of the maximum absolute element of a mean-zero multivariate normal random variable with covariance matrix Ω, which can be found by simulation. Note that $|\mathbf{T}|$ depends on the values $\{\tilde{y}_1, \tilde{p}_1, \ldots\}$. Let $\left|\hat{T}\right|$ be the optimized version of this, minimizing over $\{\tilde{y}_1, \tilde{p}_1, \ldots\}$ subject to the pairwise slope constraints, as in the main version of the test. The following result shows that a test based critical value $|Z|_{1-\alpha}$ controls size asymptotically.

Theorem 1 *Suppose the conditions of Theorem 2 hold, in addition to covariate specifications (2) and (3). Then* $\lim_{n\rightarrow\infty} \Pr\left(\left|\hat{T}\right| > |Z|_{1-\alpha}\right) \leq \alpha$.

Proof. By construction $\left|\hat{T}\right| \leq |\mathbf{T}|$, so by the convergence in distribution of $|\mathbf{T}|$ we have

$$\lim_{n\to\infty} \Pr\left(\left|\hat{T}\right| > |Z|_{1-\alpha}\right) \leq \lim_{n\to\infty} \Pr\left(|\mathbf{T}| > |Z|_{1-\alpha}\right) = \alpha.$$

■

Saturated Covariates

Suppose X_i takes on k values so that without loss of generality we can assume $X_i \in \{1, \ldots, k\}$ and that the number of observations in each covariate cell is large enough for asymptotic approximations to be appropriate. As described in the main text, denote the test statistic for the cell with $X_i = x$ as $\left|\hat{T}_x\right|$, and define the overall test statistic

$$\left|\hat{T}\right| = \max_x \left|\hat{T}_x\right|.$$

The critical value is $C_{1-\alpha} = F^{-1}(1-\alpha)$, where the test statistic's cdf is now $F(t) = \prod_{x,j} F_{x,j}(t)$, and

$$\begin{aligned} F_{x,j}(t) &= \Phi(t, t; \hat{\rho}_{x,j}) + \Phi(-t, -t; \hat{\rho}_{x,j}) - 2\Phi(-t, t; \hat{\rho}_{x,j}), \\ \hat{\rho}_{x,j} &= \frac{\hat{\sigma}^2_{Yx,j} - K^2\hat{\sigma}^2_{Dx,j}}{\sqrt{\left(\hat{\sigma}^2_{Yx,j} + K^2\hat{\sigma}^2_{Dx,j}\right)^2 - \left(2K\hat{\sigma}_{YDx,j}\right)^2}}, \end{aligned}$$

and $\hat{\sigma}^2_{Yx,j}, \hat{\sigma}^2_{Dx,j}, \hat{\sigma}_{YDx,j}$ are the sample variances and covariance of Y_i and D_i conditional on judge j and covariate cell x.

The following result shows that the test statistic described in the text controls size asymptotically.

Theorem 2 *Suppose the conditions of Theorem 2 hold conditional on* $X_i \in \{1, \ldots, k\}$ *and that* $\min_{x\in\{1,\ldots,k\}} \Pr(X_i = x) \geq \kappa$ *for some* $\kappa > 0$*. Then* $\lim_{n\to\infty} \Pr\left(\left|\hat{T}\right| > C_{1-\alpha}\right) \leq \alpha$.

Proof. Define $|T_n| = \max_x |T_{x,n}|$, where $|T_{x,n}|$ is the cell-specific test statistic evaluated at the true quantities $\{y_{j,x}, p_{j,x}\}$. Note that $F(t)$ defined above is the limiting distribution of $|T_n|$. By construction $\left|\hat{T}\right| \leq |T_n|$, so we have

$$\lim_{n\to\infty} \Pr\left(\left|\hat{T}\right| > C_{1-\alpha}\right) \leq \lim_{n\to\infty} \Pr\left(|T_n| > C_{1-\alpha}\right) = \alpha.$$

■

Exact Finite Sample Testing Procedure

The nonparametric test in the main text relies on an asymptotic approximation to the distribution of the test statistic that converges as the number of observations per judge grows. In applications of the judge fixed effects design with few observations per judge, or where conditioning on covariates necessitates running the test in small cells, a natural question is whether the approximation is accurate.

In this section we adapt the nonparametric test to the case where the number of observations per judge may be small. We derive the exact finite-sample distribution of the test statistic and show the approximation used in the main text is very accurate, even for observations per judge much fewer than is common in applications.

The main challenge to overcome when the number of observations per judge is small is that the distribution of the judge-level sample quantities $(\hat{y}_j, \hat{p}_j)$ may not be well approximated by a normal distribution. Fortunately, when the outcome is binary, as in our application, the distribution of judge-level sample quantities is exactly characterized by a multinomial distribution. When the outcome is not binary, it can be replaced by a binary indicator for the outcome exceeding some value. Without loss of generality, therefore, we assume in this section that the outcome is binary.

Define the observed quantities $W = YD$ and $V = Y(1-D)$, and let $w_j = E[W|J=j]$, and $v_j = E[V|J=j]$. Because $Y = W + V$, we have $y_j = w_j + v_j$. Treatment effects on W are bounded between zero and one, and treatment effects on V are between negative one and zero. This puts the following constraints on (p_j, v_j, w_j) within and across judges:

$$\left\{\begin{array}{c} 0 \leq p_j \leq 1, \\ 0 \leq w_j \leq p_j, \\ 0 \leq v_j \leq 1 - p_j, \end{array}\right\}_{j=1}^{J}$$

$$\left\{\left\{\begin{array}{c} 0 \leq \frac{w_j - w_{j'}}{p_j - p_{j'}} \leq 1, \\ -1 \leq \frac{v_j - v_{j'}}{p_j - p_{j'}} \leq 0, \end{array}\right\}_{j=1}^{J-1}\right\}_{j'=j+1}^{J}.$$

All the sample information for a given judge is captured in the following two-by-two table:

$$\begin{array}{cc|c|c|} & & \multicolumn{2}{c}{D} \\ & & 0 & 1 \\ \cline{3-4} Y & 0 & A_j, a_j & B_j, b_j \\ \cline{3-4} & 1 & V_j, v_j & W_j, w_j \\ \cline{3-4} & & 1 - p_j & p_j \end{array}.$$

In each cell, the capitalized quantity (e.g., A_j) represents the random (multinomial) count in that cell, and the lower case quantity represents the probability associated with that cell. Note that we have

$$\begin{aligned} a_j &= 1 - p_j - v_j \\ b_j &= p_j - w_j, \\ \hat{y}_j &= \left(V_j + W_j\right)/n_j, \\ \hat{p}_j &= \left(W_j + B_j\right)/n_j. \end{aligned}$$

Our testing approach is based on the similar statistics to those used in the main text:

$$\begin{aligned} \hat{T}_{1,j} &= \frac{\hat{y}_j - \tilde{y}_j - K\left(\hat{p}_j - \tilde{p}_j\right)}{\sqrt{\left(\tilde{\sigma}^2_{Y,j} + K^2\tilde{\sigma}^2_{D,j} - 2K\tilde{\sigma}_{YD,j}\right)/n_j}}, \\ \hat{T}_{2,j} &= \frac{\hat{y}_j - \tilde{y}_j + K\left(\hat{p}_j - \tilde{p}_j\right)}{\sqrt{\left(\tilde{\sigma}^2_{Y,j} + K^2\tilde{\sigma}^2_{D,j} + 2K\tilde{\sigma}_{YD,j}\right)/n_j}}, \end{aligned}$$

where

$$\begin{aligned} \tilde{\sigma}^2_{Y,j} &= \tilde{y}_j\left(1 - \tilde{y}_j\right), \\ \tilde{\sigma}^2_{D,j} &= \tilde{p}_j\left(1 - \tilde{p}_j\right), \\ \tilde{\sigma}_{YD,j} &= \tilde{w}_j - \tilde{y}_j\tilde{p}_j. \end{aligned}$$

As before, we take the larger of the absolute values of these statistics:

$$\left|\hat{T}_j\left(\hat{y}_j, \hat{p}_j\right)\right| = \max\left\{\left|\hat{T}_{1j}\right|, \left|\hat{T}_{2j}\right|\right\}.$$

Let the function $g_j\left(t; \tilde{w}_j, \tilde{v}_j, \tilde{p}_j\right)$ be the probability that $\hat{T}_j$ is strictly less than some value t, given that $\tilde{w}_j$, $\tilde{v}_j$, and $\tilde{p}_j$ are the true parameters; that is,

$$\begin{aligned} & g_j\left(t; \tilde{w}_j, \tilde{v}_j, \tilde{p}_j\right) = \\ & \sum_{v=0}^{n_j} \sum_{b=0}^{n_j - v} \sum_{w=0}^{n_j - v - b} 1\left(\left|\hat{T}_j\left(\left(v + w\right)/n_j, \left(w + b\right)/n_j\right)\right| < t\right) f\left(x_j\left(v, b, w\right); n_j, \phi\left(\tilde{w}_j, \tilde{v}_j, \tilde{p}_j\right)\right), \end{aligned}$$

where f is the multinomial pmf and

$$\begin{aligned} x_j\left(v, b, w\right) &= \begin{pmatrix} v \\ b \\ w \\ n_j - v - b - w \end{pmatrix}, \\ \phi\left(\tilde{w}_j, \tilde{v}_j, \tilde{p}_j\right) &= \begin{pmatrix} \tilde{v}_j \\ \tilde{p}_j - \tilde{w}_j \\ \tilde{w}_j \\ 1 - \tilde{v}_j - \tilde{p}_j \end{pmatrix}. \end{aligned}$$

As before, we take the maximum deviation across judges:

$$\left|\hat{T}\left(\{\tilde{w}_j,\tilde{v}_j,\tilde{p}_j\}\right)\right| = \max_j \left|\hat{T}_j\right|.$$

Define the p-value for the test:

$$\begin{aligned}
\hat{P} &= \max_{\{\tilde{w}_j,\tilde{v}_j,\tilde{p}_j\}} 1 - \prod_j g_j\left(\left|\hat{T}\left(\{\tilde{w}_j,\tilde{v}_j,\tilde{p}_j\}\right)\right|; \tilde{w}_j,\tilde{v}_j,\tilde{p}_j\right) \\
& \quad \text{s.t.} \left\{ \begin{array}{c} 0 \leq w_j \leq p_j, \\ 0 \leq v_j \leq 1 - p_j, \end{array} \right\}_{j=1}^{J}, \\
& \quad \left\{ \left\{ \begin{array}{c} 0 \leq \frac{w_j - w_{j'}}{p_j - p_{j'}} \leq 1, \\ -1 \leq \frac{v_j - v_{j'}}{p_j - p_{j'}} \leq 0, \end{array} \right\}_{j=1}^{J-1} \right\}_{j'=j+1}^{J}
\end{aligned}$$

The following result shows that a level α test rejects if $\hat{P} < \alpha$.

Theorem 3 *Let $\{Y_i, J_i, D_i\}_{i=1}^{n}$ be an iid sample where Y_i has compact support, J_i has support $\{1,\ldots,J\}$. If Condition 1 holds,* $\Pr\left(\hat{P} < \alpha\right) \leq \alpha$.

Proof. Define $|T| = \left|\hat{T}\left(\{w_j, v_j, p_j\}\right)\right|$ and $P = 1 - \prod_j g_j\left(|T|; w_j, v_j, p_j\right) = 1 - G\left(|T|\right)$. Note that given Condition 1, $\hat{P} \geq P$ by construction. Then

$$\begin{aligned}
\Pr\left(\hat{P} < \alpha\right) &\leq \Pr\left(P < \alpha\right) \\
&= \Pr\left(G\left(|T|\right) > 1 - \alpha\right) \\
&= 1 - \Pr\left(G\left(|T|\right) \leq 1 - \alpha\right) \\
&\leq 1 - (1 - \alpha) \\
&= \alpha,
\end{aligned}$$

The first inequality follows because $\hat{P} \geq P$ by construction. The second to last line holds with weak inequality because $|T|$ has discrete support. ■

Thus, the exact finite-sample adaptation of the nonparametric test controls size in finite samples. Why, then, do we rely on an asymptotic approximation to the test statistic's distribution in the main text? There are considerable computational advantages to the approximation. Namely, under the asymptotic approximation, the test statistic can be computed by efficient linear programming algorithms, while the exact finite sample version cannot be. The reason is that in the finite-sample version of the test statistic, the normalizing factors in the denominators depend on the choice parameters, and thus the optimization is nonlinear. Under the asymptotic approximation, the normalizing factors can be replaced by consistent estimates.

We now show that the asymptotic approximation converges quickly to the exact finite-sample distribution of the test statistic. We thus gain a sizeable computational advantage without material size distortions for the number of cases per judge typical in applications, including ours. We compute the exact cdf of the judge-level test statistic for several values of the parameters (p_j, w_j, v_j), as defined in the notation above. We set $K = 1$ throughout. In the first set of simulations, we set the parameters to match the average judge in our empirical application: $p_j = .4, w_j = .31, v_j = .39$, and we examine how the exact distribution compares to the asymptotic approximation. Figure 1 plots the exact distribution of the test statistic along with its asymptotic approximation for several values of cases per judge: 5, 10, 20, and 50. For reference, in our application there are on average about 1,000 observations per judge, and the minimum is 50. Restricting to judges with at least 20 or 50 cases is nearly universal practice in this research design. The figure shows that the approximation converges quickly to the true distribution. The discreteness of the test statistic is evident in the step-like nature of the exact distribution, but systematic distortions are minimal, especially in the upper tail of the distribution where rejection occurs. By 50 observations per judge (the minimum in our application) the differences between the exact distribution and the approximation are minimal.

It is well known that the normal approximation to the binomial distribution is most accurate when the binomial probabilities are closest to one-half. The normal approximation to the multinomial distribution here may also depend on the underlying parameters. We repeat the simulation exercise above for additional sets of parameter values, one constructed to be unfavorable to the normal approximation, with multinomial probabilities near zero or one, and one that is favorable to the normal approximation with probabilities evenly distributed. The unfavorable parameter values are $p_j = .1, w_j = .05, v_j = .8$, and the favorable values are $p_j = .5, w_j = v_j = .25$. We focus on how the exact 95th percentile of the test statistic compares to the approximation, since it is the tail percentiles that are relevant for controlling size test. Figure 2 plots the exact and approximate 95th percentiles of the test statistic $\left|\hat{T}_j\right|$ as a function of the number of cases per judge for each of the three scenarios: parameters calibrated to the empirical example, parameters unfavorable to the normal approximation, and parameters favorable to the empirical example. In all cases the exact converges to the approximation quickly, and the normal distribution offers an accurate approximation for as few as 100 cases per judge even in the least favorable scenario. Given the significant computational advantage, we recommend the test based on the approximation for use in applications.

