Online Appendix to "Target-Population Fragility in Judge IV under Average Monotonicity"

Mohamed Coulibaly Sidi Mohamed Sawadogo

May 4, 2026

E Online Appendix: Boundary-Consistent IV Sets

This appendix gives the formal construction of the boundary-consistent IV set used to quantify target ambiguity under average monotonicity. The construction treats judge propensities as fixed and varies the assignment-share vector. Changing assignment shares changes the assignment-weighted average propensity, and therefore moves the AM boundary. The exercise asks how much the IV estimate can move across AM target populations that preserve the classifications of judges whose boundary positions are not fragile.

E.1 Boundary locations and feasible classifications

The AM boundary is determined by the assignment-weighted average propensity. Let p_z denote judge z's treatment propensity and let λ_z denote the assignment share. For any assignment-share vector λ, define

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

Judge z's distance from the boundary is

$$d_z(\lambda) = p_z - \bar{p}(\lambda).$$

A judge is below the boundary if $d_z(\lambda) < 0$, and above the boundary if $d_z(\lambda) > 0$.

The sample diagnostic determines which judge classifications are held fixed. Let

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

For a cutoff $c > 0$, define the fragile set

$$\widehat{\mathcal{F}}(c) = \left\{ z : \frac{|\hat{d}_z|}{\widehat{\mathrm{se}}(\hat{d}_z)} \leq c \right\}.$$

Judges outside $\widehat{\mathcal{F}}(c)$ are treated as having boundary classifications pinned down by the sample. Judges inside $\widehat{\mathcal{F}}(c)$ are allowed to fall on either side of the boundary.

Feasible boundary locations must preserve the classifications of non-fragile judges. Define the non-fragile judges below and above the baseline boundary as

$$\mathcal{L}(c) = \left\{ z : \hat{d}_z < 0,\ z \notin \widehat{\mathcal{F}}(c) \right\},$$

and

$$\mathcal{U}(c) = \left\{ z : \hat{d}_z > 0,\ z \notin \widehat{\mathcal{F}}(c) \right\}.$$

Then define

$$\underline{b}(c) = \max_{z \in \mathcal{L}(c)} \hat{p}_z, \qquad \overline{b}(c) = \min_{z \in \mathcal{U}(c)} \hat{p}_z.$$

When $\mathcal{L}(c)$ is empty, set $\underline{b}(c) = \min_z \hat{p}_z$. When $\mathcal{U}(c)$ is empty, set $\overline{b}(c) = \max_z \hat{p}_z$. The plausible boundary interval is

$$\widehat{\mathcal{I}}(c) = [\underline{b}(c), \overline{b}(c)].$$

Lemma 1 (Feasible boundary interval) *A boundary value b preserves the sample classification of every non-fragile judge if and only if*

$$b \in \widehat{\mathcal{I}}(c).$$

Proof. A non-fragile judge $z \in \mathcal{L}(c)$ satisfies $\hat{d}_z < 0$. Preserving that classification requires $\hat{p}_z - b < 0$, or $b > \hat{p}_z$. This must hold for every $z \in \mathcal{L}(c)$, so $b \geq \underline{b}(c)$, up to the knife-edge convention at equality.

A non-fragile judge $z \in \mathcal{U}(c)$ satisfies $\hat{d}_z > 0$. Preserving that classification requires $\hat{p}_z - b > 0$, or $b < \hat{p}_z$. This must hold for every $z \in \mathcal{U}(c)$, so $b \leq \overline{b}(c)$, again up to the knife-edge convention at equality.

Combining the two restrictions gives $b \in [\underline{b}(c), \overline{b}(c)]$. Conversely, any b in this interval preserves both sets of non-fragile classifications. ■

The plausible boundary interval is a set of boundary locations, not a confidence interval for $\bar{p}$. It records which AM boundaries are compatible with the judge classifications that the diagnostic treats as pinned down. Fragile judges can change side within this interval; non-fragile judges cannot.

