Supplemental Appendix

Paper Title: Monotonicity among Judges: Evidence from Judicial Panels and Consequences for Judge IV Designs

Henrik Sigstad

A Proofs

Proof. (Proposition 1.) We have

$$\begin{aligned}
\operatorname{Cov}(p(J), Y) &= \mathrm{E}[(p(J) - \mathrm{E}[p(J)]) Y] \\
&= \mathrm{E}[(p(J) - \mathrm{E}[p(J)]) (Y(0) + D\beta)] \\
&= \mathrm{E}[(p(J) - \mathrm{E}[p(J)]) D\beta] \\
&= \mathrm{E}[\mathrm{E}[(p(J) - \mathrm{E}[p(J)]) D\beta \mid S]] \\
&= \mathrm{E}[\mathrm{E}[(p(J) - \mathrm{E}[p(J)]) S(J)\beta \mid S]] \\
&= \mathrm{E}[\mathrm{E}[(p(J) - \mathrm{E}[p(J)]) S(J) \mid S] \mathrm{E}[\beta \mid S]] \\
&= \mathrm{E}[\operatorname{Cov}(P, S(J) \mid S) \mathrm{E}[\beta \mid S]]
\end{aligned}$$

where the third and sixth equalities invoke Assumption 1, and the fourth equality uses the law of iterated expectations. Thus

$$\beta^{\mathrm{2SLS}} = \operatorname{Var}(p(J))^{-1} \operatorname{Cov}(p(J), Y) = \mathrm{E}[w_S \beta_S]$$

where $w_s \equiv \operatorname{Var}(p(J))^{-1} \operatorname{Cov}(p(J), s(J))$ and $\beta_s \equiv \mathrm{E}[\beta \mid S = s]$. □

Proof. (Proposition 2.) The monotonicity violation rates in judicial panels and individual decisions are given by $\Pr[S^p \in \mathcal{S}_v] / \Pr[S^p \in \mathcal{S}_v \cup \mathcal{S}_s]$ and $\Pr[S \in \mathcal{S}_v] / \Pr[S \in \mathcal{S}_v \cup \mathcal{S}_s]$, respectively. It is straightforward to see that the equality of these two rates is equivalent to Assumption 3:

$$\frac{\Pr[S \in \mathcal{S}_v]}{\Pr[S \in \mathcal{S}_v \cup \mathcal{S}_s]} = \frac{\Pr[S^p \in \mathcal{S}_v]}{\Pr[S^p \in \mathcal{S}_v \cup \mathcal{S}_s]}$$

$$\Leftrightarrow \frac{\Pr[S \in \mathcal{S}_v]}{\Pr[S \in \mathcal{S}_v] + \Pr[S \in \mathcal{S}_s]} = \frac{\Pr[S^p \in \mathcal{S}_v]}{\Pr[S^p \in \mathcal{S}_v] + \Pr[S^p \in \mathcal{S}_s]}$$

$$\Leftrightarrow \frac{\Pr[S^p \in \mathcal{S}_v]}{\Pr[S \in \mathcal{S}_v]} = \frac{\Pr[S^p \in \mathcal{S}_s]}{\Pr[S \in \mathcal{S}_s]}$$

□

Proof. (Proposition 3.) Under Assumption 4, the statement is equivalent to

$$\frac{\Pr\left[S^p\left(1\right) < S^p\left(2\right)\right]}{\Pr\left[S\left(1\right) < S\left(2\right) \mid T = 1\right]} = \frac{\Pr\left[S^p\left(1\right) > S^p\left(2\right)\right]}{\Pr\left[S\left(1\right) > S\left(2\right) \mid T = 1\right]}$$

Under Assumption 5, we have

$$\begin{aligned}
\frac{\Pr\left[S\left(1\right) > S\left(2\right) \mid T = 1\right]}{\Pr\left[S\left(1\right) \neq S\left(2\right) \mid T = 1\right]} &= \Pr\left[S\left(1\right) > S\left(2\right) \mid S\left(1\right) \neq S\left(2\right), T = 1\right] \\
&= \Pr\left[S\left(1\right) > S\left(2\right) \mid S\left(1\right) \neq S\left(2\right)\right] \\
&= \frac{\Pr\left[S\left(1\right) > S\left(2\right)\right]}{\Pr\left[S\left(1\right) \neq S\left(2\right)\right]}
\end{aligned}$$

$$\Rightarrow \Pr\left[S\left(1\right) > S\left(2\right) \mid T = 1\right] = \Pr\left[S\left(1\right) > S\left(2\right)\right] \frac{\Pr\left[S\left(1\right) \neq S\left(2\right) \mid T = 1\right]}{\Pr\left[S\left(1\right) \neq S\left(2\right)\right]}$$

Similarly

$$\Pr\left[S\left(1\right) < S\left(2\right) \mid T = 1\right] = \Pr\left[S\left(1\right) < S\left(2\right)\right] \frac{\Pr\left[S\left(1\right) \neq S\left(2\right) \mid T = 1\right]}{\Pr\left[S\left(1\right) \neq S\left(2\right)\right]}$$

By insertion, we get that the statement is equivalent to Assumption 3 applied to Judges 1 and 2:

$$\frac{\Pr\left[S^p\left(1\right) < S^p\left(2\right)\right]}{\Pr\left[S\left(1\right) < S\left(2\right)\right]} = \frac{\Pr\left[S^p\left(1\right) > S^p\left(2\right)\right]}{\Pr\left[S\left(1\right) > S\left(2\right)\right]}$$

□

Proof. (Proposition B.1.) *Part i).* This is a direct application of the classical result on binary instruments (*e.g.*, Theorem 1 in Imbens and Angrist (1994)) applied to the subsample of cases assigned the strictest and the most lenient judge:

$$
\begin{aligned}
&\mathrm{E}\left[Y \mid J=\bar{j}\right]-\mathrm{E}\left[Y \mid J=\underline{j}\right] \\
&= \mathrm{E}\left[DY(1)+(1-D)Y(0) \mid J=\bar{j}\right]-\mathrm{E}\left[DY(1)-(1-D)Y(0) \mid J=\underline{j}\right] \\
&= \mathrm{E}\left[D(Y(1)-Y(0)) \mid J=\bar{j}\right]-\mathrm{E}\left[D(Y(1)-Y(0)) \mid J=\underline{j}\right] \\
&= \mathrm{E}\left[\left(S\left(\bar{j}\right)-S\left(\underline{j}\right)\right)(Y(1)-Y(0))\right] \\
&= \mathrm{E}\left[Y(1)-Y(0) \mid S\left(\underline{j}\right)=0, S\left(\bar{j}\right)=1\right]\Pr\left[S\left(\bar{j}\right)>S\left(\underline{j}\right)\right]
\end{aligned}
$$

The second equality used Assumption 1 and the fourth equality used that extreme-pair monotonicity implies $\Pr\left[S\left(\bar{j}\right)<S\left(\underline{j}\right)\right]=0$. The result follows since

$$
\Pr\left[S\left(\bar{j}\right)>S\left(\underline{j}\right)\right]=\mathrm{E}\left[D \mid J=\bar{j}\right]-\mathrm{E}\left[D \mid J=\underline{j}\right]
$$

Part ii). Assume lenient-judge monotonicity. We then have

$$
\begin{aligned}
&\mathrm{E}[Y]-\mathrm{E}\left[Y \mid J=\underline{j}\right] \\
&= \mathrm{E}[DY(1)+(1-D)Y(0)]-\mathrm{E}\left[DY(1)+(1-D)Y(0) \mid J=\underline{j}\right] \\
&= \mathrm{E}[D(Y(1)-Y(0))]-\mathrm{E}\left[D(Y(1)-Y(0)) \mid J=\underline{j}\right] \\
&= \mathrm{E}\left[\mathrm{E}[D(Y(1)-Y(0)) \mid S]-\mathrm{E}\left[D(Y(1)-Y(0)) \mid J=\underline{j}, S\right]\right] \\
&= \mathrm{E}\left[\mathrm{E}[D(Y(1)-Y(0)) \mid S]-\mathrm{E}\left[S\left(\underline{j}\right)(Y(1)-Y(0)) \mid S\right]\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\left(D-S\left(\underline{j}\right)\right)(Y(1)-Y(0)) \mid S\right]\right] \\
&= \mathrm{E}\left[D(Y(1)-Y(0)) \mid S\left(\underline{j}\right)=0\right]\Pr\left[S\left(\underline{j}\right)=0\right] \\
&= \mathrm{E}\left[Y(1)-Y(0) \mid S\left(\underline{j}\right)=0, D=1\right]\mathrm{E}\left[D \mid S\left(\underline{j}\right)=0\right]\Pr\left[S\left(\underline{j}\right)=0\right]
\end{aligned}
$$

The second equality used Assumption 1 and the third equality used the law of iterated expectations. The sixth equality used that $S\left(\underline{j}\right)=1 \Rightarrow D=1$ under lenient-

judge monotonicity. Moreover

$$\begin{aligned}
\mathrm{E}[D] - \mathrm{E}\left[D \mid J = \underline{j}\right] &= \mathrm{E}[D] - \Pr\left[S\left(\underline{j}\right) = 1\right] \\
&= \mathrm{E}\left[D \mid S\left(\underline{j}\right) = 0\right] \Pr\left[S\left(\underline{j}\right) = 0\right] + \Pr\left[S\left(\underline{j}\right) = 1\right] - \Pr\left[S\left(\underline{j}\right) = 1\right] \\
&= \mathrm{E}\left[D \mid S\left(\underline{j}\right) = 0\right] \Pr\left[S\left(\underline{j}\right) = 0\right]
\end{aligned}$$

where the second equality used that $\mathrm{E}\left[D \mid S\left(\underline{j}\right) = 1\right] = 1$ under lenient-judge monotonicity. Thus

$$\frac{\mathrm{E}[Y] - \mathrm{E}\left[Y \mid J = \underline{j}\right]}{\mathrm{E}[D] - \mathrm{E}\left[D \mid J = \underline{j}\right]} = \mathrm{E}\left[Y(1) - Y(0) \mid S\left(\underline{j}\right) = 0, D = 1\right]$$

The proof of part iii) is analogous to part ii). □

Proof. (Proposition C.1.) Let everything be conditional on G. We then have for any q in the support of Q that[52]

$$\mathrm{E}[M \mid Q = q] = \mathrm{E}[M \mid Q = q, T = 1] = \mathrm{E}[M \mid T = 1] \equiv m_1$$

The first (second) equation applied Assumption C.2 (C.3). The law of iterated expectations then gives

$$\mathrm{E}[M] = \mathrm{E}[\mathrm{E}[M \mid Q]] = m_1$$

$$\Rightarrow \mathrm{E}[M \mid T = 1] = \mathrm{E}[M]$$

which is equivalent to $T \perp M$ since T and M are binary variables. □

Proof. (Proposition C.2.) Assume $d^p_{vu}/d_{vu} = d^p_{vo}/d_{vo}$. First, note that

$$\mathrm{E}[S(1) - S(2) \mid C \in \mathcal{C}_v] = \Pr[S(1) > S(2) \mid C \in \mathcal{C}_v] - \Pr[S(1) < S(2) \mid C \in \mathcal{C}_v]$$

[52] Using that $T = 1$ is assumed to be a non-zero probability event for all values of Q.

$$\begin{aligned}\Rightarrow \Pr\left[S(1) > S(2), C \in C_v\right] &= \mathrm{E}\left[S(1) - S(2) \mid C \in C_v\right] \Pr\left[C \in C_v\right] \\ &+ \Pr\left[S(1) < S(2), C \in C_v\right]\end{aligned}$$

We thus have that the share of cases violating monotonicity equals the sum of observable and unobservable monotonicity violations:

$$\begin{aligned}\Pr\left[S(1) > S(2)\right] &= \Pr\left[S(1) > S(2), C \in C_v\right] + \Pr\left[S(1) > S(2), C \notin C_v\right] \\ &= \mathrm{E}\left[S(1) - S(2) \mid C \in C_v\right] \Pr\left[C \in C_v\right] \\ &+ \Pr\left[S(1) < S(2), C \in C_v\right] + \Pr\left[S(1) > S(2), C \notin C_v\right] \\ &= d_{vo} + d_{vu}\end{aligned}$$

Similarly, $\Pr\left[S(1) < S(2)\right] = d_{so} + d_{vu}$. Define $\lambda \equiv d^p_{vu}/d^p_{vo} = d_{vu}/d_{vo}$. Assumption 3, applied to Judges 1 and 2, then becomes

$$\frac{\Pr\left[S^p(1) < S^p(2)\right]}{\Pr\left[S(1) < S(2)\right]} = \frac{\Pr\left[S^p(1) > S^p(2)\right]}{\Pr\left[S(1) > S(2)\right]}$$

$$\Leftrightarrow \frac{d^p_{so} + d^p_{vu}}{d_{so} + d_{vu}} = \frac{d^p_{vo} + d^p_{vu}}{d_{vo} + d_{vu}}$$

$$\Leftrightarrow \frac{d^p_{so} + \lambda d^p_{vo}}{d_{so} + \lambda d_{vo}} = \frac{d^p_{vo}(1+\lambda)}{d_{vo}(1+\lambda)} = \frac{d^p_{vo}}{d_{vo}}$$

$$\Leftrightarrow d^p_{vo}/d_{vo} = d^p_{so}/d_{so}$$

□

B Additional Results

B.1 Additional Tables for The Main Result

Table B.1 shows the robustness of Table 2 to using only panels deciding more than 50 and 100 cases. Table B.2 shows a version of Table 2 with medians instead of means.

Table B.1: Robustness: Only panels with more than 50 and 100 cases.

	Share of Cases Violating Monotonicity Condition								
	São Paulo Appeal Court			Brazilian Superior Court			United States Supreme Court		
	Cutoff			Cutoff			Cutoff		
	20	50	100	20	50	100	20	50	100
IA monotonicity violated	0.35	0.35	0.32	0.49	0.52	–	0.42	0.40	0.41
Conditional number of violations	2.00	2.07	1.75	2.05	2.00	–	3.30	3.35	3.25
Average monotonicity violated	0.11	0.11	0.10	0.09	0.12	–	0.04	0.03	0.02
Sum of negative 2SLS weights	0.09	0.10	0.08	0.06	0.08	–	0.01	0.01	0.01

Note: Robustness of the results in Table 2 to only including panels that decide at least 50 and 100 cases together. There are no panels in the Brazilian Superior Court deciding more than 100 cases.

Table B.2: Table 2 with Medians instead of Means

	São Paulo Appeal Court (Five judges)	Brazilian Superior Court (Five judges)	United States Supreme Court (Nine judges)
IA monotonicity violated	0.357 (0.031)	0.472 (0.077)	0.404 (0.028)
Conditional number of violations	2.0 (0.1)	2.1 (0.2)	3.2 (0.2)
Average monotonicity violated	0.077 (0.018)	0.079 (0.031)	0.030 (0.009)
Sum of negative 2SLS weights	0.059 (0.013)	0.048 (0.023)	0.008 (0.003)
Observations	1,306	76	1,516

Note: A version of Table 2 with medians across judicial panels instead of means. Each parameter (e.g., the share of cases violating IA monotonicity) is calculated separately for each judicial panel, and the table reports the median across panels. Bootstrapped standard errors in parentheses.

B.2 Alternatives to 2SLS

While average monotonicity ensures that 2SLS identifies a positively weighted sum of individual treatment effects, the weights assigned by 2SLS are not uniform. For instance, in judge IV designs, 2SLS assigns higher weights to cases with a medium propensity to be treated and lower weights to cases violating Imbens-Angrist monotonicity. As argued by Heckman and Vytlacil (2007), this particular weighted sum is unlikely to be a parameter of policy interest. More meaningful treatment parameters include

$$\begin{aligned} \text{LATE} &\equiv \mathrm{E}\left[\beta \mid S\left(\underline{j}\right)=0, S\left(\bar{j}\right)=1\right] \\ \text{LATT} &\equiv \mathrm{E}\left[\beta \mid S\left(\underline{j}\right)=0, D=1\right] \\ \text{LATUT} &\equiv \mathrm{E}\left[\beta \mid S\left(\bar{j}\right)=1, D=0\right] \end{aligned}$$

where $\underline{j}$ and $\bar{j}$ are the most lenient and the strictest judges, respectively. The LATE parameter is the local average treatment effect for defendants incarcerated by the strictest judge but not by the most lenient judge. The LATT and LATUT parameters are the (local) average treatment effects on the treated and the untreated for

a similar complier population.[53] These parameters can be identified under the following monotonicity conditions:

Definition B.1. (Monotonicity conditions)

i) *Extreme-pair monotonicity* holds if $s\left(\bar{j}\right) \geq s\left(\underline{j}\right)$ for all $s \in \mathcal{S}$.

ii) *Lenient-judge monotonicity* holds if $s\left(j\right) \geq s\left(\underline{j}\right)$ for all $s \in \mathcal{S}$ and j.

iii) *Stringent-judge monotonicity* holds if $s\left(\bar{j}\right) \geq s\left(j\right)$ for all $s \in \mathcal{S}$ and j.

