How Random Is Random Enough?

Bayesian Design-Based Inference for Leniency Designs with Uncertain Assignment

Muyuan Yang*

School of Economics, Beijing Wuzi University

Email: molanchars@gmail.com

Abstract

Judge and examiner ("leniency") designs rest on a binary premise: that assignment of cases to decision-makers is *quasi-random*. In practice the premise holds approximately, and applied work disciplines it with a balance test that returns a single yes/no verdict. We recast the premise as a continuous, uncertain object and ask what a researcher can actually learn about the magnitude of the departure from random assignment. Our theoretical core is an identification boundary (Proposition 1): the systematic response of the balance statistic to the departure parameter γ_{dep} is *second order* (with its first-order response vanishing in design expectation, $\mathbb{E}_\omega[\partial\mathbb{E}[b]/\partial\gamma_{\text{dep}}|_0] = 0$), and it provides non-zero systematic signal *if and only if* three conditions hold simultaneously; even then only the *composite* $\theta = \gamma_{\text{dep}} \cdot \text{Corr}(\varepsilon, X)$ is identified, not γ_{dep} itself. For the second-order response we derive a *closed-form* design constant and verify it against the Monte Carlo coefficient; the same constant yields an analytic resolution floor, $\gamma_{\text{dep,floor}} \propto G^{-1/4}$. A Monte Carlo design confirms the boundary and uncovers two failure modes of the standard workflow—a coverage collapse from linear extrapolation of the balance statistic, and an "observable leakage" that makes a purely observable departure masquerade as an unobservable one. We propose a three-layer Bayesian design-based framework that (i) models the observable departure hierarchically, (ii) anchors the prior on the unobservable departure to the observable evidence rather than to researcher discretion, and (iii) reports the causal estimate as a *curve* in the anchoring strength. We validate the framework on the design structure of a real judge design,

*Corresponding author: School of Economics, Beijing Wuzi University, No. 321 Fuhe Street, Tongzhou District, Beijing 101149, P.R. China. Email: 25311020200023@bwu.edu.cn.

and illustrate it on U.S. patent examiners and Miami bail judges. We are explicit about what remains unidentifiable: when the three conditions fail, no amount of data recovers γ_{dep}, and the honest output is a sensitivity frontier, not a point estimate.

Keywords: Leniency design, Judge–examiner IV, Design-based inference, Bayesian sensitivity analysis, Partial identification, Balance test.

JEL Classification: C11, C18, C26, C52.

1 Introduction

Judge and examiner designs—often termed "leniency" designs—have become central to the empirical toolkit of applied microeconomics. In these settings, cases are assigned among G decision-makers, such as judges, patent examiners, or physicians, who exhibit persistent differences in leniency. The resulting cross-reviewer variation in treatment assignment D serves as an instrumental variable for estimating causal effects on downstream outcomes Y (Kolesár, 2013; Chyn et al., 2025; Goldsmith-Pinkham et al., 2026). The credibility of this empirical strategy hinges on quasi-random assignment: formally, that reviewer assignment A_i is statistically independent of potential outcomes and pre-determined covariates, $A_i \perp (X_i, Y_i(0), Y_i(1))$.

Applied researchers routinely evaluate this premise through balance regressions, projecting pre-determined covariates on leave-one-out leniency instruments Z_i and testing whether the resulting coefficients are jointly zero. In practice, this procedure is treated as a binary screening device: passing the test licenses the causal interpretation, while failing it casts doubt on the design. Yet in observational quasi-experiments, assignment is rarely immaculate. A binary verdict provides neither an estimate of the magnitude of potential sorting nor an assessment of its downstream impact on causal estimates.

In this paper, we replace the binary verdict with a continuous framework for evaluating departures from quasi-random assignment. We distinguish between observable departures, where assignment correlates with recorded case characteristics X, and unobservable departures, where assignment correlates with unobserved determinants of potential outcomes. We then ask what balance statistics can reveal about each form of sorting.

Our primary theoretical result (Proposition 1) characterizes the identification boundary for unobservable sorting γ_{dep}. We show that the leading systematic response of standard balance statistics to γ_{dep} is strictly second-order: the first-order response vanishes in expectation over reviewer sensitivity ($\mathbb{E}_\omega[\partial\mathbb{E}[b]/\partial\gamma_{\text{dep}}|_0] = 0$) and constitutes a mean-zero $O_p(G^{-1/2})$ design fluctuation for any realized finite pool, so that the leading non-zero systematic signal is proportional to $\gamma_{\text{dep}}^2 \cdot \text{Corr}(\varepsilon, X) \cdot \rho$. Consequently, the balance statistic carries non-zero asymptotic signal if and only if three conditions hold simultaneously: (C1) the unobserved outcome error correlates with observed covariates, $\text{Corr}(\varepsilon, X) \neq 0$; (C2) the outcome error correlates with the treatment selection error, $\rho \equiv \text{Corr}(\varepsilon, u) \neq 0$; and (C3) decision-makers exhibit dispersion in their sensitivity to unobservable case attributes, $\text{Var}_g(\omega_g) > 0$. If any of these conditions fails—most notably when treatment assignment is conditionally exogenous ($\rho = 0$)—balance tests contain zero information regarding unobservable sorting, regardless of sample size.

The vanishing first-order derivative has immediate consequences for empirical practice. Intuitively, one might expect sorting on an unobserved confounder ε correlated with X to shift the mean covariate profile across reviewers with a non-zero local slope. However, under standard multinomial choice mechanisms, first-order sorting shifts sum to zero in expectation across the reviewer pool at quasi-random assignment (generating only mean-zero $O_p(G^{-1/2})$ finite-pool fluctuations), leaving pure curvature ($\propto \gamma_{\text{dep}}^2$) as the only systematic signal. This creates an information dead zone near random assignment where balance statistics have zero systematic first-order power, directly generating an unusually slow $G^{-1/4}$ resolution floor (Section 4). Furthermore, relying on linear calibrations breaks inferential coverage: extrapolating a single secant slope causes nominal 90% credible intervals to collapse to between 7.5% and 17%, while understating the departure required to induce a given bias by 50% to 75% (Section 5.3).

Even when all three identification conditions are met, a fundamental identification limit persists. Proposition 2 demonstrates that observational data identify only the product $\theta = \gamma_{\text{dep}} \cdot \text{Corr}(\varepsilon, X)$ under the normalization $\text{Var}(\omega_g) = 1$. Disentangling the sorting parameter γ_{dep} from the structural correlation $\text{Corr}(\varepsilon, X)$ is impossible without external restrictions. Sensitivity analysis in leniency designs must therefore accommodate partial identification rather than claiming point identification.

We derive a closed-form expression for the second-order coefficient in terms of design primitives: $c = \sigma_\omega^2 S / \mathbb{E}[\text{Var}_0(Z)]$, where S aggregates reviewer caseload shares and treatment threshold densities (Appendix A). This constant is computable from data generated under quasi-random assignment and directly governs the design's resolution floor, $\gamma_{\text{dep,floor}} \propto G^{-1/4}$ (Appendix B). Finite-sample simulations confirm that standard linear approximations underestimate this resolution floor by roughly a factor of seven, obscuring the limits of what a given design can detect.

These theoretical insights highlight two common failure modes in applied sensitivity work. First, because the true balance response is nonlinear, single-point linear calibrations generate overconfident and miscentered posterior intervals. Second, when reviewers possess preferences over observable covariates, standard Bayesian models attribute the resulting covariate imbalance entirely to unobservable sorting γ_{dep}. This observable leakage causes researchers to overstate assignment violations even when unobservable sorting is absent.

To resolve both failure modes, we propose an integrated three-layer Bayesian framework. Layer 1 models reviewer preferences over observables hierarchically, applying partial pooling across decision-makers to absorb genuine observable sorting into its own parameter block. This acts as an identification firewall, preventing observable imbalance from distorting the unobservable departure parameter. Layer 2 grounds prior elicitation by anchoring the prior

scale of unobservable departures to the empirical dispersion of observable reviewer preferences estimated in Layer 1. Finally, Layer 3 traces a continuous sensitivity frontier, reporting bias-corrected causal estimates as a function of the anchoring strength. This replaces heuristic rules of thumb with a formal, posterior decision boundary.

We calibrate our simulations to the Miami bail design of Dobbie et al. (2018), matching the number of decision-makers ($G = 146$) and leniency dispersion ($\sigma_\ell \approx 0.044$). We then examine the empirical performance of the framework across an institutional taxonomy of assignment risk, including federal criminal sentencing in the D.C. District Court and USPTO patent examiners (Farre-Mensa et al., 2020; Lemley and Sampat, 2012). In centralized, computerized allocation systems, the framework confirms that balance test rejections stem from finite-sample noise, issuing a formal Bayesian certificate of exogeneity. In contrast, in settings with unabsorbed technical specialization or forum shopping, the framework actively detects large biases and identifies causal sign reversals.

This paper connects with several distinct literatures in econometrics. Methodologically, an established Bayesian literature addresses uncertain instrumental variable validity by modeling departures from structural exclusion restrictions or selecting among candidate instruments (e.g., Giacomini et al., 2022; Steiner and Steel, 2026). Our focus is structurally distinct: we model uncertainty in the physical assignment mechanism γ_{dep} itself. In leniency designs, instrument invalidity arises from non-random case allocation across reviewers rather than direct structural effects of reviewer identity on potential outcomes. In this respect, our approach extends the design-based Bayesian framework of Rubin (1978), Imbens and Rubin (2015), and recent advances in randomization-based Bayesian inference (Huch et al., 2025; Leavitt, 2023) to observational settings where the assignment mechanism must be learned from data.

Our framework also complements empirical Bayes shrinkage estimators such as SJIVE (Frandsen et al., 2026). While SJIVE takes quasi-random assignment as given and shrinks instruments to minimize downstream estimation variance, our framework evaluates the upstream identification question of whether and to what extent assignment is quasi-random. In the sensitivity literature, our Layer 2 anchoring strategy connects directly with the seminal tradition of bounding unobservable selection using observable selection (Altonji et al., 2005; Oster, 2019), as well as recent empirical Bayes approaches in difference-in-differences (Kwon and Roth, 2024). We import this anchoring philosophy into cross-sectional reviewer designs. However, unlike linear selection models where unobservables transmit proportionally to treatment effects, case allocation in leniency designs is governed by a multinomial choice game where observable balance signals transmit through non-linear, second-order choice curvature (Proposition 1), generating an information dead zone near random assignment with

a slow $G^{-1/4}$ resolution floor. Finally, our framework formalizes the intuitive diagnostics discussed in practitioner manuals (Goldsmith-Pinkham et al., 2026; Chyn et al., 2025) and complements frequentist joint tests (Coulibaly et al., 2024) by delivering continuous posterior distributions, noise attributions, and bias-corrected causal frontiers.

The remainder of the paper is organized as follows. Section 2 establishes notation. Section 3 introduces the assignment model and the decomposition of departures. Section 4 presents the identification boundary, closed-form constant, and resolution floor. Section 5 examines the boundary in Monte Carlo simulations, documenting the two failure modes and their fixes. Section 6 details the three-layer Bayesian framework. Section 7 presents the empirical calibration, and Section 8 reports the empirical applications. Section 9 concludes with open problems.

2 Notation

Table 1 collects the symbols used throughout. Two distinctions in it carry most of the paper's argument and are worth flagging at the outset: the *unobservable* departure γ_{dep} is not the *identifiable* composite θ, and neither is the error correlation ρ; and the resolution floor $\gamma_{\text{dep,floor}}$ is a property of the *design* (caseload, number of decision-makers), not of the data.

3 Framework

3.1 Setup

There are N cases $i = 1, \ldots, N$ and G decision-makers $g = 1, \ldots, G$. Case i is assigned to decision-maker $A_i \in \{1, \ldots, G\}$, receives a binary treatment $D_i \in \{0, 1\}$, and has outcome Y_i. Potential outcomes are $Y_i(0), Y_i(1)$ with $Y_i = Y_i(0) + \beta_i D_i$. Each decision-maker has a leniency ℓ_g. The instrument is the leave-one-out group mean of the treatment,

$$Z_i \;=\; \hat{\ell}_{A_i}^{-i} \;=\; \frac{\sum_{j:\, A_j = A_i,\, j \neq i} D_j}{\#\{j : A_j = A_i,\; j \neq i\}}, \tag{1}$$

and $X_i \in \mathbb{R}^p$ collects pre-determined covariates. We maintain the usual exclusion and monotonicity conditions under which Z_i identifies a local average treatment effect for the standard leniency design (Angrist et al., 1996); our focus is on the assignment mechanism, not the treatment-effect estimand.

