Target-Population Fragility in Judge IV under Average Monotonicity

Mohamed Coulibaly* Sidi Mohamed Sawadogo†

May 4, 2026

Abstract

Instrumental variable designs based on random judge assignment increasingly rely on average monotonicity when strong monotonicity is hard to defend. We show that average monotonicity does not necessarily preserve a stable target population: the response types receiving positive weight switch discretely when a judge's propensity crosses the assignment-weighted average. With heterogeneous treatment effects, the IV estimand need not be locally invariant. We propose a diagnostic based on estimated judge propensities and sampling uncertainty. In Philadelphia bail data, boundary fragility is visible, and the magnitude of the estimated treatment effect depends nontrivially on the target population consistent with that fragility.

JEL Classification: C26, C21, C18

*Corresponding author. Department of Applied Economics, HEC Montréal. E-mail: mohamed.coulibaly@hec.ca.
†Department of Economics, Université de Montréal and CIREQ. E-mail: sidi.mohamed.sawadogo@umontreal.ca.

1 Introduction

Instrumental variable designs based on the random assignment of judges have become a standard source of quasi-experimental variation in applied economics. Researchers use differences in judges' treatment propensities to estimate the causal effects of incarceration, pretrial detention, disability insurance, foster care, and other consequential decisions.[1] But their interpretation becomes less clear once strong monotonicity, the requirement that a judge who is stricter on average be weakly stricter in every case, is hard to defend. A growing literature has therefore turned to average monotonicity, a weaker condition that preserves the IV estimand as a proper weighted average of treatment effects even when judges do not rank cases in the same way (Frandsen et al., 2023; Chyn et al., 2025). While that shift is useful, it leaves open a separate question that matters for interpretation: whether the estimand still refers to a stable target population.

This paper asks whether the IV estimand under average monotonicity identifies the effect for a stable target population. Under strong monotonicity, the target population is defined by behavior: it consists of individuals whose treatment status changes when they are assigned to a more treatment-prone judge rather than a less treatment-prone judge. Under average monotonicity, the set of individuals receiving positive weight depends not only on how people respond to judges, but also on where each judge's treatment propensity lies relative to the assignment-weighted average propensity. We therefore study what happens at and near that boundary. The issue therefore is whether the population to which the positive weights apply remains the same under small perturbations of the design.

We answer this question in the smallest setting rich enough to show the mechanism: a three-judge design. Each individual belongs to one of eight response types, defined by which judges would assign treatment. Under average monotonicity, whether a response type receives positive weight depends on where each judge's treatment propensity lies relative to the assignment-weighted average propensity. Two response types sit at the boundary. One is treated by the strictest and most lenient judges but not by the intermediate judge. The other is treated only by the intermediate judge. These two types are mirror images, and their weights have opposite signs. When the intermediate judge's propensity lies below the average, average monotonicity admits the first type and excludes the second. When that propensity lies above the average, the roles reverse. At the boundary, both types receive zero weight. A small perturbation of the design around that point therefore produces a discrete change in the set of response types receiving positive weight.

That discrete switch can change the IV estimand once treatment effects are heterogeneous. If the two response types that trade places at the boundary have different average treatment effects,

[1] Prominent applications include Kling (2006); Maestas et al. (2013); Doyle Jr et al. (2015); Dobbie et al. (2017, 2018), and more recent work such as Humphries et al. (2025); Garin et al. (2025).

then the IV estimand under average monotonicity is not locally invariant to perturbations in the design. Average monotonicity continues to rule out negative weights, but the composition of the weighted average changes at the boundary: one response type exits, another enters, and with them the causal parameter can shift. Under strong monotonicity, sampling uncertainty affects how precisely the estimand for a fixed target population is estimated. Under average monotonicity, small perturbations near the boundary can also change which target population the estimand refers to. Appendix B shows that the same boundary logic extends to any finite number of judges. With more judges, any judge whose propensity lies at or near the assignment-weighted average can trigger the same kind of switch.

We then turn the identification result into a boundary diagnostic for applied work. Researchers do not observe population judge propensities and must estimate them. The relevant empirical question is therefore whether the data place each judge securely on one side of the average-monotonicity boundary. We propose reporting three objects: each judge's estimated distance to the boundary, measured as the difference between the judge's estimated treatment propensity and the assignment-weighted average; that distance expressed in standard-error units; and the assignment share handled by judges close to the boundary. These objects do not test average monotonicity. They measure whether the target population implied by average monotonicity is stably classified in the sample.

We illustrate these ideas using the Philadelphia bail data studied by Stevenson (2018). In that setting, bail magistrates rotate across shifts, and the scheduled magistrate provides quasi-random variation in pretrial detention risk. Earlier work finds no decisive evidence against the standard judge design in the aggregate sample (Coulibaly et al., 2024), which makes the Philadelphia data a useful setting in which to assess boundary fragility on its own terms. The diagnostic classifies three judges as close to the average-monotonicity boundary at the conventional cutoff; together, they handle about 29 percent of cases. This does not show that the design fails or that strong monotonicity is violated. It shows that, once average monotonicity is invoked, target-population stability becomes an empirical question rather than an assumption. A boundary-consistent perturbation exercise then separates fragility from quantitative influence: the estimated effect of pretrial detention on conviction remains positive across the reported perturbations, but its magnitude moves by as much as 38 percent of the baseline estimate.

This paper contributes first to the literature on the assumptions underlying judge-IV designs. Frandsen et al. (2023) show that average monotonicity preserves a proper weighted-average interpretation of the IV estimand, and recent work has brought that weaker monotonicity condition to the center of judge and examiner designs (Chyn et al., 2025; Sigstad, 2026). We show that this interpretation does not by itself guarantee a stable target population. In that sense, our result complements the observation in Mogstad and Torgovitsky (2024) that average monotonicity

is not a purely behavioral restriction. Where that work emphasizes that the assumption depends on the distribution of the instrument and propensity profile, we show that this dependence has a concrete implication for interpretation: the positive-weight population can switch at a boundary that researchers estimate but do not control.