Proposition B.1. *(Alternatives to 2SLS).*

i) $\text{LATE} = \frac{\text{E}\left[Y|J=\bar{j}\right]-\text{E}\left[Y|J=\underline{j}\right]}{\text{E}\left[D|J=\bar{j}\right]-\text{E}\left[D|J=\underline{j}\right]}$ *if extreme-pair monotonicity holds*

ii) $\text{LATT} = \frac{\text{E}[Y]-\text{E}\left[Y|J=\underline{j}\right]}{\text{E}[D]-\text{E}\left[D|J=\underline{j}\right]}$ *if lenient-judge monotonicity holds*

iii) $\text{LATUT} = \frac{\text{E}\left[Y|J=\bar{j}\right]-\text{E}[Y]}{\text{E}\left[D|J=\bar{j}\right]-\text{E}[D]}$ *if stringent-judge monotonicity holds*

In particular, LATE is identified by the standard Wald estimand of the effect of being assigned the strictest judge compared to being assigned the most lenient judge as long as monotonicity holds between these two judges. Furthermore, LATT and LATUT are identified by the difference between the mean outcome and the expected outcomes for cases assigned to the most lenient and the strictest judge, respectively. These estimands require monotonicity to hold for all pairs of judges involving the most lenient and the strictest judge, respectively.

Sigstad (2024) shows that the Proposition B.1 estimands are equivalent to the estimands based on *marginal treatment effects* (MTE) invoked in the literature (*e.g.*, Arnold, Dobbie, and Crystal S. Yang (2018) and Bhuller, Dahl, et al. (2020)). These estimands can thus be estimated using the standard techniques in the MTE literature.

[53] The LATT and LATUT complier population includes all cases except never-takers and always-takers. The LATE complier population also ignores, for instance, response types incarcerated by some intermediate judges but not by the strictest nor the most lenient judge. I do not see a way to identify a local average treatment effect that also covers such compliers.

Table B.3: Violations of Monotonicity Conditions: Alternatives to 2SLS

	São Paulo Appeal Court (Five judges)	Brazilian Superior Court (Five judges)	United States Supreme Court (Nine judges)
Panel A: Share of cases violating monotonicity condition			
Extreme-pair Monotonicity	0.043 (0.005)	0.026 (0.020)	0.011 (0.003)
Lenient-judge Monotonicity	0.260 (0.011)	0.105 (0.034)	0.067 (0.007)
Stringent-judge Monotonicity	0.079 (0.007)	0.197 (0.052)	0.087 (0.007)
Panel B: Number of violations in cases violating monotonicity			
Lenient-judge Monotonicity	1.5 (0.04)	1.8 (0.24)	2.6 (0.16)
Stringent-judge Monotonicity	1.9 (0.11)	2.1 (0.27)	2.6 (0.16)
Panel C: Sum of Negative Weights			
LATE	0.065 (0.009)	0.038 (0.033)	0.012 (0.004)
LATT	0.184 (0.015)	0.082 (0.035)	0.041 (0.006)
LATUT	0.126 (0.016)	0.326 (0.163)	0.067 (0.008)
Observations	1,306	76	1,516

Note: Violations of monotonicity conditions corresponding to alternative IV estimands. *Share of cases violating monotonicity condition* is the share of cases where the monotonicity condition is violated. *Number of violations in cases violating monotonicity* is the mean number of judge pairs violating the condition in cases violating the condition. *Sum of negative weights* is the sum of the negative weights the estimand would assign treatment effects if monotonicity is violated as in the observed data. Bootstrapped standard errors in parentheses.

Violations of Alternative Monotonicity Conditions in Judicial Panels

In Table B.3, Panel A, I show how often the Proposition B.1 monotonicity conditions are violated in judicial panels. Extreme-pair monotonicity is violated in only 4.3 percent of São Paulo Appeal Court cases, 2.6 percent of Brazilian Superior Court cases, and 1.1 percent of US Supreme Court cases. The sums of the negative weights, reported in Panel C, imply that the bias induced by these monotonicity violations is small even under severe levels of heterogeneous effects. For instance, if the causal effect for cases satisfying average monotonicity is 0.3 and the effect for cases violating monotonicity is 0.6, the Equation 1 bias due to negative weights is 0.02 in the São Paulo Appeal Court, 0.011 in the Brazilian Superior Court, and 0.004 in the US Supreme Court.

Estimates of LATT are not as robust as LATE estimates. Lenient-judge monotonicity is violated in 26 percent of São Paulo Appeal Court cases, 11 percent of Brazilian Superior Court cases, and seven percent of US Supreme Court cases. The Proposition B.1 LATT estimand thus relies on an assumption that is around six times more likely to be violated than the condition necessary to identify LATE. Conditional on a case violating lenient-judge monotonicity, I find that the condition is violated in, on average, 1.5 out of four possible judge pairs in the São Paulo Appeal Court, 1.8 out of four possible judge pairs in the Brazilian Superior Court, and 2.6 out of eight justice pairs in the US Supreme Court. The sums of the negative weights indicate that the bias induced by negative weights is moderate if heterogeneous effects are severe. For instance, if the causal effect for cases satisfying average monotonicity is 0.3 and the effect for cases violating monotonicity is 0.6, the bias due to negative weights is 0.06 in the São Paulo Appeal Court, 0.025 in the Brazilian Superior Court, and 0.012 in the US Supreme Court. Similar results hold for LATUT: The share of cases violating stringent-judge monotonicity is eight percent in the São Paulo Appeal Court, 20 percent in the Brazilian Superior Court, and nine percent in the US Supreme Court.

B.3 Bias Due to Estimating Judge Stringencies

When estimating the rate of monotonicity violations, I use the share of cases in which a given judge votes with the prosecution as a proxy for the judge's true large-sample stringency. These *sample* stringency measures are noisy measures of the true stringencies. This section discusses the bias in my monotonicity violation estimates due to these measurement errors.

To fix ideas, consider the case of two judges, Judge A and Judge B. Denote by q_A (q_B) the probability that Judge A (Judge B) is stricter than Judge B (Judge A) in a randomly drawn case. Assume Judge A is the stricter judge: $q_A > q_B$. Then the true share of cases violating monotonicity is q_B. When I estimate the share of monotonicity violations, I rely on sample analogs $\hat{q}_A$ and $\hat{q}_B$ of q_A and q_B. If $\hat{q}_B > \hat{q}_A$, I wrongly conclude that B is the strictest judge, and use $\hat{q}_A$ instead of $\hat{q}_B$ as my estimate of the share of cases violating monotonicity. This causes me to underestimate the true number of monotonicity violations. In particular, my estimate of the share of cases violating monotonicity is

$$\hat{r} \equiv \min\{\hat{q}_A, \hat{q}_B\}$$

which has a negative bias of[54]

$$\mathrm{E}[\hat{r}] - q_B = -\Pr[\hat{q}_B > \hat{q}_A]\,\mathrm{E}[\hat{q}_B - \hat{q}_A \mid \hat{q}_B > \hat{q}_A]$$

In words, the amount of underestimation equals the probability of getting the stringency order wrong times the average difference in estimated stringencies between the two judges in that case. In my settings, this bias is small. For instance, the median distance between the stringencies of two judges with adjacent stringency

[54] To see this, note that

$$\mathrm{E}[\hat{r}] = \Pr[\hat{q}_A > \hat{q}_B]\,\mathrm{E}[\hat{q}_B \mid \hat{q}_A > \hat{q}_B] + \Pr[\hat{q}_A \leq \hat{q}_B]\,\mathrm{E}[\hat{q}_A \mid \hat{q}_A \leq \hat{q}_B]$$

and

$$q_B = \mathrm{E}[\hat{q}_B] = \Pr[\hat{q}_A > \hat{q}_B]\,\mathrm{E}[\hat{q}_B \mid \hat{q}_A > \hat{q}_B] + \Pr[\hat{q}_A \leq \hat{q}_B]\,\mathrm{E}[\hat{q}_B \mid \hat{q}_A \leq \hat{q}_B]$$

ranks in the São Paulo Appeal Court is 0.10.[55] This median judge pair—Judge Castro and Judge Coelho—decide 59 cases together. Coelho is stricter than Castro in 14% of these cases, and Castro is stricter than Coelho in 3% of the cases.[56] If these rates equal the true rates at which these two judges disagree, the bias in the estimated share of cases violating monotonicity for this judge pair is only 0.0002.[57] The bias is negligible since the probability of wrongly coding Judge Castro as strictest is low (3%), and, conditional on coding Judge Castro as strictest, the average difference between the two judges' estimated stringencies is small (0.01).

My estimates of the *share of disagreements* violating monotonicity, $q_A/(q_A+q_B)$, suffer from an additional bias due to the division by the estimated disagreement rate. This "denominator bias" is not straightforward to sign since the disagreement estimate in the denominator is not independent of the monotonicity violation estimate in the numerator. The denominator bias, and the presence of more than two judges, means that the sign of the bias in my monotonicity violation rate estimates is ultimately ambiguous.[58]

How large is the overall bias in my monotonicity violation estimates? To estimate this bias, I use the bootstrap. In particular, I draw 1,000 bootstrap samples of cases from my data and estimate the monotonicity violation rate for each sample. The bootstrap bias is then the difference between the mean of the 1,000 bootstrap estimates and my baseline estimate. The bootstrap bias is the bias in my estimator in the hypothetical case in which the true stringencies and disagreement rates equal those observed in the data.[59] Note that since the monotonicity violation rates are not everywhere differentiable in the judge stringencies, the bootstrap might be in-

[55] This median judge pair is in the Ninth Criminal Courtroom (9ª Câmara Criminal). The stringencies of the five judges in this courtroom, with the median judge pair in bold, are 0.42, 0.59, 0.71, **0.73**, and **0.83**.

[56] In the remaining 83% of the cases, the two judges vote in the same way.

[57] I obtain this bias estimate by drawing 10,000 random samples of 59 "cases" where the probability that Coelho is stricter (more lenient) than Castro in a case is 0.14 (0.03). The mean of $\hat{r}$—the estimated share of cases violating monotonicity—across these samples equals 0.0337 compared to the true rate of 0.0339.

[58] In simulations, the bias in each of the Table 2 and Table B.3 estimates can be either positive or negative, depending on the assumed true disagreement pattern across judges.

[59] For instance, in this hypothetical case, the true share of cases where Judge Coelho is stricter (more lenient) than Judge Castro is assumed to equal the sample rate of 14% (3%).

consistent.[60] Still, the bootstrap bias estimate gives an indication of the magnitude of the bias that would be expected for the sample sizes and stringency distributions in my data.

The results of this exercise are presented in Table B.4. The bias estimates are small. For instance, the bias in the Imbens-Angrist monotonicity violation rate in the São Paulo Appeal Court is estimated to be -0.003 compared to a baseline estimate of 0.346.

Table B.4: Bootstrap Bias Estimates

	São Paulo Appeal Court (Five judges)	Brazilian Superior Court (Five judges)	United States Supreme Court (Nine judges)
IA monotonicity violated	-0.0034	-0.0181	0.0074
Conditional number of violations	-0.0477	-0.0732	-0.1553
Average monotonicity violated	0.0018	0.0054	-0.0016
Sum of negative 2SLS weights	-0.0041	-0.0036	-0.0009

Note: Bootstrap bias estimates for the Table 2 estimates.

Another way to assess the small-sample bias in my estimates is to restrict the attention to judicial panels that decide a large number of cases together in which case their stringency order is precisely estimated. For instance, in panels deciding at least 100 cases, I can be highly confident of the stringency order.[61] The Table B.1 results, showing that my findings remain robust when restricting to panels that have decided at least 100 cases, thus also suggest that bias from measurement errors in judge stringencies is negligible.

[60] According to Theorem 3.1 of Fang and Santos (2019), the bootstrap is inconsistent in my setting if and only if the monotonicity violation rates are differentiable with respect to the true judge tendencies. This condition is violated whenever two judges share the same large-sample stringency levels.

[61] For example, in the São Paulo Appeal Court, in 92% of disagreements, I can reject at the 99.99% level that the two disagreeing judges have equal stringencies against the alternative that the judge estimated to be strictest is indeed strictest using a paired t-test. The monotonicity violation estimates do not materially change if I exclude disagreements where I am not confident of the stringency order at the 99.99% level.

Table B.5: Violations of Monotonicity Conditions: Leave-one-out stringency measures

	São Paulo Appeal Court (Five judges)	Brazilian Superior Court (Five judges)	United States Supreme Court (Nine judges)
IA monotonicity violated	0.358 (0.015)	0.553 (0.081)	0.428 (0.015)
Conditional number of violations	2.0 (0.1)	2.2 (0.2)	3.3 (0.1)
Average monotonicity violated	0.112 (0.011)	0.092 (0.032)	0.038 (0.005)
Sum of negative 2SLS weights	0.098 (0.011)	0.061 (0.028)	0.015 (0.003)

Note: Robustness of Table 2 to using leave-one-out stringency measures.

B.3.1 Leave-one-out and other split-sample approaches

One possible way to alleviate the downward bias in my monotonicity violation estimates is to use a split-sample approach, where the judges' stringency ordering is estimated in one sample, and the monotonicity violation rate given this stringency ordering is estimated in another sample. For instance, one can use a leave-one-out approach, where the stringency order is determined from the judges' votes in all cases except the one for which monotonicity violation is assessed. Such approaches will, however, lead us to *overestimate* the amount of monotonicity violations. Suppose Judge B is wrongly coded as strictest in the first sample. In that case, we will use $\hat{q}_A$ from the second sample as our monotonicity violation estimate—an overestimate of the true rate q_B. For example, in the leave-one-out approach, if Judge A and Judge B have the exact same observed stringencies, *all* disagreement between these two judges will be coded as monotonicity violations.

In Table B.5, I show the robustness of Table 2 to using leave-one-out stringency measures. The estimates of monotonicity violations are slightly higher, as expected. But the increases in the estimates are very small. The reason is that the leave-one-out approach only changes my conclusions about monotonicity violations for judge pairs with very similar stringencies, and I have few such instances in my data.

B.4 Do Conventional Tests Detect the Monotonicity Violations?

The judge IV literature has proposed various tests of monotonicity. In this section, I assess whether these tests are able to reject monotonicity in my settings. When running the tests, I disregard the fact that all judges vote in all cases. I thus delete the case identifier from my data set, treat each vote as an independent decision in a randomly assigned case, and run standard tests of monotonicity from the judge IV literature.[62]

Several authors have noted that monotonicity implies that judges who are strict in one case type must also be strict in other case types (Dobbie, Goldin, and Crystal S Yang 2018; Bhuller, Dahl, et al. 2020). Bhuller, Dahl, et al. (2020) test this implication by running a standard first-stage IV regression on various subsamples where the instrument—the judge's incarceration rate—is calculated based on cases outside of the subsample. Under monotonicity, these first-stage coefficients should be non-negative for all subsamples. I present results from such a test in Table B.6. For the São Paulo Appeal Court, I use the same subsamples as Bhuller, Dahl, et al. (2020): drug-related crimes, property crimes, violent crimes, and economic crimes.[63] For the US Supreme Court, I use the subsamples of Fourth, Fifth, Sixth, Eighth, and Fourteenth Amendment cases.[64] All estimates are highly statistically significant and positive. In other words, judges that are strict in, say, drug-related crimes also tend to be strict in other cases. And justices who tend to be strict when interpreting one constitutional amendment also tend to be strict when interpreting other constitutional amendments. This test is thus far from rejecting monotonicity.

Norris (2018) proposes a stronger version of this test: Assessing monotonicity for *individual judge pairs*. For instance, if Judge A is stricter than Judge B in drug-related cases, Judge A must also be stricter than Judge B in violent crime cases. The Norris (2018) test can be implemented by running the regression

[62] I thus allow the same case to appear as multiple observations—one for each judge. An alternative approach is to keep one observation for each case—the vote of a randomly chosen judge. While this approach would make the data closer to what is observed in judge IV designs, it will drastically reduce the size of the data and introduce statistical noise, making it harder to detect monotonicity violations.

[63] Using subsamples defined by specific crimes rather than broad crime types gives similar results.