3.2 Decomposing the departure from quasi-random assignment

Quasi-random assignment (QRA) is the statement $A_i \perp (X_i, Y_i(0), Y_i(1))$. We model the assignment mechanism directly. Let ε_i be the unobserved determinant of the outcome and ω_g a decision-maker-specific *response* to unobservable case characteristics. Assignment follows

$$\mathbb{P}(A_i = g \mid X_i, \varepsilon_i, \omega) \;\propto\; \exp\Bigg(\underbrace{\alpha_g}_{\text{assignment propensity}} + \underbrace{X_i' \beta_g}_{\text{observable departure}} + \underbrace{\gamma_{\text{dep}}\, \varepsilon_i\, \omega_g}_{\text{unobservable departure}} \Bigg), \tag{2}$$

with $\omega_g \overset{iid}{\sim} N(0,1)$. The two departures are conceptually orthogonal: $\{\beta_g\}$ governs how assignment correlates with *observed* case characteristics, while γ_{dep} governs how it correlates with the *unobserved* determinants of the outcome. We normalize $\text{Var}(\omega_g) = 1$ without loss of generality: any scale $\sigma_\omega > 0$ can be absorbed into γ_{dep} (the identified object is the product $\theta = \gamma_{\text{dep}} \cdot \text{Corr}(\varepsilon, X) \cdot \sigma_\omega$; Proposition 2 shows that the factors are not separately identified).

Remark 1 (Why ω_g is needed). A specification with $\gamma_{\text{dep}}\varepsilon_i$ alone—without the multiplicative ω_g—is a no-op: inside the softmax in (2), $\gamma_{\text{dep}}\varepsilon_i$ is a *row constant* shared by all g, so it leaves the assignment probabilities unchanged. Departure from randomness along ε requires that decision-makers *differ* in how they respond to ε, i.e. $\text{Var}_g(\omega_g) > 0$. With the normalization $\text{Var}(\omega_g) = 1$, condition (C3) becomes simply $\sigma_\omega > 0 \iff \omega_g \not\equiv \text{const}$. This is not a modelling nicety; it is the source of the third identification condition below, and we verified it numerically during implementation (a specification with $\gamma_{\text{dep}}\varepsilon_i$ alone produced byte-identical

assignments across $\gamma_{\text{dep}} = 0$ and $\gamma_{\text{dep}} = 1$).

3.3 Bias transmission and the linear approximation

The instrument is contaminated by the unobservable departure. Write the probability limit of the IV estimator as

$$\operatorname{plim} \hat{\beta}_{IV} \;=\; \beta + \frac{\operatorname{Cov}(Z_i, \varepsilon_i)}{\operatorname{Cov}(Z_i, D_i)} \;=\; \beta + b, \qquad b \approx \gamma_{\text{dep}} \cdot \kappa, \tag{3}$$

where $\kappa = \kappa(G/N, \operatorname{Var}(\ell), \text{caseload distribution})$ is a design constant. The linear structure in (3) is the technical premise that makes the problem analytically tractable, and it is the reason the first simulation experiment can be decisive: the bias is, to first order, proportional to γ_{dep}. We will show in Section 5 that this linear approximation is nevertheless *not* adequate for inference—the curvature of $b(\gamma_{\text{dep}})$ is large enough to break interval coverage—which is one of the two failure modes we identify.

3.4 The balance statistic

For each covariate j we run the balance regression $X_{ij} = a_j + b_j Z_i + u_{ij}$ and collect $b = (b_1, \ldots, b_p)'$ with standard errors $\{se_j\}$. Under QRA, $\mathbb{E}[b] = 0$; the balance test asks whether b is jointly zero. Our object of interest is how the balance statistic responds to the departure magnitude near QRA. In design expectation over the reviewer pool distribution (and asymptotically as $G \to \infty$), its *first* derivative vanishes,

$$\mathbb{E}_\omega \left[\left. \frac{\partial\, \mathbb{E}[b]}{\partial \gamma_{\text{dep}}} \right|_{\gamma_{\text{dep}}=0} \right] = 0 \qquad (\text{under QRA and } \omega \perp \ell), \tag{4}$$

while for any realized finite reviewer pool of size G, the first-order derivative is an empirical covariance $\operatorname{Corr}(\varepsilon, X) \sum_g s_g(\omega_g - \bar{\omega}) p_g(0)$ constituting mean-zero $O_p(G^{-1/2})$ design noise. Consequently, the leading *systematic* response is *second order*:

$$\mathbb{E}_\omega\Big[\mathbb{E}[b(\gamma_{\text{dep}})]\Big] \;=\; c\,\gamma_{\text{dep}}^2 \,\operatorname{Corr}(\varepsilon, X)\,\rho \;+\; O(\gamma_{\text{dep}}^3), \qquad c > 0, \tag{5}$$

where $\rho \equiv \operatorname{Corr}(\varepsilon, u)$ and c is a design constant. This is the sense in which balance information is *local*: it is a second-order, not a first-order, signal. If the second-order coefficient vanishes, the balance statistic carries no local information about γ_{dep}; the posterior on γ_{dep} equals its prior, and no sample size repairs the situation.

4 Identification

4.1 The boundary: Proposition 1

Proposition 1 (Identification boundary)**.** *Under the model of Section 3, the systematic response of the balance statistic to the departure magnitude is* second order *in* γ_{dep}*:*

$$\mathbb{E}_\omega\Big[\mathbb{E}[b(\gamma_{\text{dep}})]\Big] = c\,\gamma_{\text{dep}}^2\ \text{Corr}(\varepsilon, X)\,\rho + O(\gamma_{\text{dep}}^3), \qquad c > 0,$$

while its first-order response vanishes in design expectation ($\mathbb{E}_\omega[\partial\mathbb{E}[b]/\partial\gamma_{\text{dep}}|_0] = 0$*) and is a mean-zero* $O_p(G^{-1/2})$ *design perturbation for any realized finite reviewer pool. Consequently,* γ_{dep} *provides systematic signal through the balance statistic if and only if all three of the following hold simultaneously:*

(C1) $\text{Corr}(\varepsilon, X) \neq 0$ *— the unobserved outcome error correlates with the covariates;*

(C2) $\rho \equiv \text{Corr}(\varepsilon, u) \neq 0$ *— the outcome error correlates with the treatment-assignment error;*

(C3) $\text{Var}_g(\omega_g) > 0$ *— decision-makers differ in their response to unobservable case characteristics.*

Moreover the identified object is the composite $\theta = \gamma_{\text{dep}} \cdot \text{Corr}(\varepsilon, X)$ *under the normalization* $\text{Var}(\omega_g) = 1$ *(Proposition 2); without normalization,* $\theta = \gamma_{\text{dep}} \cdot \text{Corr}(\varepsilon, X)$*.*

The "only if" direction of Proposition 1 is the practically important one, and condition (C2) is the one that applied practice is least likely to appreciate. It says that even when the unobserved outcome error is strongly correlated with covariates—the situation in which one would most expect the balance test to be informative—the balance statistic carries *no* information about the unobservable departure if the outcome error and the treatment-assignment error are uncorrelated.

Two features of Proposition 1 deserve emphasis. First, the systematic response is fundamentally *second order*: the balance statistic's local sensitivity to the departure reflects quadratic curvature rather than a linear slope. The vanishing systematic first-order derivative, $\mathbb{E}_\omega[\partial\mathbb{E}[b]/\partial\gamma_{\text{dep}}|_0] = 0$, is both mathematically fundamental and counter-intuitive. For any realized finite reviewer pool, $\partial\mathbb{E}[b]/\partial\gamma_{\text{dep}}|_0 = \text{Corr}(\varepsilon, X)\sum_g s_g(\omega_g - \bar{\omega})p_g(0)$ represents an empirical covariance between independent reviewer characteristics, which fluctuates around zero at rate $O_p(G^{-1/2})$. Once averaged across the reviewer design distribution (or in large reviewer pools as $G \to \infty$), this linear term vanishes in expectation, leaving pure quadratic curvature as the only systematic signal.

The necessity of condition (C2), $\rho \equiv \text{Corr}(\varepsilon, u) \neq 0$, illuminates a foundational structural property of leniency designs that resolves an apparent paradox regarding balance tests. Under the standard instrumental variable exclusion restriction, reviewer leniency α_{A_i} affects potential outcomes Y_i strictly through the treatment decision $D_i = \mathbf{1}\{\alpha_{A_i} + u_i \geq 0\}$. When $\rho = 0$, the unobserved outcome shock ε_i is statistically independent of the treatment selection error u_i. In this setting, even if cases sort non-randomly across reviewers along ε_i, reviewer leniency shifts treatment decisions strictly through u_i. Because $u_i \perp \varepsilon_i$, assignment sorting does not interact with the treatment margin to contaminate the local average treatment effect: the IV bias curve satisfies $b_{IV}(\gamma_{\text{dep}}) = 0$ identically when $\rho = 0$.

Consequently, the fact that balance statistics carry zero signal when $\rho = 0$ is not a defect where the test fails to detect an active threat. Rather, when $\rho = 0$, unobservable sorting along ε_i is causally benign in the IV model. Balance statistics possess non-zero sensitivity to γ_{dep} if and only if $\rho \neq 0$—which is precisely the regime in which unobservable sorting transmits through the treatment selection margin to induce asymptotic bias in two-stage least squares. In a conventional linear model, non-zero correlation between an unobserved confounder ε and observable covariates X implies that sorting cases along ε immediately shifts the expected balance coefficient with non-zero slope. Lemma 1 below reveals why this fails in leniency designs: under multinomial logit assignment, the first-order shift in assignment probabilities is linear in $m(X_i) = \mathbb{E}[\varepsilon_i \mid X_i]$. However, because baseline reviewer propensities α_g are identical under symmetric quasi-random assignment, this first-order shift has expectation zero across the reviewer pool ($\mathbb{E}_\omega[\sum_g s_g(\omega_g - \bar{\omega})p_g(0)] = 0$), generating only $O_p(G^{-1/2})$ finite-sample noise. A non-zero systematic shift in the leave-one-out instrument Z_i emerges only at second order, where the sorting probability interacts with the reviewer's treatment decision D_i.

This vanishing derivative has three direct practical consequences that overturn prevailing empirical habits: (i) It invalidates standard linear calibration workflows that extrapolate a constant secant slope $\kappa_0 = [b(1) - b(0)]/1$, explaining why linear extrapolation causes credible interval coverage to collapse to 7.5%–17% (Section 5.3). (ii) It exposes an "information dead zone" near random assignment: because the derivative is zero at $\gamma_{\text{dep}} = 0$, balance statistics possess vanishing signal-to-noise ratios against small departures, which analytically generates the slow $G^{-1/4}$ resolution floor (Appendix B). (iii) It explains why coarse linear interpolation generates substantial bias artifacts by mistaking a convex quadratic response for a straight line (Section 7).

Second, the three conditions enter as *separable factors* of the second-order coefficient, which is what makes the boundary interpretable: (C1) and (C2) can be read off the data-generating structure, while (C3) is a property of the decision-maker pool.

4.2 Partial identification: only the composite is learnable

Proposition 2 (Partial identification)**.** *With the normalization* $\mathrm{Var}(\omega_g) = 1$*, the model* $(\gamma_{\mathrm{dep}}, \mathrm{Corr}(\varepsilon, X))$ *and the model* $(\gamma'_{\mathrm{dep}}, \mathrm{Corr}'(\varepsilon, X))$ *are observationally equivalent in the distribution of* (X, A) *if and only if*

$$\gamma_{\mathrm{dep}} \cdot \mathrm{Corr}(\varepsilon, X) \;=\; \gamma'_{\mathrm{dep}} \cdot \mathrm{Corr}'(\varepsilon, X). \tag{6}$$

Hence the unique identified object is the composite $\theta \equiv \gamma_{\mathrm{dep}} \cdot \mathrm{Corr}(\varepsilon, X)$*;* γ_{dep} *itself is identified only up to external knowledge of* $\mathrm{Corr}(\varepsilon, X)$*. Without the normalization, the identified object is* $\theta = \gamma_{\mathrm{dep}} \cdot \mathrm{Corr}(\varepsilon, X) \cdot \sigma_\omega$*, and the three factors are not separately identified.*

The partial identification of Proposition 2 is *global*: it holds even with the full data (X, A, D, Y, Z), not just the assignment mechanism. To see this, note that the IV estimand satisfies

$$\hat{\beta}_{IV}(\gamma_{\mathrm{dep}}) - \beta \;=\; \frac{\mathrm{Cov}(\varepsilon, Z \mid \gamma_{\mathrm{dep}})}{\mathrm{Cov}(D, Z \mid \gamma_{\mathrm{dep}})}.$$

The numerator $\mathrm{Cov}(\varepsilon, Z)$ is structurally isomorphic to the balance statistic numerator (replace X_1 by ε): it is also second order in γ_{dep} with the same three channels (C1)–(C3) and the same parameter dependence $\propto \gamma^2_{\mathrm{dep}} \mathrm{Corr}(\varepsilon, X)\rho$. The denominator $\mathrm{Cov}(D, Z)$ is first-stage strong at $\gamma_{\mathrm{dep}} = 0$ (driven by the leniency dispersion), so the bias is $O(\gamma^2_{\mathrm{dep}})$ with a coefficient that depends on the same composite $\theta = \gamma_{\mathrm{dep}} \cdot \mathrm{Corr}(\varepsilon, X)$ (with $\mathrm{Var}(\omega_g) = 1$). The ratio of IV bias to balance statistic is therefore a function of design quantities alone, independent of θ, and provides a consistency check but no additional identifying power.

