Target-Population Fragility in Judge IV under Average Monotonicity

Mohamed Coulibaly* Sidi Mohamed Sawadogo†

June 24, 2026

Abstract

Judge-IV designs increasingly invoke average monotonicity when strong monotonicity is hard to defend. We show that average monotonicity does not preserve a stable target population: the response types receiving positive weight switch discretely when a judge's treatment propensity crosses the assignment-weighted average. With heterogeneous treatment effects, the estimand remains a proper weighted average but may represent a different target population after an arbitrarily small design change. A standardized boundary-distance diagnostic flags fragile judges; in Philadelphia bail data, three of eight judges, covering a non-negligible share of cases, are fragile. Target-population stability is therefore an empirical object.

Keywords: Judge-IV designs; Average monotonicity; Target population; Complier weights; Pretrial detention
JEL Classification: C21, C26, C18

1 Introduction

Instrumental-variable designs based on the random assignment of judges have become a standard source of quasi-experimental variation, used to estimate the effects of incarceration, pretrial detention, disability insurance, and other consequential decisions.[1] Their interpretation rests on monotonicity. Strong monotonicity, the requirement that a judge

*Corresponding author. Department of Applied Economics, HEC Montréal. E-mail: mohamed.coulibaly@hec.ca.

†Département de sciences économiques and CIREQ, Université de Montréal. E-mail: sidi.mohamed.sawadogo@umontreal.ca.

[1] Prominent applications include Kling (2006); Maestas et al. (2013); Doyle Jr et al. (2015); Dobbie et al. (2017, 2018).

who is stricter on average be weakly stricter in every case, is often hard to defend. Thus, a growing literature has turned to average monotonicity, a weaker condition that preserves the IV estimand as a proper weighted average of treatment effects even when judges rank cases differently (Frandsen et al., 2023; Chyn et al., 2025). That shift fixes the sign of the weights, but it leaves a separate question open. This paper asks whether the IV estimand under average monotonicity identifies the effect for a stable target population.

We answer this question in the smallest setting that displays the mechanism, a three-judge design with eight response types. Under strong monotonicity, the target population is defined by behavior, namely the individuals whose treatment changes when they face a stricter judge. Under average monotonicity, the set of positively weighted types depends instead on where each judge's propensity lies relative to the assignment-weighted average $\bar{p}$. Two non-monotone types sit at that boundary with weights of opposite sign, so when a judge's propensity crosses $\bar{p}$, one type leaves the positively weighted population and its mirror enters. If these two types have positive shares in their respective populations and distinct average treatment effects, the identified IV parameter need not be locally invariant, even though the propensity profile changed by an arbitrarily small amount. Average monotonicity continues to rule out negative weights, but it does not pin the estimand to a population fixed by behavior alone. Where Mogstad and Torgovitsky (2024) note that the target population identified under average monotonicity depends on the instrument distribution, we show that this dependence is discontinuous: an arbitrarily small change in the propensity profile can change which response types the estimand averages over, not only their weights. The same logic extends to any finite number of judges.

The boundary result has an empirical counterpart. Because researchers estimate judge propensities rather than observe them, we propose a standardized boundary distance that measures whether the sample places each judge securely on one side of $\bar{p}$, and we illustrate it in the Philadelphia bail data of Stevenson (2018). The exercise does not reject the design or test average monotonicity; it shows that target-population stability should be reported alongside balance, first-stage, and exclusion checks (Chyn et al., 2025; Goldsmith-Pinkham et al., 2025).

2 IV Weights under Average Monotonicity

We work in the standard judge-IV design with random assignment and exclusion, where the IV estimand is a weighted average of individual treatment effects. Let $Z_i \in \{1, \ldots, J\}$ denote individual i's judge, with $\Pr(Z_i = z) = \lambda_z > 0$, and let $D_{iz} \in \{0, 1\}$ be the treatment individual i would receive if assigned to judge z. Potential outcomes are Y_{1i} and Y_{0i}, with treatment effect $\delta_i = Y_{1i} - Y_{0i}$. We maintain that judge assignment is independent of $(Y_{1i}, Y_{0i}, \{D_{iz}\}_z)$ and affects outcomes only through treatment. Writing $p_z = E[D_i \mid Z_i = z]$ for judge z's propensity and $\bar{p} = \sum_z \lambda_z p_z$ for the assignment-weighted average propensity, the estimand is $\beta^{IV} = E[\omega_i \delta_i]/E[\omega_i]$.

The weight on individual i is the covariance, across judge assignments, between judge propensities and i's potential treatment profile, and its sign governs whether IV retains a weighted-average interpretation. This weight is

$$\omega_i = \sum_{z=1}^{J} \lambda_z (p_z - \bar{p})(D_{iz} - \bar{D}_i), \qquad \bar{D}_i = \sum_{z=1}^{J} \lambda_z D_{iz}. \tag{2.1}$$

It is positive when judges who treat more often in the population are also more likely to treat individual i. When treatment effects are heterogeneous, an individual with $\omega_i < 0$ enters β^{IV} with negative weight, and the estimand can no longer be read as an average effect for a positively weighted subpopulation.

Strong monotonicity fixes the target population through behavior, whereas average monotonicity asks only that each individual's weight be nonnegative. Strong monotonicity requires $D_{iz} \geq D_{iz'}$ whenever $p_z \geq p_{z'}$. This pins the estimand to a fixed complier population but forces judges to differ along a single leniency dimension, a restriction that fails when judges rank cases differently across margins. Average monotonicity weakens this requirement to a sign condition on the individual weight.

Assumption 1 (Average Monotonicity) *For each individual i, $\omega_i \geq 0$ almost surely.*

Average monotonicity allows individual profiles that violate the common ranking of strong monotonicity while still preserving nonnegative weights, so β^{IV} remains a proper weighted average of treatment effects, as in Frandsen et al. (2023).