This paper also contributes to the practice of assessing and reporting judge-IV designs. Recent methodological guidance has focused attention on what researchers should check when using leniency designs (Chyn et al., 2025; Goldsmith-Pinkham et al., 2025). Our contribution is to show that target-population stability belongs on that list when average monotonicity is invoked. We propose a boundary diagnostic based on estimated judge propensities, their sampling uncertainty, and assignment shares. The Philadelphia bail illustration shows why this reporting practice matters: a design can remain credible, and its main estimate can remain stable in sign, even when the average-monotonicity target population is not sharply classified for a nontrivial share of cases.

The remainder of the paper proceeds as follows. Section 2 introduces the judge-IV setup and defines the IV weights under average monotonicity. Section 3 develops the boundary discontinuity result in the three-judge case. Section 4 proposes the boundary diagnostic and applies it to the Philadelphia bail data. Section 5 concludes.

2 IV Weights under Average Monotonicity

This section defines the IV weights that determine the target population in a judge leniency design. Let $Z_i \in \{1, \ldots, J\}$ denote the judge assigned to individual i, and let judge z be assigned with probability $\lambda_z > 0, \forall z = 1, ..., J$. Let $D_{iz} \in \{0, 1\}$ denote the treatment status individual i would receive if assigned to judge z. Let Y_{1i} and Y_{0i} denote potential outcomes under treatment and control, and write $\delta_i = Y_{1i} - Y_{0i}$. The observed treatment is

$$D_i = \sum_{z=1}^{J} D_{iz} \mathbf{1}\{Z_i = z\},$$

and the observed outcome is

$$Y_i = Y_{1i} D_i + Y_{0i}(1 - D_i).$$

The maintained design condition is random assignment of judges and exclusion through treatment.

Assumption 2.1 (Random Assignment and Exclusion) *Judge assignment is independent of potential outcomes and potential treatment choices:*

$$(Y_{1i}, Y_{0i}, \{D_{iz}\}_{z=1}^{J}) \perp Z_i.$$

In addition, judge assignment affects outcomes only through treatment:

$$Y_{di}(z) = Y_{di}, \qquad d \in \{0, 1\}.$$

Assumption 2.1 rules out two threats to the judge-IV design. Random assignment requires judges to receive comparable cases. Exclusion requires judge identity to affect outcomes only through the treatment decision.[2]

Under Assumption 2.1, the IV estimand can be written as a weighted average of individual treatment effects. Let

$$p_z = E[D_i \mid Z_i = z]$$

denote judge z's treatment propensity, and let

$$\bar{p} = \sum_{z=1}^{J} \lambda_z p_z$$

denote the assignment-weighted average treatment propensity. The IV estimand is

$$\beta^{IV} = \frac{E[\omega_i \delta_i]}{E[\omega_i]}, \tag{2.1}$$

where

$$\omega_i = \sum_{z=1}^{J} \lambda_z (p_z - \bar{p})(D_{iz} - \bar{D}_i), \qquad \bar{D}_i = \sum_{z=1}^{J} \lambda_z D_{iz}. \tag{2.2}$$

The weight ω_i is the covariance, across judge assignments, between judge propensities and individual i's potential treatment profile. It is positive when judges who treat more often in the population are also more likely to treat individual i. It is negative when individual i is more likely to be treated by judges with lower overall treatment propensities.

The sign of ω_i determines whether the IV estimand has a proper weighted-average interpretation. When treatment effects are heterogeneous, equation (2.1) alone does not guarantee a causal parameter with nonnegative weights. If some individuals have $\omega_i < 0$, the estimand places negative weight on their treatment effects and can no longer be read as an average treatment effect for a positively weighted subpopulation. Monotonicity restrictions are used to rule out this problem.

Strong monotonicity fixes the target population by imposing a common ranking of judges across individuals.

[2] Frandsen et al. (2023) also consider weaker average exclusion restrictions. We maintain the standard exclusion restriction throughout.

Assumption 2.2 (Strong Monotonicity) *For all individuals i and all pairs of judges z, z',*

$$D_{iz} \geq D_{iz'} \qquad \textit{whenever} \qquad p_z \geq p_{z'}.$$

Assumption 2.2 requires each individual's treatment profile to be weakly increasing in judge propensity. Once judges are ordered by how often they treat, a more treatment-prone judge must treat every individual who would be treated by a less treatment-prone judge. This restriction allows the IV estimand to be interpreted as the effect for a fixed complier-type population. It also requires judges to differ only along a single leniency dimension, a restriction that may fail when judges rank cases differently across case characteristics or decision margins.[3]

Average monotonicity relaxes the common-ranking requirement while preserving nonnegative IV weights.

Assumption 2.3 (Average Monotonicity) *For each individual i,*

$$\omega_i \geq 0 \qquad \textit{almost surely.}$$

Assumption 2.3 allows individual treatment profiles that violate strong monotonicity. It does not require treatment status to be weakly increasing in judge propensity for every individual. It only requires that, for each individual, the covariance between judge propensities and potential treatment decisions is nonnegative. Under Assumption 2.3, the IV estimand in (2.1) remains a proper weighted average of individual treatment effects, as in Frandsen et al. (2023) and related work on judge and examiner designs.

Average monotonicity preserves nonnegative weights but does not by itself define a target population from behavior alone. To see this, describe individual i by a response type $\tau = (\tau_1, \ldots, \tau_J) \in \{0,1\}^J$, where τ_z is the treatment status the individual would receive from judge z. Because $\sum_{z=1}^{J} \lambda_z(p_z - \bar{p}) = 0$, equation (2.2) implies the type-specific weight

$$\omega(\tau) = \sum_{z=1}^{J} \lambda_z(p_z - \bar{p})\tau_z. \tag{2.3}$$

Average monotonicity requires $\omega(\tau) \geq 0$ for every response type that appears in the population.

The target-population problem starts with the dependence of $\omega(\tau)$ on the design. The response type τ records how an individual would respond to each judge. But the sign of $\omega(\tau)$ also depends on the propensity profile $(p_1, \ldots, p_J)$ and on the assignment probabilities $(\lambda_1, \ldots, \lambda_J)$. Under strong monotonicity, behavioral restrictions determine which response types can receive positive

[3]The monotonicity condition follows the local average treatment effect framework of Imbens and Angrist (1994). For concerns about one-dimensional judge rankings, see Chan et al. (2022).

weight. Under average monotonicity, whether a response type receives positive weight depends on where judges lie relative to $\bar{p}$ and on how often they are assigned. Average monotonicity therefore solves the negative-weight problem without guaranteeing that the IV estimand refers to a target population fixed by individual response behavior alone (Mogstad and Torgovitsky, 2024).