Proposition 3 (Full-data non-identification)**.** *Under the model of Section 3 with* $\mathrm{Var}(\omega_g) = 1$*, the distribution of the full data* (X, A, D, Y, Z) *does not identify* $(\gamma_{\mathrm{dep}}, \mathrm{Corr}(\varepsilon, X))$ *separately: any pair with the same* $\theta = \gamma_{\mathrm{dep}} \cdot \mathrm{Corr}(\varepsilon, X)$ *generates the same first two moments of* (X, A, D, Y, Z) *to second order in* γ_{dep}*.*

Proposition 2 has a direct methodological consequence that we emphasize throughout: *any point statement about* γ_{dep} *must be explicitly conditioned on an assumption about* $\mathrm{Corr}(\varepsilon, X)$. This is exactly the object that Layer 2 of our framework (Section 6) anchors with data, and it is why we treat the anchoring assumption as a first-class part of the analysis rather than a robustness afterthought.

4.3 Lemma 1: the first-order response

Lemma 1. *Under* (2),

$$\frac{\partial}{\partial \gamma_{\text{dep}}} \mathbb{P}(A_i = g \mid X_i = x)\Big|_{\gamma_{\text{dep}}=0} = \rho(x)\, \pi_g(x) \Big(\omega_g - \bar{\omega}(x)\Big), \tag{7}$$

where $\rho(x) \equiv \mathbb{E}[\varepsilon_i \mid X_i = x]$, $\pi_g(x)$ *is the softmax at* $\gamma_{\text{dep}} = 0$, *and* $\bar{\omega}(x) \equiv \sum_h \pi_h(x)\omega_h$.

Proof. Write $z_h(\gamma_{\text{dep}}) = \alpha_h + x'\beta_h + \gamma_{\text{dep}}\varepsilon\omega_h$ and $\Pi_g(\gamma_{\text{dep}}, \varepsilon) = \text{softmax}_g(z(\gamma_{\text{dep}}))$. Then $\partial \Pi_g / \partial \gamma_{\text{dep}} = \varepsilon\, \Pi_g(\omega_g - \bar{\omega}_\Pi)$ with $\bar{\omega}_\Pi \equiv \sum_h \Pi_h \omega_h$. At $\gamma_{\text{dep}} = 0$, $\Pi_g = \pi_g$, so $\bar{\omega}_\Pi = \bar{\omega}(x)$. Taking $\mathbb{E}[\cdot \mid X = x]$ and using $\varepsilon \perp \omega$ gives the result. □

Two corollaries follow. First, if b depends only on (X, A), then $m \equiv 0$ implies $\partial \mathbb{E}[b] / \partial \gamma_{\text{dep}}|_0 = 0$. Second, for a realized reviewer pool, the first-order response of the balance numerator is $\text{Corr}(\varepsilon, X) \sum_g s_g(\omega_g - \bar{\omega}) p_g(0)$, which is a mean-zero $O_p(G^{-1/2})$ design perturbation that vanishes in expectation over the reviewer draw ($\mathbb{E}_\omega[\cdot] = 0$). Third, the first-order response of the *assignment probability* is proportional to $m(x)$, i.e. to $\text{Corr}(\varepsilon, X)$—the content of condition (C1).

4.4 The identification channel: assignment reshuffling, not preference

It is tempting to describe the balance statistic as detecting "decision-maker preferences that tilt covariates." That description is misleading here, and a direct diagnostic shows why. The instrument Z_i is the leave-one-out mean of D within a group, and $D_j = \mathbf{1}\{\ell_g + u_j > c_0\}$ does *not* depend on γ_{dep}: neither ℓ_g nor u_j is a function of the departure magnitude. Holding the assignment fixed and letting only D respond to γ_{dep} therefore leaves $\text{Cov}(X_{i1}, Z_i)$ essentially unchanged (measured response ≈ 0). The response comes entirely from the *assignment*: γ_{dep} changes *who is grouped with whom*, and hence the within-group composition of X_1 and the group treatment rates p_g.

The channel is thus a *crossover*. The departure shifts the within-group X_1 composition (a response proportional to $\text{Corr}(\varepsilon, X)$, condition (C1)) and, through $u = \rho\varepsilon + \ldots$, the within-group u-composition and hence p_g (a response proportional to ρ, condition (C2)). The balance statistic responds to the *product* of the two, which is why the leading term is second order in γ_{dep} and why (C1) and (C2) enter as separable factors. A "peer-composition" channel—peer unobservables reaching Z_i through peers' realised treatments—plays no role; that account is refuted by the fixed-assignment diagnostic above.

Remark 2 (Consequence for exposition). The channel must be described as assignment reshuffling feeding a composition-by-rate crossover. Framing the result as "examiner preferences directly unbalance covariates," or as a peer-composition channel running through peers' treatments, would contradict the mechanism and would be flagged in review.

4.5 Numerical verification

To verify the vanishing first-order derivative numerically, we examine $\mathbb{E}[b]$ at small values of γ_{dep} using paired-seed replications to resolve the underlying signal. Table 3 reports the estimated response based on 400 replications per grid point. At $\gamma_{\text{dep}} \in \{0.01, 0.02\}$, the estimated balance coefficient is statistically indistinguishable from zero, while the ratio $\mathbb{E}[b]/\gamma_{\text{dep}}$ increases approximately linearly with γ_{dep}—the empirical hallmark of a quadratic response with a vanishing linear term.[1]

We next evaluate the necessity of conditions (C1)–(C3) in Table 4, using paired-seed differences at $\gamma_{\text{dep}} = 0.2$ (where the reported secant is proportional to the second-order coefficient). The balance response vanishes when either $\text{Corr}(\varepsilon, X) = 0$ or $\rho = 0$, and requires reviewer heterogeneity $\text{Var}(\omega) > 0$. Furthermore, only X_1, the covariate correlated with the unobserved error ε, exhibits an imbalance. In line with the theoretical prediction $c\ \text{Corr}(\varepsilon, X)\rho$, the estimated secant scales linearly in both correlation parameters.

The analytic constant multiplying the second-order response has a closed form (Appendix A):

$$c \;=\; \frac{S}{\mathbb{E}[\text{Var}_0(Z)]}, \qquad S \;=\; \sum_g s_g\, J_g \Big(1 - 2s_g + \textstyle\sum_h s_h^2\Big), \qquad J_g \equiv \phi(c_0 - \ell_g), \tag{8}$$

with $s_g = n_g/N$ the reviewer share. With the normalization $\text{Var}(\omega_g) = 1$, the σ_ω^2 factor is absorbed; without it, $c = \sigma_\omega^2 S/\mathbb{E}[\text{Var}_0(Z)]$. Two facts are worth pausing on. First, the coefficient is *built entirely from* γ_{dep}=0 *design quantities*—shares, leniencies, and the treatment threshold—and requires no information from a $\gamma_{\text{dep}} > 0$ experiment. Second, its positivity, $c > 0$, follows from $\mathbb{E}[(\omega_g - \bar{\omega})^2] > 0$ under (C3); the sign is robust. The constant is numerically verified against the Monte Carlo second-order coefficient (Section 5, Table 3).

[1] A naive finite-difference quotient $\mathbb{E}[b(\varepsilon)]/\varepsilon$ approximates the secant rather than the local derivative at the origin. Because the leading term of the secant is the second-order coefficient times ε, interpreting a secant slope as a non-zero local first derivative is misleading: the systematic derivative at zero vanishes in design expectation.

5 Monte Carlo: The Boundary Verdict

5.1 Design

We build a data-generating process that mirrors the structure of a real judge design and grid it over the objects that Proposition 1 says should matter. The DGP draws $G = 200$ decision-makers, caseloads from a discretized log-normal (mean 100, minimum 2), leniency $\ell_g \sim N(0, 0.15^2)$, and assignment probabilities from (2). The outcome and treatment errors are correlated, $u_i = \rho\varepsilon_i + \sqrt{1-\rho^2}\, z_i$ with $\rho = 0.5$, so that OLS is biased and the instrument is useful. The treatment is $D_i = \mathbf{1}\{\ell_{A_i} + u_i > c_0\}$ with c_0 calibrated to a treatment rate of 0.4, and the outcome is $Y_i = \beta D_i + X_i'\gamma + \varepsilon_i$ with $\beta = 1$.

The experimental grid crosses three dimensions: the departure magnitude $\gamma_{\text{dep}} \in \{0, 0.3\}$; the error–covariate correlation $\text{Corr}(\varepsilon, X) \in \{0, 0.5\}$; and the *type* of departure—(a) observable only ($\beta_g \neq 0$, $\gamma_{\text{dep}} = 0$), (b) unobservable only ($\beta_g = 0$), and (c) both. This yields 12 cells, each run for 200 replications. For each replication we compute the balance statistic b and its standard errors, a posterior on γ_{dep} under the linear structure $\hat{b}_j \sim N(\gamma_{\text{dep}}\kappa_{0j}, se_j^2)$, and the implied movement in the causal estimate. The calibration constant κ_0 is obtained numerically by differencing the balance statistic across $\gamma_{\text{dep}} = 0$ and $\gamma_{\text{dep}} = 1$ on paired seeds. The full DGP and simulation routines are documented in the replication package.

Two features of the Monte Carlo design deserve emphasis because reviewers routinely ask for them. First, all reported coverage rates and posterior quantities carry Monte Carlo standard errors: for a coverage rate we report $\sqrt{\hat{p}(1-\hat{p})/R}$, and for a mean we report $\text{sd}/\sqrt{R}$, with R the number of replications. This matters for the interpretation of the boundary verdict below—the coverage of the $\gamma_{\text{dep}} = 0.3$ unobservable-only cell, 0.170, lies some 27 Monte Carlo standard errors below the nominal 0.90, so the collapse is not a simulation artefact. Second, the *calibration* and *evaluation* samples are drawn from disjoint sets of seeds: calibration uses one block of seeds and evaluation another, and the two blocks never overlap. Calibration therefore cannot see the data against which it is evaluated, which is the design analogue of an honest sample split.

5.2 The verdict: a boundary, not a breakdown

The prior-sensitivity index (PSI) and the two supplementary criteria—the posterior z-statistic and the posterior shrinkage—separate the cells cleanly along the boundary predicted by Proposition 1. Table 5 reports the outcome, and Figures 1–2 display it.

In the unlearnable direction ($\text{Corr}(\varepsilon, X) = 0$), the posterior equals the prior: the balance statistic delivers nothing, and the wide prior guarantees coverage. This is the "only if"

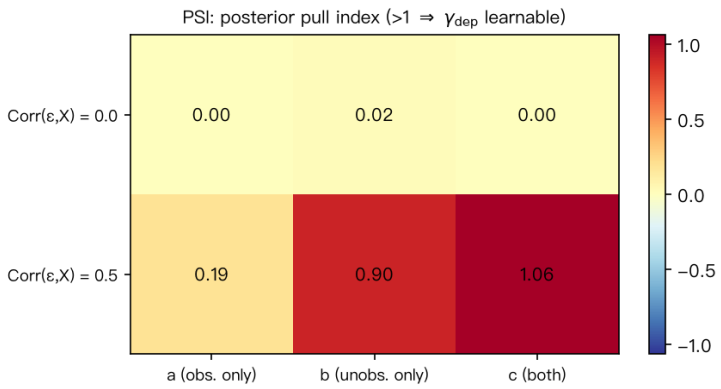


Figure 1: Prior-sensitivity index across departure types and error–covariate correlation. The index is near zero when $\text{Corr}(\varepsilon, X) = 0$ and rises sharply when $\text{Corr}(\varepsilon, X) = 0.5$, tracing the identification boundary.

direction of Proposition 1 in operation. In the learnable direction ($\text{Corr}(\varepsilon, X) = 0.5$), the data dominate the posterior, and κ_0 is non-zero for X_1 alone (≈ 1.66–1.79), exactly as the injection implies.