Average monotonicity removes negative weights but ties the admissible response types to where each judge sits relative to the average propensity, not to behavior alone. Describe individual i by a response type $\tau = (\tau_1, \ldots, \tau_J) \in \{0,1\}^J$, where τ_z is the treatment status under judge z. Because $\sum_z \lambda_z (p_z - \bar{p}) = 0$, the type-specific weight is $\omega(\tau) = \sum_z \lambda_z (p_z - \bar{p}) \tau_z$, and average monotonicity requires $\omega(\tau) \geq 0$ for every type that appears in the population. Whether a type clears this bar depends on the propensity profile $(p_1, \ldots, p_J)$ and the assignment probabilities $(\lambda_1, \ldots, \lambda_J)$, not on the response type alone (Mogstad and Torgovitsky, 2024). Average monotonicity therefore solves the negative-weight problem without fixing a target population from behavior alone, an observation we sharpen next.

3 Boundary Discontinuity

The three-judge design is the smallest setting in which average monotonicity changes the target population without changing the sign of the IV weights. Order the judges so that $p_1 \leq p_2 \leq p_3$. The two extreme judges generate ordinary compliers; three judges is the smallest design that also admits individuals whose treatment is non-monotone in judge propensity while leaving $\bar{p}$ free to move between them.

Written relative to the intermediate judge, the weight makes the sign of ω_i depend on treatment status under the extreme judges compared with the middle judge. Since $\sum_{z=1}^{3} \lambda_z(p_z - \bar{p}) = 0$,

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

Because $p_1 \leq \bar{p} \leq p_3$, the first coefficient is nonpositive and the second nonnegative, so the sign of ω_i is set by how individual i's treatment under the lenient and strict judges compares with treatment under the intermediate judge. Each individual belongs to one of eight response types $g = (D_{i1}, D_{i2}, D_{i3})$, enumerated with their weights and admissibility under strong and average monotonicity in Appendix A.

Six of the eight response types have an admissibility that does not depend on the propensity profile. Always-takers and never-takers receive zero weight; the monotone compliers $(0, 0, 1)$ and $(0, 1, 1)$ are admissible under both conditions; the reverse types $(1, 1, 0)$ and $(1, 0, 0)$ carry negative weight and are ruled out under average monotonicity. None of these classifications turns on where p_2 lies relative to $\bar{p}$.

The two non-monotone mirror types, $g_4 = (1, 0, 1)$ and $g_8 = (0, 1, 0)$, carry weights of opposite sign, fixed entirely by the sign of $p_2 - \bar{p}$. Type g_4 is treated by the lenient and strict judges but not the intermediate one; type g_8 is treated only by the intermediate judge. Strong monotonicity rules out both, since neither profile is weakly increasing in judge propensity. Their weights are

$$\omega(g_4) = -\lambda_2(p_2 - \bar{p}), \qquad \omega(g_8) = \lambda_2(p_2 - \bar{p}). \tag{3.2}$$

When $p_2 < \bar{p}$, average monotonicity admits g_4 and excludes g_8; when $p_2 > \bar{p}$, the roles reverse; when $p_2 = \bar{p}$, both receive zero weight.

The admissible set therefore changes discontinuously as p_2 crosses $\bar{p}$, and when the switching types are present and have distinct average treatment effects, the estimand need not be locally invariant. All six profile-independent types keep their sign through the crossing, so the only change in the positive-weight set is the switch between g_4 and g_8: one type exits and its mirror enters (Figure 1).

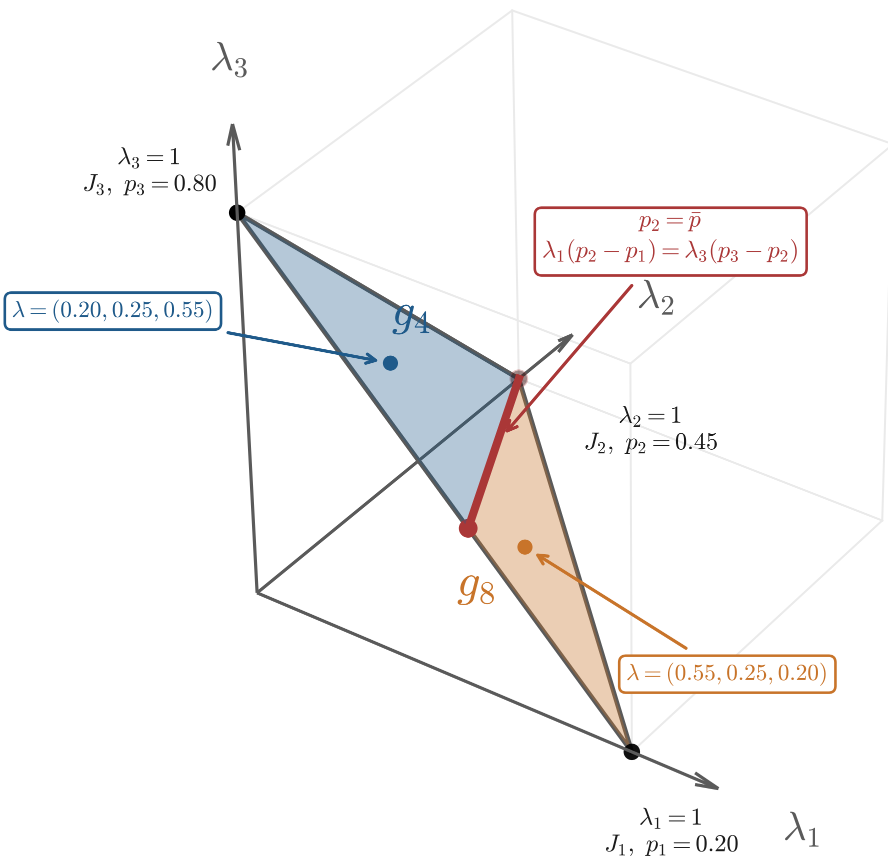

Figure 1: The Average-Monotonicity Boundary in the Assignment-Share Simplex (Three Judges)