The next section studies the three-judge setting, where this distinction changes which response types enter the positive-weight target population. When an intermediate judge's propensity crosses the assignment-weighted average propensity, one non-monotone response type enters the estimand while its mirror type exits. If the response types that trade places have different treatment effects, the IV estimand need not remain locally invariant at that boundary. Appendix B shows that the same logic extends to any finite number of judges.

3 Boundary Discontinuity

The three-judge case is the smallest setting in which average monotonicity changes the target population without changing the sign of the IV weights. Consider three judges, indexed by $z \in \{1, 2, 3\}$, ordered so that

$$p_1 \leq p_2 \leq p_3.$$

Section 2 showed that average monotonicity requires nonnegative type-specific weights, but that these weights depend on the propensity profile and assignment probabilities. The three-judge case makes this dependence explicit. It also shows why the target population can change discontinuously when one judge crosses the assignment-weighted average propensity.

The IV weight can be written relative to the intermediate judge. Since

$$\sum_{z=1}^{3} \lambda_z (p_z - \bar{p}) = 0,$$

equation (2.2) implies

$$\omega_i = \lambda_1 (p_1 - \bar{p})(D_{i1} - D_{i2}) + \lambda_3 (p_3 - \bar{p})(D_{i3} - D_{i2}). \tag{3.1}$$

Because $p_1 \leq \bar{p} \leq p_3$, the first term has a nonpositive coefficient and the second term has a nonnegative coefficient. The sign of ω_i therefore depends on how individual i's treatment status under judges 1 and 3 compares with treatment status under the intermediate judge.

Each individual belongs to one of eight response types. Let

$$g = (D_{i1}, D_{i2}, D_{i3}) \in \{0, 1\}^3$$

denote the treatment profile individual i would receive across the three judges. Table 1 reports each response type, its IV weight, and whether it is admissible under strong monotonicity and average monotonicity.

Table 1: Response Types, IV Weights, and Admissibility

Type	(D_{i1}, D_{i2}, D_{i3})	$\omega(g)$	Strong monotonicity	Average monotonicity
g_1	$(1,1,1)$	0	Yes	Yes
g_2	$(0,0,0)$	0	Yes	Yes
g_3	$(0,0,1)$	$\lambda_3(p_3 - \bar{p}) > 0$	Yes	Yes
g_5	$(0,1,1)$	$-\lambda_1(p_1 - \bar{p}) > 0$	Yes	Yes
g_6	$(1,1,0)$	$-\lambda_3(p_3 - \bar{p}) < 0$	No	No
g_7	$(1,0,0)$	$\lambda_1(p_1 - \bar{p}) < 0$	No	No
g_4	$(1,0,1)$	$-\lambda_2(p_2 - \bar{p})$	No	Depends
g_8	$(0,1,0)$	$\lambda_2(p_2 - \bar{p})$	No	Depends

Notes: The table reports response types in the three-judge case, ordered by judge propensities $p_1 \leq p_2 \leq p_3$. The signs for g_3, g_5, g_6, and g_7 are strict when $p_1 < \bar{p} < p_3$. At boundary cases, the corresponding weights may be zero. Under average monotonicity, a response type is admissible only if its weight is nonnegative.

The first six response types have the same status across all propensity profiles. Always-takers, $g_1 = (1,1,1)$, and never-takers, $g_2 = (0,0,0)$, receive zero weight. The monotone complier types, $g_3 = (0,0,1)$ and $g_5 = (0,1,1)$, receive positive weight and are admissible under both strong and average monotonicity. The reverse types, $g_6 = (1,1,0)$ and $g_7 = (1,0,0)$, receive negative weight and are ruled out under average monotonicity.

The two non-monotone response types are where average monotonicity differs from strong monotonicity. Type $g_4 = (1,0,1)$ is treated by the lowest- and highest-propensity judges but not by the intermediate judge. Type $g_8 = (0,1,0)$ is treated only by the intermediate judge. Strong monotonicity rules out both types because neither treatment profile is weakly increasing in judge propensity. Average monotonicity can admit one of them because their weights have opposite signs:

$$\omega(g_4) = -\lambda_2(p_2 - \bar{p}), \qquad \omega(g_8) = \lambda_2(p_2 - \bar{p}).$$

Thus, whether g_4 or g_8 can enter the positive-weight target population depends on the sign of $p_2 - \bar{p}$.

The sign switch creates a boundary in the target population. If $p_2 < \bar{p}$, then $\omega(g_4) > 0$ and $\omega(g_8) < 0$, so average monotonicity permits g_4 and rules out g_8. If $p_2 > \bar{p}$, then $\omega(g_8) > 0$ and $\omega(g_4) < 0$, so average monotonicity permits g_8 and rules out g_4. If $p_2 = \bar{p}$, both types receive zero weight. The positive-weight population therefore changes when the intermediate judge crosses the assignment-weighted average propensity.

Proposition 1 (Boundary discontinuity) *In the three-judge design with $p_1 \leq p_2 \leq p_3$, the set of response types consistent with average monotonicity changes discontinuously at $p_2 = \bar{p}$.*

Proof. See Appendix A.

Proposition 1 is stronger than the statement that average monotonicity is design-dependent. The dependence does not only reweight a fixed set of response types. Around the boundary $p_2 = \bar{p}$, one response type exits the admissible set and its mirror type enters. The composition of the positive-weight target population changes discretely even when the propensity profile changes by an arbitrarily small amount.

The discontinuity matters for interpretation when treatment effects differ across the switching types. Let

$$\tau_g = E[\delta_i \mid i \text{ has response type } g]$$

denote the average treatment effect for response type g. Average monotonicity implies that β^{IV} is a weighted average over response types with nonnegative weights. But when p_2 crosses $\bar{p}$, the estimand switches between a weighted average that can include g_4 and one that can include g_8. If $\tau_{g_4} \neq \tau_{g_8}$, the causal parameter identified by IV changes with the propensity profile.

Corollary 1 (Target-population instability) *Suppose that $\tau_{g_4} \neq \tau_{g_8}$, where $\tau_g = E[\delta_i \mid i \text{ has response type } g]$. Then the IV estimand under average monotonicity need not be locally invariant to perturbations in the propensity profile at points where $p_2 = \bar{p}$.*

Proof. See Appendix A.