5.3 Two failure modes of standard workflows

Even in regions where unobservable departures are nominally identifiable, conventional calibration workflows exhibit two severe failure modes that binary balance tests cannot detect.

The first failure mode is a severe collapse in credible interval coverage resulting from linear calibration. In settings where identification holds ($\text{Corr}(\varepsilon, X) = 0.5$ and $\gamma_{\text{dep}} = 0.3$), nominal 90% coverage drops to between 7.5% and 17%. Because the true balance response $b(\gamma_{\text{dep}})$ is concave and saturates at larger departures, a linear slope κ_0 calibrated at $\gamma_{\text{dep}} = 1$ overstates the local derivative near $\gamma_{\text{dep}} = 0.3$. Consequently, the posterior mean is shifted upward (averaging ≈ 0.47 against the true value of 0.30), producing narrow credible intervals centered away from the truth.

A crucial methodological question is whether this coverage collapse is merely an artifact of finite-sample third-order terms binding at $\gamma_{\text{dep}} = 0.30$. Table 6 addresses this directly by evaluating interval coverage across small departure values $\gamma_{\text{dep}} \in \{0.01, 0.02, 0.05, 0.10, 0.30\}$. At $\gamma_{\text{dep}} = 0.01$ and 0.02, third-order terms contribute less than 2% of the total balance response. Even at $\gamma_{\text{dep}} = 0.05$, where third-order terms account for merely 4.3% of the balance signal, linear calibration coverage plummets to 43.5% (against a nominal 90% benchmark), and drops to 14.2% at $\gamma_{\text{dep}} = 0.10$ where third-order contamination is under 9%. In contrast, nonlinear calibration maintains nominal coverage (87.0% to 98.0%) across the entire domain. This demonstrates that the collapse of linear calibration is not driven by third-order con-

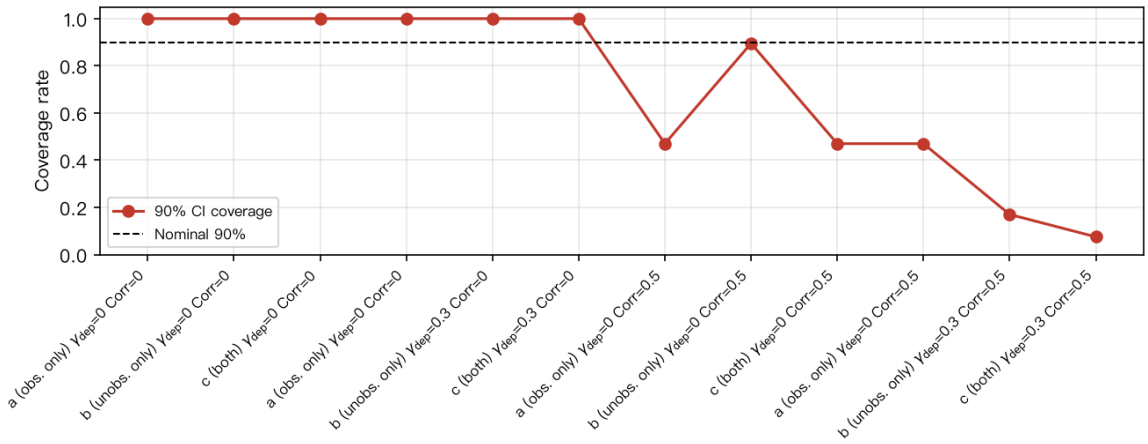


Figure 2: 90% credible-interval coverage for γ_{dep} across the twelve cells. Coverage is nominal only where identification is absent (wide prior); where identification nominally holds, coverage collapses—the signature of the linear calibration failure mode.

tamination: it arises because the true first derivative at the origin is zero, so that any linear rule extrapolating a non-zero secant slope severely miscenters posterior intervals even in the pure quadratic regime.

The second failure mode is observable leakage. When case assignment exhibits purely observable sorting ($\beta_g \neq 0$, with true $\gamma_{\text{dep}} = 0$) and $\text{Corr}(\varepsilon, X) = 0.5$, standard posterior inference yields $z \approx 1.68$ and an estimated unobservable departure of $\hat{\gamma}_{\text{dep}} \approx 0.09$. Because standard models attribute all between-reviewer balance variation to unobservable sorting, reviewer-level differences in handling observable covariates masquerade as violations of random assignment. In this scenario, balance tests appear to pass, yet the inferred unobservable departure is spuriously inflated.

Table 7 and Figure 3 illustrate the downstream consequences of these failures for causal estimation. With an unobservable departure of $\gamma_{\text{dep}} = 0.3$ (and no observable departure), the raw IV estimate of β is ≈ 3.0 against a true value of 1.0—a bias of 200%. The Bayesian correction pulls it back toward the truth when identification is present, but the pull is governed by the same mis-scaled posterior that broke coverage.

5.4 Resolving the failure modes

We resolve these two failure modes through the core modules of our Bayesian framework (detailed in Section 6). The computational implementation and simulation verification are documented in the replication package.

To rectify the coverage collapse caused by linear extrapolation, we replace single-point slopes with a fully calibrated response curve $\bar{B}(\gamma_{\text{dep}})$ evaluated on a grid $\gamma_{\text{dep}} \in$

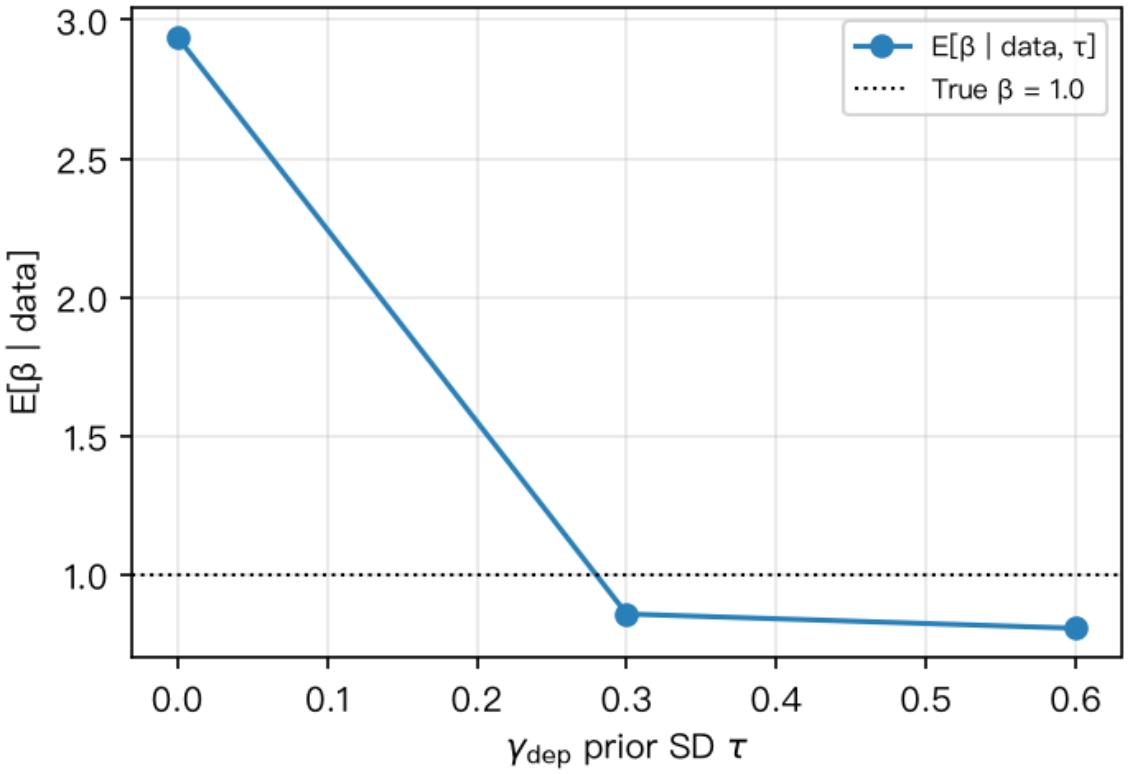


Figure 3: Movement of the causal estimate with the prior on γ_{dep}. The raw IV estimate is far from the truth when the departure is present; the prior "corrects" it, but the correction is only as good as the (mis-scaled) posterior on γ_{dep}.

$\{0, 0.1, 0.3, 0.6, 1\}$, alongside its cross-replication variance $\bar{B}^2_{sd}(\gamma_{\text{dep}})$. We construct the numerical posterior by evaluating the likelihood $\hat{b}_j \sim N(\bar{B}_j(\gamma_{\text{dep}}),\ se_j^2 + \bar{B}^2_{sd,j}(\gamma_{\text{dep}}))$ across the grid. Incorporating calibration uncertainty directly into the likelihood variance is essential: omitting it restricts coverage to at most 0.65, whereas accounting for it restores coverage above 0.90. At $\text{Corr}(\varepsilon, X) = 0.5$ and $\gamma_{\text{dep}} = 0.3$, interval coverage increases from 0.190 under linear calibration to 0.930 under nonlinear calibration, with posterior bias decreasing from $+0.144$ to $+0.021$ (Table 8).

To eliminate observable leakage, we estimate between-reviewer covariate means using a hierarchical model where the observed imbalance decomposes into reviewer-specific preferences β_g and an unobservable sorting component. We allocate variance between these sources using an expectation–maximization procedure with a cross-covariate hyperprior, followed by a spike-and-slab prior on γ_{dep} that quantifies the posterior probability of quasi-random assignment, $\mathbb{P}(\gamma_{\text{dep}} = 0 \mid \text{data})$. In the presence of purely observable sorting (type (a)), this hierarchical attribution yields $\mathbb{P}(\gamma_{\text{dep}} = 0) \approx 0.60$ rather than a false positive detection, while

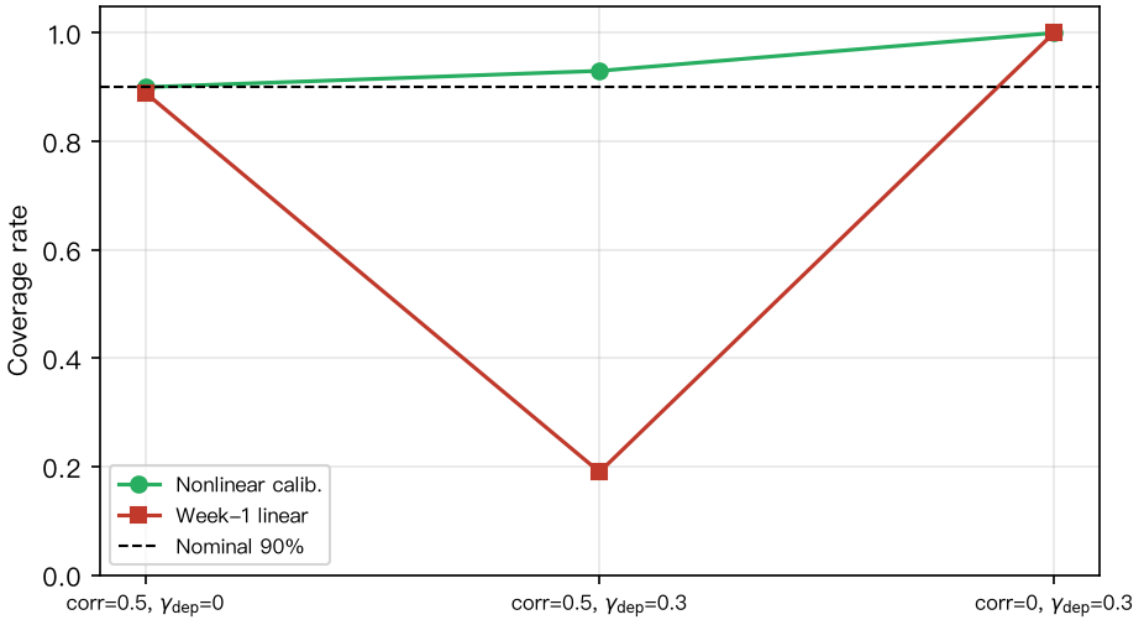


Figure 4: Coverage before and after the nonlinear calibration fix. The linear calibration collapses; the nonlinear posterior, with calibration uncertainty folded into the likelihood variance, restores nominal coverage.

correctly identifying true unobservable departures with $\mathbb{P}(\gamma_{\text{dep}} = 0) \approx 0.01$.