E.2 Feasible assignment-share perturbations

The assignment-share perturbation keeps the alternative design close to the observed one. Let

$$\Delta^{J-1} = \left\{ \lambda \in \mathbb{R}_+^J : \sum_{z=1}^{J} \lambda_z = 1 \right\}$$

denote the assignment-share simplex. For a radius $\kappa \geq 0$, define

$$\widehat{\Lambda}(c, \kappa) = \left\{ \lambda \in \Delta^{J-1} : \sum_{z=1}^{J} \lambda_z \hat{p}_z \in \widehat{\mathcal{I}}(c),\ \|\lambda - \hat{\lambda}\|_1 \leq \kappa \right\}.$$

The first restriction makes the perturbed boundary consistent with all non-fragile classifications. The second restriction limits how far the assignment shares may move from the observed design.

The radius κ indexes the scale of the sensitivity exercise. Since assignment shares sum to one, $\kappa/2$ is the maximum share of cases that can be reassigned across judges. A small κ gives a local perturbation. A larger κ gives a stress test. If $\widehat{\Lambda}(c, \kappa)$ is empty, the empirical table should report emptiness rather than force a perturbation.

E.3 IV estimates along feasible assignment profiles

The IV estimate can be evaluated at any feasible assignment-share vector. For $\lambda \in \widehat{\Lambda}(c, \kappa)$, define

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

Let $\hat{y}_z$ denote the judge-specific outcome mean, using the same residualization or covariate adjustment as the baseline IV specification. The reweighted IV estimate is

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

The denominator is

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

The denominator records the first-stage support for the reweighted contrast. Values of λ for which $|\hat{\Gamma}(\lambda)|$ is close to zero should be flagged because they can generate numerical extremes.

Define the boundary-consistent IV set as

$$\widehat{\mathcal{B}}(c,\kappa) = \left\{ \hat{\beta}(\lambda) : \lambda \in \widehat{\Lambda}(c,\kappa),\ |\hat{\Gamma}(\lambda)| \geq \gamma_{\min} \right\},$$

where $\gamma_{\min} \geq 0$ is a pre-specified numerical tolerance. Setting $\gamma_{\min} = 0$ gives the formal set.

The endpoints summarize the range of boundary-consistent estimates. Define

$$\underline{\beta}(c,\kappa) = \inf_{\lambda \in \widehat{\Lambda}(c,\kappa)} \hat{\beta}(\lambda), \qquad \overline{\beta}(c,\kappa) = \sup_{\lambda \in \widehat{\Lambda}(c,\kappa)} \hat{\beta}(\lambda),$$

with the same denominator restriction. Let

$$\hat{\beta}^0 = \hat{\beta}(\hat{\lambda})$$

denote the baseline estimate. The maximum boundary-consistent movement is

$$\widehat{\Delta}_{\max}(c,\kappa) = \max\left\{ |\underline{\beta}(c,\kappa) - \hat{\beta}^0|, |\overline{\beta}(c,\kappa) - \hat{\beta}^0| \right\}.$$

Proposition 1 (Boundary-consistent IV set) *Fix $c > 0$ and $\kappa \geq 0$. For every $\lambda \in \widehat{\Lambda}(c,\kappa)$, the induced boundary*

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

preserves the boundary classification of all non-fragile judges and allows only fragile judges to change side. Therefore,

$$\widehat{\mathcal{B}}(c,\kappa) = \left\{ \hat{\beta}(\lambda) : \lambda \in \widehat{\Lambda}(c,\kappa) \right\}$$

collects the IV estimates generated by AM target populations that are consistent with the sample's non-fragile classifications.