Notes: The figure fixes judge propensities at $p_1 = 0.20$, $p_2 = 0.45$, $p_3 = 0.80$ and varies the assignment shares $(\lambda_1, \lambda_2, \lambda_3)$ over the simplex. The red locus is the average-monotonicity boundary $p_2 = \bar{p}$, equivalently $\lambda_1(p_2 - p_1) = \lambda_3(p_3 - p_2)$. In the blue region $p_2 < \bar{p}$ and the non-monotone type $g_4 = (1, 0, 1)$ receives positive weight; in the orange region $p_2 > \bar{p}$ its mirror $g_8 = (0, 1, 0)$ receives positive weight. Crossing the boundary switches the positively weighted type discretely rather than reweighting smoothly.

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

Let $\tau_g = E[\delta_i \mid \text{type } g]$. The crossing replaces a weighted average loading on g_4 with one loading on g_8, so if the switching types carry positive mass and $\tau_{g_4} \neq \tau_{g_8}$, the identified parameter need not be locally invariant.

Corollary 1 (Target-population instability) *Across $p_2 = \bar{p}$, two average-monotonicity-consistent designs arbitrarily close in the propensity profile can place positive weight on disjoint switching types, g_4 and g_8. If these types have positive shares in their respective populations and $\tau_{g_4} \neq \tau_{g_8}$, the resulting IV estimands target different positive-weight populations and need not be locally invariant. The discontinuity is therefore in the target population, not in the fixed-population continuity of β^{IV} as a function of the propensity profile.*

The same mechanism extends to any finite number of judges. For any judge z, the type treated only by z carries weight $\lambda_z(p_z - \bar{p})$ and its componentwise complement carries $-\lambda_z(p_z - \bar{p})$, so the two switch sign exactly when p_z crosses $\bar{p}$. The Supplementary Material (Appendix B) states and proves the finite-J result; we use the three-judge case in the text because it displays the mechanism without the additional notation. Proofs of Proposition 1 and Corollary 1 are in Appendix A.

4 A Boundary Diagnostic

Once judge propensities are estimated, the relevant question is whether the sample places each judge securely on one side of $\bar{p}$, since the exact crossing $p_z = \bar{p}$ is non-generic. Let $\hat{d}_z = \hat{p}_z - \hat{\bar{p}}$ denote judge z's estimated distance to the boundary, with $\hat{\bar{p}} = \sum_j \hat{\lambda}_j \hat{p}_j$, and let $\hat{r}_z = |\hat{d}_z| / \widehat{\mathrm{se}}(\hat{d}_z)$ be its standardized value. A judge with small $\hat{r}_z$ is weakly classified, and when such judges carry nontrivial assignment mass the sign of their contribution to β^{IV}, and hence the target population, is not pinned down by the data; the statistic is a sign-classification device, not a test of average monotonicity. We recommend reporting, alongside the usual balance, first-stage, and exclusion checks, each judge's $\hat{d}_z$, its standardized value $\hat{r}_z$, and the assignment share of judges near the boundary. In the Philadelphia bail data of Stevenson (2018), three of eight judges (about 29 percent of cases) are boundary-fragile at $c = 1.96$, even in a design earlier work does not reject (Coulibaly et al., 2024); the Supplementary Material (Appendix C) gives the construction and the full illustration, and Appendix D shows that such fragility is common in many-judge designs.

5 Conclusion

Average monotonicity gives judge-IV estimands nonnegative weights, but it does not give them a fixed target population. The positive-weight response types switch discretely when a judge's treatment propensity crosses the assignment-weighted average, and when the switching types have positive population shares and distinct average treatment effects, that switch need not leave the IV estimand locally invariant. The estimand stays a proper weighted average; what average monotonicity does not guarantee is the population over which it is taken.

The accompanying diagnostic makes this stability checkable without testing average monotonicity itself: researchers who invoke average monotonicity should report where estimated judge propensities lie relative to the assignment-weighted average, how precisely those distances are estimated, and how much case mass sits near the boundary. When fragility is detected, a natural next step is a sensitivity analysis bounding how much the IV estimand can vary across the populations induced by alternative classifications of the fragile judges.

A Proofs for the Three-Judge Case

This appendix proves Proposition 1 and Corollary 1 and records the response-type table. Judges are ordered $p_1 \leq p_2 \leq p_3$ with $\lambda_z > 0$ and $\bar{p} = \sum_{z=1}^{3} \lambda_z p_z$; a response type $g \in \{0,1\}^3$ has type-specific weight $\omega(g) = \sum_{z=1}^{3} \lambda_z (p_z - \bar{p}) g_z$.

The weight admits a representation relative to the intermediate judge. Because $\sum_{z=1}^{3} \lambda_z (p_z - \bar{p}) = 0$,

$$\omega(g) = \lambda_1(p_1 - \bar{p})(g_1 - g_2) + \lambda_3(p_3 - \bar{p})(g_3 - g_2), \tag{A.1}$$

where $\lambda_1(p_1 - \bar{p}) \leq 0$ and $\lambda_3(p_3 - \bar{p}) \geq 0$ because $p_1 \leq \bar{p} \leq p_3$. Evaluating (A.1) at the eight types gives the weights in Table A.1; the non-monotone mirror types satisfy $\omega(1,0,1) = -\lambda_2(p_2 - \bar{p})$ and $\omega(0,1,0) = \lambda_2(p_2 - \bar{p})$, the only weights whose sign depends on $p_2 - \bar{p}$.