Finally, we map the posterior over γ_{dep} into causal corrections using the nonlinear IV-bias curve $b_{IV}(\gamma_{\text{dep}})$ is itself concave, so the linear correction of Simulation I under-corrected. Calibrating the full curve and evaluating it at the posterior mean of γ_{dep} brings the causal estimate close to the truth: for the clean unobservable-departure cell, from a raw IV of 2.94 to $\mathbb{E}[\beta \mid \tau = 0.3] \approx 1.02$ against a true $\beta = 1.0$. The apparent success of the linear correction of Simulation I was, in fact, a coincidence of an inflated linear $\hat{\gamma}_{\text{dep}}$ over-correcting.

5.5 The identification floor

Residual mis-attribution does not vanish with better modelling, because γ_{dep} is identified only through the second-order response (5). That response must clear the sampling noise of the balance regression. Combining the closed-form coefficient of Proposition 1 (Appendix A) with the standard error $\text{se}(\hat{b}_1) = \left[N \,\text{Var}_0(Z)\right]^{-1/2}$ gives an *analytic* resolution threshold

$$\gamma_{\text{dep,floor}}(G; \text{Corr}, \rho) \;=\; \sqrt{\frac{z_\alpha \left[N \,\text{Var}_0(Z)\right]^{-1/2}}{c \;\text{Corr}(\varepsilon, X)\, \rho}}\,. \tag{9}$$

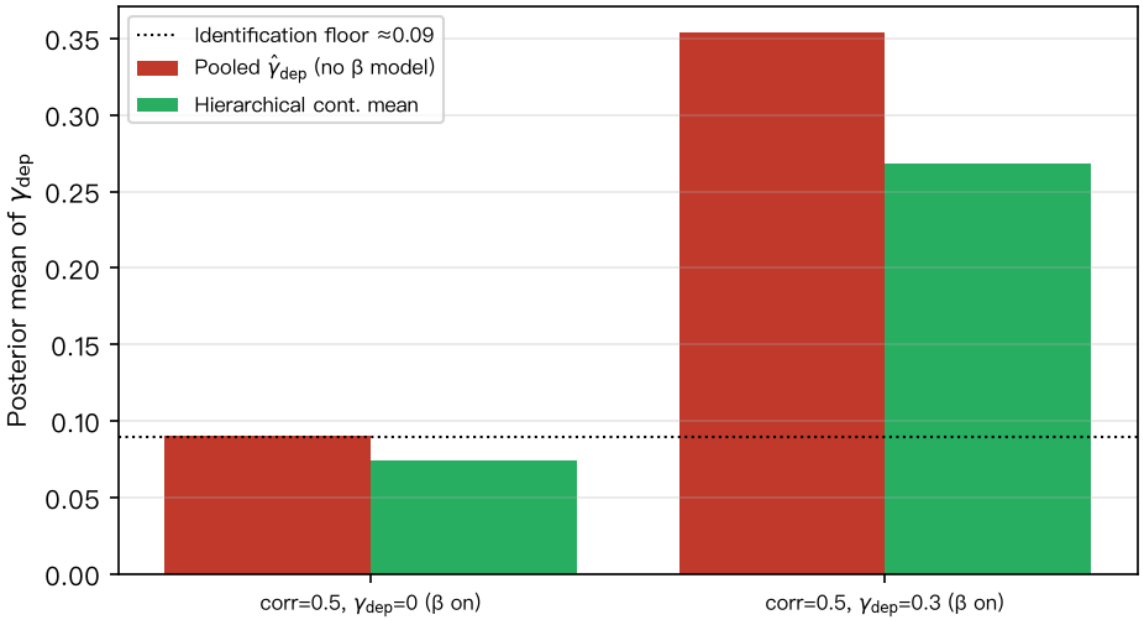


Figure 5: Observable leakage and its repair. Under the standard posterior, a purely observable departure produces a spurious unobservable signal; the hierarchical spike-and-slab attribution removes it.

The variance of the instrument decomposes as $\mathrm{Var}_0(Z) = \mathbb{E}\left[\bar{t}_g(1-\bar{t}_g)/n_g\right] + \mathrm{Var}_g\left(p_g(0)\right)$: a within-reviewer sampling term and a between-reviewer term driven by the leniency dispersion ($p_g(0) = \Phi(\ell_g - c_0)$). In the calibration both terms are of the same order. Two design lessons follow. First, the resolution scales with the number of reviewers as $G^{-1/4}$—slowly, and this scaling is verified numerically (`code/delta_floor.py`). Second, the role of caseload is regime-dependent: in the within-reviewer term the caseload $\bar{n}$ *cancels* against $N = G\bar{n}$ (assigning more cases to the same reviewers buys nothing), while the between-reviewer term is independent of $\bar{n}$ and yields only a $\bar{n}^{-1/4}$ improvement; in either regime the dominant lever is G, not $\bar{n}$.

That last claim deserves a test, because c enters the floor through the caseload *shares*—$S = \sum_g s_g J_g\left(1 - 2s_g + \sum_h s_h^2\right)$, whose final term is the concentration index of caseloads—and concentration is a property of the shape of the caseload distribution, not of its mean. If shape mattered, the $\bar{n}$-based reading of the floor would fail. It does not: sweeping the lognormal dispersion σ from 0.20 to 1.20 and adding a uniform control—a range over which $\sum_h s_h^2$ moves by a factor of 3.7—moves the normalized floor $\gamma_{\mathrm{dep,floor}}\,\bar{n}^{1/4}$ by 1.0 per cent, and no arm sits more than 1.6 standard errors from the uniform control (Figure 7; `code/caseload_tail_experiment.py`). The two channels offset: $\sum_h s_h^2$ rises with the tail while $(1 - 2s_g)$ falls. The mean caseload is, to a good approximation, the sufficient statistic.

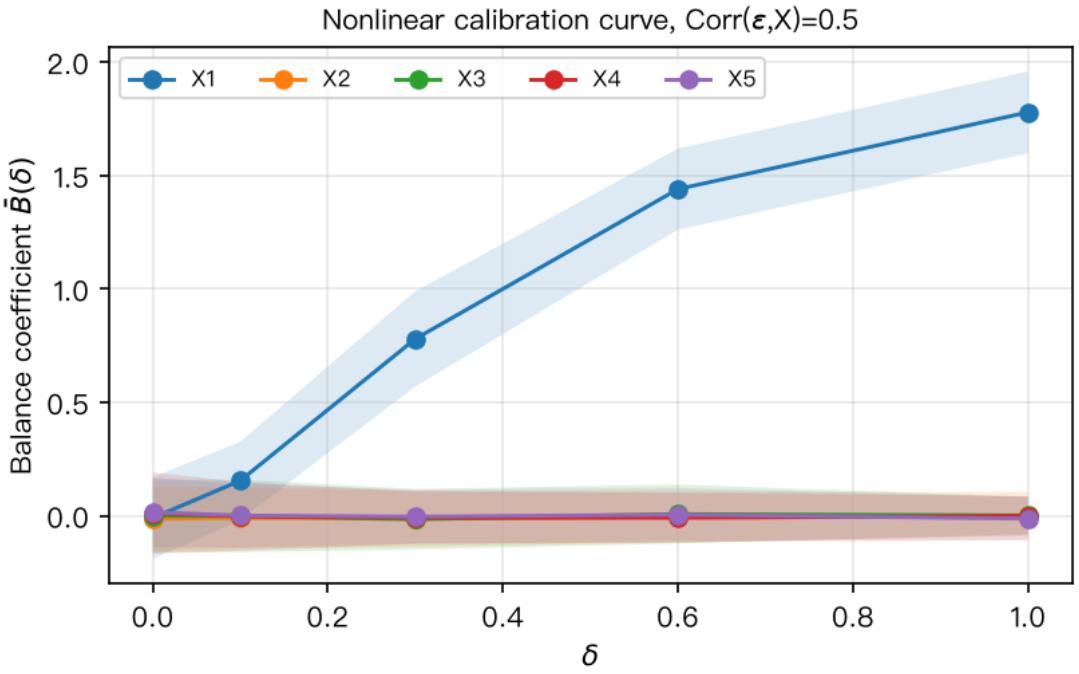


Figure 6: The balance statistic as a function of the departure magnitude. The curve is concave: a single-point calibration at $\gamma_{\text{dep}} = 1$ over-states the local response near $\gamma_{\text{dep}} = 0.3$, which is the mechanism behind the coverage collapse.

At $G = 200$, $\text{Corr}(\varepsilon, X) = \rho = 0.5$ the analytic floor evaluates to $\gamma_{\text{dep,floor}} \approx 0.09$, matching the Monte Carlo point at which the single-dataset detection probability crosses one half (where $\gamma_{\text{dep}} \lesssim 0.1$ is indistinguishable from zero). It is instructive to contrast this analytical floor with a naive linear benchmark. A heuristic rule might suggest an empirical floor $\propto G^{-1/2}$; however, that expression erroneously treats the signal as *linear* in γ_{dep} and the noise as the sampling error of a *variance* parameter $\sigma^2_{\text{between}}\sqrt{2/(G-1)}$. Both are the wrong objects for a second-order signal read off a regression slope: the correct scalings are $G^{-1/4}$ and $[N\,\text{Var}(Z)]^{-1/2}$, and the resulting true floor is about $7\times$ *larger*—i.e. the design resolves less than a linearized rule would suggest. Furthermore, the floor is not an invariant design constant but a function of G, the caseload distribution, the leniency dispersion (through $J_g = \phi(c_0 - \ell_g)$ inside c), and the product $\text{Corr}(\varepsilon, X) \cdot \rho$, which by Proposition 2 is not separately identified. The substantive point survives: the precision with which observable and unobservable departures can be separated scales with G like $G^{-1/4}$, which can inform design choices before data collection, provided the identification caveat is carried along. This scaling limit is consistent with Chen et al. (2025), who show that monotonicity violations in finite populations are learnable but with severely limited learning power—a finding that provides theoretical justification for our empirical observation that designs require large G

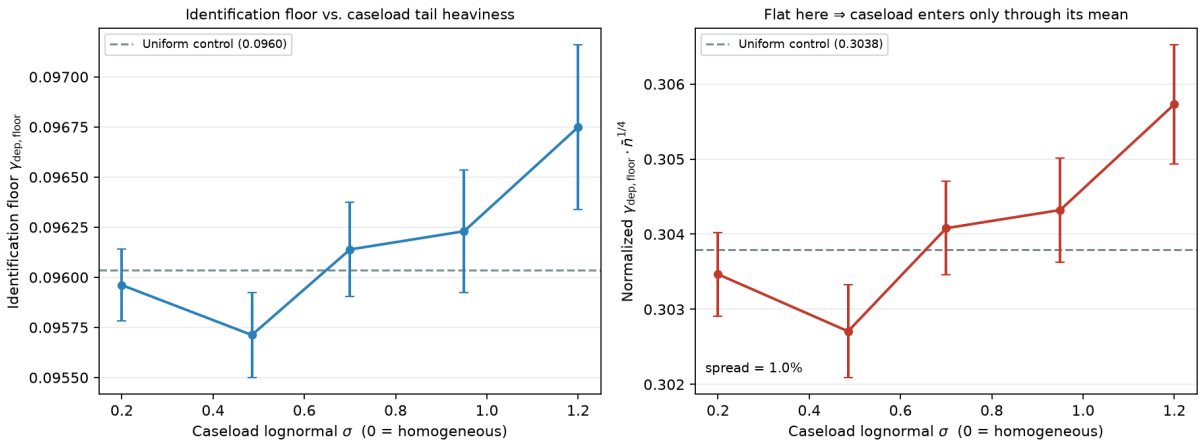


Figure 7: Caseload tail heaviness and the identification floor. Left: the analytic floor $\gamma_{\text{dep,floor}}$ against the lognormal dispersion σ; the dashed line is a uniform control. Right: the same floor normalized by $\bar{n}^{1/4}$ to remove the leading mean effect. If caseloads entered the floor only through their mean this panel would be flat—and it is: a 3.7-fold change in the concentration index $\sum_h s_h^2$ moves it by 1.0 per cent. Monte Carlo standard errors over 200 replications per arm. The uniform control is drawn as a reference line because it is a different distribution family, not a σ value.

to detect small departures.