[64] Using subsamples based on specific *clauses* of each amendment gives similar results. There are not enough observations to do a meaningful subsample analysis for the Brazilian Superior Court.

$$D_{ij} = \sum_{l=1}^{m} \alpha_l \mathbf{1}[l \leq j] + \varepsilon_{ij}$$

on subsamples where the judges $\{1, 2, \ldots, m\}$ are ordered by their overall stringencies.[65] Here, D_{ij} is an indicator for a pro-prosecution vote by judge j in case i, and ε_{ij} is an error term. The coefficient α_l captures the difference in the probability of a pro-prosecution vote between the lth most lenient judge and the $(l-1)$th most lenient judge. Under monotonicity, α_l should be non-negative for all l in all subsamples. When there are multiple panels, there are two ways of implementing this test. One approach is to define judge l as the lth most lenient judge in the panel and run a pooled regression across all panels. In Tables B.7 and B.8, I show the results from this full sample specification for the São Paulo Appeal Court and the US Supreme Court, respectively. Only two estimates are negative. For instance, the estimates indicate that the sixth most lenient justice in the US Supreme Court is more lenient than the fifth most lenient justice in Eighth Amendment cases. But the negative estimates are not statistically significant, and the tests can not reject monotonicity.

Another approach to this test is to run the test on one fixed panel. To maximize statistical power, I run this test on the panel that decides the most cases together.[66] The outcome of this test is presented in Tables B.9 and B.10. In the São Paulo Appeal Court, five out of 16 estimates are negative. But among the negative estimates, none are statistically significant at the five percent level. In the US Supreme Court, 13 out of 40 estimates are negative. Among the negative estimates, only one is statistically significant at the five percent level: Justice Stevens—who is usually stricter—is more lenient than Justice Ginsburg in Fourth Amendment cases. Taken at face value, this estimate is evidence of monotonicity violations for this judge pair. But if I take into account multiple testing—using the Wolak (1987) test of multiple

[65] Note that lenient-judge monotonicity can be tested by running the regression $D_{ij} = \alpha_1 + \sum_{l=2}^{m} \alpha_l \mathbf{1}[l = j] + \varepsilon_{ij} D_{ij}$. Stringent-judge monotonicity and extreme-pair monotonicity can be tested in similar ways. Since these conditions are weaker than Imbens-Angrist monotonicity, they are also not rejected by such tests in my data.

[66] One could run this test across all panels, not just the panels with the largest sample size. Doing this, however, will only exacerbate the problem of multiple testing without adding much in terms of statistical power—panels with fewer cases will inherently have much noisier estimates.

inequality constraints—I can not reject that all coefficients are non-negative.

Finally, I implement the joint test of monotonicity and the exclusion restriction proposed by B. Frandsen, L. Lefgren, and E. Leslie (2023). This test requires an outcome variable. To simulate an outcome variable, I follow Section 5 of B. R. Frandsen, L. J. Lefgren, and E. C. Leslie (2019) and assume a treatment effect of one for cases satisfying monotonicity and a treatment effect of -1 for cases violating monotonicity. This case can be seen as a worst-case scenario of heterogeneous effects, where the consequences of monotonicity violations are most severe. When implementing the B. Frandsen, L. Lefgren, and E. Leslie (2023) test, I need to take into account that random assignment of "cases" only holds within judicial panels.[67] Similarly, in typical judge IV applications, randomization happens only *within* courts. B. Frandsen, L. Lefgren, and E. Leslie (2023) provide two ways of dealing with multiple such "randomization units". The first approach is to assume *separable covariates*, run the test on the full sample, and control for fixed effects at the level of the randomization unit. In Table B.11, I show the results from this test. The test rejects both in the US Supreme Court and in the São Paulo Appeal Court (p-value < 0.005). This rejection, however, does not mean that we can reject monotonicity—it might be that the separable covariates assumption does not hold.[68] See Section B.4.1 below for a simple example where the test asymptotically rejects even though monotonicity is satisfied.

The second approach to implementing the B. Frandsen, L. Lefgren, and E. Leslie (2023) across randomization units is to run the test in each randomization unit and then adjust for multiple testing. In Table B.12, I show the results from implementing the B. Frandsen, L. Lefgren, and E. Leslie (2023) for the five panels deciding most cases in the São Paulo Appeal Court and in the US Supreme Court. I also report the output of the test for the panels where the test is closest to rejection. Even without adjusting for multiple testing the p-values are far from reaching conventional levels of statistical significance. Monotonicity can not be rejected.

[67] Remember, I treat each vote as an independent decision in a randomly assigned case. This hypothetical randomization is only valid within panels of judges deciding the same cases.

[68] Indeed, the separable covariates condition of B. Frandsen, L. Lefgren, and E. Leslie (2023) can not hold in these settings: The condition implicitly requires a full support assumption—that all judges appear in all randomization units—which is violated.

The graphical output of the test for the panels deciding most cases in each court is presented in Figure B.1.

Overall, even though 35 percent of São Paulo Appeal Court and 42 percent of US Supreme Court cases violate monotonicity, none of the tests employed in the literature reject monotonicity in neither court. The tests are not powerful enough to detect the type of monotonicity violations observed in these settings under the available sample sizes. Not rejecting monotonicity with these standard tests should thus not be seen as strong evidence in favor of monotonicity.

B.4.1 Fixed Effects in the Frandsen et al. (2023) Test

B. Frandsen, L. Lefgren, and E. Leslie (2023) provide two ways to implement their test across randomization units (*e.g.*, across different courts where random assignment of cases to judges only holds *within* courts). The first approach is to run the test on the full sample and control for fixed effects at the level of the randomization unit. This test relies on an additional separable covariates assumption.[69] Thus, if the test rejects, we can not conclude that monotonicity is violated. To illustrate this, I here show a simple example where the test rejects even though monotonicity is satisfied.

Assume there are two courts, Court 1 and Court 2, with two judges each. Cases are only randomly assigned within courts. Judge A_1 and Judge B_1 are in Court 1, and Judge A_2 and Judge B_2 are in Court 2. Assume the incarceration rate of Judge A_1 and A_2 equals a, and the incarceration rates of Judge B_1 and B_2 equal $b < a$. Assume monotonicity is satisfied. Thus, Judge A_1 and Judge A_2 have the same propensity to incarcerate—even after controlling for court fixed effects. However, the (residualized) expected outcomes differ between the two judges as long as the local average treatment effect (LATE) differ across the courts.[70] Thus, if LATE dif-

[69] The separable covariates assumption again relies on an implicit full support assumption which is violated unless all judges appear in all randomization units.

[70] The residualized expected outcome for Judge A_1 equals $\mathrm{E}[Y \mid j = A_1] - \mathrm{E}[Y \mid j \in \{A_1, B_1\}] = \frac{\mathrm{E}[Y|j=A_1] - \mathrm{E}[Y|j=B_1]}{2}$. This expression equals the local average treatment effect (LATE) in Court 1 multiplied by $\frac{a-b}{2}$. Similarly, the residualized expected outcome for Judge A_2 equals the LATE in Court 2 multiplied by $\frac{a-b}{2}$. Thus, the residualized expected outcomes in the two courts are equal if and only if the LATE in Court 1 equals the LATE in Court 2.

fers between the two courts, the B. Frandsen, L. Lefgren, and E. Leslie (2023) test will asymptotically reject monotonicity even though monotonicity is not violated.[71]

Table B.6: Reverse-sample tests.

	Dependent variable: Pro-prosecution vote			
Panel A: São Paulo Appeal Court				
	Drug crimes (1)	Property crimes (2)	Violent crimes (3)	Economic crimes (4)
Reverse sample pro-prosecution tendency	0.80 (0.04)	1.10 (0.04)	1.00 (0.08)	1.08 (0.10)
Observations	1,975	1,950	745	290
R^2	0.244	0.350	0.287	0.306

Panel B: US Supreme Court					
	4th Amd. (1)	5th Amd. (2)	6th Amd. (3)	8th Amd. (4)	14th Amd. (5)
Reverse sample pro-prosecution tendency	1.00 (0.03)	1.05 (0.04)	1.08 (0.04)	1.25 (0.04)	1.07 (0.04)
Observations	1,777	1,690	1,476	854	978
R^2	0.372	0.424	0.416	0.492	0.410

Note: First-stage IV regressions for subsamples of cases of different categories. The instrument, *reverse sample pro-prosecution tendency*, is the judge's rate of pro-prosecution votes in cases *not* of the indicated category. Unit of observation at the vote level. Panel fixed effects. Standard errors clustered at the case level in parentheses.

[71] The B. Frandsen, L. Lefgren, and E. Leslie (2023) test (asymptotically) rejects if two judges have the same stringencies but different average outcomes.

Table B.7: The Norris (2018) test for the São Paulo Appeal Court. Full sample approach.

	Dependent variable: Pro-prosecution vote.			
	Drug crimes	Property crimes	Violent crimes	Economic crimes
	(1)	(2)	(3)	(4)
Judge 2	0.200	0.446	0.403	0.466
	(0.041)	(0.034)	(0.064)	(0.109)
Judge 3	0.208	0.092	0.054	0.103
	(0.028)	(0.027)	(0.043)	(0.084)
Judge 4	0.089	0.125	0.148	0.138
	(0.025)	(0.024)	(0.040)	(0.060)
Judge 5	0.119	0.085	0.054	−0.001
	(0.022)	(0.020)	(0.029)	(0.063)
Observations	1,975	1,950	745	290
R^2	0.271	0.367	0.307	0.359

Note: The Norris (2018) test. The judge l estimate shows the difference in the rate of a pro-prosecution vote between the lth most lenient judge and the $(l-1)$th most lenient judge in the panel. Columns 1–4 show results for the subsamples of drug-related, property, violent, and economic crimes, respectively. Controlling for panel fixed effects. Unit of observation at the vote level. Criminal appeals in the São Paulo Appeal Court. Standard errors clustered at the case level in parentheses.

Table B.8: The Norris (2018) test for the US Supreme Court. Full sample approach.

	Dependent variable: Pro-prosecution vote.				
	4th Amd.	5th Amd.	6th Amd.	8th Amd.	14th Amd.
	(1)	(2)	(3)	(4)	(5)
Justice 2	0.045	0.127	0.006	0.063	0.037
	(0.026)	(0.033)	(0.020)	(0.033)	(0.026)
Justice 3	0.216	0.111	0.133	0.095	0.193
	(0.037)	(0.037)	(0.034)	(0.046)	(0.045)
Justice 4	0.241	0.169	0.170	0.084	0.028
	(0.041)	(0.044)	(0.041)	(0.045)	(0.054)
Justice 5	0.146	0.259	0.309	0.568	0.294
	(0.041)	(0.042)	(0.048)	(0.052)	(0.057)
Justice 6	0.040	0.069	0.042	−0.116	0.147
	(0.038)	(0.037)	(0.044)	(0.058)	(0.053)
Justice 7	−0.015	0.016	0.018	0.053	0.009
	(0.038)	(0.031)	(0.045)	(0.067)	(0.049)
Justice 8	0.121	0.085	0.061	0.147	0.083
	(0.032)	(0.027)	(0.039)	(0.045)	(0.048)
Justice 9	0.071	0.011	0.125	0.019	0.073
	(0.028)	(0.023)	(0.038)	(0.034)	(0.042)
Observations	1,777	1,690	1,476	854	978
R^2	0.406	0.427	0.425	0.527	0.417

Note: The Norris (2018) test. The Justice l estimate shows the difference in the rate of a pro-prosecution vote between the lth most lenient justice and the $(l-1)$th most lenient justice. The identity of the lth most lenient justice might change as the composition of the Court changes. Columns 1–5 show results for the subsamples of Fourth, Fifth, Sixth, Eighth, and Fourteenth Amendment cases, respectively. Controlling for panel fixed effects. Unit of observation at the vote level. US Supreme Court cases about criminal procedure. Standard errors clustered at the case level in parentheses.

Table B.9: The Norris (2018) test for the São Paulo Appeal Court. Restricted sample approach.

	Dependent variable: Pro-prosecution vote.			
	Drug crimes	Property crimes	Violent crimes	Economic crimes
	(1)	(2)	(3)	(4)
Ricardo Sale Junior	−0.097	0.857	0.703	0.700
	(0.136)	(0.046)	(0.086)	(0.159)
Willian Campos	0.290	0.044	−0.162	0.200
	(0.107)	(0.027)	(0.083)	(0.139)
Cláudio Marques	−0.032	0.022	0.216	−0.100
	(0.110)	(0.016)	(0.089)	(0.104)
Gilda Diodatti	0.097	−0.011	0.027	0.200
	(0.072)	(0.025)	(0.027)	(0.139)
Observations	155	455	185	50
R^2	0.071	0.740	0.447	0.577

Note: The Norris (2018) test for the five São Paulo Appeal Court judges deciding most cases together. In increasing stringency, these judges are Poças Leitão, Ricardo Sale Junior, Willian Campos, Cláudio Marques, and Gilda Diodatti. The estimates show the difference in the rate of pro-prosecution votes between the indicated judge and the judge ranked just below in stringency. The estimate on Ricardo Sale Junior is thus the difference between Ricardo Sale Junior and Poças Leitão, the estimate on Willian Campos is the difference between Willian Campos and Ricardo Sale Junior, and so on. Columns 1–4 show results for the subsamples of drug-related, property, violent, and economic crimes, respectively. Unit of observation at the vote level. Criminal appeals in the São Paulo Appeal Court. Standard errors clustered at the case level in parentheses.

Table B.10: The Norris (2018) test for the US Supreme Court. Restricted sample approach.

	Dependent variable: Pro-prosecution vote.				
	4th Amd.	5th Amd.	6th Amd.	8th Amd.	14th Amd.
	(1)	(2)	(3)	(4)	(5)
Stevens	−0.227	−0.063	0.000	−0.091	−0.143
	(0.115)	(0.148)	(0.108)	(0.095)	(0.279)
Souter	0.364	0.125	0.071	0.182	0.143
	(0.126)	(0.129)	(0.131)	(0.127)	(0.279)
Breyer	0.227	0.188	0.429	−0.091	0.143
	(0.115)	(0.140)	(0.142)	(0.095)	(0.279)
Kennedy	0.136	0.250	0.286	0.545	0.286
	(0.122)	(0.149)	(0.130)	(0.164)	(0.198)
Scalia	0.182	0.187	−0.143	0.364	0.143
	(0.109)	(0.169)	(0.148)	(0.159)	(0.279)
O'Connor	−0.182	−0.062	0.071	−0.273	−0.286
	(0.109)	(0.114)	(0.202)	(0.147)	(0.306)
Thomas	0.091	0.063	−0.143	0.182	0.286
	(0.093)	(0.114)	(0.213)	(0.190)	(0.306)
Rehnquist	0.045	0.063	0.286	0.000	−0.143
	(0.046)	(0.114)	(0.169)	(0.141)	(0.279)
Observations	198	144	126	99	63
R^2	0.378	0.390	0.424	0.596	0.252

Note: The Norris (2018) test for the nine justices deciding most cases together. In increasing stringency, these justices are Ginsburg, Stevens, Souter, Breyer, Kennedy, Scalia, O'Connor, Thomas, and Rehnquist. Only cases where all the nine justices vote. The estimates show the difference in the rate of pro-prosecution votes between the indicated justice and the justice ranked just below in stringency. The Stevens estimate is thus the difference between Stevens and Ginsburg, the Breyer estimate is the difference between Breyer and Stevens, and so on. Columns 1–5 show results for the subsamples of Fourth, Fifth, Sixth, Eighth, and Fourteenth Amendment cases, respectively. Unit of observation at the vote level. US Supreme Court cases about criminal procedure. Standard errors clustered at the case level in parentheses.

Table B.11: The B. Frandsen, L. Lefgren, and E. Leslie (2023) test.

	US Supreme Court	**São Paulo Appeal Court**
Fit-based p-value	0.001	0.000
Slope-based p-value	1.000	0.035
Combined p-value	0.003	0.000
Observations	13,564	6,530

Note: The B. Frandsen, L. Lefgren, and E. Leslie (2023) test under the assumption that covariates are separable. Three knots in the quadratic spline specification of the relationship between the outcome and the instrument propensity. Equal weight on the fit component and the slope component of the test in the combined p-value. The treatment effect for cases satisfying (violating) monotonicity is assumed to be one (-1). Unit of observation at the vote level.

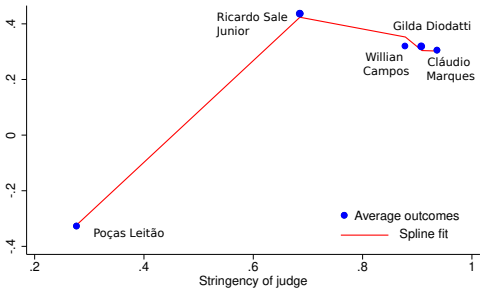
Table B.12: The B. Frandsen, L. Lefgren, and E. Leslie (2023) test.