Proof of Proposition 1. Since $\lambda_2 > 0$, the signs of $\omega(g_4) = -\lambda_2(p_2 - \bar{p})$ and $\omega(g_8) = \lambda_2(p_2 - \bar{p})$ are set by $p_2 - \bar{p}$. When $p_2 < \bar{p}$, $\omega(g_4) > 0$ and $\omega(g_8) < 0$, so average monotonicity admits g_4 and rules out g_8. When $p_2 > \bar{p}$, the signs reverse, so average monotonicity admits g_8 and rules out g_4. When $p_2 = \bar{p}$, both weights are zero. All other response types have signs that do not depend on $p_2 - \bar{p}$, so the only change in the positive-weight set as p_2 crosses $\bar{p}$ is the switch between g_4 and g_8. The positive-weight set therefore changes discontinuously at $p_2 = \bar{p}$. ■

Table A.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.

Proof of Corollary 1. Write the IV estimand under average monotonicity as a ratio over positively weighted types,

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

where $\pi_g = \Pr(G_i = g)$ and $\tau_g = E[\delta_i \mid G_i = g]$. *For a fixed population, this ratio is continuous in the propensity profile.* Every weight $\omega(g)$ is continuous in p_2, and the switching weights $\omega(g_4) = -\lambda_2(p_2 - \bar{p})$ and $\omega(g_8) = \lambda_2(p_2 - \bar{p})$ vanish as $p_2 \to \bar{p}$, so numerator and denominator are continuous and, wherever the denominator is bounded away from zero, so is β^{IV}. *The admissible support is discontinuous.* By Proposition 1 the admissible set switches g_4 for g_8 as p_2 crosses $\bar{p}$, since a positive share on the excluded mirror type would carry negative weight. Preserving average monotonicity across the boundary therefore requires two distinct populations, with identified parameters

$$\beta_- = \frac{\sum_{g\in\{g_3,g_4,g_5\}} \pi_g\, \omega(g)\, \tau_g}{\sum_{g\in\{g_3,g_4,g_5\}} \pi_g\, \omega(g)}, \qquad \beta_+ = \frac{\sum_{g\in\{g_3,g_5,g_8\}} \pi_g\, \omega(g)\, \tau_g}{\sum_{g\in\{g_3,g_5,g_8\}} \pi_g\, \omega(g)},$$

that average over disjoint switching components and need not coincide. Example 1 gives admissible shares and heterogeneous effects with $\tau_{g_4} \neq \tau_{g_8}$ for which $\beta_- \neq \beta_+$. ■

Example 1 (Target population across the boundary) *Consider two three-judge designs that differ only in whether p_2 lies just below or just above $\bar{p}$, each consistent with average monotonicity. In the first, $p_2 < \bar{p}$ and the non-monotone type present is $g_4 = (1,0,1)$; in the second, $p_2 > \bar{p}$ and it is the mirror $g_8 = (0,1,0)$. Apart from the switching type, both populations contain only always-takers and never-takers, which carry zero weight. Average monotonicity holds in each design, since the admitted non-monotone type carries positive* *weight while its mirror is absent. Set $\tau_{g_4} \neq \tau_{g_8}$. The only positively weighted type is the switching type, so the first design identifies $\beta_- = \tau_{g_4}$ and the second $\beta_+ = \tau_{g_8}$. The two designs are arbitrarily close in the propensity profile yet identify effects for disjoint groups, so $\beta_- \neq \beta_+$, even though for a single fixed population β^{IV} would vary continuously in p_2.*

References

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.

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

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.

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.

Supplementary Material for
"Target-Population Fragility in Judge IV under Average Monotonicity"

Mohamed Coulibaly Sidi Mohamed Sawadogo

This material supports the main text, whose in-paper Appendix A proves the three-judge results and records the response-type table. Appendix B extends the boundary mechanism to any finite number of judges. Appendix C derives the boundary diagnostic, constructs the Philadelphia propensities and standard errors, and reports the full empirical illustration. Appendix D reports simulations that validate the diagnostic as a finite-sample classification tool. Appendix E positions the paper in the related literature. Equation, proposition, figure, and table numbers carry the appendix letter.

B General Finite-J Extension

This appendix shows that the boundary mechanism of Section 3 is not specific to three judges. Judges are indexed by $z \in \{1, \ldots, J\}$ with $J \geq 3$, each assigned with probability $\lambda_z > 0$, and each individual has a response type $\tau = (\tau_1, \ldots, \tau_J) \in \{0,1\}^J$. The assignment-weighted average propensity is $\bar{p} = \sum_{z=1}^{J} \lambda_z p_z$.

The type-specific weight is an inner product between the response type and the propensity-deviation vector. Let $v = (\lambda_1(p_1 - \bar{p}), \ldots, \lambda_J(p_J - \bar{p}))$. Then

$$\omega(\tau) = \sum_{z=1}^{J} \lambda_z (p_z - \bar{p}) \tau_z = \langle v, \tau \rangle, \qquad \langle v, \mathbf{1} \rangle = 0,$$

where $\mathbf{1} = (1, \ldots, 1)$, since $\sum_{z=1}^{J} \lambda_z (p_z - \bar{p}) = 0$. 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 changes the positive-weight set. Fix judge j and let $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)$ be the type treated only by judge j. Its complement $\mathbf{1} - e_j$ is the type treated by every judge except j. Their weights are

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

which switch sign exactly when p_j crosses $\bar{p}$.