Relationship with Testable Implications in Kitagawa (2015)

Kitagawa (2015) develops a test for IV validity in the same heterogeneous treatment effects framework we consider. The principal difference between the two settings is that Kitagawa (2015) assumes the order of the instruments in terms of treatment propensity is known a priori, while we do not require this knowledge. Another apparent difference in the two settings is in the testable implications exploited by the tests. Kitagawa (2015) is based on the restriction that the implied densities of potential outcomes for compliers be nonnegative everywhere. Ours is based on the restriction that the implied average treatment effect among compliers be within the bounds allowed by the support of the outcome. Kitagawa (2015) shows that the complier density condition is optimal in that it exhausts the possible testable implications of instrument validity. Here we show that our condition is equivalent to Kitagawa's for a suitably defined collection of outcomes. Our testable implication therefore inherits the same optimality.

Without loss of generality, consider a single pair of judges, where $Z = 0$ indicates the judge with the lower treatment propensity, and $Z = 1$ indicates the judge with the higher propensity. The testable implication in Kitagawa (2015) is

$$\begin{aligned} P(B,1) - Q(B,1) &\geq 0, \\ Q(B,0) - P(B,0) &\geq 0 \end{aligned}$$

for all Borel sets $B \subseteq \mathcal{Y}$, where

$$\begin{aligned} P(B,d) &= \Pr(Y \in B, D = d | Z = 1), \\ Q(B,d) &= \Pr(Y \in B, D = d | Z = 0). \end{aligned}$$

Define the observed variables $W_B := 1(Y \in B) D$ and $V_B := -1(Y \in B)(1 - D)$ for Borel set B. Note that for both of these variables, the possible support of treatment effects is $\{0, 1\}$. Our complier treatment effect restriction therefore yields:

$$\begin{aligned} \frac{E[W_B|Z=1] - E[W_B|Z=0]}{E[D|Z=1] - E[D|Z=0]} &\geq 0, \\ \frac{E[W_B|Z=1] - E[W_B|Z=0]}{E[D|Z=1] - E[D|Z=0]} &\leq 1, \\ \frac{E[V_B|Z=1] - E[V_B|Z=0]}{E[D|Z=1] - E[D|Z=0]} &\geq 0, \\ \frac{E[V_B|Z=1] - E[V_B|Z=0]}{E[D|Z=1] - E[D|Z=0]} &\leq 1. \end{aligned}$$

We now show that these implications imply and are implied by Kitagawa's (2015) implications. Recalling that $E[D|Z=1] - E[D|Z=0] > 0$ by definition, the first and third of these inequalities simplify to

$$\begin{aligned} E[W_B|Z=1] - E[W_B|Z=0] &\geq 0, \\ E[V_B|Z=1] - E[V_B|Z=0] &\geq 0. \end{aligned}$$

Noting that $E\left[W_B|Z=1\right] = P\left(B,1\right)$, $E\left[W_B|Z=0\right] = Q\left(B,1\right)$, $E\left[V_B|Z=1\right] = -P\left(B,0\right)$, and $E\left[V_B|Z=0\right] = -Q\left(B,0\right)$, these inequalities are precisely Kitagawa's restrictions. Similarly, the second and fourth inequalities simplify to

$$\begin{aligned} E\left[D\left(1-1\left(Y\in B\right)\right)|Z=1\right]-\left(E\left[D\left(1-1\left(Y\in B\right)\right)|Z=0\right]\right) &\geq 0, \\ E\left[\left(1-D\right)\left(1-1\left(Y\in B\right)\right)|Z=0\right]-E\left[\left(1-D\right)\left(1-1\left(Y\in B\right)\right)|Z=1\right] &\geq 0. \end{aligned}$$

These are Kitagawa's restrictions where B is replaced by the complement of B. Thus, our restrictions applied to W_B and V_B imply and are implied by Kitagawa's restrictions applied to Borel set B and its complement.

It is conceptually straightforward to extend our test to enable it to detect violations of instrument validity at any Borel set. In the spirit of Kitagawa (2015), let $T_W\left(y,y'\right)$ be our test statistic applied to outcome $W_{[y,y']}$, where $y \leq y'$ are values in the support of Y. Let $T_V\left(y,y'\right)$ be our test statistic applied to outcome $V_{[y,y']}$. The overall test statistic is

$$T = \sup_{-\infty\leq y\leq y'\leq\infty} \max\left\{T_W\left(y,y'\right), T_V\left(y,y'\right)\right\}.$$

The critical values for this test statistic can be simulated in a straightforward manner. Like the Kitagawa (2015) test, this procedure has asymptotic power against any alternative that violates the testable implications of instrument validity, but does not require a priori knowledge of the instrument ordering. We do not fully explore this extension because it would be computationally prohibitive, but note the connection between the testable implications exploited in our approach and Kitagawa's (2015) approach.

Semiparametric Test

The nonparametric test proposed in the main text imposes minimal structure on treatment effects beyond bounds on their magnitude. This means the test can fail to reject because it implicitly allows wildly fluctuating treatment effects from judge to judge which may not be plausible. If one is willing to impose more smoothness on the relationship between treatment effects and judge propensity, one can construct a more powerful test. In this section we propose a semiparametric approach that allows researchers to impose additional structure. This results in a more powerful test, but requires additional assumptions and a larger number of cases per judge for the accurate inference than is necessary for the nonparametric test.

Like the nonparametric test developed in the main text, the semiparametric test is based on two observations that follow from Theorem 1: (1) average outcomes conditional on judge assignment should fit a continuous function of judge propensities; and (2) the slope of that continuous function should be bounded in magnitude by the width of the outcome variable's support. The semiparametric test allows researchers to posit a flexible form for the relationship between judge-level average outcomes and propensities, and then examines whether observed outcomes averaged by judge are consistent with such a function.

Figure 3 illustrates graphically the intuition behind the test. The top panel depicts a situation in which the assumptions are satisfied, so that average outcomes by judge lie on a continuous function of judge propensity, and the slope of that function is within the required bounds. The bottom panel illustrates two ways that violations of the assumptions may appear. In the first (labeled "A" on the figure), two judges have identical propensities, but different average outcomes; thus no continuous function can pass through both points. In the second (labeled "B"), two adjacent judges do not have identical propensities, but their average outcomes are sufficiently different that the slope of the curve connecting them exceeds the possible treatment effect values.

This suggests a conceptually straightforward procedure for testing the judge fixed effects design's assumptions:

1. Regress the outcome Y_i on a flexible function of the judge propensity, $\phi\left(p_{J_i}\right)$
2. Jointly test fit and slope by
   (a) Regressing the residuals from step 1, $u_i = Y_i - \phi\left(p_{J_i}\right)$, on judge indicators and testing whether the coefficients are jointly zero;
   (b) Testing whether the function's slope stays within the bounds dictated by the support of Y_i.

To implement this process in our application, we regress the indicator for whether a defendant was convicted on a flexible function of judge severity. Visually, this means fitting a function to the observed judge severity and average outcome points plotted in Figure 2 in the main text. Our test will assess whether judge assignment has significant explanatory power over outcomes beyond the predictions from the fitted function, and whether the implied treatment effects are too big anywhere based on the slope of the fitted function.

The procedure presents two complications. The first is specifying the propensity regression in step 1. The step 1 regression of the outcome on the judge propensities should be as flexible as the researcher's assumptions regarding treatment effect heterogeneity. A linear regression imposes constant treatment effects and makes the test procedure above equivalent to the usual Sargan-Hansen overidentification test (Sargan, 1958; Hansen, 1982). In many applications, including all those with a binary outcomes, constant treatment effects are unlikely or impossible. To impose less structure on treatment effect heterogeneity, Theorem 1 suggests one should choose a flexible specification that approximates Lipschitz functions well, such as polynomials or splines (Chen, 2007). Our simulations and application use b-splines (see Racine, 2018), but other bases could be used as well.

We now formalize this intuition. Let the number of terms in the chosen series be $m+1$, and let the function class in which the chosen specification lies be denoted $\mathcal{S}_m$; for example, degree-m polynomials or degree-r splines with $m-r$ knots. In the context of the judges design, the number of terms in the approximating series is limited by the number of judges; settings with a large number of judges, such as our application, allow the specification to be quite flexible.

The second complication is accounting for the estimation of the judge propensities and the step 1 residuals when performing the tests in step 2. The simplest estimator for the propensity of judge J assigned to individual i, p_{J_i}, is simply the fitted value from a regression of treatment status D_i on a vector of judge indicators $W_i = (1, 1(J_i = 1), \ldots, 1(J_i = J))'$, which amounts to the fraction treated among individuals assigned to judge j, although it may be generalized by adding controls to the first stage regression. Denote the estimated fitted values $\hat{P}_i$. The first-step residuals also depend on a linear regression coefficient: collecting the terms of the spline (or whichever basis is chosen) in the estimated propensity into the vector $\hat{S}_i$, the estimated residual for the i-th observation is

$$\hat{u}_i = Y_i - \hat{S}_i' \left(\sum_{i=1}^{n} \hat{S}_i \hat{S}_i' \right)^{-1} \sum_{i=1}^{n} \hat{S}_i Y_i.$$

The fit component of our test is based on the second-step coefficients obtained by regressing $\hat{u}_i$ on W_i:

$$\hat{\gamma} = \left(\sum_{i=1}^{n} W_i W_i' \right)^{-1} \sum_{i=1}^{n} W_i \hat{u}_i.$$

Under the conditions of Theorem 1 and the posited functional form, $\hat{\gamma}$ converges in probability to zero. Our procedure tests this via the following Wald statistic:

$$\hat{T} = n \hat{\gamma}' \hat{\Omega}^{-1} \hat{\gamma}, \tag{4}$$

where $\hat{\Omega}$ is a consistent estimator of the limiting covariance of $\sqrt{n}\hat{\gamma}$, accounting for the first-step estimates $\left(\sum_{i=1}^{n} \hat{S}_i \hat{S}_i' \right)^{-1} \sum_{i=1}^{n} \hat{S}_i Y_i$ and $\hat{P}_i$. We derive a suitable estimator for $\hat{\Omega}$, given an iid sample, in the appendix.

Given the assumptions so far, the test statistic (4) converges in distribution to a chi-squared random variable with degrees of freedom equal to the difference between the number of judges and the number of terms in the specification for ϕ, as the following theorem formalizes:

Theorem 4 *Suppose Condition 1 holds and $\phi \in \mathcal{S}_m$, where $m < J$. Suppose further that $\{Y_i, D_i, J_i\}_{i=1}^n$ comprise an iid sample and $E\left[|Y_i|^3\right] < \infty$. Then*

$$\hat{T} \underset{d}{\rightarrow} \chi^2 (J - m).$$

Proof. Define the estimated judge propensity to treat as

$$\begin{aligned}\hat{p}_{J_i} &= W_i'\hat{\alpha}, \\ \hat{\alpha} &= \left(\sum_{i=1}^{n} W_i W_i'\right)^{-1} \sum_{i=1}^{n} W_i D_i.\end{aligned}$$

Define $v_i := D_i - W_i'\alpha$ and $u_i := Y_i - S_i'\delta$, where S_i is a vector powers of $p_{J_i} := W_i'\alpha$, judge J_i's (population) propensity to treat, and δ is the vector of coefficients from the population regression of Y_i on S_i. Write $S_i'\delta := f(\lambda, W_i)$, where $\lambda = (\alpha', \delta')'$. Letting

$$\hat{\delta} = \left(n^{-1}\sum_{i=1}^{n} \hat{S}_i\hat{S}_i'\right)^{-1} n^{-1}\sum_{i=1}^{n} \hat{S}_i Y_i,$$

we can write $\hat{u}_i = Y_i - f\left(\hat{\lambda}, W_i\right)$ where $\hat{\lambda} = \left(\hat{\alpha}', \hat{\delta}'\right)'$, which has limiting behavior as follows:

$$\sqrt{n}\left(\hat{\lambda} - \lambda\right) = n^{-1/2}\sum_{i=1}^{n}\left(\begin{array}{c} Q_W^{-1} W_i v_i \\ Q_S^{-1} S_i u_i \end{array}\right) + o_p(1),$$

where for some random vector A_i we adopt the notation $Q_A := E\left[A_i A_i'\right]$. By a mean value expansion we can write $\hat{u}_i = u_i - \nabla\left(\tilde{\lambda}, W_i\right)'\left(\hat{\lambda} - \lambda\right)$, where $\nabla(\lambda, W_i)$ is the Jacobian of $f(\lambda, W_i)$ with respect to λ,

$$\nabla(\lambda, W_i) = \left(\begin{array}{c} W_i\Delta_i'\delta \\ S_i \end{array}\right)$$

and

$$\hat{\Delta}_i = \left(\frac{dS_0(\hat{p}_i)}{dp}, \ldots, \frac{dS_m(\hat{p}_i)}{dp}\right)'.$$

The estimator on which the test statistic is based can therefore be expanded as:

$$\begin{aligned}\sqrt{n}\hat{\gamma} &= \sqrt{n}\left(n^{-1}\sum_{i=1}^{n} W_i W_i'\right)^{-1} n^{-1}\sum_{i=1}^{n} W_i\hat{u}_i \\ &= Q_W^{-1}\left(n^{-1/2}\sum_{i=1}^{n} W_i u_i - r_i\right) + o_p(1),\end{aligned}$$

where

$$r_i = E\left[W_i\left(\begin{array}{c} W_i\Delta_i'\delta \\ S_i \end{array}\right)'\right]\left(\begin{array}{c} Q_W^{-1} W_i\left(D_i - W_i'\alpha\right) \\ Q_S^{-1} S_i u_i \end{array}\right),$$

a consistent estimator for which is

$$\hat{R}_i = \left(n^{-1}\sum_{j=1}^{n} W_j\left(\begin{array}{c} W_j\Delta_j'\hat{\delta} \\ \hat{S}_j \end{array}\right)'\right)\left(\begin{array}{c} \hat{Q}_W^{-1} W_i\left(D_i - \hat{p}_i\right) \\ \hat{Q}_S^{-1}\hat{S}_i\hat{u}_i \end{array}\right). \qquad (5)$$

By the central limit theorem we therefore have

$$\sqrt{n}\hat{\gamma} \underset{d}{\rightarrow} N\left(0, \Omega\right),$$

where

$$\Omega = Q_W^{-1} Var\left(W_i u_i - r_i\right) Q_W^{-1}$$

is consistently estimated by

$$\hat{\Omega} = \left(n^{-1}\sum_{i=1}^{n} W_i W_i'\right)^{-1}\left(n^{-1}\sum_{i=1}^{n}\left(W_i\hat{u}_i - \hat{R}_i\right)\left(W_i\hat{u}_i - \hat{R}_i\right)'\right)\left(n^{-1}\sum_{i=1}^{n} W_i W_i'\right)^{-1},$$

The quadratic form

$$n\hat{\gamma}'\hat{\Omega}^{-1}\hat{\gamma}$$

is therefore asymptotically a chi-squared random variable with degrees of freedom equal to the rank of $\hat{\Omega}^{-1}$, in this case $k - m$. ■

Performing the fit component of the test means computing the test statistic and obtaining the associated p-value from the appropriate chi-squared distribution.