Panel A: São Paulo Appeal Court						
Panel	1	2	3	4	5	Panel closest to rejecting
Fit-based p-value	-	-	-	-	-	-
Slope-based p-value	0.84	0.93	1.00	1.00	0.99	0.15
Combined p-value	1.00	1.00	1.00	1.00	1.00	0.30
Observations	1,125	455	445	435	390	115
Panel B: US Supreme Court						
Panel	1	2	3	4	5	Panel closest to rejecting
Fit-based p-value	0.70	0.13	0.59	0.19	0.38	0.88
Slope-based p-value	0.70	1.00	1.00	0.77	1.00	0.09
Combined p-value	1.00	0.25	1.00	0.39	0.75	0.18
Observations	1,638	1,557	1,377	1,170	828	603

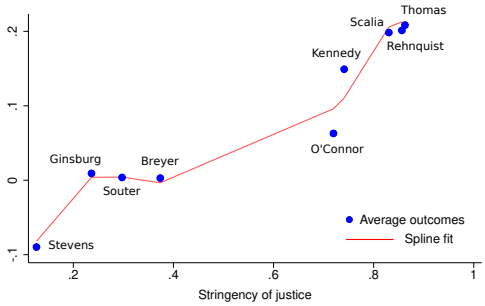
Note: The B. Frandsen, L. Lefgren, and E. Leslie (2023) test for the five panels deciding the most cases together in the São Paulo Appeal Court (Panel A) and the US Supreme Court (Panel B). Three knots in the quadratic spline specification of the relationship between the outcome and the instrument propensity. Equal weight on the fit component and the slope component of the test in the combined p-value. The treatment effect for cases satisfying (violating) monotonicity is assumed to be one (-1). Unit of observation at the vote level. The fit-based p-value for the São Paulo Appeal Court is missing since there are only five judges in the panel (allowing for a perfect fit). The panel closest to rejecting is the panel with the lowest combined p-value among all panels deciding at least 20 cases.

Figure B.1: The B. Frandsen, L. Lefgren, and E. Leslie (2023) test.

(a) São Paulo Appeal Court



(b) The US Supreme Court



Note: The graphical output of the B. Frandsen, L. Lefgren, and E. Leslie (2023) test for the panels deciding most cases together in the São Paulo Appeal Court and the US Supreme Court. The treatment effect for cases satisfying (violating) monotonicity is assumed to be one (-1). Three knots in the quadratic spline specification of the relationship between the outcome and the instrument propensity.

C More on Panel Effects

C.1 Peer Effects in the São Paulo Appeal Court

Judges might decide cases differently in panels than alone. Indeed, there is a large literature documenting that members of judicial panels tend to be influenced by their peers (Revesz 1997; Peresie 2004; Sunstein et al. 2007; Cox and Miles 2008; Boyd, Epstein, and Martin 2010; Kastellec 2013). In this section, I document similar peer effects in the São Paulo Appeal Court.

To get a quick sense of peer effects in the São Paulo Appeal Court, consider the following numbers: The method of Fischman (2014) applied to the first vote in all cases imples that two judges in the São Paulo Appeal Court should disagree in at least 7% of cases.[72] But when they vote together in the same panel, pairs of judges disagree in only 0.6% of cases. These numbers suggest substantial dissent aversion.

Another way to document panel effects is to exploit that, in three-judge appeals, the panel composition is randomly determined. For instance, consider the 15th criminal courtroom (15ª Câmara de Direito Criminal). In increasing order of seniority, this courtroom consists of the judges Gilda Diodatti, Cláudio Marques, Ricardo Sale Junior, Willian Campos, and Poças Leitão. As explained in Section 3, each case is randomly assigned to one judge (the *relator*), who writes a preliminary opinion and votes first. Then, the two judges with seniority below vote. For instance, in cases assigned to Judge Leitão, Judge Campos is the second-voting judge, and Judge Sale Junior is the third-voting judge. The identity of the randomly assigned first-voting judge thus generates random variation in the panel composition: For instance, Judge Sale Junior is randomly assigned to sit with either Judge Leitão and Judge Campos (as a third-voting judge), Judge Campos and Judge Marques (as a second-voting judge), and Judge Marques and Judge Diodatti (as a first-voting judge). Under no peer effects, the identity of the other panel members should not

[72] Fischman (2014) noted that if Judge A votes pro-prosecution in 50% of cases, Judge B votes pro-prosecution in 60% of cases, and cases are randomly assigned, the two judges must disagree in at least 10% of cases. I focus on first votes since—as shown later in this section—they are close to being unaffected by peer effects. To avoid overestimating stringency differences, I follow Copus and Hübert (Forthcoming) and use half the sample to determine the stringency order and the remaining sample to estimate stringency differences given this ordering.

matter for a judge's vote. To test this prediction, I run the following regression:

$$d_{ji} = \alpha_{jt(i)} + \beta p_{-j,i} + \varepsilon_{ji} \tag{3}$$

Here, d_{ji} is an indicator for judge j voting for the prosecution in case i, $\alpha_{jt(i)}$ is a judge-by-year fixed effect, and $p_{-j,i}$ is the sum of the *stringencies* of the two other members of the panel. A judge's *stringency* is measured by the judge's rate of pro-prosecution votes in cases where this judge votes first, leaving out the current case.[73] To avoid including substitute judges, I only keep observations from panels that decide at least 100 cases together in the given year.

Under no panel effects, $\beta = 0$. This hypothesis is strongly rejected by the results of estimating Equation 3, reported in Panel A of Table C.1. The Column 3 estimate indicates that a 10 percentage points increase of a peer's stringency increases the probability of a pro-prosecution vote by 3.5 percentage points. The randomization test in Table C.2 suggests that this result is not driven by panels with stricter members being assigned different cases: Consistent with randomization, there is no statistically significant correlation between the stringency of peers and observable case characteristics.

Which of the three judges on the panel is most influential? To answer this question, I exploit that judge absences—for instance, due to health reasons—generate further variation in panel compositions. For instance, in February 2019, Poças Leitão was away. This absence meant that Judge Sale Junior became the third judge in cases assigned to Judge Diodatti—a stricter judge than Judge Leitão.[74] To assess whether Judge Sales Junior is affected by the identity of the first-voting judge, we can thus compare his votes as a third-voting judge in February to his votes as a third-voting judge in the rest of the year. Formally, I run the following regressions:

$$d_{1i} = \alpha_{j_1(i)t(i)} + p_{2i} + p_{3i} + \varepsilon_{1i} \tag{4}$$

$$d_{2i} = \alpha_{j_2(i)t(i)} + p_{1i} + p_{3i} + \varepsilon_{2i} \tag{5}$$

[73] This measure is a good proxy for a judge's true stringency since—as shown later in this section—judges are only weakly influenced by their peers when they vote first.

[74] In parts of the month, Judge Leitão was replaced by a substitute judge, creating further variation in panel compositions.

Table C.1: Peer Effects

Panel A: Overall Peer Effects			
	Dep. Var.: Pro-prosecution vote		
	(1)	(2)	(3)
Stringency of peers	0.492	0.346	0.349
	(0.018)	(0.033)	(0.034)
Judge × Year FE		✓	✓
Case controls			✓
Observations	906,729	906,729	906,729
R^2	0.142	0.193	0.236
Panel B: Peer Effects By Vote Order			
	Dep. Var.: Pro-prosecution vote		
	First judge	Second judge	Third judge
	(1)	(2)	(3)
First judge stringency		0.759	0.815
		(0.058)	(0.040)
Second judge stringency	0.073		0.061
	(0.024)		(0.024)
Third judge stringency	0.028	0.055	
	(0.018)	(0.020)	
Case controls	✓	✓	✓
Judge × Year FE	✓	✓	✓
Observations	368,474	368,477	368,517
R^2	0.259	0.254	0.250

Note: Stringency of peers is the sum of the stringencies of the two other judges in the panel. A judge's stringency is measured as the judge's rate of pro-prosecution votes in cases where the judge votes first. Case controls are broad crime categories interacted with whether the prosecutor appealed. Panel A uses only panels of judges that decide at least 100 cases in the given year. Unit of observation at the vote level in Panel A and at the case level in Panel B. São Paulo criminal appeals decided by three-judge panels. Standard errors clustered at the judge level in parentheses.

Table C.2: Randomization Test

	Dependent variable:			
	Stringency of Peers		Pro-Prosecution Vote	
	Coef.	(se)	Coef.	(se)
Drug-Related Crime	0.001	(0.001)	−0.0002	(0.008)
Violent Crime	0.0001	(0.001)	0.024	(0.004)
Economic Crime	0.0002	(0.001)	−0.059	(0.008)
Property Crime	0.0005	(0.001)	−0.007	(0.005)
Prosecutor Appealed	0.001	(0.001)	−0.315	(0.009)
Drug-Related Crime × Prosecutor Appealed	−0.0002	(0.002)	0.042	(0.011)
Violent Crime × Prosecutor Appealed	−0.002	(0.001)	−0.063	(0.008)
Economic Chime × Prosecutor Appealed	0.001	(0.002)	0.022	(0.009)
Property Crime × Prosecutor Appealed	0.0003	(0.001)	0.075	(0.007)
F-statistic	1.37		280.95	
p-value of F-statistic	0.208		0	
Observations	906,729		906,729	
R^2	0.918		0.230	

Note: São Paulo criminal appeals decided by three-judge panels. Unit of observation at the vote level. *Stringency of peers* is the sum of the stringencies of the two other judges in the panel. A judge's stringency is measured as the judge's rate of pro-prosecution votes in cases where the judge votes first. The excluded crime type is "other crimes". Including only observations with panels of judges that decide at least 100 cases together in the given year. Standard errors clustered at the judge level.

Table C.3: Randomization Test

	Dependent variable:					
	First Judge Stringency		Second Judge Stringency		Third Judge Stringency	
Drug-Related Crime	0.0001	−0.001	0.001	0.001	0.0003	−0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Violent Crime	−0.0004	−0.00001	0.001	0.0001	−0.001	−0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Economic Crime	0.00004	0.001	0.001	0.001	−0.001	−0.0002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Property Crime	−0.0004	−0.001	0.001	0.001	−0.0002	−0.0005
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Prosecutor Appealed	0.002	0.003	0.0001	−0.001	−0.002	−0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Drug-Related Crime	−0.001	0.001	−0.001	0.0004	0.003	0.002
× Prosecutor Appealed	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Violent Crime	−0.002	−0.001	−0.001	0.0003	0.001	0.001
× Prosecutor Appealed	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
Economic Crime	−0.003	−0.003	0.001	0.0002	0.003	0.0002
× Prosecutor Appealed	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)
Property Crime	−0.002	−0.002	−0.001	0.0001	0.002	0.002
× Prosecutor Appealed	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
First Judge × Year FE			Yes		Yes	
Second Judge × Year FE	Yes					Yes
Third Judge × Year FE		Yes		Yes		
F-statistic	2.12	2.00	0.73	1.07	2.13	1.24
p-value of F-statistic	0.032	0.044	0.683	0.391	0.031	0.277
Observations	368,520	368,520	368,517	368,517	368,477	368,477

Note: São Paulo criminal appeals decided by three-judge panels. Unit of observation at the case level. A judge's stringency is measured as the judge's rate of pro-prosecution votes in cases where the judge votes first. The excluded category is "other crimes". Standard errors clustered at the judge level in parentheses.

$$d_{3i} = \alpha_{j_3(i)t(i)} + p_{1i} + p_{2i} + \varepsilon_{3i} \tag{6}$$

Here, d_{ki} is an indicator for the kth judge voting for the prosecution in case i, p_{ki} is the stringency of the kth judge in case i, and $\alpha_{j_k(i)t(i)}$ are kth-judge-by-year fixed effects. Equation 4 thus assesses how a given first-voting judge responds to being in panels with different second-voting and third-voting judges throughout a given year.

This specification relies on the following identification assumption: The changes in the stringencies of peers generated by judge absences are uncorrelated with case characteristics. Random assignment of cases is not sufficient for this identification assumption to hold since the composition of cases could change over the year. The condition would be violated if, for instance, the most stringent judges tended to be more absent in months with weaker appeals. The condition can, however, be tested on observable case characteristics. In Table C.3, I show the results of such a test. Indeed, changes in the stringencies of peers generated by the absences are not statistically related to observed case characteristics.

The results from estimating Equations 4–6 in Panel B of Table C.1 indicate that the first-voting judge is the most influential judge. For instance, Column 2 indicates that a ten percentage point increase in the stringency of the first-voting judge increases the rate of pro-prosecution votes by the second-voting judge by 7.6 percentage points. In comparison, increasing the stringency of the second-voting judge by ten percentage points only increases the rate of pro-prosecution votes by the first-voting judge by only 0.7 percentage points. The third-voting judge appears to be the least influential judge. These results suggest that the first-voting judges tend to vote according to their own preferences while the second and third judges tend to vote according to the first judge's vote, perhaps due to dissent aversion.

C.2 Assumption 3 and the Fischman (2011) Model

In this section, I show that there are reasonable parameter values in the Fischman (2011) model of panel effects under which Assumption 3 is satisfied. To see this, consider the calibrated parameter values (p. 796) for the three judges deciding most cases in Fischman (2011)'s data: Judge Brunetti, Judge Fernandez, and Judge

Schroeder. Under these parameter values, the model predicts that, in individual decisions, the judges disagree in ways satisfying (violating) monotonicity in 22% (6%) of cases.[75] But when deciding cases together in a panel, they disagree in ways satsifying (violating) monotonicity in 0.75% (0.26%) of cases. Assumption 3 is thus very close to being satisfied:

$$\frac{\Pr\left[S^p \in \mathcal{S}_v\right]}{\Pr\left[S \in \mathcal{S}_v\right]} = \frac{0.26\%}{6\%} = 0.043 \approx 0.034 = \frac{0.75\%}{22\%} = \frac{\Pr\left[S^p \in \mathcal{S}_s\right]}{\Pr\left[S \in \mathcal{S}_s\right]}$$

With minor changes in the parameter values, Assumption 3 can be made to hold exactly. This example shows that there are reasonable parameterizations of the Fischman (2011) model under which Assumption 3 holds. Whether Assumption 3 holds in practice is ultimately an empirical question, which I address in Section 5.3.

C.3 Testing Assumption 4

In Section C.1, I documented strong peer effects among the three first-voting judges in São Paulo criminal appeals. In particular, if Judge A and Judge B vote first, Judge B has a very strong tendency to follow A's vote. Assumption 4 in Section 5.3 assumed there is no such influence when Judge A and B are the two last-voting judges in five-judge cases. I test Assumption 4 here.

In particular, I run similar peer effects regressions as in Section C.1 to test whether the fifth judge's vote is influenced by the stringency of the fourth-voting judge.[76] Column 1 of Table C.4 shows the result of regressing the fifth judge's vote on the fourth judge's stringency in five-judge cases controlling for year fixed

[75] These numbers are estimated by drawing 1,000,000 cases from the model.

[76] The last two judges could, of course, also be influenced by the three first-voting judges. But such influence would be less concerning for the Section 5.3 analysis since it can be thought of as "external" influence affecting both judges. Such influence would be similar to how precedents—decisions by other judges in *other* cases—influence all judges, a type of peer influence that exists also in settings where judges decide cases individually. Empirically identifying the influence of the three first-voting judges would also be challenging since these judges determine which cases end up in the sample of five-judge cases. I thus focus on the fourth judge's influence on the fifth judge. Note that there might also be peer effects in the other direction: the fourth judge being influenced by the fifth judge. But the Section C.1 evidence suggests that the influence of later-voting judges on earlier-voting judges is very limited.