Proof. By definition, $\lambda \in \widehat{\Lambda}(c,\kappa)$ implies

$$\hat{\bar{p}}(\lambda) \in \widehat{\mathcal{I}}(c).$$

By Lemma 1, any boundary in $\widehat{\mathcal{I}}(c)$ preserves the classification of every non-fragile judge. Thus the only judges whose boundary classification can differ from the baseline are those in $\widehat{\mathcal{F}}(c)$. Evaluating $\hat{\beta}(\lambda)$ over all such assignment-share vectors gives the IV estimates generated by boundary-consistent AM target populations. ■

E.4 Connection to response-type weights

The boundary-consistent set has the same target-population logic as the theory in the main text. Let g index response types, with type share π_g, treatment profile $\{D_z(g)\}_{z=1}^J$, and treatment effect τ_g. Under assignment profile λ, the AM weight on type g is

$$\omega_g(\lambda) = \sum_{z=1}^{J} \lambda_z \{p_z - \bar{p}(\lambda)\} D_z(g).$$

The corresponding population IV estimand is

$$\beta(\lambda) = \frac{\sum_g \pi_g \omega_g(\lambda) \tau_g}{\sum_g \pi_g \omega_g(\lambda)}.$$

When fragile judges can change side, some response types can change sign in $\omega_g(\lambda)$. The boundary-consistent set is the empirical analogue of evaluating this estimand across assignment profiles that preserve non-fragile classifications.

The sensitivity set measures target ambiguity rather than sampling uncertainty. A narrow set means that the IV estimate is stable across the boundary-consistent AM target populations allowed by (c, κ). A wide set means that the IV estimate depends on boundary classifications that the sample does not strongly pin down.

E.5 Computation

The endpoint problem is a fractional optimization problem over a simplex with linear restrictions. The feasible set is defined by

$$\lambda_z \geq 0, \qquad \sum_{z=1}^{J} \lambda_z = 1,$$

$$\underline{b}(c) \leq \sum_{z=1}^{J} \lambda_z \hat{p}_z \leq \bar{b}(c),$$

and

$$\|\lambda - \hat{\lambda}\|_1 \leq \kappa.$$

The L_1 restriction can be written with nonnegative variables u_z^+ and u_z^-:

$$\lambda_z - \hat{\lambda}_z = u_z^+ - u_z^-, \qquad u_z^+ \geq 0, \qquad u_z^- \geq 0,$$

with

$$\sum_{z=1}^{J}(u_z^+ + u_z^-) \leq \kappa.$$

The computation should report denominator diagnostics and feasibility diagnostics. These include the baseline denominator, endpoint denominators, the number of feasible solutions or draws, whether the feasible set is empty, and whether any denominator tolerance binds. A grid over the scalar boundary

$$b = \sum_{z=1}^{J} \lambda_z \hat{p}_z$$

can be used as a numerical check by solving the endpoint problem conditional on fixed values of $b \in \widehat{\mathcal{I}}(c)$.

E.6 Recommended output and interpretation

The empirical output should report the fragile set, the plausible boundary interval, the baseline estimate, the lower and upper endpoints, the width, and the maximum movement from the baseline. For each c and κ, the table should include

$$|\widehat{\mathcal{F}}(c)|, \qquad \widehat{\mathcal{I}}(c), \qquad \hat{\beta}^0, \qquad \underline{\beta}(c,\kappa), \qquad \overline{\beta}(c,\kappa), \qquad \widehat{\Delta}_{\max}(c,\kappa).$$

Endpoint denominators should be reported with the same table or in a companion table.

The boundary-consistent IV set should not be described as a confidence interval. It holds fixed the estimated judge propensities and varies assignment shares within a transparent neighborhood. It answers the question: how much can the IV estimate move across AM target populations that the sample cannot rule out?

F Online Appendix: Full Philadelphia Replication Tables

This appendix reports the full Philadelphia diagnostic and boundary-consistent 2SLS tables. The specification is the maximum-control specification described in the empirical section. The excluded instruments are the eight judge indicators. The endogenous variable is pretrial detention. The outcome is conviction.

The Philadelphia replication tables separate boundary proximity from quantitative movement. Table 1 shows that Judges 1, 5, and 7 are close to the AM boundary at $c = 1.96$. Table 2 shows that the sign of the 2SLS estimate remains positive across the reported boundary-consistent sets. Table 3 reports denominator diagnostics for the endpoint estimates.