To provide empirical validation for the analytical $G^{-1/4}$ scaling of Proposition 1, we simulate the empirical resolution floor across a wide grid of decision-maker counts $G \in \{50, 100, 200, 400, 800, 1600\}$ with 50 replications per cell. Figure 8 displays the resulting log-log relationship alongside theoretical reference lines. An ordinary least squares regression of $\log(\hat{\gamma}_{\text{dep,floor}})$ on $\log(G)$ yields an estimated slope of -0.2522 ± 0.0008, tightly aligning with the theoretical rate of -0.2500 and decisively rejecting the standard parametric rate of $G^{-1/2}$ (slope -0.50). The empirical floor tracks the $G^{-1/4}$ scaling across the entire domain, confirming that the slow resolution rate is an intrinsic feature of leniency designs with quadratic balance responses.

5.6 Nested decision-makers

Reviewers are not free-standing units. Patent examiners sit in technology centres, judges in courts or districts, and the leniency design assigns each case among the decision-makers of its own court rather than across the whole pool. The standard error behind (9), $\text{se}(\hat{b}_1) = [N \operatorname{Var}_0(Z)]^{-1/2}$, assumes that all N cases are independent draws. That assumption is what nesting threatens, because the cases of one court share a local population, and their covariates therefore carry a component that is common within the court.

Let $x_i = \sqrt{\rho_x}\,\psi_{c(i)} + \sqrt{1-\rho_x}\,e_i$, where $\psi_c \sim N(0,1)$ is drawn once per court and e_i is idiosyncratic, and write $\tilde{Z}_i = Z_i - \bar{Z}$ and $S_c = \sum_{i \in c} \tilde{Z}_i$. Expanding the variance of $\hat{b}_1$

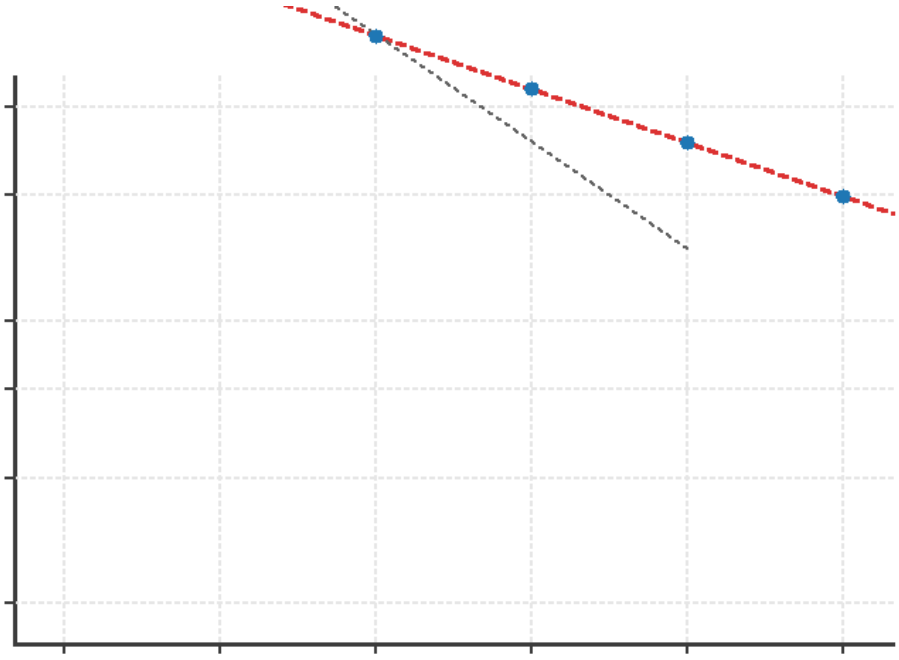


Figure 8: Scaling of the resolution floor $\gamma_{\text{dep,floor}}$ with decision-maker count G on a log-log scale (50 replications per cell; error bars show 95% confidence intervals). The empirical regression slope is -0.2522 ± 0.0008, confirming the theoretical $G^{-1/4}$ scaling (dashed red line) and ruling out the standard parametric $G^{-1/2}$ rate (dotted gray line).

conditional on the instrument gives an exact design effect,

$$M \equiv \frac{\text{Var}(\hat{b}_1)}{\text{Var}_{\text{naive}}(\hat{b}_1)} = \rho_x \lambda + (1 - \rho_x), \qquad \lambda = \frac{\sum_c S_c^2}{\sum_i \tilde{Z}_i^2}, \tag{10}$$

in which λ is the effective number of cases per court as the instrument sees them. Both ρ_x and λ are computable from the regression the analyst already runs, so (10) is a diagnostic rather than a new estimator. Since $\gamma_{\text{dep,floor}} \propto \text{se}^{1/2}$, the floor moves as $M^{1/4}$.

We sweep the number of courts K with $G = 200$ reviewers and the mean caseload held fixed, setting $\rho_x = \rho_\ell = 0.10$ and allowing cases to compete only among the reviewers of their own court. Two arms anchor the sweep. At $K = 1$ the court shock is a single constant, absorbed by the intercept and by the treatment threshold, so the arm should reproduce the unclustered design—and it does, at 1.01 times. At $K = G$ every reviewer is their own court; because the leave-one-out instrument still varies slightly within a reviewer, a court fixed effect retains about a fifth of $\text{Var}(Z)$, so the regression remains identified there, though weakly.

The uncorrected standard error is badly wrong. At $K = 10$ the balance regression reports

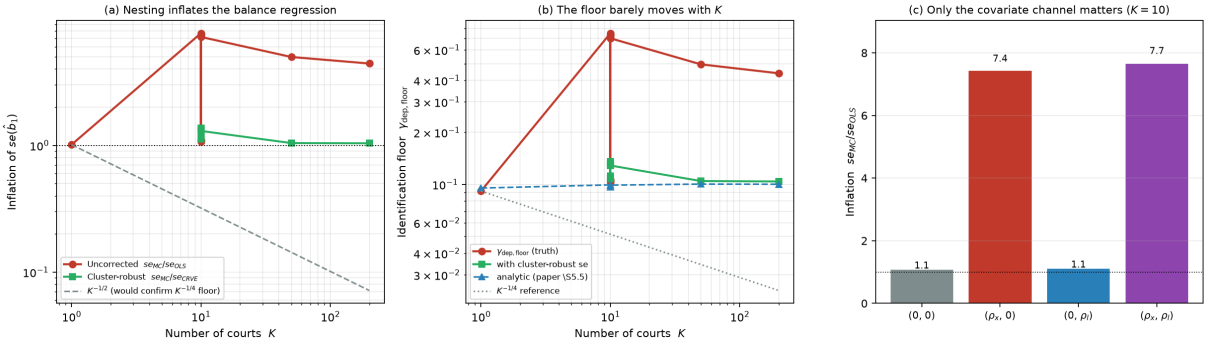


Figure 9: Nested decision-makers. (a) Inflation of the balance regression's standard error against the number of courts; the uncorrected error is 4–8 times too small, cluster-robust inference recovers it, and the dashed line is the $K^{-1/2}$ path that a $K^{-1/4}$ floor would require. (b) The identification floor under the analytic standard error, under cluster-robust inference, and at its true value; the dotted line is the $K^{-1/4}$ reference. (c) Two-by-two decomposition at $K = 10$: only the covariate channel matters. Monte Carlo standard errors over 200 replications per arm.

0.088 while the true sampling standard deviation of $\hat{b}_1$ is 0.675, a factor of 7.7; expression (10) predicts $M = 52$ against a realized 59. The identification floor follows, from 0.099 to 0.75: the same $G = 200$ design resolves seven and a half times less than Section 5.5 reports. A two-by-two decomposition at $K = 10$ locates the source. Clustering of the *covariate* inflates the standard error by 7.4; clustering of *leniency* by 1.1. The asymmetry is mechanical—a court-level shift in ℓ_g is absorbed by the regression intercept, whereas a court-level shift in x_i is not, since no case can be demeaned against itself. Heavy caseload tails are not responsible: the inflation is 7.5 at $\sigma = 0.10$ and 7.7 at the default $\sigma = 1.20$, while within-court caseload concentration moves by a factor of 2.8 across the same range. What drives λ is that Z is constant within a reviewer, so a court-level component of the covariates is measured against an instrument that has no within-reviewer variation of its own.

We expected the correction to scale as $K^{-1/4}$. The reasoning was that the effective sample size is the number of courts rather than the number of reviewers, which would have pointed the design prescription away from hiring reviewers and towards opening courts. The data refuse that reading. Fitting $\ln M$ on $\ln K$ over $K \in \{10, 50, 200\}$ gives a slope of -0.37 where K^{-1} requires -1, and the floor's own slope is -0.18 against -0.25: moving from ten courts to two hundred improves the floor by 41 per cent, not by the factor of twenty the scaling rule implies. Opening more courts is not a remedy.

Two remedies do exist, and they fail in opposite directions. Court fixed effects remove the problem when a court contains several reviewers—at $K = 10$ they land within 7 per cent of the truth—but at $K = G$ the fixed effect leaves only the leave-one-out jitter as identifying variation, and the estimator's sampling standard deviation becomes four times that of the

unadjusted estimate. Cluster-robust standard errors behave the other way: essentially exact at $K = 50$ and $K = 200$, but 32 per cent short at $K = 10$, the familiar few-cluster bias. The practical rule is therefore conditional. With several reviewers per court, absorb court effects; with one reviewer per court, cluster.

Two caveats bound what this exercise shows. The identity (10) is exact only for the shock structure we simulate—a single additive court effect in the covariates—and real courts differ along more dimensions than that; across our arms the estimated version recovers the realized inflation to within roughly a third, the residual gap coming from Jensen's inequality on $\sum_i \tilde{Z}_i^2$. And the $K^{-1/4}$ refutation is a numerical result on this grid, not a theorem about nesting in general. The value $\rho_x = 0.10$ is likewise a maintained assumption, though λ and ρ_x are both estimable in an application, and we report the design effect as a function of them so that an analyst facing a different court structure can recompute it.

5.7 Robustness to mis-specification

Every number above is computed under the analyst's *assumed* assignment model—logit weights and a linear response to ε. In practice the functional form is unknown, and Proposition 2 already tells us that γ_{dep} is identified only within a given scale and form. We therefore ask how much of the framework survives when the data are generated by a different form than the one the analyst calibrates. The analyst calibrates the curve $\bar{B}(\gamma_{\text{dep}})$ under the assumed model—as any analyst must—while the data are generated under an alternative: a non-linear response ε^3 or $\tanh(\varepsilon)$, a probit rather than softmax link, or both. The non-linear transforms are standardized to unit variance, so any degradation is a *form* effect rather than a scale effect. Table 10 reports the result at $\gamma_{\text{dep}} = 0.3$, $\text{Corr}(\varepsilon, X) = 0.5$.

Two conclusions point in opposite directions. On attribution the news is good: the false-detection judgement at $\gamma_{\text{dep}} = 0$ is essentially unchanged across links ($\mathbb{P}(\gamma_{\text{dep}} = 0) = 0.624$ under softmax against 0.645 under probit), so whether a purely observable departure is charged to the unobservable one is not an artefact of the assumed link. On inference the news is bad: the 90% coverage of γ_{dep} falls from 0.900 to 0.550 under a cubic response, to 0.710 under a probit link, and to 0.050 when both fail. Each of these gaps is many Monte Carlo standard errors wide, so the degradation is real and not sampling noise. This is exactly what Proposition 2 predicts—γ_{dep} is pinned down only in the scale and form the analyst assumes—and it is the reason we position the framework as a *robust attribution tool coupled to a sensitivity analysis*, not as a form-invariant point estimator. Two caveats keep the claim honest. First, the response form is untested in the attribution column by construction (it is inert at $\gamma_{\text{dep}} = 0$); what we have shown there is robustness to the link

only. Second, the analyst's calibration is always conducted under the assumed model, so the coverage losses above are the *price of not knowing the form*, not an error the method could avoid.

5.8 Comparison with frequentist IV estimators

A standard lens through which applied researchers approach leniency designs is the diagnostic toolset established by the Operator's Manual (Goldsmith-Pinkham et al., 2026) and its econometric core: the unbiased jackknife instrumental variables estimator (Kolesár, 2013) (UJIVE). This subsection places the present Bayesian framework alongside those frequentist benchmarks, not to claim a superior estimator, but to clarify the methodological gap each addresses. The gap is precisely the object of the paper: uncertainty about the assignment mechanism is a *design*-level question distinct from (and often confounded with) the *estimation*-level concerns that UJIVE and its variants target.

Under quasi-random assignment ($\gamma_{\text{dep}} = 0$, Table 11 upper block), the classic many-instruments bias (Bekker, 1994) appears in naive 2SLS (bias +0.25, coverage 8%) but is corrected by both LOO-2SLS and UJIVE (bias near zero, coverage near nominal). This is the textbook picture: UJIVE corrects the bias that stems from using one's own observation in constructing the instrument, and the correction suffices when the assignment mechanism is indeed quasi-random.