The slope component of the test examines whether the slopes of the function relating outcomes to judge propensities lie between $-K$ and K, recalling that K is the width of the outcome variable's support. The function relating average outcomes given judge assignment to judge propensities is specified as

$$\phi\left(p\right) = \delta_0 S_0\left(p\right) + \cdots + \delta_m S_m\left(p\right),$$

where $S_0, \ldots, S_m$ are elements of a polynomial series, spline series, or whichever basis is chosen for ϕ. When ϕ is specified as a quadratic b-spline, the maximum slope occurs at one of the knots, $\{t_0 = 0, t_1, \ldots, t_{m-2}, t_{m-1} = 1\}$. The slope at the l-th knot is given by

$$\phi'\left(t_l\right) = \frac{2}{t_{l+1} - t_{l-1}}\left(\delta_{l+1} - \delta_l\right), \quad l = 0, \ldots, m-1,$$

where we define $t_{-1} = t_0 = 0$ and $t_{m-1} = t_m = 1$. The restriction on the slope of ϕ corresponds to the following set of inequality constraints:

$$\left\{-K \leq \frac{2}{t_{l+1} - t_{l-1}}\left(\delta_{l+1} - \delta_l\right) \leq K\right\}_{l=0}^{m-1}.$$

Given estimates $\hat{\delta} = \left(\sum_{i=1}^{n}\hat{S}_i\hat{S}_i'\right)^{-1}\sum_{i=1}^{n}\hat{S}_i Y_i$ and the corresponding variance matrix that accounts for the estimation of $\hat{P}_i$, we implement the moment inequality testing procedure proposed by Andrews and Soares (2010). This procedure first performs generalized moment selection to eliminate inequalities that are far from binding, and then constructs a modified method of moments (MMM) test statistic to test the remaining inequalities. The following appendix section describes the details of the implementation.

Finally, we combine the fit component and slope component of the test via a Bonferroni procedure to produce a single joint test. In the "just identified" case when there are only two judges, the fit component of the test will have no power. As the number of judges grows, the specification of ω becomes more flexible and the number of inequalities being tested in the slope component grows, causing the slope component of the test to lose power as the fit component gains power.[1]

In Table 1 we present the results of our test in the context of our empirical example. Again, we control for the same set of covariates as in our baseline specification. We see that our semiparametric test rejects the null hypothesis on the full sample for various numbers of knots in the spline function. Indeed, we reject the null hypothesis even when our assumed function form is quite flexible with 15 knot points. It is only when we increase the number of knot points to 20 that we fail to reject the null hypothesis.[2]

Extensions

The proposed tests have power against alternatives that manifest themselves in shifts in the mean of Y_i, but will not have power against alternatives where other features of Y_i are changed but not the mean. The tests naturally extend to have power against shifts in other features of the distribution, as well by replacing the outcome with a set of indicator variables of the form $1\left(Y_i \leq y_j\right)$ for a grid of $\{y_j\}$ values. For the semiparametric version specifically, one can also replace the mean regression with a set of quantile regressions of Y_i with a grid of quantile values $\tau_j \in (0,1)$. The test then consists of jointly testing the hypothesis that the coefficients on the instrument dummies are zero across all quantile regressions. In this case and the dummy dependent variable alternative above, the test is carried out using a variance matrix accounting for the estimation of $\hat{p}\left(J_i\right)$, analogous to the procedure described in the main text. These extensions allow the test to have power against a wider array of alternatives, although at the expense of more computational burden and perhaps

[1] We recommend putting equal weight on each of the test components, as we do in our application. However, it is possible to adjust the weighting scheme. If we denote the p-value from the fit component of the test as p_f and the p-value from the slope-component of the test as p_s, then a joint level-α test rejects if either $p_f < \omega\alpha$ or $p_s < (1-\omega)\,\alpha$, for some weight $\omega \in [0,1]$. Equivalently, one can define a joint p-value as $\min\left\{p_f/\omega, p_s/\left(1-\omega\right)\right\}$ and reject if the joint p-value is less than α. The choice of ω governs the direction of power between the fit component and the slope component, with values near one directing more power to the fit component of the test.

[2] If one suspects that monotonicity violations are violated across observable groups but hold within these groups, one can test jointly for monotonicity within the groups. Assuming independence of the subsamples, one can add up the chi-squared test statistics and degrees of freedom after running the test on all subsamples defined by the relevant observables to get the joint test statistic and its chi-squared degrees of freedom.

a lack of specific power against alternatives where only the mean is shifted.

Generalized Moment Selection Implementation

The slope component of the semiparametric test implements the moment inequality testing procedure proposed by Andrews and Soares (2010). This procedure is based on the following modified method of moments (MMM) test statistic:

$$\hat{M} = \sum_{l=0}^{m-1} \left(\left[\frac{K - \hat{\phi}'(t_l)}{s.e.\left(\hat{\phi}'(t_l)\right)} \right]_-^2 + \left[\frac{K + \hat{\phi}'(t_l)}{s.e.\left(\hat{\phi}'(t_l)\right)} \right]_-^2 \right),$$

where $[x]_- = x1(x<0)$,

$$\hat{\phi}'(t_l) = \frac{2}{t_{l+1} - t_{l-1}} \left(\hat{\delta}_{l+1} - \hat{\delta}_l \right),$$

$$s.e.\left(\hat{\phi}'(t_l)\right) = n^{-1/2} \frac{2}{t_{l+1} - t_{l-1}} \left(\hat{\Sigma}_{l+1,l+1} + \hat{\Sigma}_{l,l} - 2\hat{\Sigma}_{l+1,l} \right)^{1/2},$$

and $\hat{\Sigma}$ is a consistent estimator of the variance matrix of $\hat{\delta} = \left(\sum_{i=1}^n \hat{S}_i \hat{S}_i' \right)^{-1} \sum_{i=1}^n \hat{S}_i Y_i$ that takes into account estimation of $\hat{P}_i$:

$$\hat{\Sigma} = \hat{Q}_S^{-1} \left(\sum_{i=1}^n \left(\hat{S}_i \hat{u}_i - \hat{\Delta}_i W_i' \hat{Q}_W^{-1} W_i \hat{v} \right) \left(\hat{S}_i \hat{u}_i - \hat{\Delta}_i W_i' \hat{Q}_W^{-1} W_i \hat{v} \right)' \right) \hat{Q}_S^{-1}.$$

Under the regularity conditions described in Andrews and Soares (2010), the distribution of the MMM test statistic can be approximated by the distribution of

$$\hat{M}^* = \sum_{l \in \mathcal{L}^-} [Z_l^*]_-^2 + \sum_{l \in \mathcal{L}^+} [-Z_l^*]_-^2,$$

where Z^* is an m-element multivariate normal random variable with unit variances and correlation matrix corresponding to the asymptotic variance of

$$\left([0_{m\times 1} : I_m] - [I_m : 0_{m\times 1}] \right) \hat{\delta},$$

and the moments selected by the generalized moment selection are given by:

$$\mathcal{L}^- = \left\{ l : \frac{K - \hat{\phi}'(t_l)}{s.e.\left(\hat{\phi}'(t_l)\right)} \leq \sqrt{\ln n} \right\},$$

and

$$\mathcal{L}^{+} = \left\{ l : \frac{K + \hat{\phi}'(t_l)}{s.e.\left(\hat{\phi}'(t_l)\right)} \leq \sqrt{\ln n} \right\}.$$

The p-value from the slope component of the test can be found to arbitrary precision by simulating many multivariate draws, constructing $\hat{M}^*$ for each draw, and computing the fraction of draws for which $\hat{M}^* \geq \hat{M}$.

Simulations

Nonparametric version

We examine the finite-sample performance of the nonparametric test by applying it to data generated from several different processes that mimic our empirical example. In our simulated data, J judges are endowed with propensity to treat p_j, $j = 1, \ldots, J$. Each judge handles n_j defendants. The propensities p_j are drawn from the following distribution, calibrated to match our empirical example:

$$p_j \sim \underline{p} + \left(\bar{p} - \underline{p}\right) B\left(a, b\right),$$

where $\underline{p} = .2$ is the lowest propensity observed in the data, $\bar{p} = .7$ is the highest propensity observed in the data, and $B\left(a, b\right)$ is a beta-distributed random variable with parameters $a \approx 6.67$ and $b \approx 4.44$ chosen to match the observed mean and interquartile range of propensities. The caseloads n_j were drawn from the following distribution, also calibrated to match the empirical example:

$$n_j \sim \underline{n} + NB\left(r, p\right),$$

where $\underline{n}$ is the minimum caseload cutoff (50 in the baseline case) and $NB\left(r, p\right)$ is a negative binomial random variable with parameters $r = .59$ and $p = 1 - \left(\bar{n} - \underline{n}\right) / \left(\bar{n} - \underline{n} + r\right)$, where $\bar{n}$ is the average number of cases per judge (around 1,000 in the baseline case). The negative binomial parameters r and p were chosen this way to match the mean and IQR of caseloads in the baseline case with our empirical example. For individual i assigned to judge j, treatment status was assigned as $D_i = 1\left(V_i < p_j\right)$, where V_i is a uniformly distributed random variable.

In the first set of simulations we examine the finite-sample size of our test in a scenario where the exclusions and monotonicity conditions are satisfied. To this end the outcome was defined as $Y_i = 1\left(U_i < \beta_0 + \beta_1 p_j\right)$, where U_i is a uniformly distributed random variable, $\beta_0 = .644$, and $\beta_1 = .14$ to match the two-stage least squares estimates from our empirical example. We vary the number of judges J from 10 to 256 (the number in our empirical example), the average number of cases per judge $\bar{n}$ from 50 to 1,000, and the minimum number of cases per judge from zero to 50. For each set of parameters we generate 999 simulated datasets, apply our nonparametric test, and record the rejection rate. In investigating the size of the test we make a modification to our nonparametric test to avoid mechanically reducing the rejection rate. The fact that our testing procedure searches over parameter values to minimize the test statistic builds in some finite-sample conservativity mechanically. We remove this mechanical conservativity by calculating the test statistic relative to the true parameter values, without optimizing. This allows us to directly assess the accuracy of the large sample approximations in finite samples. Naturally, we do not remove the conservativity when we assess power.

The first set of simulations shows that the test maintains accurate size as we vary the number of judges and the average number of casees per judge. Table 2 reports the rejection rate as a function of the number of judges from $J = 10$ to $J = 256$, which is the number of judges in our empirical example, and average cases per judge from 50 to $1,000$ (the average in our empirical example). We set the minimum number of cases per judge at 50, the value in our empirical example. The table shows that the rejection rate hovers close to the nominal size, .05, across the entire range, with the minimum rejection rate .0465 and the maximum .0700.

The next set of simulations shows that setting a minimum number of cases per judge is important to maintaining correct size of the test. Figure 4 plots the rejection rate as a function of the minimum number of cases per judge ranging from zero (no minimum) to 50. We set the average cases per judge to 1,000 and the number of judges to 256 to match the empirical example. The figure shows that when no minimum is imposed the test can overreject. The rejection rate is at about 15 percent when no minimum is imposed, but falls to near the nominal level as the minimum increases to above 20. The overrejection for cases where the minimum caseload is very small shows that the large sample approximation to the distribution of the test stastistic is poor when there are few cases per judge. This result provides support for the common practice of including only judges with at least a minimum number of cases per judge, as in our example.

The preceding sets of simulations show that our test maintains accurate size in realistic data generating processes. We now investigate the finite-sample power of our test. In these simulations we also use a data generating process calibrated to our empirical example. We modify the process described above to incorporate violations of the exclusion and monotonicity conditions. We do this by introducing a judge-specific additive effect, γ_j, that has the following distribution:

$$\gamma_j \sim N\left(0, \sigma_\gamma^2\right).$$

We then generate the outcome as

$$Y_i = \beta_0 + \beta_0 p_j + \gamma_j + \varepsilon_i,$$

where ε_i is normally distributed error term with mean zero and standard deviation .46 to match the residual variance in our empirical example. Setting $\sigma_\gamma = 0$ corresponds to a setting where the assumptions are satisfied. Increasing σ_γ violates the instrument validity assumptions to a greater and greater degree. By simulating the rejection rate for different values of σ_γ we can assess the power of the test. We calibrate the range for σ_γ to our empirical example by splitting our sample within each judge, regressing the outcome on a cubic polynomial in the estimated propensity, and averaging the resulting residuals at the judge level. The square root of the covariance between the judge averages across the two samples is around .025, and provides a reasonable magnitude for σ_γ in our simulations. We vary σ_γ between zero and .2 to provide a range around this calibrated value.

The first power simulations show that power increases when there are more cases per judge. Figure 5 plots the rejection rate by the degree of violation (σ_γ) for several different levels of average caseload per judge, ranging from 50 to 1,000. The simulations set the minimum caseload and number of judges to 50 and 20. For 500 and 1,000 cases per judge (which corresponds to our empirical example) the rejection rate increases from near zero when $\sigma_\gamma = 0$ to one for values above .1 or so. Power increases more slowly when there are fewer cases per judge.

The next power simulations show that power improves when there are more judges. Figure 6 plots the rejection rate by the degree of violation (σ_γ) for 2, 10, and 20 judges. The simulations set the minimum number of cases per judge to 50 and the average number of cases per judge to 1,000 to match the empirical example. When there are at least 10 judges, power increases rapidly and reaches 100 percent at or before $\sigma_\gamma = .2$. The increase is somewhat more rapid when there are 20 judges than when there are 10. When there are only 2 judges, however, the increase in power is much slower, and does not reach 100 percent for the values of σ_γ considered. It should be noted that in practice the number of judges in judge fixed effects designs is nearly always at least of the order of 10 or 20 if not much greater (as in our empirical example), where our simulations show that power is excellent.

The next set of power simulations show that imposing a minimum number of cases per judge improves power. Figure 7 plots the rejection rate by the degree of violation (σ_γ) for one, 30, and 50 minimum cases per judge. The simulations set the average number of cases per judge to 1,000 (as in our empirical example) and the number of judges to 20. For minimum caseload equal to 30 or 50 the power curves are essentially identical, increasing quickly to 100 percent. When there is no minimum caseload, however, power increases more slowly and tops out around 60 percent. Thus, both size and power of the test is greatly improved by imposing a minimum number of cases of at least 30.

Finally, we show that the test's power increases when the bound on the possible magnitude of treatment effects is tightened. Figure 8 plots the rejection rate by the degree of violation (σ_γ) for $K = 1, .6, .2$. The simulations set the average cases per judge to 1,000, the minimum cases per judge to 50, and the number of judges to 20. Power increases much more quickly for smaller values of K. For example, for $K = .2$, power is between 70 and 80 percent for σ_γ near .02-.03, the value calibrated to our empirical example.