Table 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 treatment propensity $\hat{p}_z$ is the regression-adjusted judge-specific propensity from the first-stage detention 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.

Table 2: Boundary-consistent 2SLS sets in the Philadelphia bail data

c	κ	Fragile judges	$\widehat{\mathcal{I}}(c)$	$\hat{\beta}^0$	$\underline{\beta}$	$\overline{\beta}$	Width	$\widehat{\Delta}_{\max}$	Feasible draws
1.96	0.10	3	[0.4017, 0.4186]	0.1859	0.1762	0.1992	0.0230	0.0133	200
1.96	0.25	3	[0.4017, 0.4186]	0.1859	0.1549	0.2059	0.0510	0.0310	200
1.96	0.50	3	[0.4017, 0.4186]	0.1859	0.1158	0.2281	0.1123	0.0702	200

Notes: The table reports plug-in boundary-consistent 2SLS sets for the maximum-control specification. The baseline estimate is $\hat{\beta}^0 = 0.1859$. The set $[\underline{\beta}, \overline{\beta}]$ is computed over assignment-share profiles λ satisfying three restrictions: λ lies in the simplex, the implied boundary $\sum_z \lambda_z \hat{p}_z$ lies in the plausible boundary interval $\widehat{\mathcal{I}}(c)$, and $\|\lambda - \hat{\lambda}\|_1 \leq \kappa$. Since assignment shares sum to one, $\kappa/2$ is the maximum share of cases that can be reassigned across judges. The endpoints are approximated by random feasible search with a fixed seed. The current working-paper implementation uses 200 feasible draws for each row.

Table 3: Endpoint diagnostics for the boundary-consistent 2SLS sets

c	κ	Endpoint	$\hat{\beta}(\lambda)$	$\hat{\bar{p}}(\lambda)$	$\|\lambda - \hat{\lambda}\|_1$	First-stage scale	$\hat{\Gamma}_B(\lambda)$
1.96	0.10	Lower	0.1762	0.4105	0.0757	75,746	0.000157
1.96	0.10	Upper	0.1992	0.4130	0.0946	76,559	0.000197
1.96	0.25	Lower	0.1549	0.4087	0.2040	74,947	0.000133
1.96	0.25	Upper	0.2059	0.4133	0.2083	77,182	0.000160
1.96	0.50	Lower	0.1158	0.4087	0.3202	75,456	0.000103
1.96	0.50	Upper	0.2281	0.4136	0.4815	77,338	0.000098

Notes: The table reports diagnostics for the lower and upper endpoint estimates in Table 2. The first-stage scale is the weighted scale of the fitted endogenous regressor used in the 2SLS computation. The boundary denominator is

$$\hat{\Gamma}_B(\lambda) = \sum_z \lambda_z \{\hat{p}_z - \hat{\bar{p}}(\lambda)\}\hat{p}_z.$$

The endpoint denominators do not collapse toward zero, so the reported endpoint estimates are not driven by a vanishing first-stage contrast.

G Online Appendix: Extra Literature Discussion

This appendix positions the paper relative to the judge-IV and heterogeneous-IV literatures. The main text keeps the literature 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. 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 studies whether it delivers a stable target population.

The second related literature studies IV estimands under treatment-effect heterogeneity. **?** show that monotonicity gives a local average treatment effect interpretation in the binary instrument case. Mogstad and Torgovitsky (2024) emphasize that heterogeneous-IV interpretations can be sensitive to instrument structure and that average monotonicity is not purely behavioral. The present paper uses that observation to identify a specific discontinuity: under average monotonicity, the positive-weight response types can 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 exclusion and monotonicity implications in judge designs. 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 observed data. The boundary diagnostic asks a different question: conditional on using average monotonicity, does the estimated propensity profile stably classify the AM target population?

The diagnostic complements standard empirical checks rather than replacing them. Balance tests, first-stage reporting, exclusion discussions, 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 propensity profile.