Proposition B.1 (General boundary discontinuity)**.** *Fix a judge j and let $\mathcal{A}(v) = \{\tau \in \{0,1\}^J : \langle v, \tau \rangle \geq 0\}$ denote the set of response types with nonnegative average-monotonicity weight. At any propensity profile satisfying $p_j = \bar{p}$, the set-valued map $v \mapsto \mathcal{A}(v)$ changes discontinuously: for two arbitrarily close 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 $\omega(e_j) < 0$ and $\omega(\mathbf{1} - e_j) > 0$, so $e_j \notin \mathcal{A}(v)$ while $\mathbf{1} - e_j \in \mathcal{A}(v)$. If $p_j > \bar{p}$, then $\omega(e_j) > 0$ and $\omega(\mathbf{1} - e_j) < 0$, so $e_j \in \mathcal{A}(v)$ while $\mathbf{1} - e_j \notin \mathcal{A}(v)$. At $p_j = \bar{p}$, both types have zero weight. The admissible set therefore changes discretely when judge j crosses the boundary, and since such profiles can be chosen arbitrarily close to the boundary, $v \mapsto \mathcal{A}(v)$ is discontinuous at $p_j = \bar{p}$. □

The finite-J estimand inherits the same target-population problem. With $\pi_\tau = \Pr(G_i = \tau)$ and $\gamma_\tau = E[\delta_i \mid G_i = \tau]$,

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

If e_j and $\mathbf{1} - e_j$ have positive population shares and different average effects, crossing the boundary changes which treatment-effect component enters the estimand. The scope for such switches grows with the number of judges: any judge whose propensity lies near $\bar{p}$ can generate a pairwise switch, and several judges near the boundary can be weakly classified at once. This is why the diagnostic of Section 4 reports judge-level distances, assignment shares, and a design-level minimum distance.

C The Boundary Diagnostic: Propensities, Standard Errors, and Sign Classification

This appendix derives the boundary diagnostic of Section 4 and reports the full Philadelphia illustration. The diagnostic is a sign-classification statistic for each judge's position relative to the average-monotonicity boundary, not a test of average monotonicity.

C.1 Population object and Wald interpretation

The population object is the judge-specific distance from the boundary, $d_z = p_z - \bar{p}$ with $\bar{p} = \sum_{j=1}^{J} \lambda_j p_j$. Its sign determines whether judge z lies above or below the boundary, and under average monotonicity that sign governs which mirror response types can receive positive weight. The sample analogue is $\hat{d}_z = \hat{p}_z - \hat{\bar{p}}$ with $\hat{\bar{p}} = \sum_{j=1}^{J} \hat{\lambda}_j \hat{p}_j$, and the standardized diagnostic is $\hat{r}_z = |\hat{d}_z| / \widehat{\mathrm{se}}(\hat{d}_z)$. Direction is given by $\mathrm{sign}(\hat{d}_z)$ and distance from the boundary by $\hat{r}_z$.

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

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

Proof. By definition $\hat{r}_z \leq c$ is equivalent to $|\hat{d}_z| \leq c\,\widehat{\mathrm{se}}(\hat{d}_z)$, which holds if and only if $\hat{d}_z - c\,\widehat{\mathrm{se}}(\hat{d}_z) \leq 0 \leq \hat{d}_z + c\,\widehat{\mathrm{se}}(\hat{d}_z)$, that is, zero lies in the displayed Wald interval. □

The Wald interpretation links the diagnostic to target-population ambiguity. When $\hat{r}_z \leq c$, the sample does not separate judge z from the boundary at cutoff c, and because average-monotonicity weights depend on the sign of d_z, a weakly pinned sign means a weakly pinned positive-weight classification.

C.2 Local-to-boundary sign error

A local-to-boundary approximation gives the diagnostic its sign-error reading. Consider a sequence of designs indexed by n, with population distance $d_{z,n}$, estimator $\hat{d}_{z,n}$, and standard error $\sigma_{z,n} = \mathrm{se}(\hat{d}_{z,n})$. Suppose

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

with ρ_z finite and nonzero. Then

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

The probability of sign misclassification is therefore governed by distance to the boundary in standard-error units, which is what $\hat{r}_z$ estimates. The design-level summary is the minimum standardized distance $\hat{r}_{\min} = \min_z \hat{r}_z$: a high value means every judge is separated from the boundary, while a low value means at least one judge is weakly classified. The minimum is an alarm statistic only and does not replace the full diagnostic vector $\{\hat{d}_z, \hat{r}_z, \hat{\lambda}_z\}_{z=1}^{J}$, since the identity of the judge attaining $\hat{r}_{\min}$ can itself be unstable when several judges lie near the boundary.

C.3 Philadelphia propensities and standard errors

The Philadelphia diagnostic uses the maximum-control specification of the Stevenson replication. The first-stage outcome is pretrial detention, the excluded instruments are the eight judge indicators, and the controls include charge categories, demographics, prior-record variables, offense indicators, and court-calendar controls. The judge propensities $\hat{p}_z$ are regression-adjusted detention margins, computed by setting the judge indicator to judge z and holding the control distribution fixed at the sample distribution, and $\hat{\lambda}_z$ is the empirical assignment share. The boundary distance is $\hat{d}_z = \hat{p}_z - \hat{\bar{p}}$ with $\hat{\bar{p}} = \sum_{z=1}^{J} \hat{\lambda}_z \hat{p}_z$.

The standard errors treat empirical assignment shares as fixed. Let $\widehat{V}_p$ denote the estimated covariance matrix of the adjusted judge margins $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_J)'$, and define $A = I_J - \mathbf{1}\hat{\lambda}'$ with $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_J)'$. Then the vector of boundary distances is $\hat{d} = A\hat{p}$, its delta-method covariance is $\widehat{V}_d = A\widehat{V}_p A'$, and the standard error for judge z is $\widehat{\mathrm{se}}(\hat{d}_z) = \sqrt{e_z' \widehat{V}_d e_z}$, where e_z is the z-th unit vector. A judge is boundary-fragile at cutoff c when $\hat{r}_z \leq c$, and Section 4 uses $c = 1.96$.