Additional data generating processes

In this section we examine the finite-sample performance of the test in additional data generating processes. Like those above, these simulations mimic the random assignment of defendants to one of J judges, but with n defendants assigned per judge. We endow judge $j \in \{1, \ldots, J\}$ with a severity (probability of treatment):

$$p_j := p_a + (j-1)(1-p_a-p_n) / (J-1).$$

Thus, a fraction p_a of defendants are always treated and a fraction p_n never treated regardless of judge assignment. We generate a binary outcome, $Y_i \in \{0, 1\}$ so that the expected value of Y_i given assignment to a judge with propensity p_j is

$$E[Y_i|p_j] = \frac{1-(1-\lambda)(p_n+p_a)}{1-(p_n+p_a)} p_j - \frac{\lambda}{1-(p_n+p_a)} p_a.$$

When $\lambda = 0$, the slope of this function is one, corresponding to no violation of the assumptions. When λ exceeds zero, the slope of this function exceeds one, reflecting a violation of the exclusion restriction. Thus, λ governs the degree of departure from the instrument assumptions. We show how the nonparametric test's power as a function of λ depends on the number of judges J and the number of cases per judge, n. We set $p_a = p_n = .2$ in the simulations.

Figure 9 plots the nonparametric test's rejection rate as a function of λ for 2, 10, and 20 judges, setting the number of cases per judge at $n = 100$. The rejection rate is near zero for small values of λ, but increases sharply as the degree of violation increases, reaching 100 percent around $\lambda = .8$. The power of the test does not differ dramatically for different numbers of judges.

Figure 10 plots the nonparametric test's rejection rate as a function of λ for 30, 100, and 1000 cases per judge, setting the number of judges at $J = 10$. The figure shows that power varies dramatically by the number of cases per judge. Power is near zero for 30 cases per judge until λ exceeds .5, and power never exceeds .1. When cases per judge is 1000, however, power reaches 100 percent when λ is .3 or greater.

Semiparametric version

The section illustrates how the semiparametric version of the proposed test performs in terms of finite-sample size and power.

The data generating process for the first set of simulations is calibrated to the empirical application in the main text, just as the nonparametric test simulations presented above. We set the minimum number of cases per judge to 50 and the number of judges to 256, to match our empirical application. The average number of cases per judge is 100. We generate datasets where the degree of violation of the assumptions is parameterized by the standard deviation of judge direct effects, ranging from zero (assumptions satisfied) to .2, as in the simulations above. Figure 11 plots the rejection rate as a function of the degree of violation. At the far left of the plot, where the assumptions are satisfied, the rejection rate is just below the nominal level of .05, showing that the test has correct size. As the violations become more severe, the rejection rate increases steeply to 100 percent at about .04.

We also explore the performance of the semiparametric version of the test in an alternative data generating process that mimics a setting with J judges, to whom individuals are assigned with uniform probablity:

$$J_i \sim U\left\{1, \ldots, J\right\}.$$

A judge's propensity to assign treatment is given by:

$$p\left(J_i\right) = \theta J_i / \left(J\right).$$

The outcome is generated as

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i,$$

and treatment is determined by

$$D_i = 1\left(\Phi\left(-\nu_i\right) \leq p\left(J_i\right)\right),$$

where

$$v_i = \rho\varepsilon_i + \sqrt{1-\rho^2}\eta_i$$

and

$$\left(\varepsilon_i, \eta_i\right) \sim N\left(0, I_2\right).$$

In this setup the parameter θ governs the strength of the instruments and ρ determines the degree of treatment endogeneity. Note that this setup satisfies Condition 1. The main simulation results set $\omega = 1$, which directs power to the fit component of the test. Further simulation results below show how the test performs under different choices for ω.

The first set of simulations examines how the test's size depends on the number of observations. We set the simulation parameters as $J = 10, \theta = 1, \beta_0 = \beta_1 = 1$, and $\rho = .5$. We consider sample sizes of $n \in \{500; 1,000; 2,000; 5,000; 10,000\}$, and for each sample size draw 999 samples from the data generating process described above and perform the test with nominal size $\alpha = .05$, recording the rejection rate for each sample size. The simulations show that the test has very close to nominal size even for modest sample sizes. Figure 12 plots the rejection rate as a function of the sample size. The horizontal line is at .05. The simulated rejection rate is very near the nominal level throughout the range of sample sizes.

The next set of simulations explores the test's power to detect a violation of the exclusion restriction, Condition 1a, which can arise if judges have direct effects on outcomes other than through treatment. The data generating process is as described above, except judges now have direct effects on the outcomes:

$$Y_i = \beta_0 + \beta_1 D_i + \sum_{j=1}^{J} \gamma_j 1\left(J_i = j\right) + \varepsilon_i,$$

where the individual judge effects γ_j are drawn from a normal distribution with mean zero and standard deviation that varies from zero (corresponding to no violation) to 1 (severe violation). We set $n = 1000$ for this set of simulations. The simulations show that the proposed test's power increases rapidly with the severity of the violation. Figure 13 plots the rejection rate as a function of the standard deviation of the direct judge effects. At the far left the rejection rate is very near .05, reproducing the result that the test has correct size when the assumptions are satisfied. As the standard deviation of the direct judge effects grows, the rejection rate increases rapidly. Power exceeds 90 percent when the standard deviation is 0.2 and is essentially 100 percent for standard deviations above 0.3.

The next set of simulations illustrates the test's power to detect violations of the monotonicity assumption, Condition 1c, which can occur if judges do not implicitly agree on the order in which defendants should be treated. To allow for monotonicity violations in the simulations, we introduce heterogeneity in defendants and judges. We introduce an additional set of J judges (indexed $J + 1, \ldots, J$) who order most defendants identically to the first J judges, but order a fraction $\phi < .5$ of defendants, whom we call defiers, in the opposite order. Since violations of monotonicity only lead to bias when treatment effects vary, we set defiers' treatment effect to $-\beta_1$. Let the binary variable F_i with $\Pr\left(F_i = 1\right) = \phi$ indicate whether a defendant is a defier. Treatment assignment and outcomes are then determined as follows:

$$\begin{aligned} Y_i &= \begin{cases} \beta_0 - \beta_1 D_i + \varepsilon_i & , \quad F_i = 1 \\ \beta_0 + \beta_1 D_i + \varepsilon_i & , \quad \text{otherwise} \end{cases}, \\ D_i &= \begin{cases} 1\left(\Phi\left(-\nu_i\right) \leq 1 - p\left(J_i\right)\right) & , \quad F_i = 1 \text{ and } J_i \geq J + 1 \\ 1\left(\Phi\left(-\nu_i\right) \leq p\left(J_i\right)\right) & , \quad \text{otherwise} \end{cases}. \end{aligned}$$

The simulations show the proposed test has good power to detect violations of monotonicity of this sort. Figure 14 plots the test's rejection rate as a function of the fraction of defiers ϕ. At the far left ($\phi = 0$, corresponding to no violation) the test rejects at a rate near $\alpha = .05$ as expected. As the fraction ϕ increases and the violation of monotonicity becomes more severe, the rejection rate increases rapidly. Power exceeds 80 percent when the fraction of defiers is greater than .3.

We also run simulations that allow us to compare how our test performs relative to the the test described in Kitagawa (2015). The Kitagawa test assumes a priori knowledge of the instruments' order with respect to the probability of treatment. In the context of judge assignment, this assumption is problematic because the judge propensities to treat are estimated rather than directly observed. To assess the size of the Kitagawa test, we run simulations with four judges (the test quickly becomes computationally burdensome as the number of judges increases) with population propensities $.25, .495, .5$, and $.505$ using three different samples sizes ($n \in \{5,000; 10,000; 100,000\}$). The monotonicity assumption in this scenario means individuals fall into one of five compliance categories: always-takers, never-takers, and one of three complier groups. We set the treatment effect equal to zero, and set always-takers' outcomes to $Y_i = 0$, judge 1 compliers' outcomes to $Y_i = 1$, judge 2 compliers' outcomes to $Y_i = 2$, judge 3 compliers' outcomes to $Y_i = 3$, and never-takers' outcomes to $Y_i = 4$. Figure 15 plots the rejection rate from the Kitagawa test as a function of the sample size, along with the rejection rate from our semiparametric test. With a nominal test size of 5%, rejection rates for the Kitagawa test are 10.3% for a sample size of 5,000, and grow to 22.2% for $n = 10,000$ and 27.9% with a sample size of 100,000. In comparison, rejection rates for our test in the same simulations are 5.2%, 5.1% and 4.7%. The results show that the Kitagawa test substantially overrejects and that the distortion does not decrease with the sample size over the range considered. For a large enough sample size, of course, and given a data generating process, estimation error in the propensities will become negligible and the Kitagawa test will have correct size. But for any given sample size, there is a data generating process for which the Kitagawa test will fail to control size; that is, the Kitagawa test is not uniformly asymptotically valid when propensities are estimated, as illustrated in our simulations.

Finally, we show how the test performs under different choices for ω. To demonstrate this, we run two sets of simulations: one with a small number of judges ($J = 2$), in which we will see that small values for ω are best, and another with a large number of judges ($J = 20$) in which larger values for ω are best. The outcome variable is binary, as in our application below, with expected value conditional on judge assignment given by

$$\Pr\left(Y_i = 1 | J_i\right) = \beta_0 + \beta_1 J_i/k + \sum_{j=0}^{J} \gamma_j 1\left(J_i = j\right),$$

and treatment propensity given by

$$\Pr\left(D_i = 1 | J_i\right) = \alpha_0 + \alpha_1 J_i/k.$$

As above, the γ_j terms represent violations of the exclusion restriction when they are nonzero; in this set of simulations they are drawn from normal distribution with standard deviation .2. This simulation setup also allows the assumptions to be violated when $\beta_1 > \alpha_1$, as this would imply an average treatment effect greater than one, which is impossible for a binary outcome. This corresponds to a violation of the slope condition. In this simulation setup we set $\beta_1 = .3$ and $\alpha_1 = .2$. Using a simulated sample size of $n = 1,000$, we perform our test for several choices of ω between zero and one, and examine how the test's power depends on ω in the few-judge case ($J = 2$) and the many-judge case ($J = 20$). The simulation results show that in the few-judge case, power is greatest when $\omega = 0$, since the fit component of the test has no power in this case. The upper panel of Figure 16 shows that the test's power is over 80 percent when $\omega = 0$, and drops to zero when $\omega = 1$. The situation is reversed in the many-judge case. The lower panel of Figure 16 shows that power is very poor (around 10 percent) when $\omega = 0$, and increases for higher values of ω. Typical instances of the judge fixed effects design, including our application below, involve relatively many judges. These simulation results suggest choosing ω to be high in these cases. In the application we set $\omega = 1$.

References

Donald W. K. Andrews and Gustavo Soares. Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, 78(1):119–157, 2010. doi: 10.3982/ECTA7502.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. In James J. Heckman and Edward E. Leamer, editors, *Handbook of Econometrics, Volume 6, Part B*, chapter 76, pages 5549–5632. Elsevier B.V., 2007.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, July 1982.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994. ISSN 00129682. URL `http://www.jstor.org/stable/2951620`.

Toru Kitagawa. A test for instrument validity. *Econometrica*, 83(5): 2043–2063, 2015. ISSN 1468-0262. doi: 10.3982/ECTA11974. URL `http://dx.doi.org/10.3982/ECTA11974`.

Jeffrey S. Racine. A primer on regression splines. unpublished manuscript, May 2018.

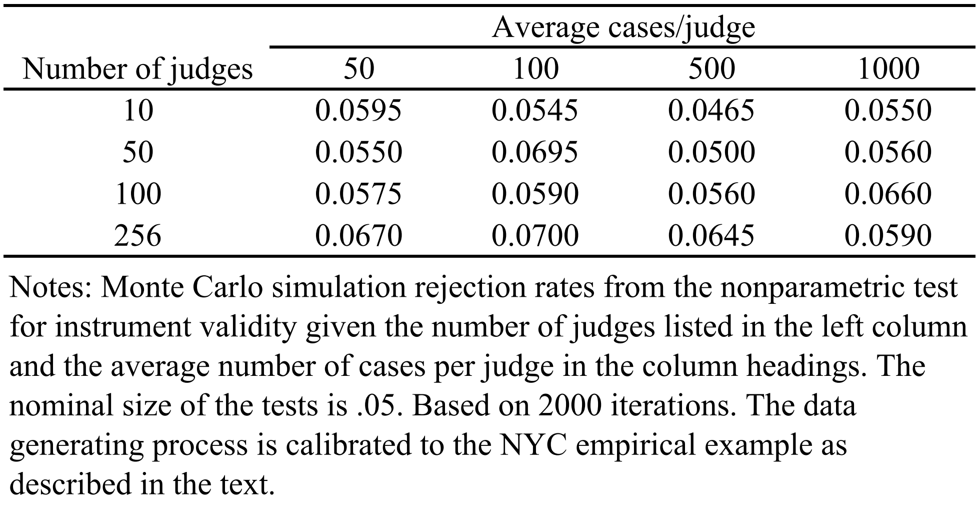
J. D. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3):393–415, 1958. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1907619`.

Table 1: Test Results

	5 knots	10 knots	15 knots	20 knots
Test statistic	440	368	315	244
d.f.	(248)	(243)	(238)	(233)
P-value	[0.000]	[0.000]	[0.001]	[0.601]

Note: This table displays the test statistics, degrees of freedom, and associated p-values from the proposed semiparametric testing procedure. Each column shows results using a different number of knots in the spline function.

Table 2: Simulated rejection rate by number of judges and average cases per judge



	Average cases/judge			
Number of judges	50	100	500	1000
10	0.0595	0.0545	0.0465	0.0550
50	0.0550	0.0695	0.0500	0.0560
100	0.0575	0.0590	0.0560	0.0660
256	0.0670	0.0700	0.0645	0.0590

Notes: Monte Carlo simulation rejection rates from the nonparametric test for instrument validity given the number of judges listed in the left column and the average number of cases per judge in the column headings. The nominal size of the tests is .05. Based on 2000 iterations. The data generating process is calibrated to the NYC empirical example as described in the text.