The Philadelphia illustration should therefore be read as a diagnostic exercise, not as a new test of Stevenson's design. Earlier work finds no decisive evidence against the aggregate Philadelphia judge design using the sharp test. The boundary diagnostic shows that, even in such a setting, some judges can be close to the AM boundary. The empirical contribution is to make target-population stability visible once average monotonicity is invoked.

The paper's main distinction is between nonnegative weights and a stable target population. Average monotonicity solves the negative-weight problem. It does not, by itself, guarantee that the positive-weight population is fixed by individual response behavior alone. That distinction is the reason the boundary result matters for empirical interpretation.

H Online Appendix: Simulations

This appendix describes the simulation exercises for the boundary diagnostic and the boundary-consistent IV set. The diagnostic is a finite-sample classification tool. It asks whether the data place each judge securely on one side of the AM boundary. The boundary-consistent IV set is a target-ambiguity tool. It asks how much the IV estimate can move across AM target populations that remain compatible with the classifications the sample cannot rule out.

The simulations focus on three objects implied by the theory. First, they verify that low values of $\hat{r}_z$ correspond to high rates of sign misclassification. Second, they evaluate whether the plausible boundary interval contains the true baseline boundary in repeated samples. Third, they evaluate whether the estimated boundary-consistent IV set contains the true baseline IV estimand and how wide that set is.

H.1 Simulation design

The data-generating process starts from a finite judge design. There are J judges. Judge z has treatment propensity

$$p_z = E[D \mid Z = z],$$

and assignment share

$$\lambda_z = P(Z = z), \qquad \sum_{z=1}^{J} \lambda_z = 1.$$

The baseline AM boundary is

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

and judge z's population distance from the boundary is

$$d_z^0 = p_z - \bar{p}^0.$$

The treatment equation is generated through latent response types. Each individual belongs to

a response type g, with treatment profile

$$D(g) = \{D_1(g), \ldots, D_J(g)\}, \qquad D_z(g) \in \{0, 1\}.$$

Let π_g denote the share of type g, with $\sum_g \pi_g = 1$. Judge-level propensities satisfy

$$p_z = \sum_g \pi_g D_z(g).$$

In each replication, individuals are drawn from the type distribution, judges are assigned according to λ or fixed to match pre-specified judge counts, and treatment is assigned as

$$D_i = D_{Z_i}(g_i).$$

Potential outcomes are generated at the response-type level. In continuous-outcome designs,

$$Y_i(0) = \alpha_{g_i} + \varepsilon_i, \qquad Y_i(1) = Y_i(0) + \tau_{g_i}.$$

In binary-outcome designs, outcomes are generated from type-specific probabilities

$$q_{0g} = P(Y_i(0) = 1 \mid g_i = g), \qquad q_{1g} = P(Y_i(1) = 1 \mid g_i = g),$$

with $\tau_g = q_{1g} - q_{0g}$. The observed outcome is

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

The baseline population IV estimand is fixed within each simulation design. It is

$$\beta^0 = \frac{\sum_g \pi_g \omega_g(\lambda) \tau_g}{\sum_g \pi_g \omega_g(\lambda)},$$

where

$$\omega_g(\lambda) = \sum_{z=1}^{J} \lambda_z \{p_z - \bar{p}(\lambda)\} D_z(g), \qquad \bar{p}(\lambda) = \sum_{z=1}^{J} \lambda_z p_z.$$

Sampling variation affects $\hat{p}_z$, $\hat{d}_z$, $\hat{r}_z$, and the estimated IV set, not the population estimand.