Corollary 1 gives the interpretation problem. Under strong monotonicity, the complier-type target population is pinned down by individual response behavior. Under average monotonicity, the target population also depends on where judges sit relative to $\bar{p}$. Near a boundary, the same behavioral environment can support different positive-weight populations depending on small changes in the propensity profile. The IV estimand then remains a proper weighted average, but it is not necessarily an average for a fixed complier-type group.

The three-judge argument extends to any finite number of judges. For any judge z, the response type treated only by judge z has weight $\lambda_z(p_z - \bar{p})$, while its componentwise complement has weight $-\lambda_z(p_z - \bar{p})$. These two types switch signs when p_z crosses $\bar{p}$. Appendix B gives the finite-J statement and proof. The main text uses the three-judge case because it displays the boundary mechanism without the notation required for the general case.

4 Empirical Relevance: A Boundary Diagnostic in Philadelphia Bail

The boundary result becomes empirically relevant when estimated judge propensities place judges near the average-monotonicity boundary. Sections 2 and 3 treat the propensity profile as a popula-

tion object. In applications, researchers estimate that profile. If a judge lies close to the assignment-weighted average propensity, the sample may weakly classify that judge as above or below the boundary. Under average monotonicity, that classification determines which mirror response types can receive positive IV weight.

The diagnostic measures each judge's distance from the boundary in standard-error units. Let $\hat{p}_z$ denote the estimated treatment propensity of judge z, and let $\hat{\lambda}_z$ denote the empirical assignment share. Define

$$\hat{\bar{p}} = \sum_{j=1}^{J} \hat{\lambda}_j \hat{p}_j, \qquad \hat{d}_z = \hat{p}_z - \hat{\bar{p}}.$$

The standardized boundary distance is

$$\hat{r}_z = \frac{|\hat{d}_z|}{\widehat{\mathrm{se}}(\hat{d}_z)}. \tag{4.1}$$

For a cutoff c, define the set of boundary-fragile judges as

$$\widehat{\mathcal{F}}(c) = \{z : \hat{r}_z \leq c\}. \tag{4.2}$$

It is worth mentioning that statistic is not a test of average monotonicity. It measures whether the sample places each judge securely on one side of the boundary that determines average-monotonicity weights.

The Philadelphia bail data combine quasi-random magistrate rotation with enough cases to estimate judge-specific boundary distances. We use the aggregate sample from Stevenson (2018). The outcome is conviction, the endogenous treatment is pretrial detention, and the excluded instruments are the eight judge indicators. The maximum-control specification includes defendant characteristics, prior-record variables, offense indicators, and court-calendar controls. The sample contains 331,971 cases and eight judges. Appendix D describes how judge propensities and standard errors are constructed.

The diagnostic classifies three Philadelphia judges as close to the average-monotonicity boundary. Figure 1 shows that Judges 1, 5, and 7 are boundary-fragile at $c = 1.96$. Judge 7 is closest to the boundary, with $\hat{d}_7 = 0.0009$ and $\hat{r}_7 = 0.44$. Judges 1 and 5 lie below the boundary, with $\hat{r}_1 = 1.45$ and $\hat{r}_5 = 1.69$. Together, these three judges handle about 29 percent of cases. The minimum standardized distance is therefore not driven by a judge who receives little case mass.

Boundary fragility is a target-population diagnostic, not a rejection of the judge-IV design. The fact that a judge lies close to the average-monotonicity boundary does not imply that judge assignment is nonrandom. It does not imply that exclusion fails. It also does not show that strong monotonicity is violated. It says something narrower: conditional on invoking average monotonic-

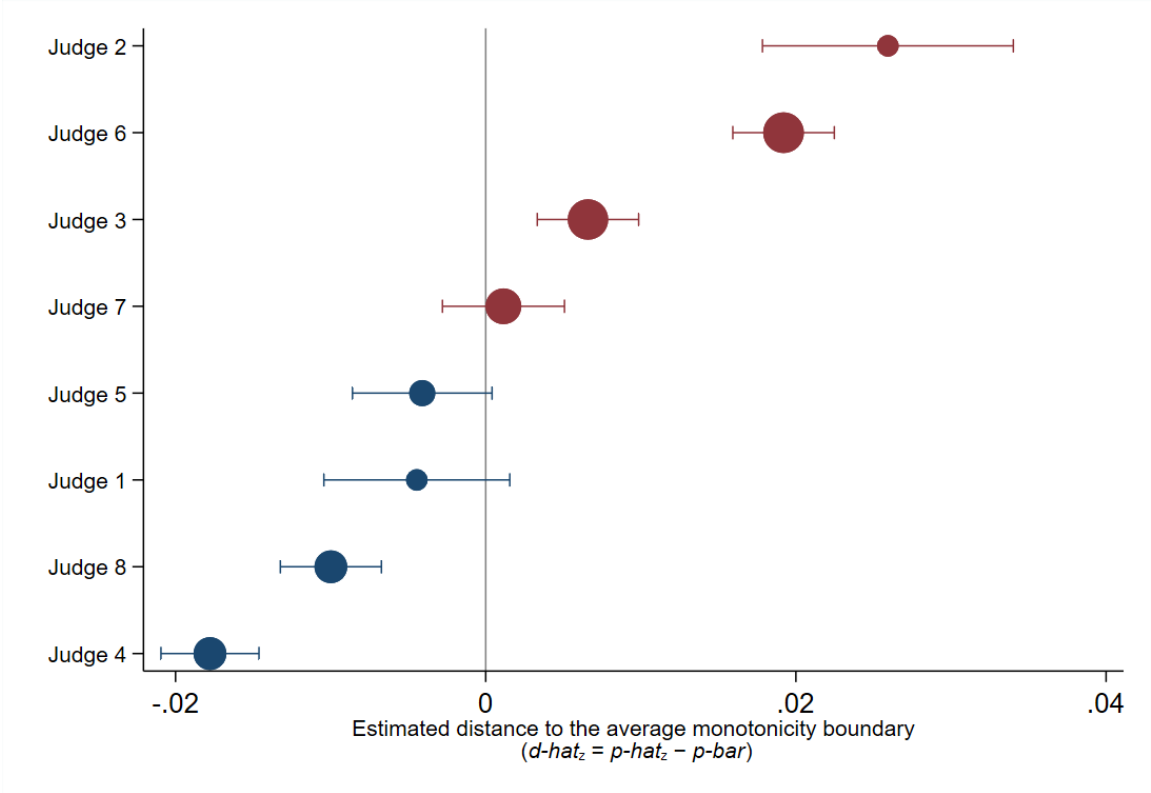

Figure 1: Judge Distances to the Average-Monotonicity Boundary in Philadelphia Bail