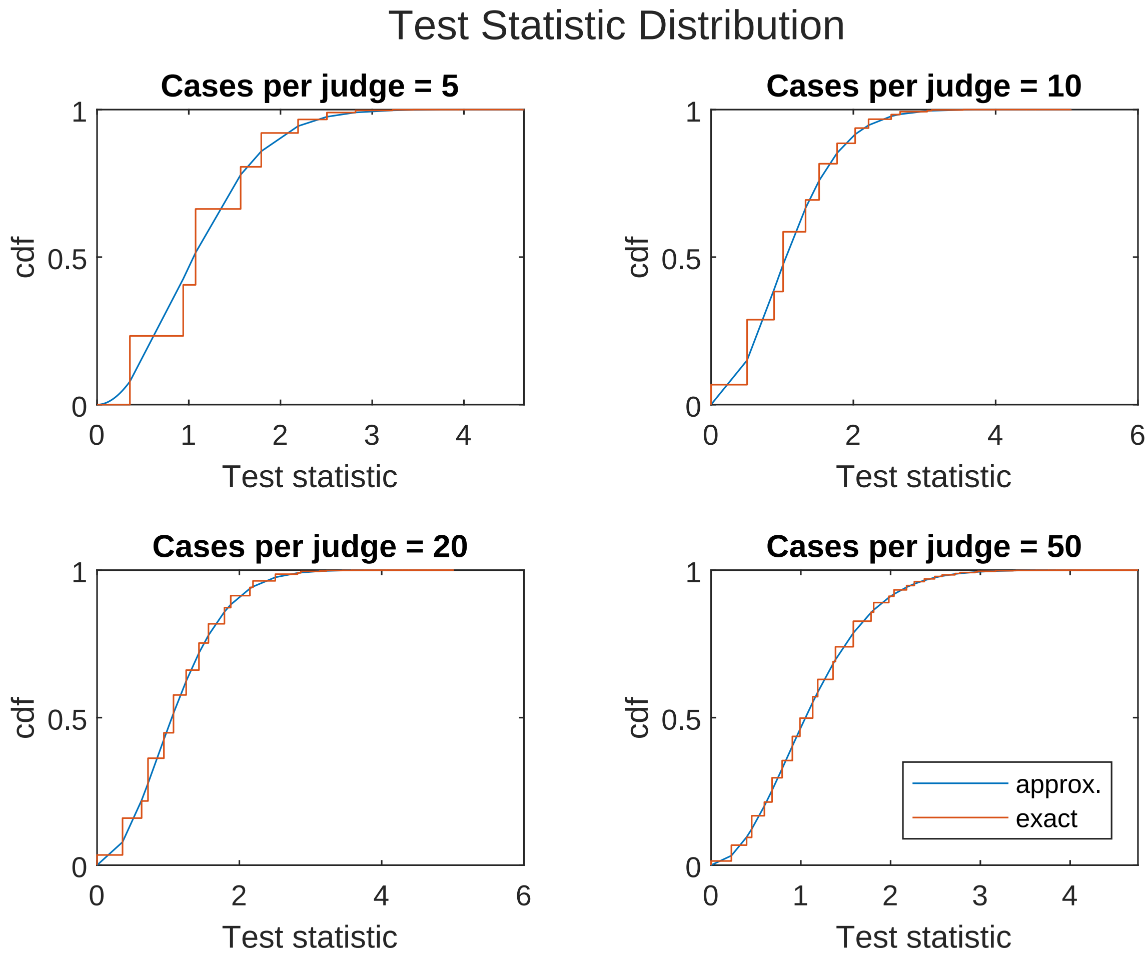


Figure 1: Exact and approximate cumulative distribution functions for the judge-level test statistic $\left|\hat{T}_j\right|$. The parameters, using notation defined in the Appendix, are as follows: $K = 1$, $p_j = .4$, $w_j = .31$, $v_j = .39$, and n_j is as indicated in each panel's title.

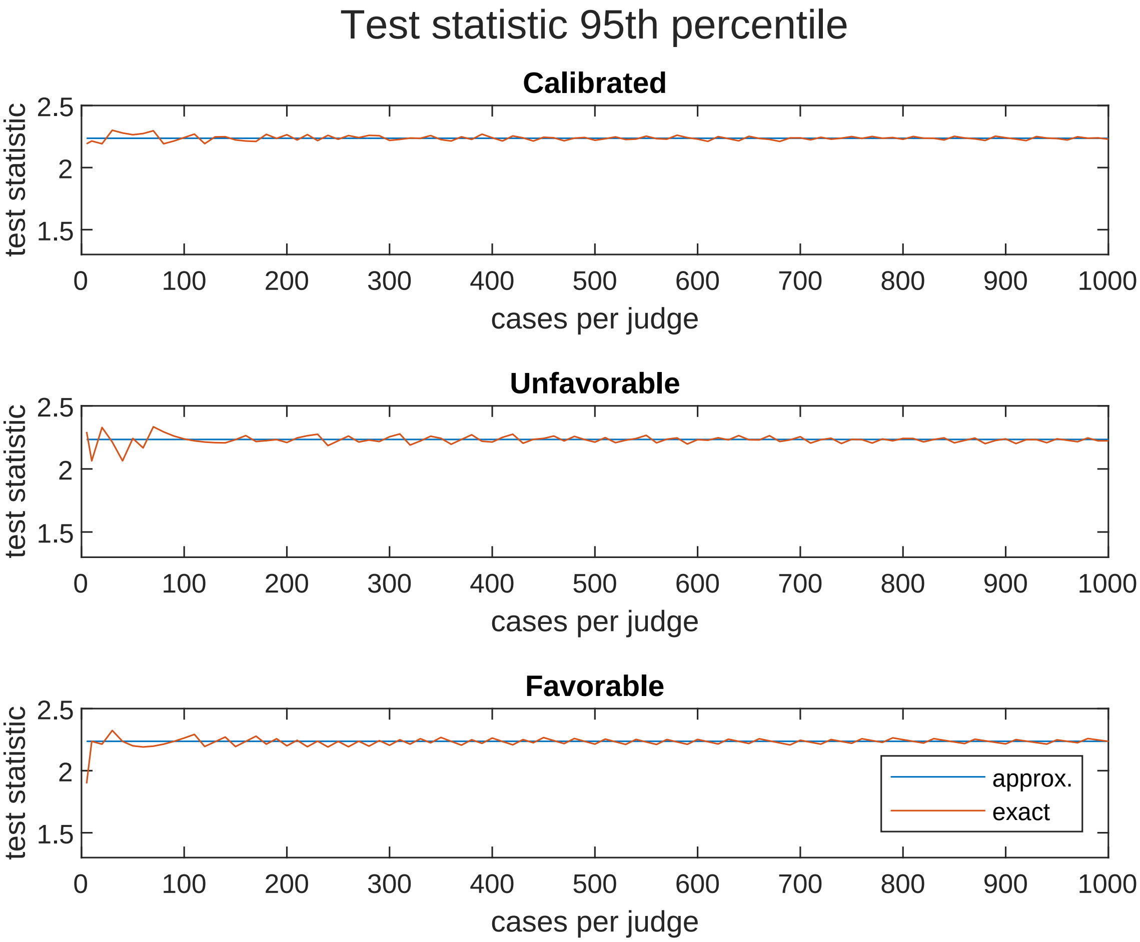


Figure 2: Exact and approximate 95th percentiles of the judge-level test statistic, $\left|\hat{T}_j\right|$, as a function of the number of cases per judge, n_j. The top panel is calibrated to the empirical example and sets $p_j = .4$, $w_j = .31$, $v_j = .39$. The middle panel sets parameters less favorable to the normal approximation: $p_j = .1$, $w_j = .05$, $v_j = .9$. The bottom panel sets parameters more favorable to the normal approximation: $p_j = .5$, $w_j = .25$, $v_j = .25$.

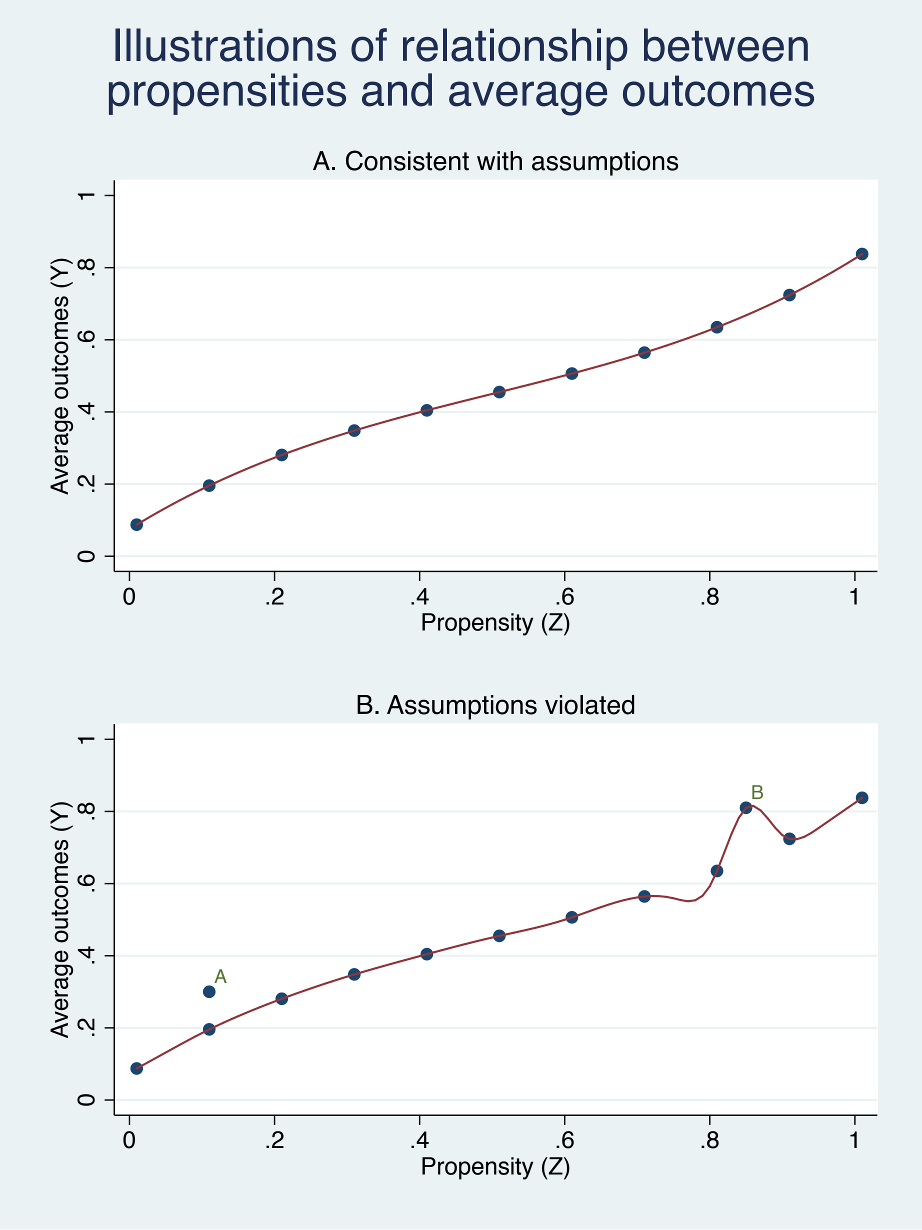


Figure 3: Illustrations of hypothetical relationships between true judge propensities to assign treatment and expected outcomes. Each dot represents a single judge. The pattern in Panel A is consistent with the exclusion restriction and monotonicity, because all the dots lie on a continuous function whose slope is nowhere larger in magnitude than the largest possible treatment effects, given a binary outcome. The pattern in Panel B could only arise if one or more of the assumptions were violated. The judge labeled “A” has exactly the same propensity as another judge, but different expected outcomes. The judge labeled “B” lies on a segment of the curve whose slope is larger than one, implying an impossibly large treatment effect.

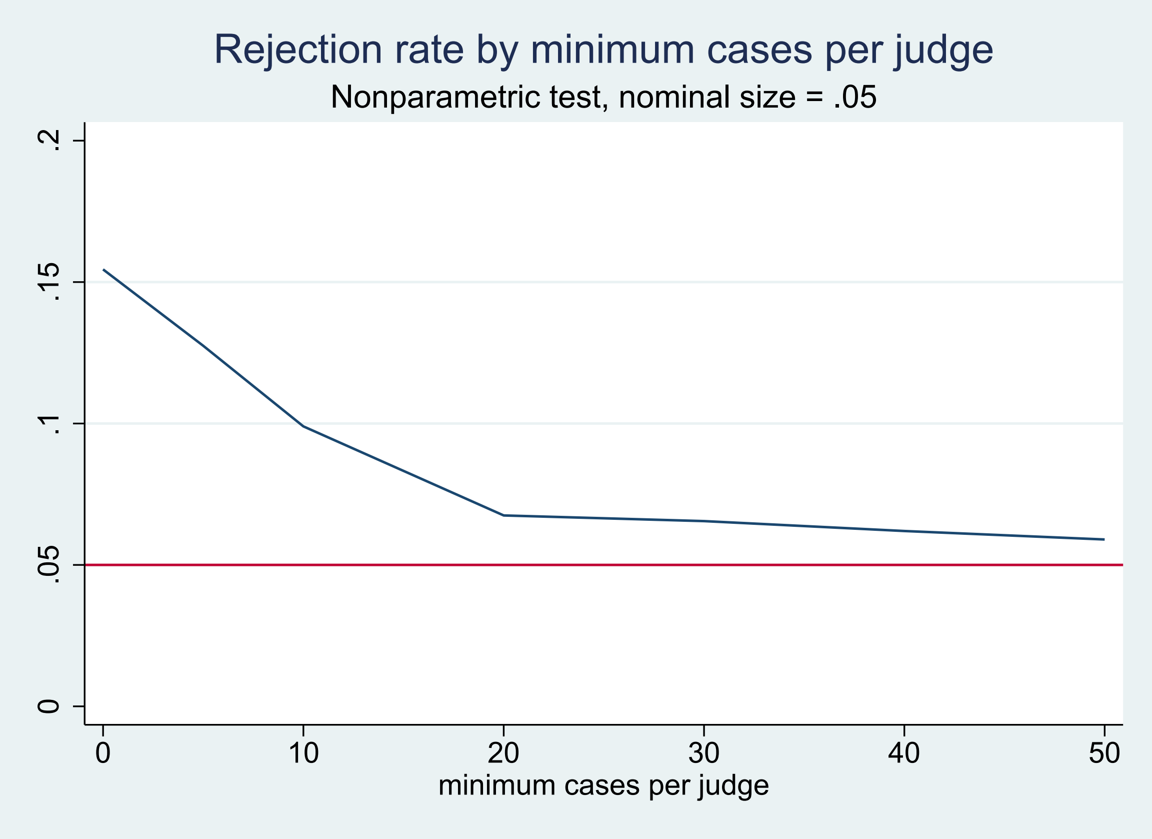


Figure 4: Monte Carlo simulation rejection rates from the nonparametric test for instrument validity as a function of the minimum cases per judge. The nominal size of the tests is .05. Based on 999 iterations. The data generating process is calibrated to the NYC empirical example as described in the text.