Table C.1 reports the full judge-level diagnostics. 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, with combined assignment share 0.291. Judge 7 is closest to the boundary, with $\hat{d}_7 = 0.0009$ and $\hat{r}_7 = 0.44$.

Table C.1: Judge-level boundary diagnostics in the Philadelphia bail data

Judge	Fragile	Cases	$\hat{\lambda}_z$	$\hat{p}_z$	$\hat{d}_z$	$\widehat{\mathrm{se}}(\hat{d}_z)$	$\hat{r}_z$	Side
1	Yes	21,523	0.065	0.4071	-0.0044	0.0031	1.45	Below
2	No	13,087	0.039	0.4378	0.0262	0.0041	6.37	Above
3	No	54,272	0.163	0.4186	0.0070	0.0017	4.21	Above
4	No	56,585	0.170	0.3936	-0.0179	0.0016	11.15	Below
5	Yes	33,690	0.101	0.4077	-0.0039	0.0023	1.69	Below
6	No	55,038	0.166	0.4304	0.0188	0.0017	11.28	Above
7	Yes	41,475	0.125	0.4124	0.0009	0.0020	0.44	Above
8	No	56,301	0.170	0.4017	-0.0098	0.0017	5.94	Below
Average propensity $\hat{\bar{p}}$					0.4116			
Minimum raw distance $\min_z \lvert\hat{d}_z\rvert$					0.0009			
Minimum standardized distance $\hat{r}_{\min}$					0.44			
Boundary-fragile judges under $c = 1.96$					1, 5, 7			
Total assignment share of fragile judges					0.291			
Total cases					331,971			

Notes: A judge is classified as boundary-fragile when $\hat{r}_z \leq 1.96$. The propensity $\hat{p}_z$ is the regression-adjusted judge-specific detention propensity from the first-stage equation. Standard errors for $\hat{d}_z$ are computed by the delta method using the covariance matrix of the adjusted judge margins, treating empirical assignment shares as fixed.

C.4 Empirical illustration

We apply the diagnostic to the Philadelphia bail data of Stevenson (2018), where bail magistrates rotate across shifts and the scheduled magistrate provides quasi-random variation in pretrial detention. The outcome is conviction, the treatment is pretrial detention, and the excluded instruments are the eight judge indicators, over 331,971 cases. This setting is useful precisely because earlier work finds no decisive evidence against the design in the aggregate sample (Coulibaly et al., 2024), so boundary fragility can be assessed on its own terms rather than as a symptom of a failed design.

Three judges are boundary-fragile at $c = 1.96$ and together handle about 29 percent of cases. Figure C.1 plots $\hat{d}_z$ with 95 percent confidence intervals against the boundary at zero. Judges 1, 5, and 7 fall inside the fragile set: Judge 7 is closest, with $\hat{d}_7 = 0.0009$, $\hat{r}_7 = 0.44$, and $\hat{\lambda}_7 = 0.125$, so the minimum standardized distance is not driven by a judge with negligible case load; Judges 1 and 5 lie just below the boundary, with $\hat{r}_1 = 1.45$ and $\hat{r}_5 = 1.69$.

Boundary fragility is a target-population diagnostic, not a rejection of the design. A judge near the boundary does not imply that assignment is nonrandom, that exclusion fails, or that strong monotonicity is violated. It instead says that once average monotonicity is invoked, the positive-weight population is not sharply classified for judges near the boundary. When those judges carry nontrivial case mass, the sign of their contribution to β^{IV} is not pinned down by the sample, so a cautious reading attaches the estimate to a design-dependent population rather than a fixed complier group.

D Simulations

This appendix reports simulations that validate the boundary diagnostic as a finite-sample classification tool. The diagnostic asks whether the data place each judge securely on one side of the average-monotonicity boundary. The simulations check three things implied by the theory: that low $\hat{r}_z$ corresponds to high rates of sign misclassification, that fragility becomes more common as the number of judges grows, and that the diagnostic-based boundary interval does not exclude the true boundary in finite samples.

D.1 Design

The data-generating process starts from a finite judge design. Judge z has treatment propensity $p_z = E[D \mid Z = z]$ and assignment share $\lambda_z = P(Z = z)$, with $\sum_z \lambda_z = 1$, baseline boundary $\bar{p}^0 = \sum_z \lambda_z p_z$, and population distance $d_z^0 = p_z - \bar{p}^0$. Each individual belongs