H.2 Sample objects

Each replication estimates judge-level treatment and outcome means. Define

$$\hat{p}_z = \frac{1}{n_z} \sum_{i:Z_i=z} D_i, \qquad \hat{y}_z = \frac{1}{n_z} \sum_{i:Z_i=z} Y_i, \qquad \hat{\lambda}_z = \frac{n_z}{n}.$$

The estimated boundary and judge-specific distance are

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

The simulation uses the same standardized diagnostic as the empirical application. The diagnostic is

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

In the simulation without covariates, the standard error is computed from the influence-function representation of $\hat{d}_z$. Let

$$\hat{\psi}_{iz} = \frac{\mathbf{1}\{Z_i = z\}}{\hat{\lambda}_z}(D_i - \hat{p}_z) - (D_i - \hat{\bar{p}}).$$

Then

$$\widehat{\mathrm{se}}(\hat{d}_z) = \left[\frac{1}{n}\widehat{\mathrm{Var}}(\hat{\psi}_{iz})\right]^{1/2}.$$

The baseline sample IV estimate is the judge-level covariance ratio. It is

$$\hat{\beta}^0 = \frac{\sum_{z=1}^{J} \hat{\lambda}_z(\hat{p}_z - \hat{\bar{p}})\hat{y}_z}{\sum_{z=1}^{J} \hat{\lambda}_z(\hat{p}_z - \hat{\bar{p}})\hat{p}_z}.$$

H.3 Boundary fragility and boundary-consistent sets

The simulation constructs the same fragile set and plausible boundary interval as the empirical application. For a cutoff $c > 0$,

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

The non-fragile below-boundary and above-boundary sets are

$$\widehat{\mathcal{L}}(c) = \{z : \hat{d}_z < 0,\ z \notin \widehat{\mathcal{F}}(c)\},$$

and

$$\widehat{\mathcal{U}}(c) = \{z : \hat{d}_z > 0,\ z \notin \widehat{\mathcal{F}}(c)\}.$$

The plausible boundary interval is

$$\widehat{\mathcal{I}}(c) = [\underline{b}(c), \overline{b}(c)],$$

where

$$\underline{b}(c) = \max_{z \in \widehat{\mathcal{L}}(c)} \hat{p}_z, \qquad \overline{b}(c) = \min_{z \in \widehat{\mathcal{U}}(c)} \hat{p}_z,$$

with the empty-set conventions used in Appendix E.

The first coverage object asks whether the true baseline boundary lies inside the plausible boundary interval. It is

$$\mathbf{1}\{\bar{p}^0 \in \widehat{\mathcal{I}}(c)\}.$$

This object checks whether the sample-implied set of plausible boundary locations contains the actual boundary from the data-generating process.

The boundary-consistent IV set is computed from feasible assignment-share perturbations. For radius $\kappa \geq 0$, define

$$\widehat{\Lambda}(c, \kappa) = \left\{ \lambda' \in \Delta^{J-1} : \sum_{z=1}^{J} \lambda'_z \hat{p}_z \in \widehat{\mathcal{I}}(c),\ \|\lambda' - \hat{\lambda}\|_1 \leq \kappa \right\}.$$

For each feasible λ', compute

$$\hat{\beta}(\lambda') = \frac{\sum_{z=1}^{J} \lambda'_z \{\hat{p}_z - \hat{\bar{p}}(\lambda')\} \hat{y}_z}{\sum_{z=1}^{J} \lambda'_z \{\hat{p}_z - \hat{\bar{p}}(\lambda')\} \hat{p}_z}.$$

The endpoint estimates are $\underline{\beta}(c, \kappa)$ and $\overline{\beta}(c, \kappa)$, with denominator diagnostics reported separately.

The second coverage object asks whether the true baseline IV estimand lies inside the estimated set. It is

$$\mathbf{1}\left\{\beta^0 \in [\underline{\beta}(c, \kappa), \overline{\beta}(c, \kappa)]\right\}.$$

The width is

$$W(c, \kappa) = \overline{\beta}(c, \kappa) - \underline{\beta}(c, \kappa),$$

and the maximum movement from the baseline estimate is

$$\widehat{\Delta}_{\max}(c, \kappa) = \max\left\{|\underline{\beta}(c, \kappa) - \hat{\beta}^0|, |\overline{\beta}(c, \kappa) - \hat{\beta}^0|\right\}.$$

H.4 Simulation outcomes

The simulations report judge-level sign classification. For each judge, the output includes

$$\Pr\left(\operatorname{sign}(\hat{d}_z) \neq \operatorname{sign}(d_z^0)\right),$$

and the conditional sign-misclassification rate

$$\Pr\left(\operatorname{sign}(\hat{d}_z) \neq \operatorname{sign}(d_z^0) \mid \hat{r}_z \leq c\right).$$

The diagnostic performs its intended role when sign errors are concentrated among judges with low $\hat{r}_z$.