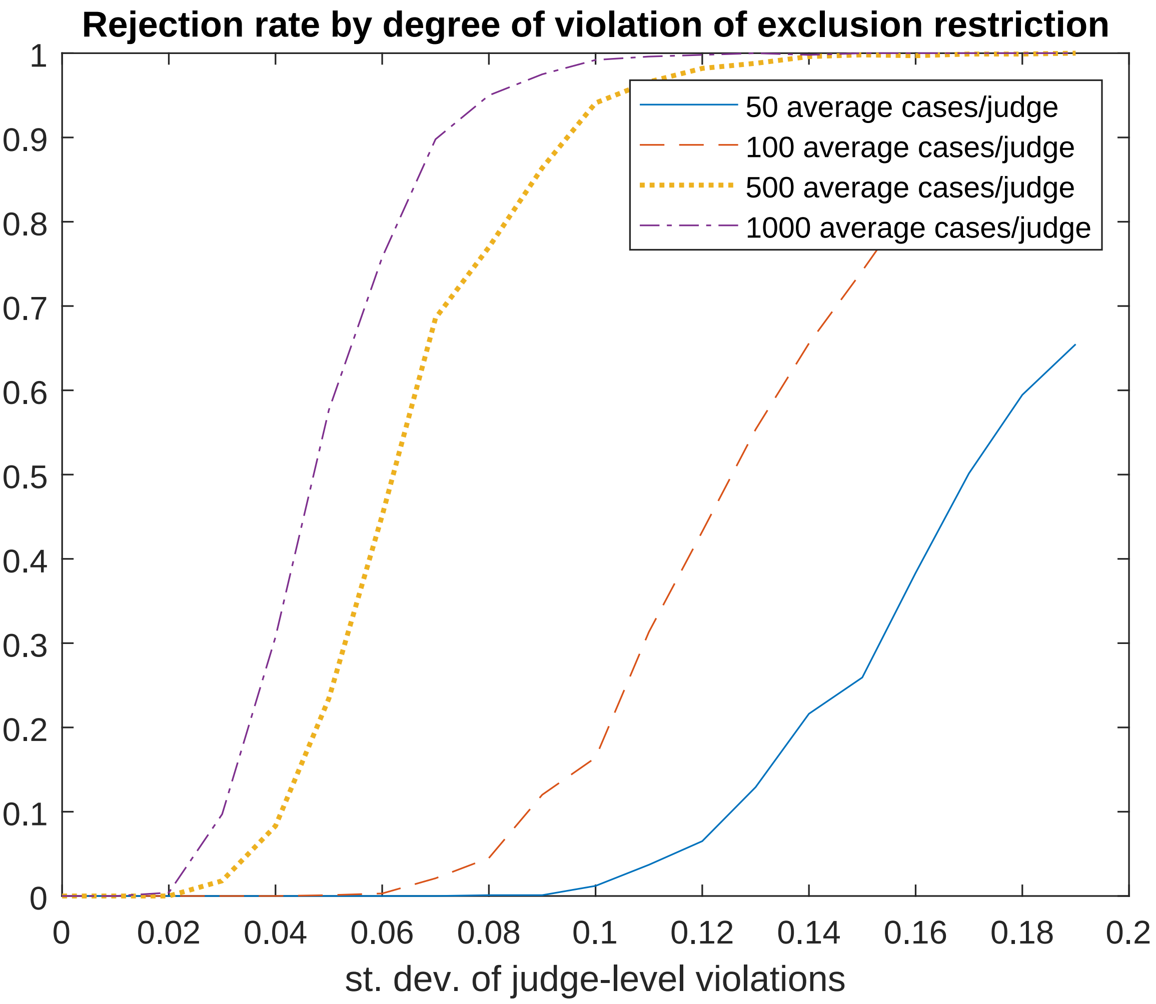


Figure 5: Monte Carlo simulation rejection rates from the nonparametric test for instrument validity as a function of the degree of violation of the assumptions measured by the standard deviation of the judge effects. Each curve corresponds to the average caseload indicated. The nominal size of the tests is .05. Based on 999 iterations. The data generating process is calibrated to the NYC empirical example as described in the text.

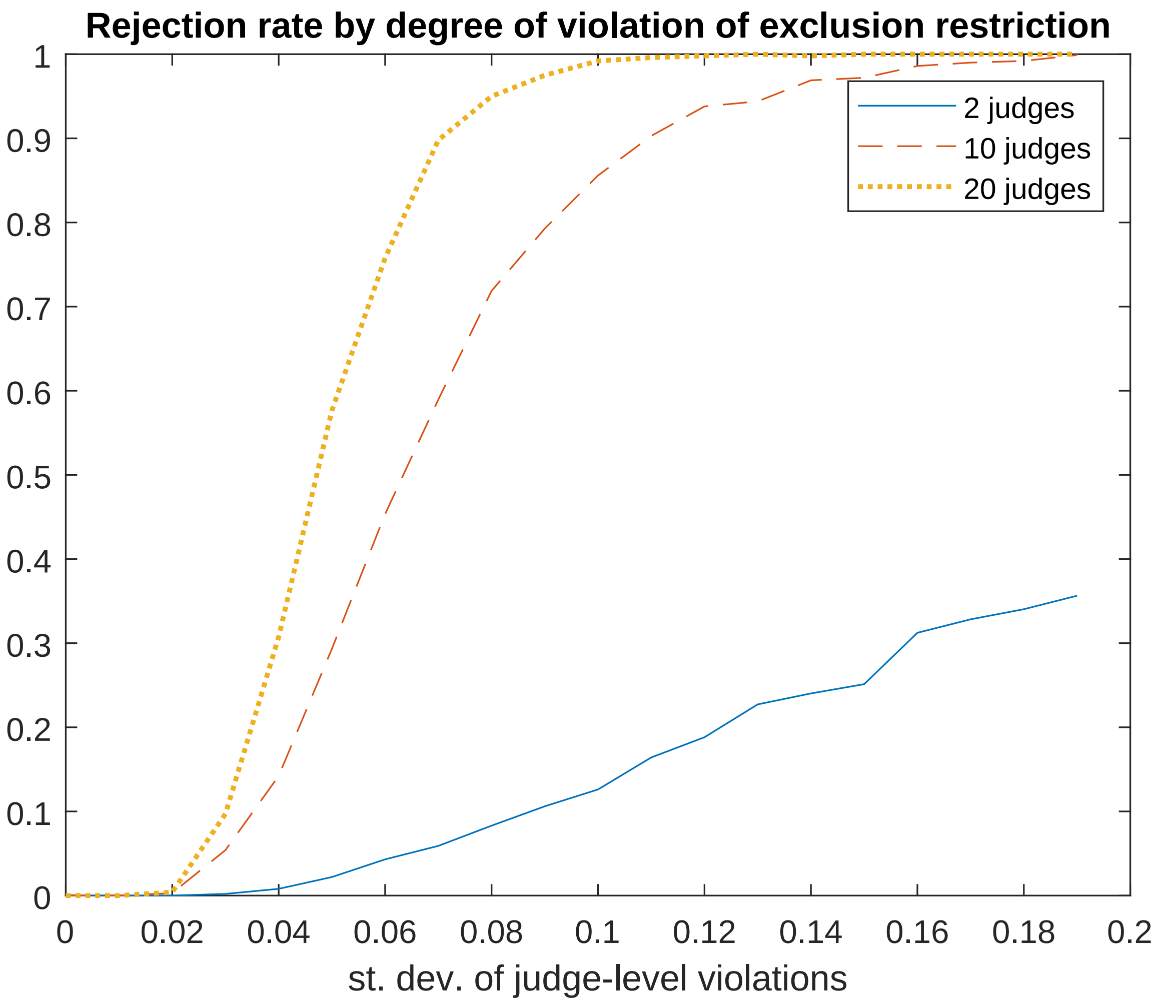


Figure 6: Monte Carlo simulation rejection rates from the nonparametric test for instrument validity as a function of the degree of violation of the assumptions measured by the standard deviation of the judge effects. Each curve corresponds to the number of judges indicated. The nominal size of the tests is .05. Based on 999 iterations. The data generating process is calibrated to the NYC empirical example as described in the text.

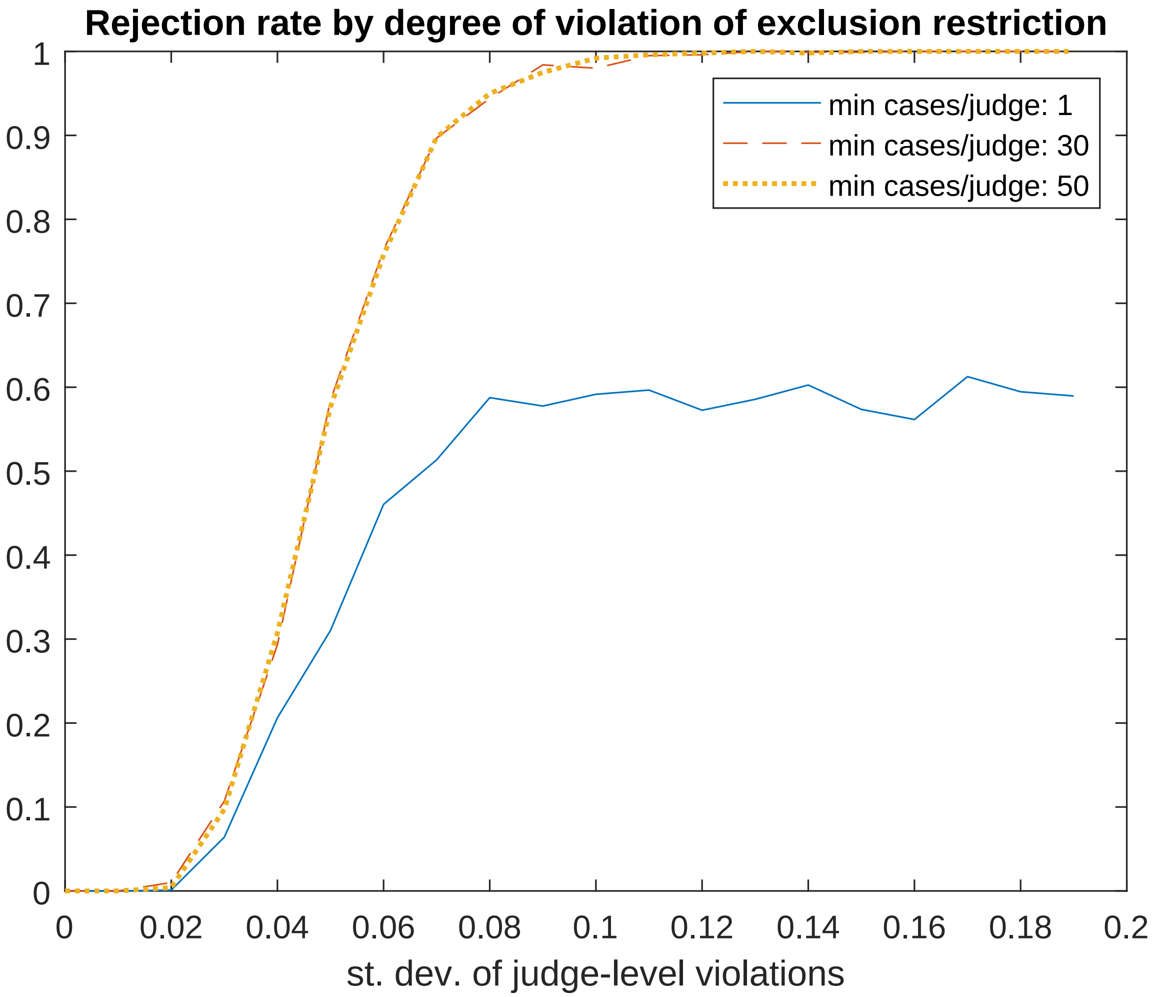


Figure 7: Monte Carlo simulation rejection rates from the nonparametric test for instrument validity as a function of the degree of violation of the assumptions measured by the standard deviation of the judge effects. Each curve corresponds to the minimum caseload indicated. The nominal size of the tests is .05. Based on 999 iterations. The data generating process is calibrated to the NYC empirical example as described in the text.

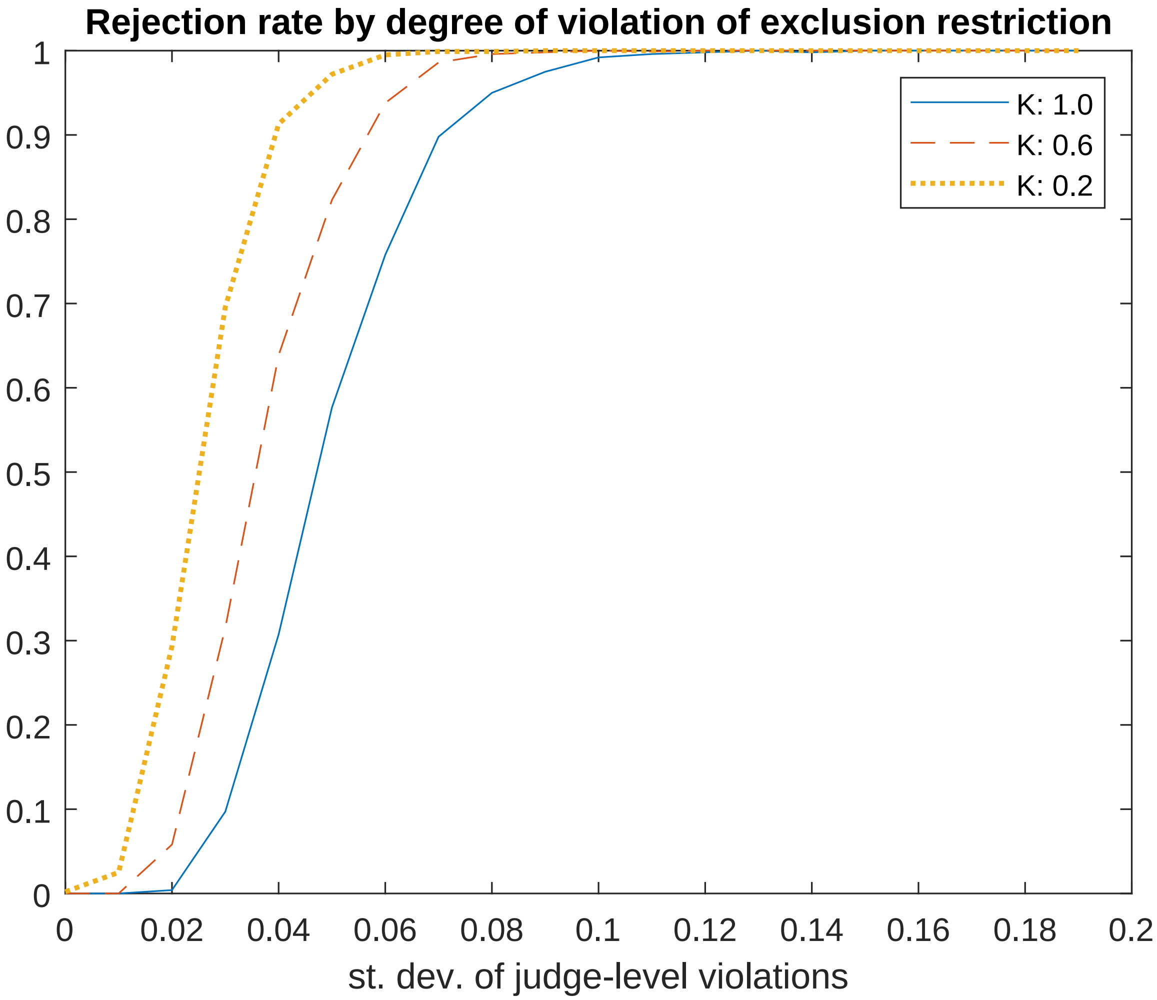


Figure 8: Monte Carlo simulation rejection rates from the nonparametric test for instrument validity as a function of the degree of violation of the assumptions measured by the standard deviation of the judge effects. Each curve corresponds to the value of K indicated. The nominal size of the tests is .05. Based on 999 iterations. The data generating process is calibrated to the NYC empirical example as described in the text.