Table C.4: Peer Effects Between The Fourth- and Fifth-Voting Judge

	Dep. Var.: Fifth judge pro-prosecution vote		
	(1)	(2)	(3)
Fourth-judge stringency	0.011	0.019	0.036
	(0.066)	(0.067)	(0.082)
Year FE	Yes	Yes	Yes
Controlling for crime type	No	Yes	Yes
Controlling for stringency of first three judges	No	No	Yes
Observations	1,261	1,261	1,261

Note: São Paulo criminal appeals decided by five judges. The outcome variable is whether the fifth-voting judge votes in favor of the prosecution. The crime types are economic crimes, drug-related crimes, violent crimes, property crimes, and other crimes. *Fourth-judge stringency* is the fourth-voting judge's rate of pro-prosecution votes in cases where that judge votes first. Standard errors in parentheses.

effects.[77]

Consistent with no peer effects, the estimated coefficient is small and statistically insignificant. This estimate should, however, be interpreted with caution. While initial appeals are randomly assigned to first-voting judges, the selection into five-judge cases depends on the first three judges. If the stringency of the fourth-voting judge is correlated with the types of cases the preceding three judges tend to "transform" into five-judge cases, the peer effect estimate might be biased. Indeed, in Table C.5, Column 1, I show that five-judge cases about property crimes tend to have more lenient fourth-voting judges than other cases.[78] In an effort to control for this selection, I add controls for the stringencies of the three first-voting judges. When rerunning the balance test in Column 2, I can not reject that five-judge cases

[77]The identifying assumption is that the fourth-voting judge's stringency is uncorrelated with other factors determining the fifth judge's vote conditional on the controls. This specification is weaker than the Section C.1 specifications (Equations 4–6) since, due to insufficient variation in the data, I am unable to control for fifth-voting judge fixed effects. But the identifying assumption can still be tested on observable case characteristics (see Table C.5).

[78]The Table C.3 balance test also included whether the appeal was filed by a prosecutor. This covariate is not included in Table C.5 since who filed the initial appeal is not easily extracted from five-judge cases.

Table C.5: Randomization Test for Peer Effects Between Fourth and Fifth Judge

	Dep. Var.: Fourth judge stringency	
	(1)	(2)
Drug-related crimes	0.005	0.014
	(0.16)	(0.013)
Violent crimes	-0.011	0.010
	(0.20)	(0.017)
Economic crimes	-0.033	-0.008
	(0.030)	(0.024)
Property crimes	-0.065	-0.006
	(0.016)	(0.013)
Year FE	Yes	Yes
Controlling for stringency of first three judges	No	Yes
F-statistic	7.07	0.84
p-value of F-statistic	0.00	0.50
Observations	1,261	1,261
R^2	0.10	0.41

Note: São Paulo criminal appeals decided by five judges. The outcome variable is the stringency of the fourth-voting judge. A judge's stringency is measured as the judge's rate of pro-prosecution votes in cases where the judge votes first. Standard errors in parentheses.

are as good as randomly assigned to fourth-voting judges conditional on the stringencies of the three first-voting judges. In Columns 2 and 3 of Table C.4, I run updated peer effect regressions while controlling for crime type and the three first-voting judges' stringencies, respectively. The estimate changes only marginally. In these more credible specifications, I can also not reject the hypothesis of no peer effects.

The point estimates suggest very low levels of peer effects. Specifically, the Column 3 estimate indicates that a ten percentage point increase in the stringency of the fourth-voting judge raises the likelihood of a pro-prosecution vote by the fifth-voting judge by only 0.36 percentage points. In contrast, peer effects between the two first-voting judges are several magnitudes larger. The estimate in Table C.1 shows that a ten percentage point increase in the stringency of the first-voting judge increases the likelihood of a pro-prosecution vote by the second-voting judge by 7.6 percentage points. Put differently, the point estimates suggest that voting after a dissent eliminates approximately 95% of peer effects ($1 - 0.36/7.6$). As noted in Footnote 43, such small remaining peer effects would not affect the conclusion that Assumption 3 is likely close to being satisfied.

C.4 Testing Assumption 5

Assumption 5 requires that the selection of cases into the five-judge sample is uncorrelated with monotonicity violation rates. I test this assumption here.

To derive a formal test, let's first state the general version of Assumption 5.[79] Redefine T to indicate whether there is a dissenting opinion and a majority vote against the defendant among the randomly assigned three first-voting judges.[80] Let G be whether the fourth- and fifth-voting judges would disagree with each other if they get a chance to vote, and let M be whether such a disagreement violates monotonicity.[81] The general version of Assumption 5 is then

[79] Assumption 5 is stated only for one pair of judges—"Judge 1" and "Judge 2". Assumption C.1 below generalizes Assumption 5 to cases with any fourth- and fifth-voting judges. It is straightforward to show that a similarly generalized version of Proposition 3 holds under Assumption C.1.

[80] *I.e.*, the condition under which the case enters the five-judge sample.

[81] Assumption 5 was stated in terms of the fourth- and fifth-voting judges' decisions absent panel effects. But under Assumption 4 these decisions are identical to how they would decide as fourth-

Assumption C.1 (Selection orthogonal to monotonicity)**.**

$$T \perp M \mid G$$

Since the three first-voting judges are randomly assigned, whether a given case enters the five-judge sample is non-deterministic. To model this, let X be a vector representing all case characteristics relevant to the judges' decisions, and let $q(x) > 0$ denote the probability that a case with characteristics $X = x$ enters the five-judge sample.[82] Whether the case actually ends up in the sample depends on the randomly assigned first three judges and their idiosyncrasies.[83] Define $Q \equiv q(X)$. Assume that, conditional on Q, the selection into the five-judge sample is independent of the direction in which the fourth- and the fifth-voting judges would disagree:

Assumption C.2 (Selection orthogonal to monotonicity conditional on Q)**.**

$$T \perp M \mid G, Q$$

This is a relatively mild assumption. Random judge assignment ensures that T is independent of observable and unobservable case characteristics once we condition on Q. Assumption C.2 could still be violated if the monotonicity violation rates of fourth- and fifth-voting judges are systematically correlated with the three first-voting judges' tendency to generate five-judge cases. But any such correlation is likely minimal since the fourth and fifth judges are exogenously determined by the seniority order, with the exact voting order often dictated by small differences in seniority.[84] Empirically, I find no evidence of such a correlation when using my

and fifth-voting judges.

[82] All cases are thus assumed to have a non-zero probability of entering the five-judge sample. If the probability is exactly zero for some cases, the analysis below applies only to the subset of cases with a non-zero probability.

[83] Other idiosyncratic factors, such as the time of day and judges' attentiveness, can be thought of as additional sources of quasi-random variation in selection.

[84] See Section 3. The median gap between appointment dates for judge pairs in the same courtroom with consecutive seniority ranks is just 1.2 years. A quarter of these pairs were appointed less than four months apart. Newly appointed judges are assigned to courtrooms based on vacancies created by the retirement of existing judges. There is little overall correlation between seniority and monotonicity violations (see Section D.3).

estimated monotonicity violation rates.[85]

Under Assumption C.2, we can evaluate Assumption C.1 by testing whether Q is uncorrelated with monotonicity violations in the sample of five-judge cases where the last two judges disagree:

Assumption C.3 (Montonicity orthogonal to selection probability)**.**

$$M \perp Q \mid G, T = 1$$

Proposition C.1. *Assumptions C.2 and C.3 imply Assumption C.1.*

Assumption C.3 requires that the probability that a case enters the five-judge sample is independent of the direction in which judges disagree. This assumption can be tested by running the following regression on the sample of five-judge cases where the fourth and the fifth judge disagree:

$$m_i = \alpha + \beta q_i + \varepsilon_i \tag{7}$$

where m_i and q_i are realizations of M and Q, respectively. Assumption C.3 requires that $\beta = 0$.

Since Q cannot be directly observed, I estimate it using observable characteristics. A key advantage of my setting for this exercise is access to highly detailed case information, including the full trial court decision from Tribunal de Justiça de São Paulo (2022) and a summary of the appeal arguments.[86] These features capture nearly all case characteristics that are observable to the judges. Unobservable characteristics that significantly influence selection are thus unlikely.[87]

[85] In an OLS regression, a one standard deviation increase in the tendency of the three first-voting judges to generate five-judge cases is associated with a 1.3 percentage point lower estimated monotonicity violation rate for the last two judges. This correlation is not statistically significant.

[86] The median trial judge's decision is 10,000 characters long and contains a detailed description of the facts of the case and the judge's justification of their decision. The median summary of the appeal arguments is 1,600 characters long and contains the key arguments made by both the defendant's lawyer and the prosecutor. This summary is extracted from the opinion of the first-voting judge.

[87] The key unobservable characteristics are the lawyers' arguments that are not included in the first-voting judge's summary. But it seems unlikely that arguments deemed irrelevant to the case by the first-voting judge would be an important predictor of selection into the five-judge sample. In any case, such unobservables would only pose a problem if they also are strongly correlated with mono-

I estimate Q using a gradient-boosted decision tree model (XGBoost) with the following features:[88] narrow crime types, 300-dimensional document embeddings of the trial court decision and appeal arguments, trial court opinion length (in characters), indicators for missing documents, and indicators for whether the prosecutor appealed and whether the trial judge convicted the defendant.[89] I train the model on 40% of all São Paulo criminal appeals ($N = 147,732$) and predict Q in the remaining 60% test sample. Equation 7 is then estimated on the five-judge cases in the test sample. I also estimate a model where Q is proxied by the share of cases with the same crime type that enters the five-judge sample.

The results are presented in Table C.6.[90] For the specification using all features, in Column 5, a one standard deviation increase in the propensity of entering the five-judge sample is associated with a 1.3 percentage point higher monotonicity violation rate. Consistent with Assumption C.3, this correlation is small and statistically insignificant. Cases more likely to select into the five-judge sample do not seem to violate monotonicity at very different rates than cases less likely to enter the sample.

To assess how the estimated correlation would affect the monotonicity violation rate, I calculate a reweighted monotonicity violation rate, where a case predicted to enter the five-judge sample with probability $Q = q$ receives weight proportional to $1/q$. The resulting rate gives the monotonicity violation rate in the

tonicity violations—which also seems unlikely given that (as documented in Table C.6) there is no such correlation between observable characteristics determining selection into the five-judge sample and monotonicity violations. Unobservable characteristics that moderately affect selection or are only moderately correlated with monotonicity violations will not change the broad conclusions of the paper. For example, suppose the defendant's lawyer delivers an exceptionally strong performance during the oral hearing in 10% of cases, increasing the likelihood of the case being included in the five-judge sample by 40%. Even if these cases violate monotonicity at a rate 50% higher than other cases, the resulting bias in the monotonicity violation rate is still only $10\% \times 40\% \times 50\% = 2\%$ (i.e., 0.0024 if the true monotonicity violation rate is 0.12).

[88] I select hyperparameters using 3-fold cross-validation with ROC AUC as the scoring metric.

[89] For the document embeddings, I use a TF-IDF weighted average of FastText word embeddings. This approach assigns higher importance to case-specific terms—those frequently appearing in certain decisions but not across all cases—while preserving semantic relationships between words. I select the 5,000 most important tokens.

[90] Note that the accuracies of the predictions of selection into the five-judge sample are only moderate, with an AUC of 0.73 for the full model. This is expected since an important explanator of the selection, the randomly assigned three first-voting judges, is not included.

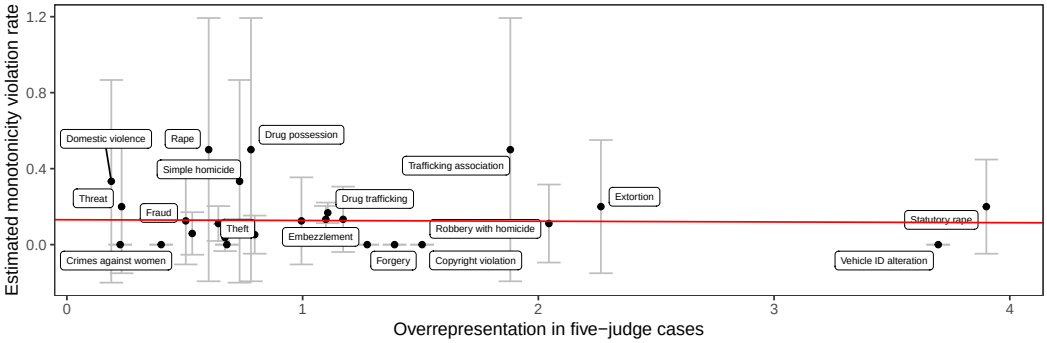
Table C.6: Testing Assumption 5

		Dep. Var.: Monotonicity violation				
	Baseline	(1)	(2)	(3)	(4)	(5)
Five-judge propensity (normalized)		0.001	0.003	-0.018	0.021	0.013
		(0.015)	(0.014)	(0.013)	(0.017)	(0.022)
Observations		557	320	320	320	320
AUC-ROC of prediction			0.61	0.60	0.58	0.73
Reweighted monotonicity violation rate	0.12	0.12	0.12	0.13	0.10	0.11
XGBoost			Yes	Yes	Yes	Yes
Features used						
Narrow crime type		Yes	Yes	Yes	Yes	Yes
Additional case data				Yes	Yes	Yes
Trial court decision embeddings					Yes	Yes
Appeal arguments embeddings						Yes

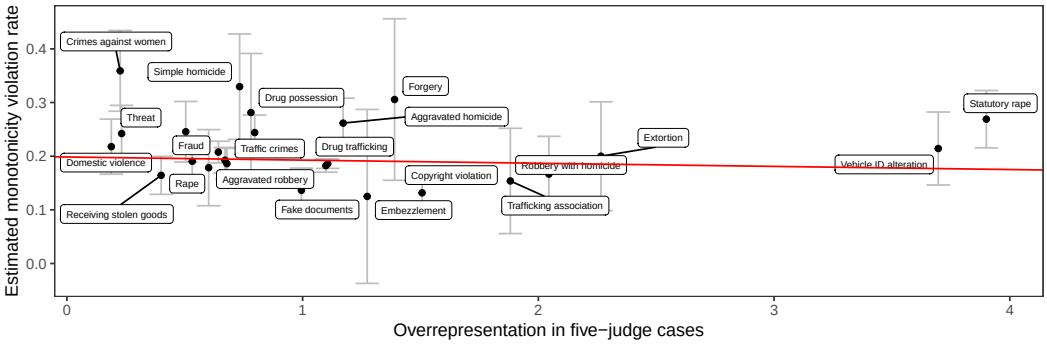
Note: Results from estimating Equation 7 using different estimates of Q. São Paulo criminal appeals decided by five-judge panels where the fourth and the fifth judge disagree. The *five-judge propensity* is the case's predicted probability of entering the five-judge sample based on the indicated observable characteristics. Columns 2–5 estimate this propensity using XGBoost, trained in a 40% holdout sample with hyperparameters selected via 3-fold cross-validation. The AUC-ROC metric assesses the accuracy of these predictions. Column 1 estimates the propensity by the proportion of cases within the same narrow crime category that enter the five-judge sample. The dependent variable is whether, between the fourth and fifth judges, the otherwise most lenient judge is the strictest in the case. The *reweighted monotonicity violation rate* is the inverse-propensity-weighted share of disagreements between the fourth and the fifth judge that violate monotonicity. *Baseline* shows the unweighted monotonicity violation rate. *Additional case data* are trial court opinion length (in characters) and indicators for whether the prosecutor appealed and whether the defendant was convicted by the trial judge. *Trial court decision* and *appeal argument embeddings* are TF-IDF weighted averages of 300-dimensional FastText word embeddings. Indicators for missing documents are included in Columns 4 and 5. Bootstrapped standard errors in parentheses.

Figure C.1: Monotonicity Violations and Selection Into Five-Judge Cases

(a) Monotonicity violations between the fourth and the fifth judge



(b) Monotonicity violations among the first three judges



Note: The relationship between estimated monotonicity violations and the overrepresentation in five-judge cases across crimes in São Paulo criminal appeals. Crimes that appear in more than 1,000 cases. The *five-judge overrepresentation* of a crime is the share of five-judge cases discussing the crime divided by the share of all cases discussing the crime. The *estimated monotonicity violation rate* is the share of disagreements in cases about the given crime where the otherwise strictest judge is most lenient. Figure (a) estimates this rate among the two last-voting judges in five-judge appeals. Figure (b) estimates the rate among the three first-voting judges. 95% confidence intervals. The red lines show the Equation 7 estimated linear fits.

full—unselected—sample of criminal appeals, if Assumption C.2 holds and Q is correctly estimated. The reweighted monotonicity violation rate is 0.11, marginally lower than the unweighted rate at 0.12.[91] This result suggests that the monotonicity violation rate is close to unaffected by selection effects. I obtain similar results for the other ways of estimating Q in Columns 1–4.