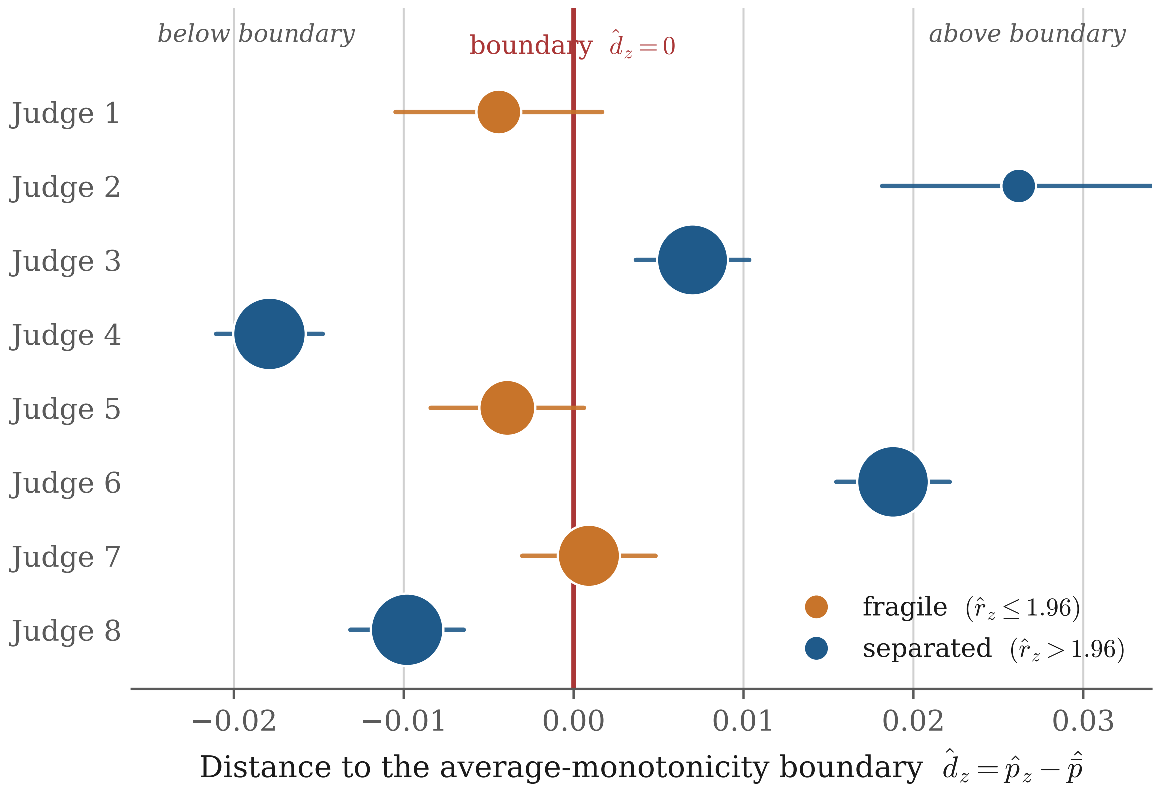

Figure C.1: Judge Distances to the Average-Monotonicity Boundary in Philadelphia Bail

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 $\hat{\lambda}_z$.

to a response type g with treatment profile $D(g) = \{D_1(g), \ldots, D_J(g)\}$ and type share π_g, so that $p_z = \sum_g \pi_g D_z(g)$. In each replication individuals are drawn from the type distribution, judges are assigned according to λ, and treatment is $D_i = D_{Z_i}(g_i)$. Potential outcomes are drawn at the response-type level as independent Bernoulli variables, $Y_i(0) \sim \text{Bernoulli}(q_{0g})$ and $Y_i(1) \sim \text{Bernoulli}(q_{1g})$ with $\tau_g = q_{1g} - q_{0g}$, and the observed outcome is $Y_i = Y_i(0) + D_i\{Y_i(1) - Y_i(0)\}$. The baseline population estimand β^0 is fixed within each design; sampling variation enters only through the estimated objects $\hat{p}_z$, $\hat{d}_z$, $\hat{r}_z$.

D.2 Sample objects

Each replication estimates judge-level treatment and outcome means, $\hat{p}_z = n_z^{-1} \sum_{i:Z_i=z} D_i$, $\hat{y}_z = n_z^{-1} \sum_{i:Z_i=z} Y_i$, and $\hat{\lambda}_z = n_z/n$. The estimated boundary and distance are $\hat{\bar{p}} = \sum_z \hat{\lambda}_z \hat{p}_z$ and $\hat{d}_z = \hat{p}_z - \hat{\bar{p}}$, and the standardized diagnostic is $\hat{r}_z = |\hat{d}_z| / \widehat{\text{se}}(\hat{d}_z)$. In the covariate-free simulation the standard error uses the influence-function representation of $\hat{d}_z$,

$$\hat{\psi}_{iz} = \frac{\mathbf{1}\{Z_i = z\}}{\hat{\lambda}_z}(D_i - \hat{p}_z) - (D_i - \hat{\bar{p}}), \qquad \widehat{\text{se}}(\hat{d}_z) = \left[\tfrac{1}{n}\widehat{\text{Var}}(\hat{\psi}_{iz})\right]^{1/2}.$$

To calibrate the difficulty of the classification problem, we report the design signal-to-noise ratio $\rho_z = |d_z^0| / \text{sd}(\hat{d}_z)$, which the local theory of Appendix C.2 maps into a sign-error rate of approximately $\Phi(-\rho_z)$.

D.3 Designs

We study five designs, summarized in Table D.1. The two stylized three-judge designs isolate the boundary mechanism by placing one judge close to the boundary, one satisfying strong monotonicity and one satisfying average monotonicity without strong monotonicity. The Philadelphia-like design uses the empirical assignment shares and propensities from the application. The two bail designs are calibrated many-judge settings in the spirit of common judge and examiner applications. The many-judge designs are useful because the probability that at least one judge lies close to the boundary rises with the number of judges, even when no single judge is placed there by construction.

D.4 Results

The standardized boundary distance is informative about sign classification. Table D.2 groups judge-replication observations by $\hat{r}_z$ and reports the frequency with which the sample sign of $\hat{d}_z$ differs from the population sign of d_z^0. Sign errors concentrate among judges with

Table D.1: Population features of the simulation designs

Design	J	n	$\bar{p}^0$	$\min_z \lvert d_z^0 \rvert$	$\min_z \rho_z$
Stylized SM	3	10,000	0.498	0.0020	0.204
Stylized AM	3	10,000	0.498	0.0020	0.204
Philadelphia-like	8	20,000	0.4116	0.0008	0.091
Bail Philadelphia-like	9	50,000	0.560	0.0049	0.553
Bail Miami-like	60	50,000	0.560	0.0001	0.010

Notes: The stylized designs isolate the boundary mechanism in three-judge settings. The Philadelphia-like design uses the empirical assignment shares and treatment propensities from the application. The bail designs are calibrated many-judge designs. The statistic $\rho_z = |d_z^0|/\mathrm{sd}(\hat{d}_z)$ is the population signal-to-noise ratio for boundary classification.

low standardized distance and essentially disappear for large $\hat{r}_z$. In the Miami-like design, judges with $\hat{r}_z < 0.5$ are misclassified in 43 percent of samples, while judges with $\hat{r}_z > 4$ are never misclassified, and the same pattern holds across the Philadelphia-like designs.