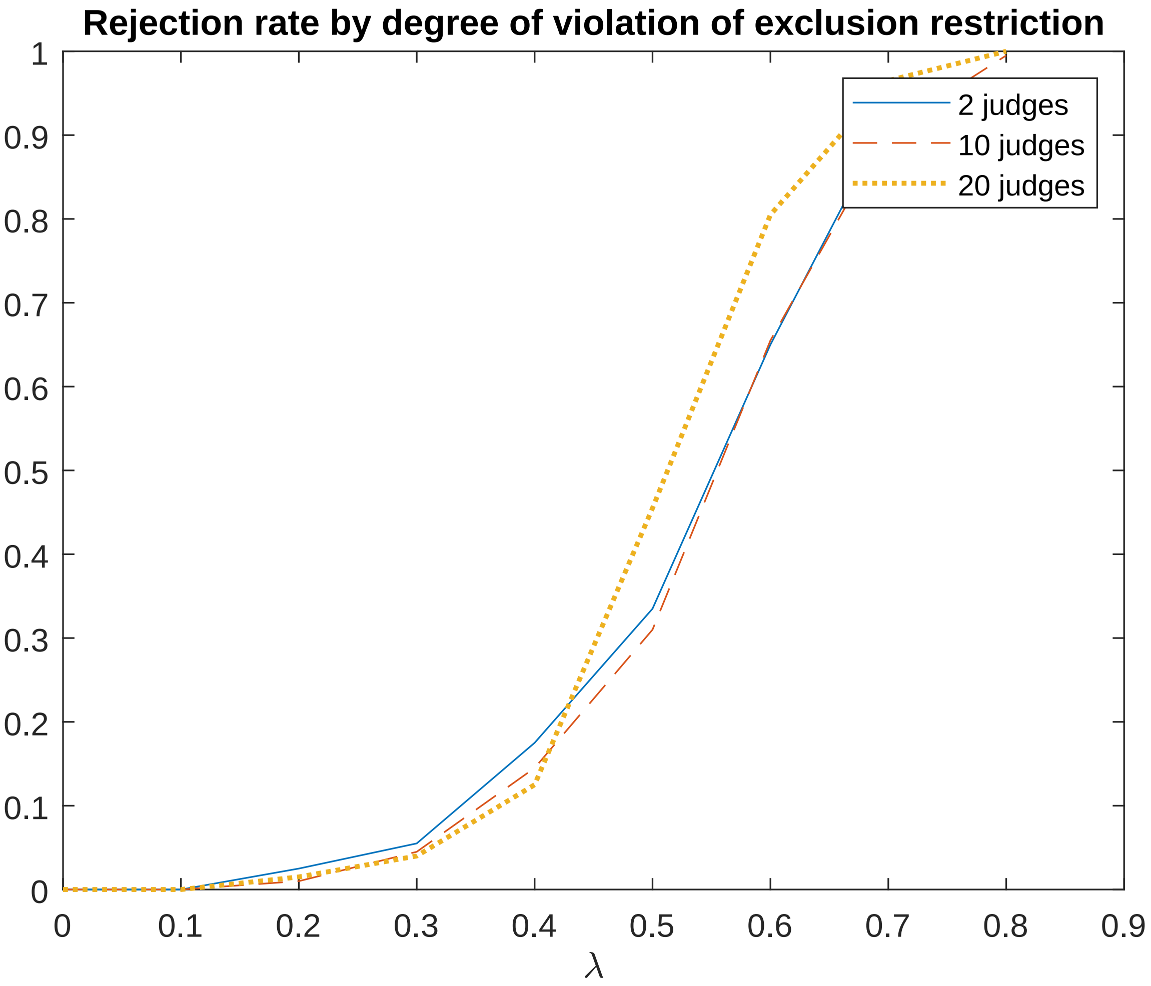


Figure 9: Monte Carlo simulation rejection rates from the nonparametric test for instrument validity as a function of the degree of violation of the assumptions, λ (x-axis) for the number of judges indicated. The nominal size of the tests is .05. Based on 999 iterations.

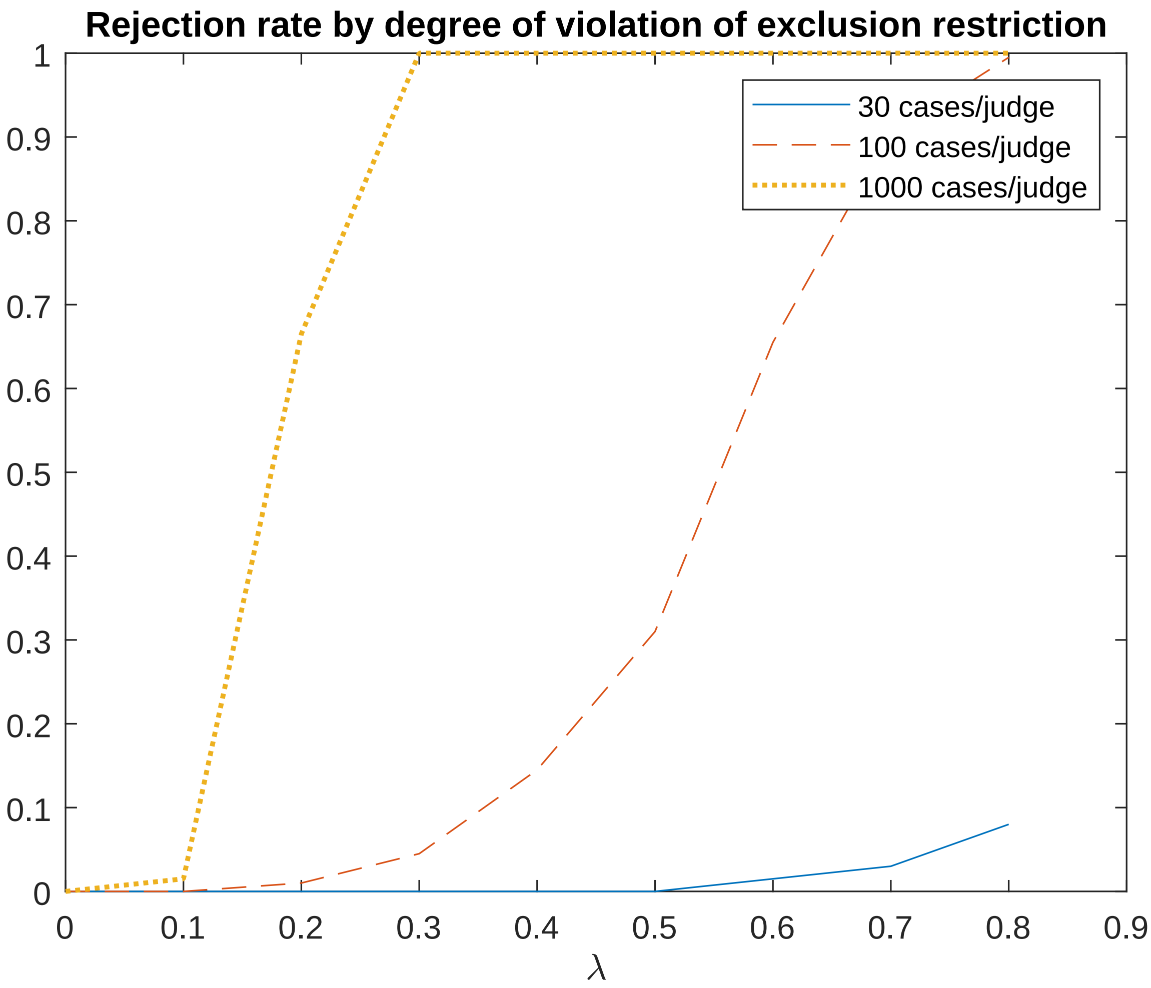


Figure 10: Monte Carlo simulation rejection rates from the nonparametric test for instrument validity as a function of the degree of violation of the assumptions, λ (x-axis) for the number of cases per judge indicated. The nominal size of the tests is .05. Based on 999 iterations.

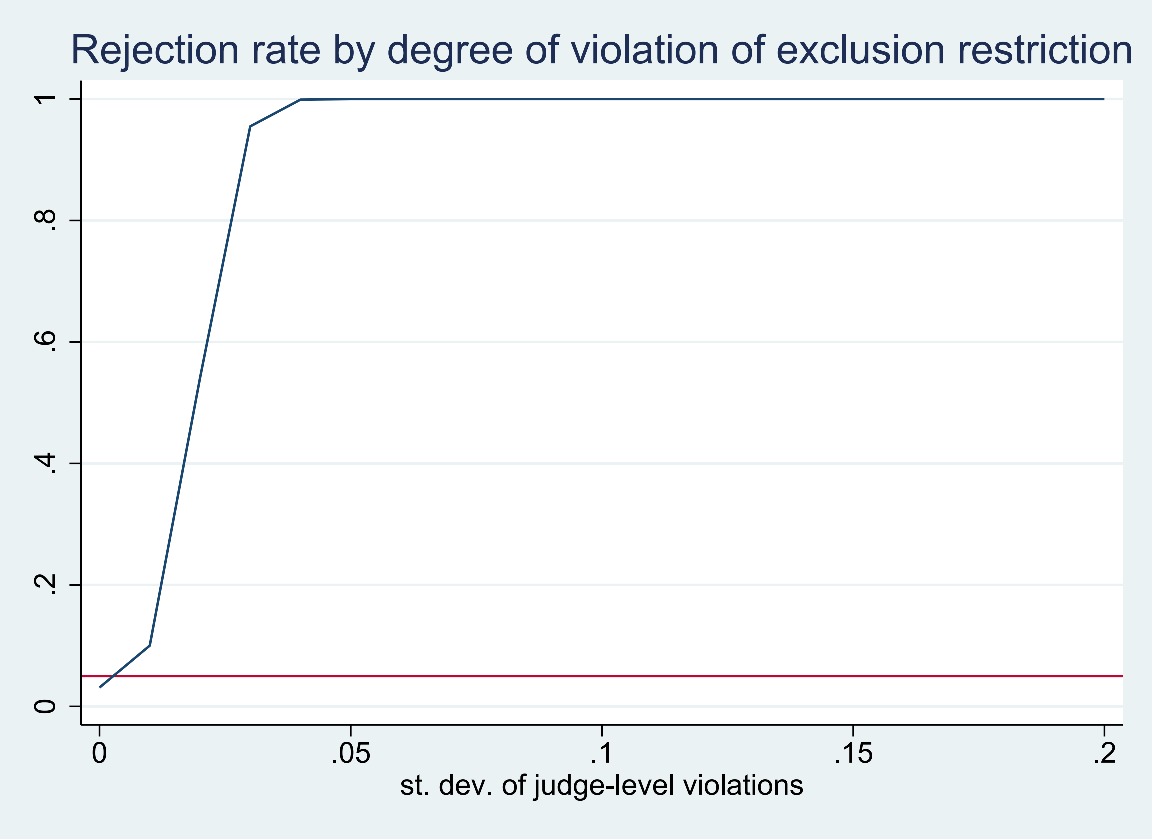


Figure 11: Monte Carlo simulation rejection rates from the semiparametric test for instrument validity as a function of the severity of the exclusion restriction violation, as measured by the standard deviation of the direct judge effects (x-axis). Data generating process calibrated to the empirical application as described in the appendix text. The nominal size of the tests is .05. Based on 999 iterations.

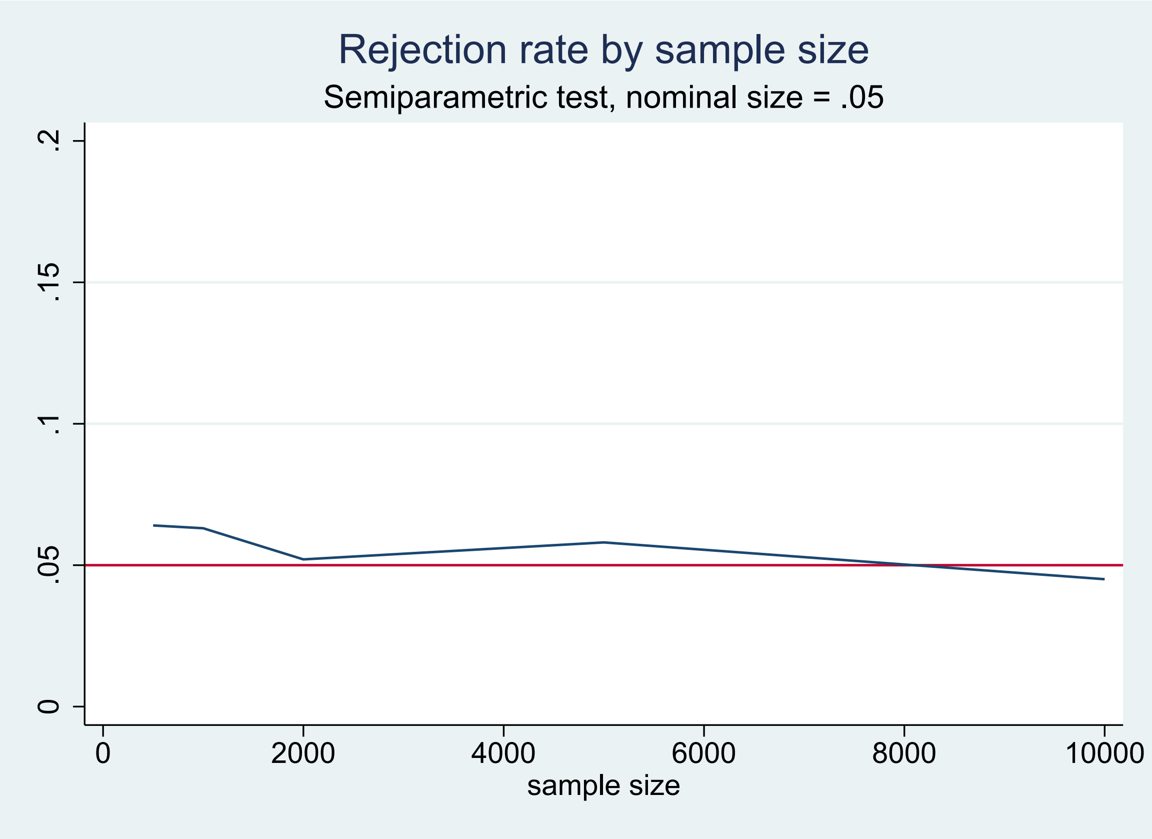


Figure 12: Monte Carlo simulation rejection rates from the semiparametric test for instrument validity as a function of the sample size (x-axis). The nominal size of the tests is .05. Based on 999 iterations.