The results from the Column 1 exercise is presented graphically in Figure C.1. The x-axis shows the *five-judge overrepresentation* for crimes appearing in at least 1,000 cases, estimated by dividing the share of five-judge cases involving the crime by the share of *all* cases involving the crime.[92] Statutory rape is the most overrepresented, appearing nearly four times more often in five-judge cases than in all cases, while domestic violence is the most underrepresented, appearing five times less often. The y-axis of Figure C.1a shows the estimated monotonicity violation rate for each crime based on the votes of the fourth and fifth judges. Consistent with Assumption 5, there appears to be no systematic relationship between a crime's five-judge overrepresentation and its monotonicity violation rate. For example, the monotonicity violation rates for statutory rape and domestic violence cases are statistically indistinguishable. The same holds true when I estimate monotonicity violation rates using the votes of the first three judges in Figure C.1b.[93]

Overall, the evidence presented in this section suggests that Assumption 5 is close to satisfied. While Assumption 5 might not hold exactly, small deviations from it would not substantially affect the broad conclusions of the paper.[94]

C.5 Testing Assumption 3 on Observable Case Characteristics

Monotonicity can not be directly measured in standard judge IV settings where cases are randomly assigned to individual judges. But one can obtain a lower bound

[91] The unweighted rate at 0.12 differs from the Table 3 estimate because that table reports a weighted mean across judges.

[92] The five-judge overrepresentation is a linear transformation of the five-judge propensity that is easier to interpret.

[93] Using the first three judges increases precision at the cost of potential bias due to panel effects.

[94] For instance, if the cases that enter the five-judge sample violate monotonicity at a rate 10% lower than the overall rate, a monotonicity violation rate of 0.12 in the five-judge sample implies a monotonicity violation rate of $0.12/0.9 = 0.13$ in the full sample.

of monotonicity violations by exploiting observable case characteristics (Norris 2018). For instance, assume Judge 1 (Judge 2) incarcerates in 50% (70%) of violence cases and 60% (50%) of drug-related cases in individual decisions. If Judge 2 is overall the strictest judge, then monotonicity must be violated in at least 10% of drug-related cases. Similarly, the two judges must disagree in a way that satisfies monotonicity in at least 20% of violence cases.

If Assumption 3 holds, we would expect it to also hold on such "observable" disagreements. For instance, assume that, when deciding together in a *panel*, Judge 1 (Judge 2) incarcerates in 55% (65%) of violence cases. Due to panel effects, their stringency difference in violence cases is halved. If panel effects reduce disagreements violating monotonicity at the same rate as disagreements satisfying monotonicity (Assumption 3), we would expect the stringency difference to be halved also in drug-related cases.[95] In this section, I test whether such an "observable" analog to Assumption 3 holds in my data.

To formalize this test, let the random variable $C \in \mathcal{C}$ denote an observable "case type" (*e.g.*, crime type).[96] Denote by $\mathcal{C}_v \subset \mathcal{C}$ the case types violating monotonicity.[97] Define *observable disagreements violating monotonicity* by

$$d_{vo} \equiv \mathrm{E}\left[S(1) - S(2) \mid C \in \mathcal{C}_v\right] \Pr\left[C \in \mathcal{C}_v\right]$$

This parameter—discussed by Norris (2018)—uses observable case characteristics to obtain a lower bound of monotonicity violations. It is a lower bound for two reasons. First, Judge 1 might be stricter than Judge 2 also in some $C \notin \mathcal{C}_v$ cases. Second, Judge 2 might be stricter than Judge 1 also in some $C \in \mathcal{C}_v$ cases, in which case the stringency difference $\mathrm{E}\left[S(1) - S(2) \mid C \in \mathcal{C}_v\right]$ underestimates the share of $C \in \mathcal{C}_v$ cases that violate monotonicity.[98] The magnitude of such "unobservable"

[95] This could be obtained by, for instance, Judge 1 (Judge 2) incarcerating in 57.5% (52.5%) of drug-related cases.

[96] Such a case type could be determined by any vector of fixed observable characterstics of the case. But due to statistical power issues, I will stick to relatively broad case categories in the empirical tests.

[97] The case types where the overall strictest judge is more lenient.

[98] In the motivating example, if Judge 2 is strictest in 5% of drug-related cases, Judge 1 must be strictest in 15% of drug-related cases for their stringency difference in drug-related cases to be 10%.

monotonicity violations is given by:[99]

$$d_{vu} \equiv \Pr\left[S\left(1\right) > S\left(2\right), C \notin C_v\right] + \Pr\left[S\left(2\right) > S\left(1\right), C \in C_v\right]$$

For instance, assume monotonicity holds conditional on crime type and defendant gender, but gender is unobserved. Then d_{vu} are the monotonicity violations along the unobserved gender dimension. Finally, define *observable disagreements satisfying monotonicity* by

$$d_{so} \equiv \mathrm{E}\left[S\left(2\right) - S\left(1\right) \mid C \notin C_v\right] \Pr\left[C \notin C_v\right]$$

Let d_{vo}^p, d_{vu}^p, and d_{so}^p be the corresponding variables when the judges decide cases in a panel (replacing S with S^p). The observable analog to Assumption 3 is then:

Assumption C.4 (Panel effects reduce observable disagreements satisfying and violating monotonicity at the same rate.)**.**

$$\frac{d_{vo}^p}{d_{vo}} = \frac{d_{so}^p}{d_{so}}$$

Assumption C.4 holding can be thought of as suggestive evidence in favor of Assumption 3. It is only *suggestive* since panel effects might influence unobservable disagreements at a different rate. Formally:

Proposition C.2. *Assume $d_{vu}^p/d_{vu} = d_{vo}^p/d_{vo}$. Then Assumption 3 holds for Judges 1 and 2 if and only if Assumption C.4 holds.*

The condition $d_{vu}^p/d_{vu} = d_{vo}^p/d_{vo}$—panel effects reducing observable and unobservable monotonicity violations at the same rate—while a natural condition, is hard to test. But unless the influence of panel effects on unobservable monotonicity violations is very different, Assumption C.4 holding would at least suggest that Assumption 3 is not grossly violated.[100]

[99] As shown in the proof of Proposition C.2, $d_{vu} = \Pr\left[S\left(1\right) > S\left(2\right)\right] - d_{vo}$.

[100] For instance, if Assumption C.4 holds but d_{vu}^p/d_{vu} deviates 10% from d_{vo}^p/d_{vo}, the rates at

To test Assumption C.4, I exploit that first-voting judges in the São Paulo Appeal Court are close to unaffected by panel effects (see Section C.1). The votes of first-voting judges are thus a good proxy for how these judges would decide cases had they decided them individually. Since the first-voting judge is randomly determined, I can thus estimate the stringency difference between judge pairs in all case types both in a setting where they decide cases "individually" and when they decide together in the same judicial panel. For instance, I can estimate the "individual-decisions" stringency difference between Judge 1 and Judge 2 in violence cases by comparing their tendencies to favor the prosecution as first-voting judges in violence cases. To measure the influence of panel effects, I can compare this stringency difference to the stringency difference between the same two judges in violence cases they *decide together as a panel*.

As observable case characteristics, I use broad crime categories—drug-related crimes, economic crimes, property crimes, violent crimes, and other crimes—and whether the prosecutor appealed.[101] I exclude judge pairs where one judge decides less than 1,000 cases.[102] To avoid overestimating the share of cases violating monotonicity, I use a split-sample approach:[103] First, I estimate the judges' stringency order in one-half of the data (*training sample*). Then, I estimate the share of cases violating monotonicity, given this stringency order, in the remaining half (*hold-out sample*). For instance, suppose that, in the training sample, Judge 2 is overall stricter than Judge 1 but more lenient in drug-related cases. The estimated share of drug-related cases violating monotonicity is then Judge 1's stringency minus Judge 2's stringency in drug-related cases in the hold-out sample.[104] While the naive approach tends to overestimate monotonicity violations, the split-sample approach is

which panel effects reduce disagreements violating and satisfying monotonicity can be shown to also differ by around 10%. I discuss the consequences of such moderate deviations from Assumption 3 in the next subsection.

[101] I obtain similar results using only broad crime categories in Table C.8.

[102] Judges deciding few cases induce noise, which tends to bias downwards the disagreement and monotonicity violation estimates.

[103] Copus and Hübert (Forthcoming) use a similar approach to bound disagreements across judges.

[104] If Judge 2 is strictest also in drug-related cases in the training sample, the share of drug-related cases violating monotonicity is estimated to be zero. I exclude observations where the two judges are equally strict in the training sample.

Table C.7: Influence of Panel Effects on Observable Monotonicity Violations

	Individual decisions	Panel decisions
	(1)	(2)
Observable disagreements	0.12	0.0068
satisfying monotonicity (d_{so})	(0.0015)	(0.0004)
Observable disagreements	0.0048	0.00023
violating monotonicity (d_{vo})	(0.0006)	(0.00012)
Observable monotonicity violation rate	0.039	0.033
	(0.004)	(0.014)

Note: Lower bound estimates of disagreements and monotonicity violations based on observable case characteristics. Using broad crime types—violent crimes, property crimes, drug-related crimes, economic crimes, and other crimes—and whether the prosecutor appealed as case characteristics. The stringency order of the judges in each case type is estimated in half the sample (training sample). The lower bound disagreement rates based on this stringency order are estimated in the remaining half (hold-out sample). If Judge A is stricter than Judge B in a given case category in the training sample, the *observable disagreement* in such cases is Judge A's stringency minus Judge B's stringency in the hold-out sample. If Judge A is overall stricter (more lenient) than Judge B in the training sample, these observable disagreements count as *observable disagreements satisfying (violating) monotonicity*. The *observable monotonicity violation rate* equals the ratio between observable monotonicity violations and observable disagreements. Keeping only judges with at least 1,000 cases in the training sample. Column 1 is based on first-voting judges in São Paulo criminal appeals. Column 2 is based on votes by judge pairs appearing in the same panel in São Paulo criminal appeals. Bootstrapped standard errors clustered at the case level in parentheses.

conservative.[105]

I estimate *observable disagreements satisfying monotonicity*—the lower bound of the share of cases where a given judge pair disagrees in a way that satisfies monotonicity—in a similar way. Finally, I estimate the lower bound of the monotonicity violation *rate* as the estimated observable monotonicity violations divided by the sum of the estimated observable disagreement rates.[106] To assess how panel effects influence observable monotonicity violations, I compare these estimates to similar estimates of observable disagreements among judges appearing in the same panel.[107]

The results are presented in Table C.7. In individual decisions—as proxied by the decisions of first-voting judges in Column 1—observable disagreements satisfying (violating) monotonicity are estimated to sum to 12% (0.48%) of all cases, giving an observable monotonicity violation rate of 3.9%.

In Column 2, I consider the same judges when they appear together in a panel. Due to panel effects, measured disagreements are substantially reduced: Observable disagreements satisfying monotonicity are reduced to 0.7%, and observable monotonicity violations to 0.02%. But panel effects are estimated to reduce observable disagreements satisfying and violating monotonicity at very similar rates: $d^p_{so}/d_{so} = 0.0068/0.12 = 0.056$ and $d^p_{vo}/d_{vo} = 0.00023/0.0048 = 0.048$. Assumption C.4 thus appears to be nearly satisfied. As a consequence, the observable monotonicity violation rate ends up being very similar in panels as in individual decisions. Unless panel effects influence unobservable monotonicity violations very differently, these results suggest that Assumption 3 is also close to being satisfied.

[105] For instance, if Judge 2 and Judge 1 are equally strict in drug-related cases, the naive approach leads, in expectation, to a positive estimated share of drug-related cases violating monotonicity for this judge pair: Due to statistical noise, Judge 1 will in 50% of samples appear stricter than Judge 2. But the split-sample estimator is unbiased when the judges are equally strict in drug-related cases and downward biased otherwise.

[106] See Appendix C.6 for why this ratio identifies a lower bound of the monotonicity violation rate. The direction of the estimator's finite sample bias is, however, unclear.

[107] For instance, for Judge 2 and Judge 1, I first collect all appellate decisions where these two judges are both part of the panel. I then delete the case identifier and apply the same method as when calculating observable monotonicity violations in the single-judge decisions.

Table C.8: Robustness of Table C.7 to Using Only Broad Crime Categories

	Individual decisions (1)	Panel decisions (2)
Observable disagreements satisfying monotonicity (d_{so})	0.11	0.0064
	(0.0015)	(0.0004)
Observable disagreements violating monotonicity (d_{vo})	0.0018	0.00015
	(0.0005)	(0.00010)
Observable monotonicity violation rate	0.016	0.023
	(0.004)	(0.013)

Note: Robustness of Table C.7 to using only broad crime types—and not whether the prosecutor appealed—as case characteristics.

C.6 The "Observable" Monotonicity Violation Rate

In this section, I explain why the observable monotonicity violation rate in Section 5.3 is a lower bound of the true monotonicity violation rate. Let $\tilde{m}$ and $\tilde{d}$ be the observable share of cases violating monotonicity and the observable disagreement rate, respectively. Denote the true rates by m and d. First, note that any unobserved monotonicity violation must be counteracted by unobserved disagreements satisfying monotonicity. For instance, in the example of Section 5.3, if Judge A and Judge B convict in, respectively, 50% and 60% of drug-related cases, the observable share of drug-related cases violating monotonicity is 10%. If there are 5% more drug-related cases violating monotonicity (where Judge B is stricter than Judge A), there must also be 5% disagreements in drug-related cases satisfying monotonicity (where Judge A is stricter than Judge B). Otherwise, their stringency difference in drug-related cases will no longer be 10 percentage points. Thus if $m = \tilde{m} + a$ for $a \geq 0$, we must have $d = \tilde{d} + 2a$. It is then easy to see that the ratio $\tilde{m}/\tilde{d}$ is a lower bound of m/d:

$$\frac{\tilde{m}}{\tilde{d}} \leq \frac{m}{d} = \frac{\tilde{m}+a}{\tilde{d}+2a} \Leftrightarrow 0 \leq a\left(\tilde{d}-2\tilde{m}\right)$$

where $\tilde{d} \geq 2\tilde{m}$ is trivially satisfied.

D External Validity

To what extent do my monotonicity violation estimates reflect those in typical judge IV settings? This section examines three key threats to external validity:

1. Judges in my three main settings decide only questions of law.
2. Cases in appellate courts may not be representative of those in trial courts, where the judge IV is typically applied.
3. Appellate judges tend to be older and more politicized than trial judges.

I then explore how monotonicity varies across different subsets of cases and judges, identifying factors that may help predict monotonicity violations in new settings.

D.1 External Validity: Questions of Fact

In the settings discussed in the main part of the paper, the judges decide only questions of law. To test whether monotonicity violations are different when judges also decide questions of *fact*, I here consider two settings where judges also discuss questions of fact: São Paulo criminal trials decided by panels and Norwegian criminal appeals.

São Paulo Criminal Trials Decided By Panels. First, I exploit that in Brazil, criminal offenses committed by certain public officials are decided by a panel of judges already at the trial stage. In São Paulo, criminal cases involving mayors, state legislators, judges, prosecutors, and some other offices are tried by a panel of 25 appeal judges (Cavalcante Filho and Lima 2017).[108] I collect all such cases in the period 2011–2023. Only five of these cases were non-unanimous—two against state legislators and three against prosecutors. The large number of judges per case, however, allows for a reasonably precise estimate of monotonicity violations between pairs of judges. I only keep disagreements between judges that disagree at

[108] Due to abstentions, the number of judges voting in a given case varies between 21 and 24 judges in my sample.

Table D.1: Additional settings

Court	Types of cases	Time coverage	Judges per panel	Cases	Judges
São Paulo Appeal Court	Criminal trials	2011–2023	21–24	5	37
Norwegian Appeal Courts	Criminal appeals	1993–2019	Two–Three	96	64

Note: Only cases with dissents.

least twice—a monotonicity violation can not be detected for judge pairs disagreeing only once. I end up with 169 disagreements between 65 judge pairs in my sample. The benefit of this setting is that the cases are actual criminal trials. The drawback is that the cases are few and not necessarily representative of criminal cases in judge IV studies.

Norwegian Criminal Appeals. My final setting is Norwegian criminal appeals. Full criminal appeals in Norway are essentially retrials of the case. The parties can produce new evidence, there is a full oral hearing, and the sentence can be altered due to issues of both facts and law. A Norwegian criminal appeal thus closely resembles a criminal trial, a setting where the judge IV design is routinely applied. Full appeals are decided by a panel of two to three professional judges and a varying number of lay judges.[109]

A drawback of the Norwegian setting is that the judges do not decide cases in fixed panels: Judges from across the court are randomly assigned to each new appeal. Thus, unlike the other settings, I rarely observe the same panel deciding multiple cases. I do, however, observe *pairs* of judges deciding multiple cases together, allowing me to measure the rate of monotonicity violations between *pairs*

[109]Norwegian criminal appeals were decided by a panel of three professional judges and four lay judges up until 2018. After 2018, the appeals were decided by two professional judges and five lay judges. Before 2018, appeals related to criminal cases with statutory sentence length above six years were decided by three professional judges and a jury of ten lay judges. I disregard lay judges in my analysis, as monotonicity violations involving lay judges are almost impossible to detect—lay judges decide cases infrequently and are thus unlikely to be observed disagreeing multiple times with the same judge. For a more thorough discussion of the Norwegian criminal appeal process, see Bhuller and Sigstad (Forthcoming).