Notes: The figure plots $\hat{d}_z = \hat{p}_z - \hat{\bar{p}}$ for each judge, where $\hat{p}_z$ is the regression-adjusted treatment propensity and $\hat{\bar{p}}$ is the assignment-weighted average propensity. Horizontal bars report 95 percent confidence intervals. The vertical line at zero marks the average-monotonicity boundary. Judges to the left of zero lie below the boundary; judges to the right lie above it. Point sizes are proportional to empirical assignment shares $\hat{\lambda}_z$.

ity, the positive-weight population is not sharply classified for judges near the boundary. When those judges handle nontrivial case mass, the target population attached to the IV estimand can depend on small changes in the estimated propensity profile.

The boundary-consistent exercise quantifies how much the IV estimate moves when only fragile classifications are allowed to change. We hold the estimated judge propensities fixed and perturb the judge-assignment shares so that non-fragile judges remain on their baseline side of the boundary. The perturbation is restricted to an L_1 neighborhood of the observed assignment shares. This exercise is not a confidence interval and not a new estimator. It is a plug-in sensitivity calculation for the target population implied by average monotonicity. The Online Appendix gives the full construction.

Table 2: Boundary Fragility and Boundary-Consistent Perturbations

Object	Value
Boundary-fragile judges at $c = 1.96$	1, 5, 7
Assignment share of fragile judges	0.291
Baseline 2SLS estimate	0.1859
Boundary-consistent set, $\kappa = 0.10$	$[0.1762,\ 0.1992]$
Boundary-consistent set, $\kappa = 0.25$	$[0.1549,\ 0.2059]$
Boundary-consistent set, $\kappa = 0.50$	$[0.1158,\ 0.2281]$
Maximum movement from baseline	0.0702

Notes: The table reports plug-in boundary-consistent perturbations of the maximum-control 2SLS estimate. The fragile set uses cutoff $c = 1.96$. The radius κ restricts the L_1 distance between perturbed and observed assignment shares. Since assignment shares sum to one, $\kappa/2$ is the maximum share of cases that can be reassigned across judges. The boundary-consistent set is not a confidence interval. It reports how the 2SLS estimand moves across assignment-share profiles that preserve non-fragile boundary classifications.

The Philadelphia exercise separates boundary proximity from quantitative influence on the IV estimate. Table 2 shows that the estimated detention effect remains positive across the reported perturbations, but its magnitude moves nontrivially. At $\kappa = 0.10$, the boundary-consistent set is close to the baseline estimate. At $\kappa = 0.50$, the set ranges from 0.1158 to 0.2281, and the maximum movement from the baseline is 0.0702, about 38 percent of the baseline estimate. The conclusion that pretrial detention increases conviction is stable in sign in this exercise, while the magnitude depends on the target population consistent with boundary fragility.

The diagnostic implies a reporting practice for applications that invoke average monotonicity. Researchers should report each judge's estimated distance from the assignment-weighted average propensity, the standardized boundary distance, and the assignment share handled by boundary-fragile judges. These objects do not replace balance tests, exclusion arguments, or monotonicity

tests. They answer a different question: whether the positive-weight population under average monotonicity is stably classified in the estimated design.

5 Conclusion

Average monotonicity gives judge-IV estimands nonnegative weights, but it does not necessarily give them a fixed target population. We show that the positive-weight response types under average monotonicity can switch discretely when a judge's treatment propensity crosses the assignment-weighted average propensity. With heterogeneous treatment effects, that switch means the IV estimand need not be locally invariant near the boundary. The estimand can still be interpreted as a proper weighted average of treatment effects. The qualification is that the weighted average is taken over a design-dependent target population, not necessarily over a complier group pinned down by behavior alone.

The empirical implication is a reporting requirement for judge-IV applications that invoke average monotonicity. Researchers should report where estimated judge propensities lie relative to the assignment-weighted average, how precisely those boundary distances are estimated, and how much case mass is handled by boundary-fragile judges. The Philadelphia bail illustration shows why this matters: boundary fragility is visible for a nontrivial share of cases, and boundary-consistent perturbations move the magnitude of the estimated treatemd effect. This does not reject the design, and it does not test average monotonicity. It shows that, once average monotonicity is used to interpret judge-IV estimates, target-population stability is an empirical question.

A Proofs of the Three-Judge Results

This appendix proves the two formal results stated in Section 3. Throughout, judges are indexed by $z \in \{1, 2, 3\}$, ordered so that $p_1 \leq p_2 \leq p_3$, with assignment probabilities $\lambda_z > 0$. The assignment-weighted average propensity is

$$\bar{p} = \sum_{z=1}^{3} \lambda_z p_z.$$

The three-judge case uses the type-specific weight

$$\omega(g) = \sum_{z=1}^{3} \lambda_z (p_z - \bar{p}) g_z,$$

where $g = (g_1, g_2, g_3) \in \{0, 1\}^3$ denotes a response type.

The weight formula can be written relative to the intermediate judge. Since

$$\sum_{z=1}^{3} \lambda_z (p_z - \bar{p}) = 0,$$

we have

$$\omega(g) = \lambda_1 (p_1 - \bar{p})(g_1 - g_2) + \lambda_3 (p_3 - \bar{p})(g_3 - g_2).$$

Because $p_1 \leq \bar{p} \leq p_3$, the term $\lambda_1 (p_1 - \bar{p})$ is nonpositive and the term $\lambda_3 (p_3 - \bar{p})$ is nonnegative.