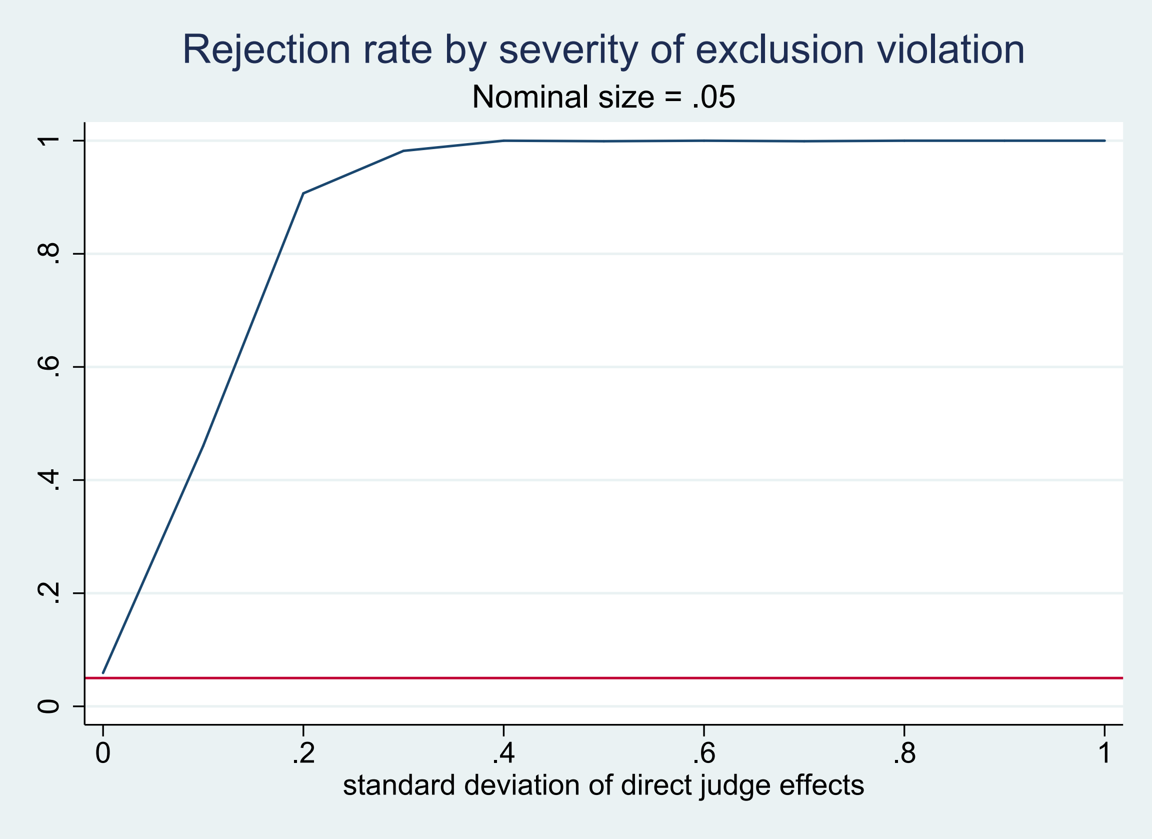


Figure 13: Monte Carlo simulation rejection rates from the semiparametric test for instrument validity as a function of the severity of the exclusion restriction violation, as measured by the standard deviation of the direct judge effects (x-axis). The nominal size of the tests is .05. Based on 999 iterations.

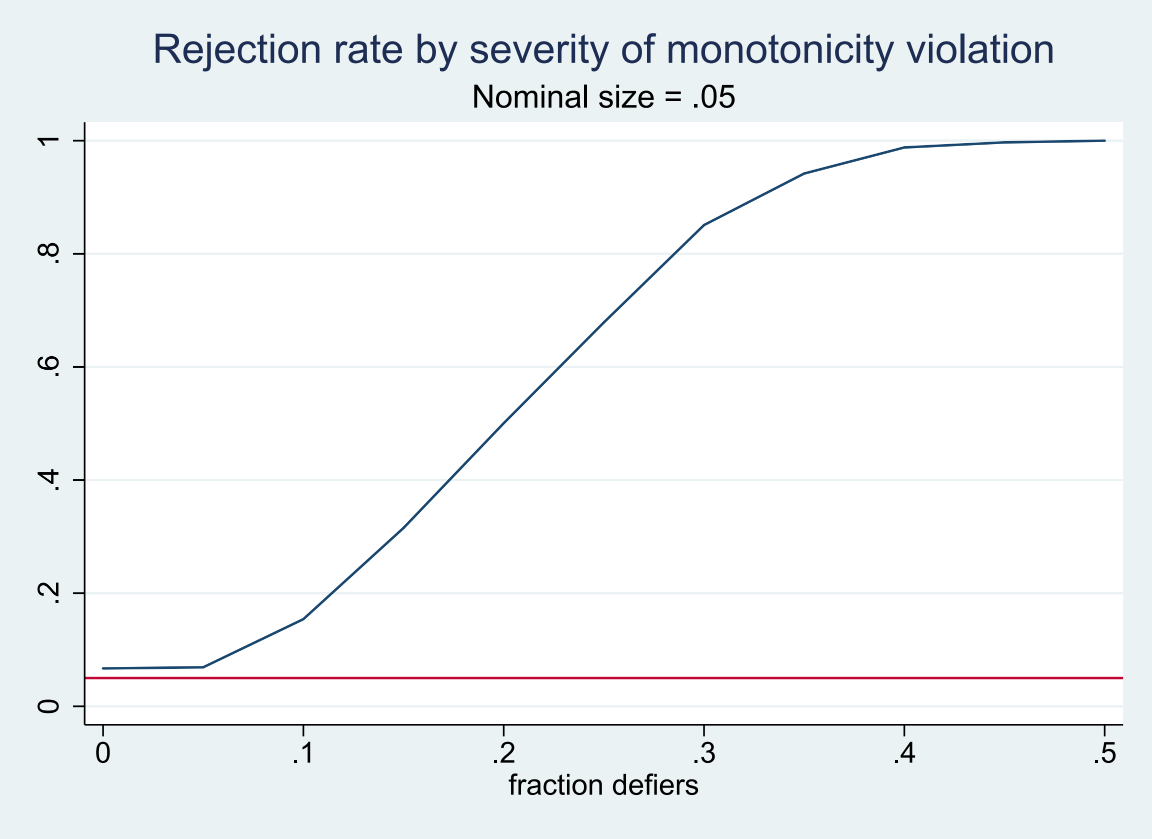


Figure 14: Monte Carlo simulation rejection rates from the semiparametric test for instrument validity as a function of the severity of the monotonicity violation, as measured by the fraction of defendants for whom judges disagree on the ordering. The nominal size of the tests is .05. Based on 999 iterations.

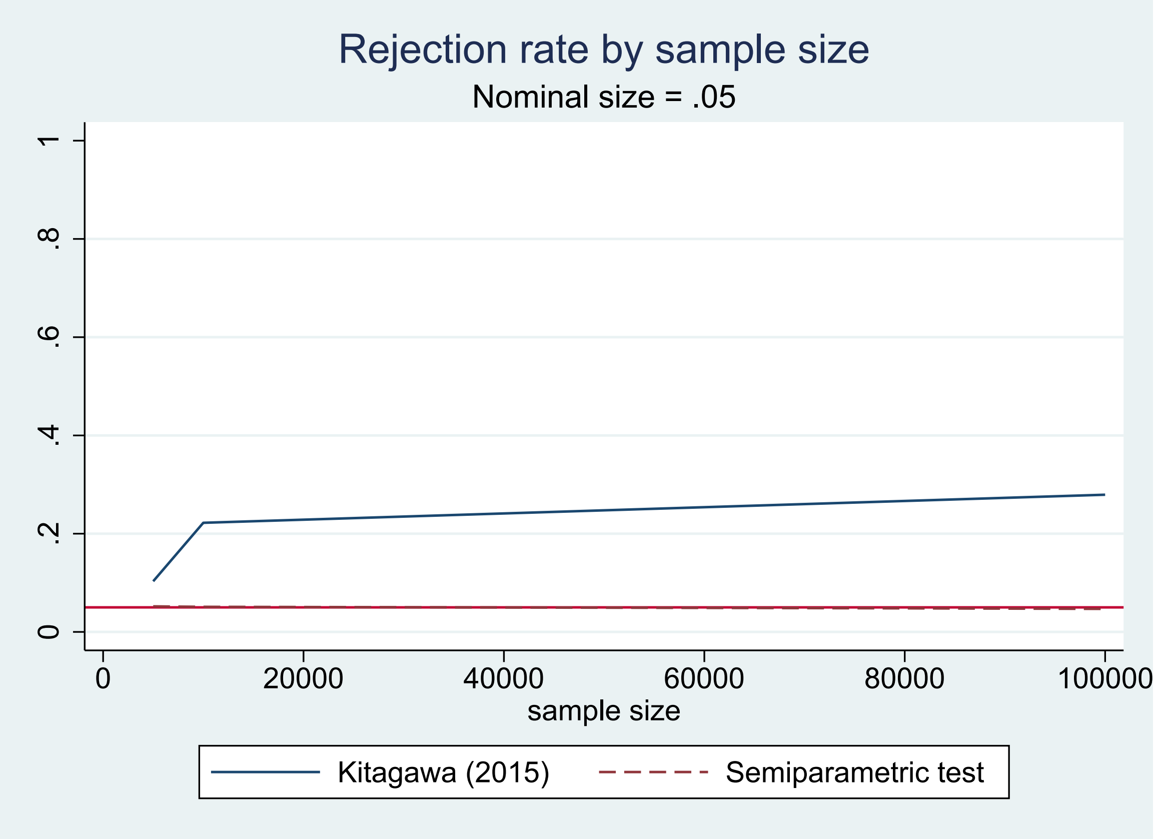


Figure 15: Monte Carlo simulation rejection rates from the Kitagawa (2015) test and the semiparametric test for instrument validity as a function of the sample size. Data-generating process based on four judges satisfies LATE conditions, as described in text. The nominal size of the tests is .05. Based on 999 iterations.

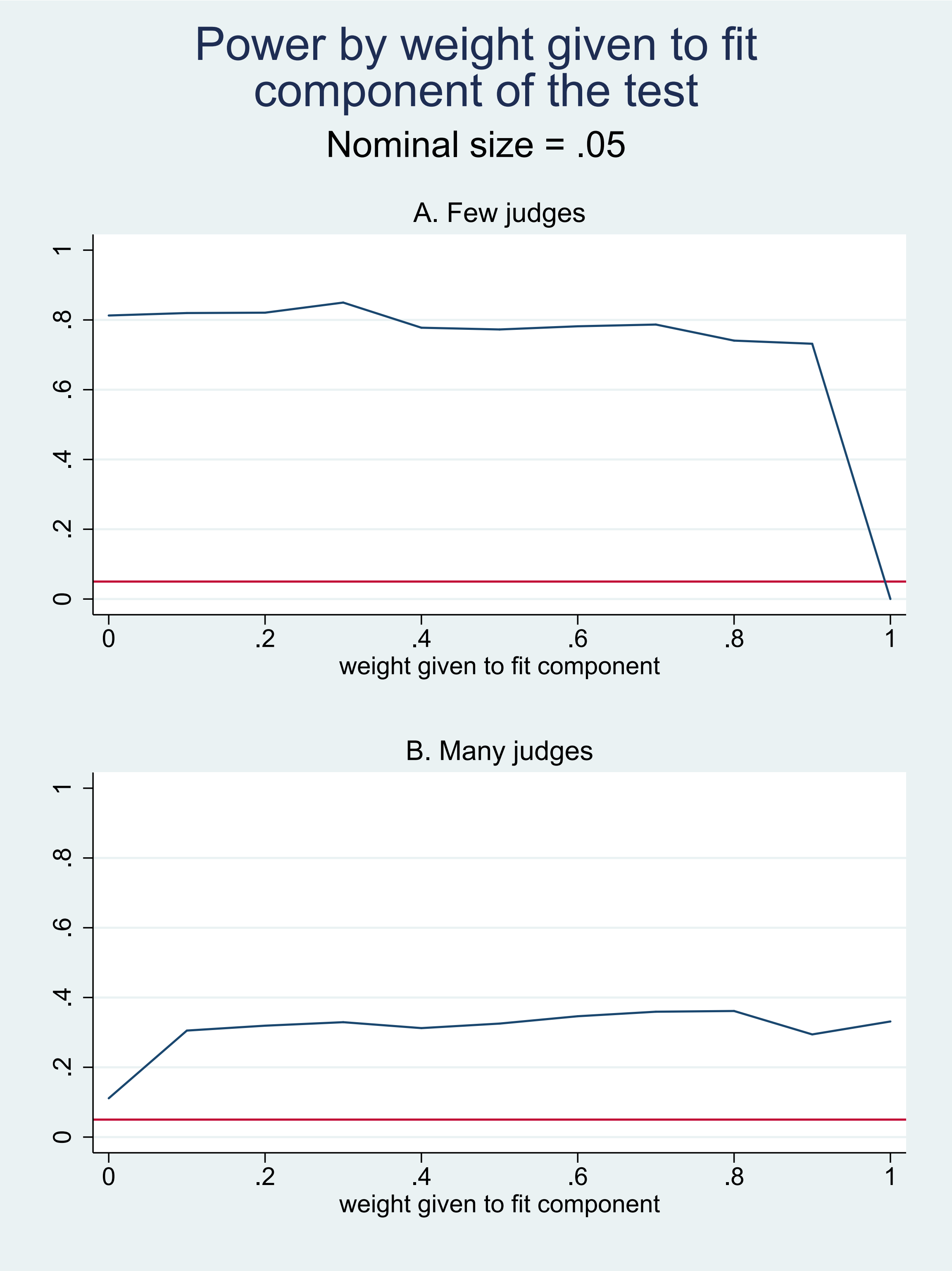


Figure 16: Monte Carlo simulation rejection rates from the semiparametric test for instrument validity as a function of the weight given to the fit component of the test. The upper panel sets $J = 2$. The lower panel sets $J = 20$. The nominal size of the tests is .05. Based on 999 iterations.