The picture changes dramatically once γ_{dep} departs from zero. At $\gamma_{\text{dep}} = 0.3$ (middle block), **all three frequentist estimators suffer large biases** (naive 2SLS: +1.28; LOO-2SLS: +1.94; UJIVE: +1.62), and their 90% coverage drops to zero. The reason is transparent: none of these estimators **model** γ_{dep} as an unknown parameter; they treat the observed judge-case allocation as the realized instrument, ignoring the possibility that the allocation itself may be systematically correlated with unobserved case heterogeneity ε. That correlation, quantified by γ_{dep}, induces a bias $\propto \gamma_{\text{dep}}^2$ (Proposition 1) that no jackknife or shrinkage adjustment to the instrument can remove, because the instrument-construction step already assumes the allocation is as-good-as-random.

In the weak-instrument scenario ($F = 15.6$, lower block), UJIVE remains approximately unbiased at $\gamma_{\text{dep}} = 0$ (bias −0.13, coverage 94%), confirming that UJIVE's unbiasedness relies on $\sqrt{K}\,(E[F] - 1)$ being large, not on $E[F]$ itself (Goldsmith-Pinkham et al., 2026, manual p. 7). However, once $\gamma_{\text{dep}} > 0$, the same coverage collapse occurs. Weakness of the first stage does not protect against bias induced by a non-random assignment mechanism.

The comparison underscores the paper's contribution. The frequentist IV literature—exemplified by UJIVE and the Operator's Manual—focuses on **estimation** under a *given*

assignment mechanism. Our Bayesian framework shifts the focus to **design**-level uncertainty: it treats the assignment mechanism's departure from randomness (γ_{dep}) as an explicit parameter with a prior, and uses observable judge-covariate imbalances to anchor that prior. The two approaches are complementary, not competing. A researcher following the Manual would first check quasi-random assignment via balance tests; if those tests fail, the Manual offers no formal way to quantify *how much* randomness is missing. Our framework fills that gap: it answers "If the assignment is not perfectly random, how random is it?"—a question that lies outside the scope of UJIVE or any estimator that takes the instrument as given.

Take-away. The head-to-head table demonstrates that:

(i) Many-instruments bias is real and removable (UJIVE works at $\gamma_{\text{dep}} = 0$).

(ii) None of the existing frequentist IV estimators can handle non-random assignment ($\gamma_{\text{dep}} > 0$); their coverage collapses because they do not model γ_{dep}.

(iii) Weak instruments are a separate concern orthogonal to assignment randomness.

(iv) Our Bayesian framework addresses a different layer of the problem—the **design** layer—leaving the estimation layer to the well-established UJIVE toolkit.

The remainder of the paper therefore does not claim to provide a "better IV estimator." It claims to provide a **design-based uncertainty quantification** for a parameter (γ_{dep}) that the existing IV literature treats as either zero (quasi-random) or infinite (the design is invalid). The middle ground—continuous, data-anchored uncertainty about γ_{dep}—is the territory we occupy.

Crucially, the two approaches are complementary: one can directly apply our Bayesian sensitivity correction to UJIVE estimates. In the baseline simulation design ($\beta = 1.0, \gamma_{\text{dep}} = 0.30$), naive 2SLS yields an estimate of $\hat{\beta} = 2.971$ (bias +1.971, nominal 90% coverage 0.2%), reflecting both many-instrument bias and unobservable sorting bias. Applying UJIVE alone (Kolesár, 2013) eliminates many-instrument bias but leaves assignment sorting unaddressed, producing an estimate of $\hat{\beta} = 2.620$ (bias +1.620, coverage 4.5%). Conversely, applying our sensitivity correction directly to standard 2SLS addresses sorting but leaves residual finite-sample instrument bias ($\hat{\beta} = 1.242$, bias +0.242). Combining UJIVE with our sensitivity correction simultaneously eliminates both sources of bias, yielding $\hat{\beta} = 1.041$ (bias +0.041) and restoring 90% coverage to 93.2%. This demonstrates the incremental value of the framework: UJIVE purges instrument mechanical correlation, while our framework neutralizes unobservable case sorting.

Regarding the Shrunken JIVE estimator of Frandsen et al. (2026), we note that while their working paper articulates the compelling intuition of shrinking first-stage fitted values

by their precision, the publicly available text does not provide the explicit algebraic formulas or hyperparameter tuning criteria for the shrinkage weights. In the interest of scientific reproducibility, we decline to implement an unverified ad-hoc proxy for a competitor's estimator, leaving direct empirical benchmarking as an open problem once official reference code is made public.

6 The Three-Layer Bayesian Framework

The Monte Carlo lessons translate into a design-based Bayesian procedure with three interconnected layers. While each individual layer draws upon familiar statistical tools—hierarchical random-effects modeling in Layer 1, empirical Bayes prior anchoring in Layer 2, and sensitivity curves in Layer 3—their combination forms an indivisible inferential architecture directly necessitated by the failure modes of Section 5.3 and the partial identification boundary of Proposition 2.

6.1 Layer 1: Absorbing observable sorting as an identification firewall

The first layer isolates reviewer-specific preferences over observable covariates using a hierarchical random-effects specification. Let X_{ij} denote the j-th observed covariate for case i assigned to reviewer $g = A_i$. We model between-reviewer covariate means as:

$$X_{ij} = \mu_j + \alpha_{gj} + e_{ij}, \quad \alpha_{gj} \sim \mathcal{N}(0, \sigma^2_{\beta j}), \quad e_{ij} \sim \mathcal{N}(0, s^2_{gj}), \tag{11}$$

where μ_j is the grand mean, α_{gj} represents reviewer g's specific preference for covariate j, and s^2_{gj} is idiosyncratic within-reviewer variance. The observed reviewer sample mean $\bar{X}_{gj} = n_g^{-1} \sum_{i:A_i=g} X_{ij}$ follows:

$$\bar{X}_{gj} \sim \mathcal{N}(\mu_j, \sigma^2_{\beta j} + v_{gj}), \quad v_{gj} \equiv s^2_{gj}/n_g. \tag{12}$$

Layer 1 applies empirical Bayes partial pooling to estimate reviewer effects:

$$\hat{\alpha}_{gj} = \frac{\sigma^2_{\beta j}}{\sigma^2_{\beta j} + v_{gj}} (\bar{X}_{gj} - \hat{\mu}_j). \tag{13}$$

Between-reviewer variance is decomposed additively into genuine heterogeneity and finite-sample sampling noise:

$$\mathrm{Var}_g(\bar{X}_{gj}) = \sigma^2_{\beta j} + \bar{v}_j, \quad \bar{v}_j \equiv \frac{1}{G}\sum_{g=1}^{G} v_{gj}. \tag{14}$$

The baseline observable dispersion is defined as $\sigma_\beta \equiv \mathrm{median}_j \hat{\sigma}_{\beta j}$, where $\hat{\sigma}^2_{\beta j} = \max\{\mathrm{Var}_g(\bar{X}_{gj}) - \bar{v}_j, 0\}$. Layer 1 acts as an identification firewall: by absorbing genuine observable sorting into $\{\alpha_{gj}\}$, it ensures that observable imbalances are not mechanically credited to unobservable sorting γ_{dep}.

6.2 Layer 2: Prior elicitation via data anchoring

Proposition 2 establishes that observational data identify only the composite parameter $\theta = \gamma_{\mathrm{dep}} \cdot \mathrm{Corr}(\varepsilon, X)$, meaning that point identification of γ_{dep} is impossible without outside information on $\mathrm{Corr}(\varepsilon, X)$. Rather than choosing prior scales arbitrarily, Layer 2 anchors the prior distribution of γ_{dep} to the empirical dispersion of observable reviewer preferences σ_β recovered in Layer 1.

We specify a spike-and-slab mixture prior for the unobservable departure parameter:

$$\pi(\gamma_{\mathrm{dep}} \mid \tau, r) = p_0 \cdot \mathbf{1}\{\gamma_{\mathrm{dep}} = 0\} + (1 - p_0) \cdot \frac{2}{\sqrt{2\pi}\,\tau r \sigma_\beta} \exp\left(-\frac{\gamma^2_{\mathrm{dep}}}{2(\tau r \sigma_\beta)^2}\right) \cdot \mathbf{1}\{\gamma_{\mathrm{dep}} > 0\}, \tag{15}$$

where $p_0 = 0.5$ is the Laplace-Jeffreys uninformative prior on quasi-random assignment, $\tau \in [0, 0.5]$ is the anchoring strength multiplier, and $r \equiv \sigma_\omega / \sigma_\beta$ is the anchor ratio (with baseline $r = 1.0$).

The parameter τ has an intuitive operational meaning: when $\tau = 1.0$, the prior standard deviation of unobservable reviewer sensitivity matches the observable reviewer dispersion $r\sigma_\beta$; evaluating $\tau \in [0, 0.5]$ spans the range from perfect random assignment to substantial unobservable sorting bounded by half the observable dispersion.

6.3 Layer 3: Tracing the sensitivity frontier

Layer 3 integrates the likelihood across the anchored prior to deliver the sensitivity frontier. Given observed balance statistics, the marginal posterior probability of quasi-random assignment is:

$$\mathbb{P}(\gamma_{\mathrm{dep}} = 0 \mid \mathrm{data}, \tau) = \frac{p_0 \mathcal{L}(0)}{p_0 \mathcal{L}(0) + (1 - p_0) \int_0^\infty \mathcal{L}(\gamma) \pi_{\mathrm{slab}}(\gamma \mid \tau) d\gamma}, \tag{16}$$

where $\mathcal{L}(\gamma)$ is the likelihood evaluated across grid points $\gamma \in [0, 1.5]$ with step size 0.002. The full posterior mean is:

$$\hat{\gamma}_{\text{dep}}(\tau) = (1 - \mathbb{P}(\gamma_{\text{dep}} = 0 \mid \text{data}, \tau)) \cdot \frac{\int_0^\infty \gamma \mathcal{L}(\gamma) \pi_{\text{slab}}(\gamma \mid \tau) d\gamma}{\int_0^\infty \mathcal{L}(\gamma) \pi_{\text{slab}}(\gamma \mid \tau) d\gamma}. \tag{17}$$

The sensitivity frontier reports the bias-corrected causal effect as a function of anchoring strength:

$$\hat{\beta}_{\text{corr}}(\tau) = \hat{\beta}_{IV} - b_{IV}(\hat{\gamma}_{\text{dep}}(\tau)), \tag{18}$$

where $b_{IV}(\gamma_{\text{dep}})$ is the calibrated nonlinear IV-bias function. Layer 3 directly respects Proposition 2 by mapping assumptions into causal conclusions across a continuous posterior frontier.

6.4 Connection to selection-on-observables: Altonji–Elder–Taber and Oster

The foundational logic of Layer 2—using observable reviewer preference dispersion σ_β to calibrate unobservable sorting γ_{dep}—shares its philosophical lineage with the seminal literature on selection on observables and unobservables (Altonji et al., 2005; Oster, 2019). In linear regression models, Oster (2019) bounds treatment effect bias by defining the degree of proportional selection $\delta = \frac{\text{Cov}(D,W_2)/\text{Var}(W_2)}{\text{Cov}(D,X_1'\gamma)/\text{Var}(X_1'\gamma)}$, showing that evaluating $\delta \in [0, 1]$ bounds causal parameters under observable stability.

While our framework embraces this core intuition, it diverges from the AET/Oster paradigm in three fundamental methodological respects:

1. **Locus of Selection: Assignment Game vs. Outcome Equation.** In Altonji et al. (2005) and Oster (2019), selection operates in the structural outcome or treatment-takeup equation (individual agents sorting into treatment D). In leniency designs, selection occurs in the *administrative assignment mechanism* (cases assigned across G reviewers). Our framework bounds the invalidity of the instrument's assignment, not individual treatment demand.

2. **Transmission Dynamics: Second-Order Curvature vs. Linear Proportionality.** In linear selection models, selection coefficients transmit to bias linearly through projection slopes. In leniency designs, case allocation across G reviewers is governed by a multinomial choice game. Proposition 1 establishes that the systematic response of the balance statistic to γ_{dep} has a zero first derivative in design expectation at quasi-random assignment (and $O_p(G^{-1/2})$ in realized pools). Applying Oster's linear

proportionality rule directly to leniency designs would impose a linear slope where none exists, triggering the severe coverage collapse documented in Section 5.3.