The eight response-type weights follow from this expression. For the always-taker and never-taker types,

$$\omega(1,1,1) = 0, \qquad \omega(0,0,0) = 0.$$

For the two monotone complier types,

$$\omega(0,0,1) = \lambda_3 (p_3 - \bar{p}) \geq 0, \qquad \omega(0,1,1) = -\lambda_1 (p_1 - \bar{p}) \geq 0.$$

For the two reverse types,

$$\omega(1,1,0) = -\lambda_3 (p_3 - \bar{p}) \leq 0, \qquad \omega(1,0,0) = \lambda_1 (p_1 - \bar{p}) \leq 0.$$

For the two non-monotone mirror types,

$$\omega(1,0,1) = \lambda_1 (p_1 - \bar{p}) + \lambda_3 (p_3 - \bar{p}) = -\lambda_2 (p_2 - \bar{p}),$$

and

$$\omega(0,1,0) = -\lambda_1 (p_1 - \bar{p}) - \lambda_3 (p_3 - \bar{p}) = \lambda_2 (p_2 - \bar{p}).$$

These calculations verify the entries in Table 1.

Proof of Proposition 1. The discontinuity follows from the two mirror types. Let

$$g_4 = (1,0,1), \qquad g_8 = (0,1,0).$$

Their weights are

$$\omega(g_4) = -\lambda_2 (p_2 - \bar{p}), \qquad \omega(g_8) = \lambda_2 (p_2 - \bar{p}).$$

Since $\lambda_2 > 0$, the signs of both weights are determined by the sign of $p_2 - \bar{p}$.

If $p_2 < \bar{p}$, then

$$\omega(g_4) > 0, \qquad \omega(g_8) < 0.$$

Average monotonicity can admit g_4 but rules out g_8. If $p_2 > \bar{p}$, then

$$\omega(g_4) < 0, \qquad \omega(g_8) > 0.$$

Average monotonicity can admit g_8 but rules out g_4. If $p_2 = \bar{p}$, then

$$\omega(g_4) = \omega(g_8) = 0.$$

Both types receive zero weight.

All other response types have signs that do not depend on $p_2 - \bar{p}$. Therefore the only change in the positive-weight set as p_2 crosses $\bar{p}$ is the switch between g_4 and g_8. One type exits and its mirror type enters. The positive-weight set therefore changes discontinuously at $p_2 = \bar{p}$. ■

Proof of Corollary 1. The IV estimand under average monotonicity can be written as

$$\beta^{IV} = \frac{\sum_{g:\omega(g)>0} \pi_g \omega(g) \tau_g}{\sum_{g:\omega(g)>0} \pi_g \omega(g)},$$

where $\pi_g = \Pr(G_i = g)$ and $\tau_g = E[\delta_i \mid G_i = g]$. Zero-weight types do not affect the estimand.

Consider two propensity profiles arbitrarily close to a boundary point where $p_2 = \bar{p}$. On the side where $p_2 < \bar{p}$, the positive-weight set can include g_4 and excludes g_8. On the side where $p_2 > \bar{p}$, the positive-weight set can include g_8 and excludes g_4. Thus the two nearby estimands average over different response-type components:

$$\beta_- = \frac{\sum_{g\in\{g_3,g_4,g_5\}} \pi_g \omega(g) \tau_g}{\sum_{g\in\{g_3,g_4,g_5\}} \pi_g \omega(g)}$$

on one side of the boundary, and

$$\beta_+ = \frac{\sum_{g\in\{g_3,g_5,g_8\}} \pi_g \omega(g) \tau_g}{\sum_{g\in\{g_3,g_5,g_8\}} \pi_g \omega(g)}$$

on the other side.

If the switching types have positive population shares and $\tau_{g_4} \neq \tau_{g_8}$, then the two nearby estimands need not coincide. One estimand can load on the average treatment effect for g_4, while the other can load on the average treatment effect for g_8. Hence there is no neighborhood of the boundary in which the IV estimand is guaranteed to remain locally invariant. This proves the corollary. ■

B General Finite-J Extension

This appendix shows that the boundary mechanism is not specific to three judges. Let judges be indexed by $z \in \{1, \ldots, J\}$, with $J \geq 3$. Each judge is assigned with probability $\lambda_z > 0$, and each individual has a response type

$$\tau = (\tau_1, \ldots, \tau_J) \in \{0,1\}^J.$$

Let

$$\bar{p} = \sum_{z=1}^{J} \lambda_z p_z$$

denote the assignment-weighted average treatment propensity.

The type-specific weight is an inner product between the response type and the propensity-deviation vector. Define

$$v = \left(\lambda_1(p_1 - \bar{p}), \ldots, \lambda_J(p_J - \bar{p})\right).$$

Then

$$\omega(\tau) = \sum_{z=1}^{J} \lambda_z(p_z - \bar{p})\tau_z = \langle v, \tau \rangle.$$

Because

$$\sum_{z=1}^{J} \lambda_z(p_z - \bar{p}) = 0,$$

we also have

$$\langle v, \mathbf{1} \rangle = 0, \qquad \mathbf{1} = (1, \ldots, 1).$$

Every response type has a componentwise complement with the opposite weight. For any $\tau \in \{0,1\}^J$,

$$\omega(\mathbf{1} - \tau) = \langle v, \mathbf{1} - \tau \rangle = \langle v, \mathbf{1} \rangle - \langle v, \tau \rangle = -\omega(\tau).$$

This complementarity is the finite-J version of the g_4-g_8 relationship in the three-judge case.

A single judge crossing the boundary is enough to change the positive-weight set. Fix judge j, and let e_j denote the response type treated only by judge j:

$$e_j = (0, \ldots, 0, 1, 0, \ldots, 0).$$

Its componentwise complement, $\mathbf{1} - e_j$, is the type treated by every judge except judge j. Their weights are

$$\omega(e_j) = \lambda_j(p_j - \bar{p}), \qquad \omega(\mathbf{1} - e_j) = -\lambda_j(p_j - \bar{p}).$$

These two types switch signs exactly when p_j crosses $\bar{p}$.

Proposition 2 (General boundary discontinuity) *Fix a judge j. Let*

$$\mathcal{A}(v) = \{\tau \in \{0,1\}^J : \langle v, \tau \rangle \geq 0\}$$

denote the set of response types with nonnegative AM weight. At any propensity profile satisfying $p_j = \bar{p}$, the set-valued map $v \mapsto \mathcal{A}(v)$ changes discontinuously. In particular, for two arbitrarily close propensity profiles, one with $p_j < \bar{p}$ and one with $p_j > \bar{p}$, the admissible sets differ by at least the two types e_j and $\mathbf{1} - e_j$.