Table D.2: Share of Judge-Pair Disagreements Violating Monotonicity.

Setting	Monotonicity Violation Rate	(s.e.)	N	Bootstrap bias estimate
São Paulo Criminal Appeals				
With Five Judges	0.16	(0.006)	5,888	-0.0044
Three First Judges	0.20	(0.004)	10,889	-0.0081
Brazilian Superior Court	0.20	(0.023)	388	-0.0087
US Supreme Court	0.10	(0.005)	24,823	-0.0009
São Paulo Criminal Trials	0.15	(0.031)	357	-0.0214
Norwegian Appeal Courts	0.11	(0.006)	110	-0.0130

Note: Unit of observation at the case-by-judge-pair level. Keeping only observations where the two judges disagree. The *monotonicity violation rate* is the sum of all disagreements violating monotonicity divided by the total number of disagreements. A judge pair violates monotonicity in a case if the judge who is most often the stricter judge in cases the pair decides together is most lenient in this particular case. Bootstrapped standard errors clustered at the case level. The two last rows estimate the bootstrap bias with an adapted bootstrap showing the bias under the assumption that the true monotonicity violation rate for each judge pair equals the overall estimate. The standard bootstrap implicitly assumes that judge pairs not observed violating monotonicity in the data *never* violate monotonicity, which is unreasonable when we observe judge pairs disagreeing in only a few cases.

of judges.

I rely on the raw text of the decision in all Norwegian criminal appeals receiving a full hearing between 1993 and 2019, available from the Lovdata Foundation ($N = 37,473$).[110] From the raw text, I extract the prison length voted for by each judge using regular expressions. There are 420 cases where at least one pair of judges disagree about sentencing. As for the São Paulo criminal trials, I only keep disagreements between judges that disagree at least twice. I end up with 52 judge pairs and 96 cases in my sample. I manually verify all observations in my final sample to avoid falsely coding a case as violating monotonicity.

[110] About half of Norwegian criminal appeals are rejected in an initial screening. I exclude such cases.

Results. In Table D.2, I report the estimated share of disagreements that violate monotonicity across all settings. The estimated violation rates are 11 percent for Norwegian criminal appeals and 15 percent for São Paulo criminal trials. Due to the limited number of observations per judge, these estimates are moderately downward biased. The bootstrap bias estimates in the last column suggest that this bias is approximately 1–2 percentage points.

For comparison, the estimated monotonicity violation rates for judge pairs are 16 percent in São Paulo five-judge appeals, 20 percent in the Brazilian Superior Court, and 10 percent in the US Supreme Court. For completeness, I also report that the pairwise violation rate for the first three votes in all São Paulo criminal appeals is 20 percent.

These results suggest that monotonicity is violated to a similar degree in the two settings that most closely resemble criminal trials as in the other settings.

D.2 External Validity: Selection of Cases in Appeal Courts

Not all cases are appealed. The monotonicity violation rate among appealed cases may thus not be representative of the overall monotonicity violation rate. To assess whether my monotonicity violation estimates are influenced by the selection of cases into appeal courts, I use data on all available decisions from São Paulo criminal *trials* in the period 2018–2021 (Tribunal de Justiça de São Paulo 2022).[111] Focusing on cases decided on the merits, I obtain a sample of 282,405 trial cases. Using the cases' unique identifiers, I can determine whether a given case was appealed.

A large share of cases (39%) are appealed, placing a limit on selection effects. But the appealed cases are not a random sample of trial cases. In Table D.3, Columns 1–2, I report the results from regressing whether a trial case is appealed on observable case characteristics. Compared to the average criminal trial, appealed cases tend to be more serious (longer potential punishments) and more complex (more characters in the trial judge's opinion). Columns 1 and 2 of Table

[111] I remove cases decided before 2018, since they substantially less likely to appear in the data on appellate decisions, suggesting incomplete data. I exclude cases decided after 2021, since many of these cases are still waiting for their appeals to be decided.

D.4 compare summary statistics among trial cases and appealed cases, showing a similar pattern.

Table D.3: The Selection of Cases into Appeal Courts: Regressions

	Dep. Var.: Case Appealed	
	(1)	(2)
Trial Court Conviction	0.241	0.264
	(0.003)	(0.003)
Log(Characters in Trial Court Decision)	0.150	0.133
	(0.002)	(0.002)
Minimum Statutory Years of Prison		0.029
		(0.003)
Maximum Statutory Years of Prison		0.001
		(0.001)
Drug-Related Crime	0.108	−0.036
	(0.003)	(0.007)
Violent Crime	0.018	−0.058
	(0.004)	(0.010)
Economic Crime	0.062	0.048
	(0.004)	(0.005)
Property Crime	0.050	−0.007
	(0.002)	(0.003)
Mean Dependent Variable	0.39	0.39
Observations	282,405	228,110
R^2	0.101	0.110

Note: São Paulo criminal trials 2018–2021. Intercept omitted. Standard errors in parentheses.

Suppose more serious or complex cases have different monotonicity violation rates than other cases. In that case, my monotonicity violation estimates might not well approximate monotonicity violations in typical criminal cases studied in judge IV designs.

Table D.4: The Selection of Cases Into Appeal Courts: Summary Statistics

	Trial cases (1)	Appeals (2)
Panel A: Case Characteristics		
log(Characters in trial court decision)	9.3	9.5
Max. statutory years prison	6.5	7.4
Min. statutory years prison	2.2	2.5
Convicted in trial court	0.83	0.95
Panel B: Type of Crime		
Robbery	0.184	0.231
Aggravated Theft	0.150	0.148
Theft	0.123	0.102
Drug Trafficking	0.092	0.124
Traffic Offenses	0.066	0.053
Receiving Stolen Property	0.059	0.052
Firearms	0.041	0.039
Threat	0.033	0.024
Fraud	0.032	0.033
Domestic Violence	0.026	0.014
Aggravated Robbery	0.017	0.022
Use of Fake Documents	0.014	0.013
Bodily Injury	0.011	0.006
Embezzlement	0.010	0.010
Tax Crimes	0.007	0.008
Observations	282,405	110,961

Note: São Paulo criminal trials 2018–2021. Column 1 includes all cases. Column 2 includes only appealed cases. Panel B contains the top 15 most common crimes.

Table D.5: Monotonicity Violations By Case Characteristics

	Monotonicity Violation Rate				
	Case Characteristic Quintile				
Case Characteristic	1	2	3	4	5
Minimum Statutory Prison Length	0.11	0.08	0.14	0.16	0.13
	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)
Characters in the Trial Judge's Opinion	0.13	0.17	0.10	0.16	0.16
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
Predicted Probability of Appeal	0.13	0.14	0.13	0.13	0.18
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)

Note: The share of disagreements violating monotonicity among judge pairs in the São Paulo Appeal Court by case characteristics. The *minimum statutory prison length* is the minimum years of prison the defendant must be sentenced to if convicted of the alleged crime. *Characters in the trial judge's opinion* is the number of characters in the trial judge's decision in the case. The *predicted probability of appeal* is obtained from regressing an indicator for the trial decision being appealed on narrow crime categories, an indicator for a trial court conviction, and log(number of characters in the trial judge's opinion). São Paulo criminal appeals decided by three-judge panels with trial court decisions between 2018–2021. Standard errors in parentheses.

In Table D.5, I thus estimate monotonicity violations according to the seriousness and complexity of the case. Reassuringly, I find similar monotonicity violation rates across quintiles of case seriousness and complexity. Differences across quintiles are mostly statistically insignificant and show no clear pattern. For instance, the monotonicity violation rate is 0.11 for the least serious quintile and 0.13 for the most serious quintile.

In the last row, I report monotonicity violations by quintiles of predicted probability of appeal.[112] These predictions are obtained from regressing an indicator for appeal on narrow crime categories, a trial court conviction indicator, and log(number of characters in the trial court decision) in the sample of all trial cases. I find similar monotonicity violation rates in cases with a high ex-ante probability of being appealed as in cases with a low such probability.[113] These analyses suggest that the selection of trial cases into appeal courts does not substantially influence my monotonicity violation estimates.

D.3 External Validity: Judges

Judges in appeal courts are often politically appointed senior judges. Are the monotonicity violations among these judges representative of the monotonicity violations among the more junior and less politicized judges in lower courts where the judge IV design is typically applied?

To address this question, I exploit that in the São Paulo Appeal Court one-fifth of judges are politically appointed, with the remaining being appointed from among trial court judges by the court administration. I also use data on the birth date of all the judges and the year they entered the appeal court from Tribunal de Justiça de São Paulo (2024).

[112]The idea behind this exercise is that cases with a low ex-ante probability of appeal are likely more "marginal" appeals that are more similar to unappealed cases than cases with a high ex-ante probability of appeal. Seeing only small differences in monotonicity violation rates across these two types of cases thus indicates that monotonicity violation rates in appealed and unappealed cases are similar.

[113]While the quintile with the highest predicted appeal probability has a somewhat higher monotonicity violation rate, the monotonicity violation rate is essentially identical across the other quintiles. The predicted appeal probability is 27% (62%) for the lowest (highest) quintile.

Table D.6: Pairwise Monotonicity Violations By Judge Characteristics

	Without Controlling for Stringency Differences					Controlling for Stringency Differences				
Panel A: Political Appointment										
Judge 1 politically	Judge 2 politically appointed									
appointed	Yes		No			Yes		No		
Yes	0.23					0.22				
No	0.18		0.20			0.27		0.28		
Panel B: Judge Age										
Age quintile	Age quintile of judge 2									
of judge 1	1	2	3	4	5	1	2	3	4	5
1 (47–60 years)	0.11					0.21				
2 (60–64 years)	0.20	0.20				0.27	0.29			
3 (64–68 years)	0.17	0.21	0.21			0.24	0.22	0.18		
4 (68-72 years)	0.16	0.19	0.18	0.14		0.23	0.25	0.17	0.11	
5 (72–76 years)	0.17	0.20	0.20	0.22	0.27	0.30	0.24	0.30	0.26	0.32
Panel C: Judge Experience										
Exp. quintile	Experience quintile of judge 2									
of judge 1	1	2	3	4	5	1	2	3	4	5
1 (0–5 years)	0.11					0.32				
2 (5–8 years)	0.15	0.16				0.31	0.19			
3 (8–12 years)	0.24	0.19	0.20			0.26	0.20	0.24		
4 (12–15 years)	0.19	0.17	0.27	0.22		0.35	0.23	0.28	0.28	
5 (15–23 years)	0.21	0.15	0.31	0.28	0.25	0.24	0.24	0.23	0.28	0.32

Note: Share of disagreements violating monotonicity for judge pairs in the São Paulo Appeal Court. The left part of the table shows the mean monotonicity violation rate across all cases decided by judge pairs of the indicated category. The right part of the table shows weighted means where weights are given by the inverse propensity of the judge pair's stringency difference being of the given decile among judge pairs of the indicated category. São Paulo criminal appeals decided by three-judge panels.

In the left panel of Table D.6, I present the variation in raw monotonicity violation rates across judge pairs based on the judges' age, experience, and type of appointment. Monotonicity is estimated to be higher than average among politically appointed judges and lower than average for the youngest and the most recently appointed appeal judges. These differences are, however, driven by the fact that the stringency dispersion is lower (higher) than average for politically appointed judges (young and inexperienced judges).[114] In the right panel of Table D.6, I account for these stringency dispersion differences using inverse probability weighting.[115] After accounting for these differences, there are small differences in monotonicity violations across judge groups. The judges with the least experience are mostly recently promoted trial judges.[116] The fact that their monotonicity violation rate is comparable to the overall rate suggests that the monotonicity violations might be similar among trial court judges and appellate judges.

D.4 Monotonicity Across Case Types

The extent to which the monotonicity violation rates in this paper generalize may depend on case types. For instance, monotonicity might be violated to a different degree in a judge IV study on the effects of drug trafficking convictions than in a study on traffic offense convictions. Table D.7 shows how monotonicity violations for judge pairs vary across case types. Reassuringly, the monotonicity violation rates are quite similar: among the ten most common crimes in the São Paulo Appeal Court, eight have violation rates between 13% and 18%.

But some case types have notably lower violation rates: economic crimes (8%)

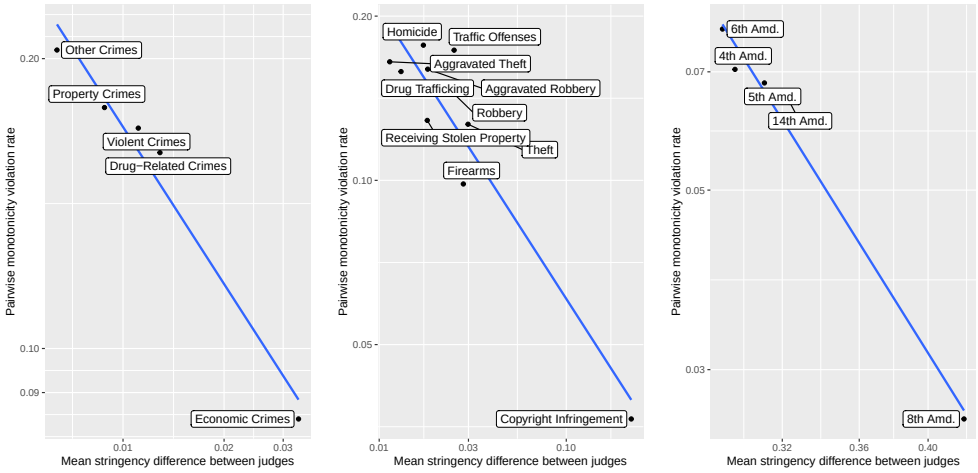
[114] As shown in Table D.8, monotonicity violations tend to decrease as stringency differences increase. That the stringency dispersion is higher than average among recently appointed appeal judges could, for instance, be explained by a learning story. It might take time for recently appointed judges to adapt to the norms of a new court (Abrams et al. 2022).

[115] In particular, I group all judge pairs into deciles according to the difference in stringencies between the two judges in the pair. Then, for instance, a politically appointed judge pair belonging to the fifth decile is weighted by the inverse of the share of politically appointed judge pairs that belong to the fifth decile. This way, I calculate the monotonicity violation rate in a pseudo-population of politically appointed judge pairs whose stringency difference distribution resembles the overall distribution.

[116] The politically appointed judges are appointed from among prosecutors and lawyers. The monotonicity violation rate between former trial judges in the bottom experience quintile is 0.25.

Figure D.1: Stringency Differences and Monotonicity Violations

(a) São Paulo Appeal Court (b) São Paulo Appeal Court (c) US Supreme Court



Note: The relationship between stringency differences and monotonicity violations across case types. The monotonicity violation rate in, for instance, drug-related cases is calculated as follows: If Judge A and Judge B disagrees in 20 drug-related cases and Judge B is stricter than Judge A in 16 of these, monotonicity is considered to be violated in the four cases where Judge A is strictest. The *pairwise monotonicity violation rate* is the sum of all monotonicity violations across all judge pairs divided by the sum of all disagreements among cases of the indicated type. The stringency difference between Judge A and Judge B in drug-related cases is the absolute difference in the rates of pro-prosecution votes in drug-related cases they decide together. The *mean stringency difference between judges* is the average of this difference across cases. Figures (a) and (b) use all São Paulo criminal appeals decided by three-judge panels. Figure (c) uses all US Supreme Court cases about civil procedure, including unanimous cases. Log-transformed scales.

and copyright infringement cases (4%) in São Paulo, and Eighth Amendment cases (3%) in the US Supreme Court. What might explain these lower rates? It turns out that these case types also have the largest stringency differences between judges. As shown in Figure D.1, much of the variation in monotonicity violation rates across case types can be explained by stringency differences. This relationship is expected: For instance, if Judge 1 has a much higher stringency than Judge 2, the share of disagreements where Judge 2 is stricter than Judge 1 must necessarily be low.[117]

[117]For instance, if Judge 1 (Judge 2) votes with the prosecution in 80% (10%) of copyright infringement cases, the share of disagreements violating monotonicity can at most be $0.1/(0.8+0.1) = 0.11$. As shown in Section D.5, there is a strong negative association between stringency differences and monotonicity violations across judge pairs.