The simulations report design-level fragility. The output includes

$$\hat{r}_{\min} = \min_z \hat{r}_z, \qquad |\widehat{\mathcal{F}}(c)|,$$

and

$$\mathbf{1}\left\{\exists z : \operatorname{sign}(\hat{d}_z) \neq \operatorname{sign}(d_z^0)\right\}.$$

These objects show how judge-level boundary uncertainty aggregates across the design.

The simulations report boundary-interval coverage and IV-set coverage. The output includes

$$\Pr(\bar{p}^0 \in \widehat{\mathcal{I}}(c)),$$

$$\Pr\left(\beta^0 \in [\underline{\beta}(c,\kappa), \overline{\beta}(c,\kappa)]\right),$$

$$E[W(c,\kappa)], \qquad E[\widehat{\Delta}_{\max}(c,\kappa)].$$

These objects measure how boundary classification uncertainty translates into the scale of the IV estimate.

H.5 Calibration of boundary distance

Some simulations vary the true distance of a judge from the AM boundary. The population distance is

$$d_z^0 = p_z - \bar{p}^0.$$

There is no standard error attached to this population object. To calibrate the difficulty of the classification problem, the simulation uses the design signal-to-noise ratio

$$\rho_z = \frac{|d_z^0|}{\mathrm{sd}(\hat{d}_z)}.$$

This object is a feature of the sampling design. It says how large the true boundary distance is relative to the sampling variability of its estimator.

The local diagnostic theory maps this ratio into sign classification. In local designs,

$$\Pr\left(\mathrm{sign}(\hat{d}_z) \neq \mathrm{sign}(d_z^0)\right) \approx \Phi(-\rho_z).$$

Designs with small ρ_z should generate frequent sign errors; designs with large ρ_z should not. The empirical diagnostic remains $\hat{r}_z$.

H.6 Stylized and calibrated designs

The stylized simulations isolate the boundary mechanism. They use small-J designs in which one or more judges can be placed close to the AM boundary. The simulations consider both designs satisfying strong monotonicity and designs satisfying average monotonicity without strong monotonicity. In the latter designs, non-monotone response types are allowed, but their shares are chosen so that AM weights remain nonnegative in the population.

The calibrated simulations use empirical features of common judge and examiner IV settings in Maestas et al. (2013) and Dobbie et al. (2018). They vary J, the number of cases per judge, the mean treatment rate, the dispersion of judge propensities, and the scale of treatment effects.

The calibrated designs are useful for many-judge settings. With many judges, the probability that at least one judge lies close to the boundary can be large even when no single judge is placed near the boundary by construction. The simulation therefore reports both judge-level and design-level diagnostics:

$$E[\hat{r}_z], \qquad E[\hat{r}_{\min}], \qquad E[|\widehat{\mathcal{F}}(c)|], \qquad \Pr(\text{any sign error}).$$

H.7 Interpretation of the simulation

The simulations support the logic of the diagnostic rather than a new estimator. The population boundary distance d_z^0 is fixed within a design. Sampling uncertainty enters through $\hat{d}_z$. The statistic $\hat{r}_z$ measures how strongly the sample classifies the sign of d_z^0.

The boundary-consistent IV set translates classification uncertainty into the scale of the IV

estimate. If

$$[\underline{\beta}(c,\kappa), \overline{\beta}(c,\kappa)]$$

is narrow, the IV conclusion is stable across AM target populations that the sample cannot rule out. If it is wide, the conclusion depends on boundary classifications that are weakly pinned down.

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.

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.

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.