Proof. If $p_j < \bar{p}$, then

$$\omega(e_j) < 0, \qquad \omega(\mathbf{1} - e_j) > 0.$$

Thus $e_j \notin \mathcal{A}(v)$ while $\mathbf{1} - e_j \in \mathcal{A}(v)$. If $p_j > \bar{p}$, then

$$\omega(e_j) > 0, \qquad \omega(\mathbf{1} - e_j) < 0.$$

Thus $e_j \in \mathcal{A}(v)$ while $\mathbf{1} - e_j \notin \mathcal{A}(v)$. At $p_j = \bar{p}$, both types have zero weight. Therefore the admissible set changes discretely when judge j crosses the boundary. Since such profiles can be chosen arbitrarily close to the boundary, the map $v \mapsto \mathcal{A}(v)$ is discontinuous at $p_j = \bar{p}$. ■

The finite-J estimand inherits the same target-population problem. Let $\pi_\tau = \Pr(G_i = \tau)$ and $\gamma_\tau = E[\delta_i \mid G_i = \tau]$. The IV estimand can be written as

$$\beta^{IV}(v) = \frac{\sum_{\tau:\langle v,\tau\rangle>0} \pi_\tau \langle v, \tau \rangle \gamma_\tau}{\sum_{\tau:\langle v,\tau\rangle>0} \pi_\tau \langle v, \tau \rangle}.$$

If the types e_j and $\mathbf{1} - e_j$ have positive population shares and different average treatment effects, then crossing the boundary can change which treatment-effect component enters the estimand. The value of the change depends on assignment shares, type shares, and treatment-effect heterogeneity. The result only says that local invariance is not guaranteed under average monotonicity.

The scope for boundary switches grows with the number of judges. With J judges, any judge whose propensity lies at or near $\bar{p}$ can generate a pairwise switch between e_j and $\mathbf{1} - e_j$. If several judges lie near the boundary, several mirror pairs can be weakly classified at the same time. This is why the empirical diagnostic in Section 4 reports judge-level distances, assignment shares, and a design-level minimum distance.

C Minimal Diagnostic Derivation

This appendix gives the formal interpretation of the boundary diagnostic used in Section 4. The population object is the judge-specific distance from the AM boundary,

$$d_z = p_z - \bar{p}, \qquad \bar{p} = \sum_{j=1}^{J} \lambda_j p_j.$$

The sign of d_z determines whether judge z lies above or below the AM boundary. The sample analogue is

$$\hat{d}_z = \hat{p}_z - \hat{\bar{p}}, \qquad \hat{\bar{p}} = \sum_{j=1}^{J} \hat{\lambda}_j \hat{p}_j.$$

The standardized diagnostic is

$$\hat{r}_z = \frac{|\hat{d}_z|}{\widehat{\mathrm{se}}(\hat{d}_z)}.$$

The diagnostic is a sign-classification statistic. It is not a test of average monotonicity. It asks whether the data place judge z securely on one side of $p_z = \bar{p}$. Direction is given by $\mathrm{sign}(\hat{d}_z)$. Distance from the boundary is given by $\hat{r}_z$.

Lemma 1 (Wald interpretation) *For any cutoff $c > 0$,*

$$\hat{r}_z \le c \quad \Longleftrightarrow \quad 0 \in \left[\hat{d}_z - c\,\widehat{\mathrm{se}}(\hat{d}_z), \hat{d}_z + c\,\widehat{\mathrm{se}}(\hat{d}_z)\right].$$

Proof. By definition,

$$\hat{r}_z \le c \quad \Longleftrightarrow \quad |\hat{d}_z| \le c\,\widehat{\mathrm{se}}(\hat{d}_z).$$

This is equivalent to

$$\hat{d}_z - c\,\widehat{\mathrm{se}}(\hat{d}_z) \le 0 \le \hat{d}_z + c\,\widehat{\mathrm{se}}(\hat{d}_z).$$

Thus zero lies inside the displayed Wald interval. ■

The Wald interpretation connects the diagnostic to target-population ambiguity. If $\hat{r}_z \le c$, the sample does not strongly separate judge z from the boundary at cutoff c. Since AM weights depend on the sign of d_z, a weakly pinned sign means that the corresponding AM positive-weight classification is weakly pinned as well.

A local-to-boundary approximation gives the sign-error interpretation. Consider a sequence of designs indexed by n. Let $d_{z,n}$ denote the population boundary distance for judge z, let $\hat{d}_{z,n}$ denote its estimator, and let

$$\sigma_{z,n} = \mathrm{se}(\hat{d}_{z,n}).$$

Suppose

$$\frac{\hat{d}_{z,n} - d_{z,n}}{\sigma_{z,n}} \xrightarrow{d} \mathcal{N}(0,1), \qquad \frac{d_{z,n}}{\sigma_{z,n}} \to \rho_z,$$

where ρ_z is finite and nonzero. Then

$$\Pr\Big(\operatorname{sign}(\hat{d}_{z,n}) \neq \operatorname{sign}(d_{z,n})\Big) \to \Phi(-|\rho_z|).$$

Thus the probability of sign misclassification is governed by distance to the boundary in standard-error units.

The design-level diagnostic is the minimum standardized distance:

$$\hat{r}_{\min} = \min_{1 \le z \le J} \hat{r}_z.$$

This statistic records whether at least one judge is close to the boundary. For any cutoff $c > 0$,

$$\hat{r}_{\min} > c \quad \Longleftrightarrow \quad \hat{r}_z > c \quad \text{for every } z.$$

A high $\hat{r}_{\min}$ means that all judges are separated from the boundary in standard-error units. A low $\hat{r}_{\min}$ means that at least one judge is weakly classified.

The minimum statistic is only an alarm statistic. It does not replace the full diagnostic vector

$$\{\hat{d}_z, \hat{r}_z, \hat{\lambda}_z\}_{z=1}^{J}.$$

The identity of the judge attaining $\hat{r}_{\min}$ can itself be unstable when several judges lie near the boundary. For interpretation, the reader needs the distance, the standardized distance, and the assignment share of each fragile judge.