Table D.7: Monotonicity Violations By Crime Type

Crime Type	Monotonicity Violation Rate	Std. Error	Observations
Panel A: Broad Categories			
Drug-Related Crimes	0.16	(0.004)	7,600
Property Crimes	0.18	(0.004)	7,408
Other Crimes	0.20	(0.009)	2,214
Economic Crimes	0.08	(0.006)	2,012
Violent Crimes	0.17	(0.009)	1,594
Panel B: Narrow Categories			
Drug Trafficking	0.16	(0.004)	7,582
Aggravated Robbery	0.16	(0.006)	3,690
Aggravated Theft	0.16	(0.010)	1,346
Copyright Infringement	0.04	(0.005)	1,314
Theft	0.13	(0.011)	868
Firearms	0.10	(0.013)	528
Traffic Offenses	0.17	(0.017)	508
Robbery	0.15	(0.016)	502
Receiving Stolen Property	0.13	(0.021)	264
Homicide	0.18	(0.024)	260
Panel C: US Supreme Court			
Fourth Amendment	0.07	(0.003)	6,388
Fifth Amendment	0.06	(0.003)	6,060
Sixth Amendment	0.08	(0.004)	5,560
Eighth Amendment	0.03	(0.003)	3,232
Fourteenth Amendment	0.06	(0.004)	3,578

Note: Monotonicity violations for judge pairs. The monotonicity violation rate in, for instance, drug-related cases is calculated as follows: If Judge A and Judge B disagrees in 20 drug-related cases and Judge B is stricter than Judge A in 16 of these, monotonicity is considered to be violated in the four cases where Judge A is strictest. The monotonicity violation *rate* is the sum of all monotonicity violations across all judge pairs divided by the sum of all disagreements. *Observations* is the number of disagreements for the given case type. São Paulo criminal appeals decided by three-judge panels in Panels A–B and non-unanimous US Supreme Court cases about criminal procedure in Panel C.

This result suggests that researchers can use stringency differences between judges to predict monotonicity violations in new settings. I discuss how to make such predictions in Section D.5.[118]

Table D.7 also suggests that more homogeneous cases tend to have slightly lower monotonicity violation rates. In Panel B, which focuses on narrow crime categories, violation rates are somewhat lower than in Panel A, which covers broader categories. These differences are, however, minor.

D.5 Factors Associated with Monotonicity Violations

In this section, I use my data to assess which observable characteristics of a group of judges are associated with higher rates of monotonicity violations. I first create hypothetical groups of judges based on all possible subsets of judges from the judicial panels in my data. For instance, for the São Paulo Appeal Court panel consisting of the judges Leitão, Sale Junior, Campos, Marques, and Diodatti, I create the subsets {Campos, Marques, Diodatti}, {Leitão, Diodatti}, and all other possible subpanels consisting of two, three, or four of the judges. I also include the full five-judge panel. Then, I calculate measures of monotonicity violations for each judge subset in cases they decide together. Finally, I regress measures of monotonicity violations on characteristics of the judge subsets. As characteristics, I include the number of judges and the variance and standard deviation of the judges' stringencies. As monotonicity violation measures, I consider the share of cases violating IA monotonicity and the sums of negative 2SLS weights.

Results for the São Paulo Appeal Court and the US Supreme Court are presented in Table D.8.[119] IA monotonicity tends to be violated at a higher rate as the number of judges increases. With more judges, there are more judge pairs for which monotonicity could be violated. But the sum of negative 2SLS weights decreases. Intuitively, as the number of judges increases, monotonicity violations between some

[118] Note that the stringency differences in Figure D.1 are muted due to panel effects. In Section D.5, I seek to transform these differences into "individual-decision" differences that can be used to predict monotonicity violations in settings where cases are randomly assigned to individual judges.

[119] The Brazilian Superior Court is excluded due to the limited number of judge subsets, which results in highly imprecise estimates.

judge pairs are more likely to be offset by disagreements *satisfying* monotonicity between other judge pairs, ensuring that the overall 2SLS weight remains positive.

Groups of judges with greater stringency standard deviations tend to have lower non-negative weights. This relationship is concave, as indicated by the positive estimated coefficients on stringency variance (standard deviation squared). These results suggest that researchers worried about monotonicity violations in judge IV design can—perhaps as a robustness exercise—exclude randomization units (*e.g.*, courts) with few judges or a low stringency dispersion. Note, however, that since judge IV estimands already place low weight on such randomization units, excluding these units would likely be of limited importance in practice.[120]

In Table D.9, I show how the monotonicity violation measures vary with the number of judges in the subpanel and their stringency differences. To obtain measures of stringency differences comparable to those in judge IV settings where cases are randomly assigned to individual judges, I engage in the following exercise in the São Paulo Appeal Court: I first calculate each judge's "single-judge" stringency by the rate of pro-prosecution votes in cases where this judge votes first.[121] I then regress the panel-based stringency variance on the "single-judge" stringency variance across subpanels. I use the stringency variance predicted from this regression as the subpanel's stringency variance.[122] For the US Supreme Court, where I do not have access to such "single-judge" stringencies, I calculate the stringency variance of a subpanel using all cases about criminal procedure, including unanimous cases.[123] The predicted standard deviation of the stringencies is the square root of the predicted variance. For each bin, I report the monotonicity violation measures for the US Supreme Court and São Paulo Appeal Court subpanels side by side, separated by "/".

Consistent with Table D.8, the share of cases violating IA monotonicity (the sum of negative 2SLS weights) is increasing (decreasing) in the number of judges

[120]For instance, 2SLS with court fixed effects is a weighted sum of court-specific 2SLS estimates where the weights are proportional to the courts' stringency variance and the number of cases in the court. The number of cases is again typically highly correlated with the number of judges.

[121]As shown in Section C.1, the first vote is close to unaffected by panel effects.

[122]Directly using the single-judge stringency variance introduces considerable noise.

[123]In a judge IV setting, unanimous cases are included in the stringency measures. Due to panel effects, these stringency differences are likely underestimates of the true differences.

and decreasing in the stringency standard deviation. Note that there is considerable agreement between the two settings, which lends some support to using Table D.9 as a rough guide to external validity. For instance, in a setting with four judges per court and stringency standard deviation in the range 0.15–0.20, one might expect IA monotonicity to be violated in roughly 20% of cases and the sum of negative 2SLS weights to be around 0.05–0.10.[124]

The potential value of Table D.9 becomes evident in its ability to explain, to a large extent, the differences in monotonicity violations reported in Table 2 between the US Supreme Court and the São Paulo Appeal Court. Among subpanels of five US Supreme Court justices with a stringency standard deviation between 0.15 and 0.20—approximately the median for the São Paulo Appeal Court—the IA monotonicity violation rate is 0.29, and the sum of negative 2SLS weights is 0.08. These figures closely resemble the corresponding values of 0.35 and 0.093 reported for the São Paulo Appeal Court in Table 2.

The fact that the US Supreme Court has a lower sum of negative 2SLS weights than the São Paulo Appeal Court can thus largely be attributed to a combination of having more judges and greater polarization (i.e., larger differences in stringency).[125] The estimates for the US Supreme Court in Table D.9 would have predicted the monotonicity violation rate in the São Paulo Appeal Court quite accurately. Notably, the only required parameters—the number of judges and their stringency differences—are easily obtainable in any new setting. The reliability of such predictions in other contexts can, of course, not be guaranteed. I leave the evaluation of Table D.9's usefulness as a guide to external validity to future research.

[124] If the number of judges and stringency levels vary across courts, a weighted average can be computed. For the sum of negative 2SLS weights, the appropriate weights are the courts' stringency variances, as these correspond to the weights assigned by 2SLS with court fixed effects.

[125] The median stringency standard deviation is 0.15 in the São Paulo Appeal Court and 0.22 in the US Supreme Court.

Table D.8: Factors Associated with Monotonicity Violations

	São Paulo Appeal Court		US Supreme Court	
	Share of cases violating IA monotonicity (1)	Sum negative 2SLS weights (2)	Share of cases violating IA monotonicity (3)	Sum negative 2SLS weights (4)
Number of judges	0.084	-0.055	0.030	-0.009
	(0.012)	(0.014)	(0.008)	(0.004)
Stringency standard deviation	-1.240	-3.305	-0.381	-1.872
	(0.393)	(1.171)	(0.463)	(0.656)
Stringency variance distance squared	0.894	3.438	-0.523	2.206
	(0.482)	(1.493)	(0.477)	(0.960)
Mean Dep. Var	0.21	0.19	0.19	0.079
Observations	511	511	3 612	3 612
R^2	0.46	0.25	0.31	0.28

Note: OLS estimates showing the association between characteristics of a group of judges and the degree of monotonicity violations among this judge group. The unit of observation is a subset of judges from a judicial panel. All potential subsets of judges, including the full panel. *Share of cases violating IA monotonicity* is the share of cases with disagreement that violate IA monotonicity. *Sum of negative 2SLS weights* is the sum of the negative weights the estimand would assign treatment effects if monotonicity is violated as in the subpanel. *Stringency standard deviation (variance)* is the standard deviation (variance) in stringencies across judges in the group. Intercept omitted. *Mean Dep. Var* is the mean of the dependent variable. Excluding groups of judges where the stringency difference between the strictest and the most lenient judges is below 0.1. São Paulo criminal appeals decided by five-judge panels and non-unanimous US Supreme Court cases about criminal procedure. Standard errors clustered at the panel level in parentheses.

Table D.9: Number of Judges, Stringency Differences, and Monotonicity

Number	Standard deviation of stringencies				
of judges	0.05–0.10	0.10–0.15	0.15–0.20	0.20–0.25	0.25–0.30
Panel A: Share of cases violating IA monotonicity					
2	0.09 / 0.21	0.09 / 0.14	0.09 / 0.10	0.08 / 0.06	0.047 / 0.028
3	0.23 / 0.35	0.21 / 0.30	0.17 / 0.16	0.12 / 0.07	0.083 / 0.006
4	0.32 / 0.46	0.27 / 0.41	0.23 / 0.18	0.17 / 0.06	0.111 /
5	0.38 /	0.35 / 0.50	0.29 / 0.20	0.20 /	0.135 /
6	0.39 /	0.42 /	0.34 /	0.23 /	0.168 /
7	0.44 /	0.43 /	0.39 /	0.26 /	0.208 /
8		0.20 /	0.44 /	0.29 /	0.252 /
9			0.50 /	0.23 /	0.282 /
Panel B: Sum of negative 2SLS weights					
2	0.20 / 0.83	0.16 / 0.30	0.15 / 0.14	0.11 / 0.07	0.061 / 0.031
3	0.22 / 0.58	0.12 / 0.20	0.12 / 0.08	0.06 / 0.04	0.046 / 0.003
4	0.19 / 0.36	0.09 / 0.15	0.08 / 0.06	0.05 / 0.01	0.026 /
5	0.16 /	0.07 / 0.14	0.08 / 0.05	0.04 /	0.016 /
6	0.14 /	0.06 /	0.07 /	0.03 /	0.009 /
7	0.12 /	0.04 /	0.07 /	0.02 /	0.008 /
8		0.01 /	0.07 /	0.02 /	0.011 /
9			0.07 /	0.03 /	0.047 /

Note: Mean monotonicity violation measures across bins of subpanels. In the notation "X / Y", "X" ("Y") is the measure in the US Supreme Court (São Paulo Appeal Court). *Stringency standard deviation* is the standard deviation of the judges' stringencies. In the US Supreme Court, I use stringencies based on all cases about criminal procedure, including unanimous cases. In the São Paulo Appeal Court, I transform the standard deviation in stringencies as measured in panel decisions as follows: First, I regress the panel-based stringency variance on the variance of the judges' stringencies as first-voting judges. Then, I calculate the square root of the variance predicted from this regression. *Share of cases violating IA monotonicity* is the share of cases with disagreement that violate IA monotonicity. *Sum of negative 2SLS weights* is the sum of the negative weights 2SLS would assign treatment effects if monotonicity is violated as in the subpanel. Excluding subpanels where the stringency difference between the strictest and the most lenient judge is below 0.1.

References

Abrams, David et al. (2022). "When in Rome... on local norms and sentencing decisions". In: *Journal of the European Economic Association* 20.2, pp. 700–738. DOI: 10.1093/jeea/jvab038.

Arnold, David, Will Dobbie, and Crystal S. Yang (2018). "Racial Bias in Bail Decisions". In: *The Quarterly Journal of Economics* 133.4, pp. 1885–1932. DOI: 10.1093/qje/qjy012.

Bhuller, Manudeep, Gordon B Dahl, et al. (2020). "Incarceration, recidivism, and employment". In: *Journal of Political Economy* 128.4, pp. 1269–1324. DOI: 10.1086/705330.

Bhuller, Manudeep and Henrik Sigstad (Forthcoming). "Feedback and Learning: The Causal Effects of Reversals on Judicial Decision-Making". In: *Review of Economic Studies*. DOI: 10.1093/restud/rdae073.

Boyd, Christina L, Lee Epstein, and Andrew D Martin (2010). "Untangling the causal effects of sex on judging". In: *American Journal of Political Science* 54.2, pp. 389–411. DOI: 10.1111/j.1540-5907.2010.00437.x.

Cavalcante Filho, João Trindade and Frederico Retes Lima (2017). "Foro, Prerrogativa e Privilégio: Quais e Quantas Autoridades Têm Foro no Brasil?" In: *Direito Público* 13.76.

Copus, Ryan and Ryan Hübert (Forthcoming). "Measuring How Much Judges Matter for Case Outcomes". In: *Journal of Law and Courts*.

Cox, Adam and Thomas J Miles (2008). "Judging the voting rights act". In: *Colum. L. Rev.* 108, p. 1.

Dobbie, Will, Jacob Goldin, and Crystal S Yang (2018). "The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges". In: *American Economic Review* 108.2, pp. 201–40. DOI: 10.1257/aer.20161503.

Fang, Zheng and Andres Santos (2019). "Inference on directionally differentiable functions". In: *The Review of Economic Studies* 86.1, pp. 377–412. DOI: 10.1093/restud/rdy049.

Fischman, Joshua B (2011). "Estimating preferences of circuit judges: A model of consensus voting". In: *The Journal of Law and Economics* 54.4, pp. 781–809. DOI: 10.1086/661512.

— (2014). "Measuring inconsistency, indeterminacy, and error in adjudication". In: *American Law and Economics Review* 16.1, pp. 40–85. DOI: 10.1093/aler/aht011.

Frandsen, Brigham, Lars Lefgren, and Emily Leslie (2023). "Judging Judge Fixed Effects". In: *American Economic Review* 113.1, pp. 253–77. DOI: 10.1257/aer.20201860.

Frandsen, Brigham R, Lars J Lefgren, and Emily C Leslie (2019). *Judging Judge Fixed Effects*. Tech. rep. 25528. National Bureau of Economic Research. DOI: 10.3386/w25528.

Heckman, James J and Edward J Vytlacil (2007). "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments". In: *Handbook of Econometrics*. Vol. 6. Elsevier, pp. 4875–5143. DOI: 10.1016/S1573-4412(07)06071-0.

Imbens, Guido W. and Joshua D. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects". In: *Econometrica* 62.2, pp. 467–475. DOI: 10.2307/2951620.

Kastellec, Jonathan P (2013). "Racial diversity and judicial influence on appellate courts". In: *American Journal of Political Science* 57.1, pp. 167–183. DOI: 10.1111/j.1540-5907.2012.00618.x.

Norris, Samuel (2018). "Judicial errors: Evidence from refugee appeals". In: *University of Chicago, Becker Friedman Institute for Economics Working Paper* 75.

Peresie, Jennifer L (2004). "Female judges matter: Gender and collegial decision-making in the federal appellate courts". In: *Yale Law Journal* 114, p. 1759.

Revesz, Richard L (1997). "Environmental regulation, ideology, and the DC circuit". In: *Va. L. Rev.* 83, p. 1717. DOI: 10.2307/1073657.

Sigstad, Henrik (2024). *Marginal Treatment Effects and Monotonicity*.

Sunstein, Cass R et al. (2007). *Are judges political?: an empirical analysis of the federal judiciary*. Brookings Institution Press.

Tribunal de Justiça de São Paulo (2022). *Consulta de Julgados de 1º Grau*. https://esaj.tjsp.jus.br/cjpg/ [Accessed: 2022].

— (2024). *O Tribunal de Justiça de São Paulo e seus Desembargadores*. https://www.tjsp.jus.br/Download/Biblioteca/Curriculum/Curriculum.pdf [Accessed: April 7, 2024].

Wolak, Frank A (1987). "An exact test for multiple inequality and equality constraints in the linear regression model". In: *Journal of the American Statistical Association* 82.399, pp. 782–793. DOI: 10.1080/01621459.1987.10478499.