Table D.2: Sign misclassification by standardized boundary distance

Design	$[0, 0.5)$	$[0.5, 1)$	$[1, 1.645)$	$[1.645, 1.96)$	$[1.96, 2.58)$	$[2.58, 4)$	$[4, \infty)$
Stylized SM	0.513	0.334	0.146	0.053	0.030	0.000	0.000
Stylized AM	0.392	0.284	0.096	0.050	0.020	0.000	0.000
Philadelphia-like	0.414	0.250	0.104	0.043	0.020	0.003	0.000
Bail Philadelphia-like	0.422	0.246	0.101	0.040	0.018	0.003	0.000
Bail Miami-like	0.426	0.243	0.100	0.039	0.019	0.003	0.000

Notes: Each entry reports the frequency with which the sample sign of $\hat{d}_z$ differs from the population sign of d_z^0, among judge-replication observations in the corresponding $\hat{r}_z$ bin. Sign errors are common when $\hat{r}_z$ is small and essentially disappear when $\hat{r}_z$ is large.

Boundary fragility is a design-level phenomenon in many-judge settings. Table D.3 reports the minimum standardized distance, the average number of fragile judges, the assignment share they handle, the probability of at least one sign error, and the frequency with which the plausible boundary interval contains the true boundary. The many-judge designs are markedly more fragile. In the Miami-like design with 60 judges, the average minimum standardized distance is close to zero, about 37 judges are fragile at $c = 1.96$, and at least one sign error occurs in every replication. This is not a symptom of poorly estimated individual judges; with many propensities around a common assignment-weighted average, some judge is likely to lie close to the boundary.

The simulations support the logic of the diagnostic. The population boundary distance d_z^0 is fixed within a design, sampling uncertainty enters through $\hat{d}_z$, and $\hat{r}_z$ measures how strongly the sample classifies the sign of d_z^0. Low values identify judges for whom sign mistakes are common; high values identify judges whose classification is stable. The calibrated

Table D.3: Design-level diagnostic performance

Design	$E[\hat{r}_{\min}]$	$E[\lvert\widehat{\mathcal{F}}(1.96)\rvert]$	Fragile share	Pr(any sign error)	$\Pr(\bar{p}^0 \in \widehat{\mathcal{I}})$
Stylized SM	0.796	0.950	0.317	0.372	1.000
Stylized AM	0.851	0.918	0.306	0.324	1.000
Philadelphia-like	0.259	5.774	0.679	0.778	1.000
Bail Philadelphia-like	0.640	2.060	0.229	0.513	1.000
Bail Miami-like	0.053	37.280	0.621	1.000	1.000

Notes: A judge is fragile when $\hat{r}_z \leq 1.96$. The fragile share is the average assignment share of fragile judges. The plausible boundary interval $\widehat{\mathcal{I}}(1.96)$ is the set of boundary locations that preserve the classification of every non-fragile judge. The final column reports the frequency with which it contains the true population boundary $\bar{p}^0$; this is not a coverage theorem, but it shows that the diagnostic-based interval does not exclude the true boundary in these designs.

designs show why this matters beyond the three-judge example: in many-judge settings fragility can arise even without targeting a single judge, which makes the diagnostic especially relevant in examiner and judge designs with many decision-makers. The diagnostic does not test average monotonicity and does not replace balance tests, first-stage checks, or exclusion discussions. It answers a different question: once average monotonicity is used to interpret the design, are the boundary classifications stably pinned down in the data?

E Additional Literature Discussion

This appendix positions the paper in the judge-IV and heterogeneous-IV literatures. The main text keeps the discussion short because the contribution is a boundary result, not a survey of examiner designs.

The first related literature studies the identifying assumptions behind judge and examiner designs. Frandsen et al. (2023) show that average monotonicity can preserve a proper weighted-average interpretation of the IV estimand even when strong monotonicity is too demanding. Chyn et al. (2025) and Goldsmith-Pinkham et al. (2025) provide methodological guidance for implementing examiner and leniency designs, and Sigstad (2026) measures monotonicity violations directly in judicial panels and studies when judge-IV estimates remain interpretable despite violations of strong monotonicity. This paper takes average monotonicity as the maintained fallback and asks whether it delivers a stable target population.

The second related literature studies IV estimands under treatment-effect heterogeneity. Imbens and Angrist (1994) show that monotonicity yields a local average treatment effect interpretation with a binary instrument. Mogstad and Torgovitsky (2024) emphasize that heterogeneous-IV interpretations can be sensitive to instrument structure and that average monotonicity is not purely behavioral. We use that observation to identify a specific discontinuity: under average monotonicity, the positive-weight response types switch when a judge propensity crosses the assignment-weighted average.

The paper also relates to work on testing judge-IV assumptions. Frandsen et al. (2023) propose a nonparametric test for the exclusion and monotonicity implications of judge designs, and Coulibaly et al. (2024) propose sharp testable implications and apply them to the Philadelphia bail setting. Those tests ask whether the identifying assumptions are rejected by the data. The boundary diagnostic asks a different question: conditional on using average monotonicity, does the estimated propensity profile stably classify the target population? The two are complements. Balance tests, first-stage reporting, exclusion arguments, and monotonicity tests address whether the design can identify a causal parameter; the boundary diagnostic addresses which population receives positive weight under average monotonicity and whether that population is stable around the estimated profile.

References

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.

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

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

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

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.