There is no population standard error attached to d_z. Sampling uncertainty enters through $\hat{d}_z$. The statistic $\hat{r}_z$ should therefore be read as a sample diagnostic: it measures how strongly the sample classifies the sign of d_z, not whether the population parameter is itself uncertain.

D Philadelphia Propensity and Standard-Error Construction

This appendix describes how the Philadelphia boundary distances in Section 4 are constructed. The application uses the maximum-control specification from the Stevenson replication file. The outcome in the first-stage equation is pretrial detention. The excluded instruments are the eight judge indicators. The controls include charge categories, demographics, prior-record variables, offense indicators, and court-calendar controls. The empirical assignment share of judge z is

denoted by $\hat{\lambda}_z$.

The judge propensities are regression-adjusted treatment margins. Let $\hat{p}_z$ denote the adjusted predicted detention rate when the judge indicator is set to judge z, holding the control distribution fixed at the sample distribution. This produces the vector

$$\hat{p} = (\hat{p}_1, \ldots, \hat{p}_J)'.$$

The assignment-weighted average propensity is

$$\hat{\bar{p}} = \sum_{z=1}^{J} \hat{\lambda}_z \hat{p}_z.$$

The judge-specific boundary distance is then

$$\hat{d}_z = \hat{p}_z - \hat{\bar{p}}.$$

The standard errors treat empirical assignment shares as fixed. Let $\widehat{V}_p$ denote the estimated covariance matrix of the adjusted judge treatment margins $\hat{p}$. Define

$$A = I_J - \mathbf{1}\hat{\lambda}',$$

where $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_J)'$ and $\mathbf{1}$ is a $J \times 1$ vector of ones. Then the vector of boundary distances can be written as

$$\hat{d} = A\hat{p}.$$

The delta-method covariance matrix is

$$\widehat{V}_d = A\widehat{V}_p A'.$$

The standard error for judge z's boundary distance is

$$\widehat{\mathrm{se}}(\hat{d}_z) = \sqrt{e_z' \widehat{V}_d e_z},$$

where e_z is the z-th unit vector.

The standardized boundary distance uses this delta-method standard error. For each judge,

$$\hat{r}_z = \frac{|\hat{d}_z|}{\widehat{\mathrm{se}}(\hat{d}_z)}.$$

A judge is classified as boundary-fragile at cutoff c if $\hat{r}_z \leq c$. Section 4 uses $c = 1.96$.

The Philadelphia table reports four objects for each judge. It reports the number of cases handled by the judge, the empirical assignment share $\hat{\lambda}_z$, the adjusted propensity $\hat{p}_z$, and the boundary distance $\hat{d}_z$ with its delta-method standard error. It then reports $\hat{r}_z$ and the implied side of the boundary. The average propensity in the maximum-control specification is

$$\hat{\bar{p}} = 0.4116.$$

At $c = 1.96$, Judges 1, 5, and 7 are boundary-fragile, and their combined assignment share is 0.291.

The boundary-consistent IV sets reported in Section 4 use these same adjusted propensities. They hold $\hat{p}$ fixed and vary the assignment-share vector λ. Changing λ changes the implied boundary

$$\hat{\bar{p}}(\lambda) = \sum_{z=1}^{J} \lambda_z \hat{p}_z.$$

The full endpoint search for the boundary-consistent sets is reported in the online appendix. The in-paper appendix records only the construction of the propensities and standard errors needed to read the diagnostic.

References

Chan, David C, Matthew Gentzkow, and Chuan Yu, "Selection with variation in diagnostic skill: Evidence from radiologists," *The Quarterly Journal of Economics*, 2022, *137* (2), 729–783.

Chyn, Eric, Brigham Frandsen, and Emily Leslie, "Examiner and judge designs in economics: a practitioner's guide," *Journal of Economic Literature*, 2025, *63* (2), 401–439.

Coulibaly, Mohamed, Yu-Chin Hsu, Ismael Mourifié, and Yuanyuan Wan, "A sharp test for the judge leniency design," Technical Report, National Bureau of Economic Research 2024.

Dobbie, Will, Jacob Goldin, and Crystal S Yang, "The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges," *American Economic Review*, 2018, *108* (2), 201–240.

_ , Paul Goldsmith-Pinkham, and Crystal S Yang, "Consumer bankruptcy and financial health," *Review of Economics and Statistics*, 2017, *99* (5), 853–869.

Frandsen, Brigham, Lars Lefgren, and Emily Leslie, "Judging judge fixed effects," *American Economic Review*, 2023, *113* (1), 253–277.

Garin, Andrew, Dmitri Koustas, Carl McPherson, Samuel Norris, Matthew Pecenco, Evan K Rose, Yotam Shem-Tov, and Jeffrey Weaver, "The impact of incarceration on employment, earnings, and tax filing," *Econometrica*, 2025, *93* (2), 503–538.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár, "Leniency Designs: An Operator's Manual," Technical Report, National Bureau of Economic Research 2025.

Humphries, John Eric, Aurelie Ouss, Kamelia Stavreva, Megan T Stevenson, and Winnie Van Dijk, "Conviction, incarceration, and recidivism: Understanding the revolving door," *The Quarterly Journal of Economics*, 2025, *140* (4), 2907–2962.

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

Jr, Joseph J Doyle, John A Graves, Jonathan Gruber, and Samuel A Kleiner, "Measuring returns to hospital care: Evidence from ambulance referral patterns," *Journal of Political Economy*, 2015, *123* (1), 170–214.

Kling, Jeffrey R, "Incarceration length, employment, and earnings," *American Economic Review*, 2006, *96* (3), 863–876.

Maestas, Nicole, Kathleen J Mullen, and Alexander Strand, "Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt," *American economic review*, 2013, *103* (5), 1797–1829.

Mogstad, Magne and Alexander Torgovitsky, "Instrumental variables with unobserved heterogeneity in treatment effects," in "Handbook of Labor Economics," Vol. 5, Elsevier, 2024, pp. 1–114.

Sigstad, Henrik, "Monotonicity among judges: Evidence from judicial panels and consequences for judge iv designs," *American Economic Review*, 2026, *116* (1), 189–208.

Stevenson, Megan T, "Distortion of justice: How the inability to pay bail affects case outcomes," *The Journal of Law, Economics, and Organization*, 2018, *34* (4), 511–542.