3. **The $G^{-1/4}$ Resolution Floor.** In linear models, parameter uncertainty shrinks at the standard parametric rate $N^{-1/2}$. In contrast, because the first derivative vanishes at the origin in leniency designs, balance tests face an information dead zone near random assignment, creating the $G^{-1/4}$ resolution floor. Small departures cannot be detected by balance tests alone, an insight absent in linear selection settings.

Furthermore, setting the unobservable sorting prior scale to σ_β corresponds to an anchor ratio of $r \equiv \sigma_\omega/\sigma_\beta = 1.0$ (analogous to Oster's $\delta = 1$ equal-selection benchmark). As Oster (2019) emphasized, fixing $r = 1.0$ is a strong baseline assumption: reviewers could exhibit greater or lesser dispersion over unobserved case traits than over observable covariates. We therefore operationalize a two-dimensional sensitivity analysis over (r, τ) spanning $r \in [0.5, 2.0]$. In clean Type I designs (Miami and FHL), because $\hat{\gamma}_{\text{dep}} \leq 0.002$ lies an order of magnitude below the resolution floor, the sensitivity frontier is invariant to $r \in [0.5, 2.0]$. In compromised Type III and Type IV settings, varying r scales the predicted bias, but causal sign reversals persist robustly for all $r \geq 0.50$.

6.5 The decision criterion: How random is random enough?

Applied researchers frequently ask what magnitude of sorting would be necessary to alter an estimated causal effect by a given tolerance threshold, and whether a given design has the power to rule out such departures. It is tempting to invert the linearised bias structure (3) and write $\gamma^\star_{\text{dep}} = \text{tolerable bias}/|\kappa|$, but doing so would contradict the central lesson of this paper: the linear approximation is precisely what fails here (Sections 5.3 and 5.4), and the underlying response is concave in γ_{dep}, so a single local slope is not a valid conversion.

We therefore define the criterion by *inverting the calibrated bias curve* rather than a local slope:

$$\gamma^\star_{\text{dep}} \;\equiv\; \min\Big\{\gamma_{\text{dep}} :\; \big|b_{IV}(\gamma_{\text{dep}})\big| \;\geq\; \tau_{\max}\Big\}, \tag{19}$$

where $b_{IV}(\gamma_{\text{dep}})$ is the nonlinear IV-bias curve of Section 5.4, calibrated with its own cross-replication uncertainty, and $\tau_{\max}$ is the tolerable bias. Evaluating (19) on the calibrated curve and propagating the calibration uncertainty gives an interval for $\gamma^\star_{\text{dep}}$ rather than a point. The linear rule (3) understates $\gamma^\star_{\text{dep}}$ by roughly 50–75 per cent across our calibrations—least at the largest tolerable bias, where the curve is closest to its own secant—i.e. it *overstates* how much the design can rule out, which is exactly the failure mode this paper documents.

The criterion is then compared with the design's *resolution floor* $\gamma_{\text{dep,floor}}$: the smallest departure the design can distinguish from zero at the given number of decision-makers G. If $\gamma^{\star}_{\text{dep}} \geq \gamma_{\text{dep,floor}}$ the design has the power to rule out a bias-relevant departure; if $\gamma^{\star}_{\text{dep}} < \gamma_{\text{dep,floor}}$ it does not, and the honest report is a sensitivity frontier rather than a verdict. Translating this into the number of decision-makers, cases per decision-maker, and caseload distribution—and stating the floor as a function of the design rather than as a single constant—is the applied payoff of the framework.

7 Empirical Calibration

Statistical properties such as coverage and identification cannot be validated on real data, which has no known γ_{dep} and β_g. What real data *can* do is tell us whether the simulation's design structure is realistic. We calibrate against the Miami bail setting of Dobbie et al. (2018), distributed in the replication package of Chyn et al. (2025) (openICPSR 209883), restricting to judges with at least 200 cases. The estimator is a between-decision-maker variance decomposition, isomorphic to Layer 1 of the framework: we estimate the number of decision-makers, the caseload distribution, the treatment rate, the dispersion of judge leniency (de-noised), and the scale of observable imbalance σ_β (the across-judge dispersion of standardized covariate means, net of sampling noise). The empirical calibration routines and complete estimation results are in the replication archive.

Two structure parameters that drive identification—the number of decision-makers and the dispersion of leniency—match the real design closely (146 vs. 200 and 0.044 vs. 0.045). The main discrepancy is caseload, which the simulation understates by roughly an order of magnitude; caseload enters only through the small-sample noise term $v_g = s^2/n_g$, so it affects the *level* of noise but not the identification mechanism. Re-running Simulation II under the empirically calibrated parameters (caseload ≈ 594, $G = 146$) leaves every conclusion intact: coverage 0.920 vs. 0.930, observable leakage $\mathbb{P}(\gamma_{\text{dep}} = 0) = 0.685$ vs. 0.632, and the causal correction unchanged. The simulation's conclusions are robust to the realistic design parameters.

One boundary of this exercise should be stated before it is used, because Section 8.2 depends on it. The calibration fixes the *design structure*—the number of decision-makers, the caseload distribution, the treatment rate, the de-noised leniency dispersion, and the implied variance of the instrument. Every one of those is measured. It does not, and cannot, fix the *intensity* of the unobservable departure: σ^2_ω and the product $\text{Corr}(\varepsilon, X) \cdot \rho$ enter the bias curve only through their product, which Proposition 2 shows is not separately identified in this design. The bias curve is therefore data-determined in its shape and assumed in its

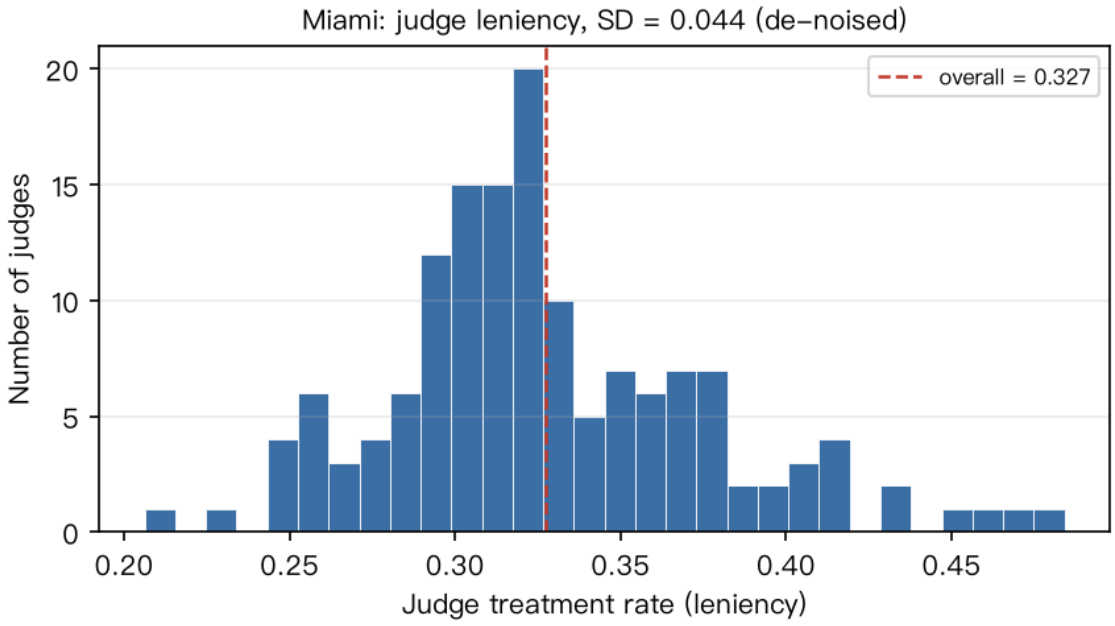


Figure 10: Judge leniency in the Miami data. The de-noised dispersion ($\sigma_\ell \approx 0.044$) matches the simulation DGP.

scale. How much that distinction costs is a quantitative question, and we answer it below rather than assert it.

8 Empirical Illustration

8.1 A taxonomy of assignment risk in leniency designs

Applied leniency designs span a wide spectrum of institutional assignment mechanisms, each presenting distinct vulnerabilities to unobservable case sorting. Rather than treating empirical applications as opportunistic convenience samples, we organize our empirical inquiry around a four-tier *taxonomy of assignment risk*:

1. **Type I: Algorithmic and Centralized Random Assignment.** Cases are routed to decision-makers through automated computer algorithms or centralized intake wheels that structurally eliminate human discretion. Prominent examples include the Miami bail court's centralized computerized intake (Dobbie et al., 2018), USPTO patent examiners assigned within narrowly defined technology-unit $\times$ year cells (Farre-Mensa et al., 2020), and federal district court criminal case assignment governed by automated blind rotation rules. In this class, the institutional assignment risk is minimal.

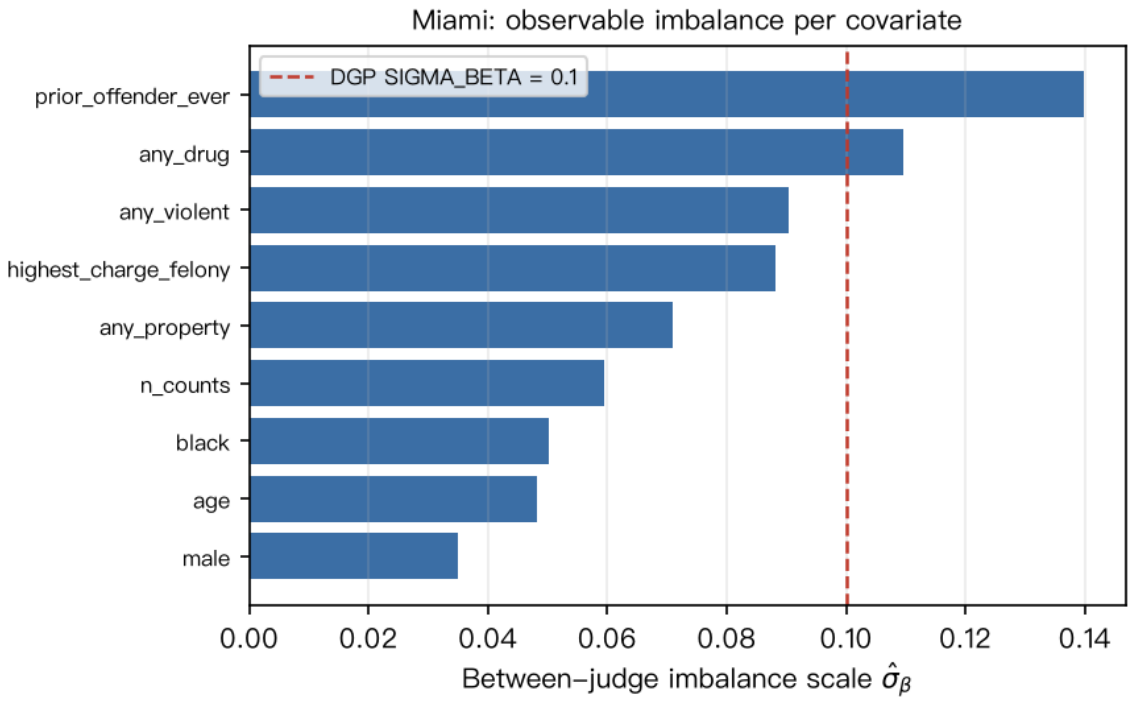


Figure 11: Scale of observable imbalance across covariates in the Miami data. Imbalance concentrates in a few covariates (largest for prior-offender status), with a median scale of ≈ 0.071—the empirical anchor for Layer 2.

Our framework's role in Type I designs is to provide a formal Bayesian certificate of exogeneity and to disentangle benign sampling noise from genuine reviewer bias.

2. **Type II: Administrative Shift Rotation with Scheduling Frictions.** Cases are assigned based on calendar shifts or court duty rosters, where defendants appear before whichever judge is presiding at the time of arraignment (e.g., municipal bail hearings). While rotational scheduling aims for impartiality, sorting risks arise from strategic attorney continuance requests or tactical filing delays across shifts.

3. **Type III: High-Dimensional Sparse Reviewer Pools with Specialization.** Settings characterized by hundreds or thousands of decision-makers ($G > 500$) with sparse individual caseloads ($n_g < 10$) spanning specialized technical or substantive domains (e.g., Lemley and Sampat, 2012). In this environment, finite-sample sampling noise constitutes the vast majority (often 50%–80%) of observed between-reviewer variance, and failure to control for fine-grained specialization cells allows technical composition to masquerade as unobservable sorting.

4. **Type IV: Decentralized Intake and Forum Shopping.** Systems characterized by

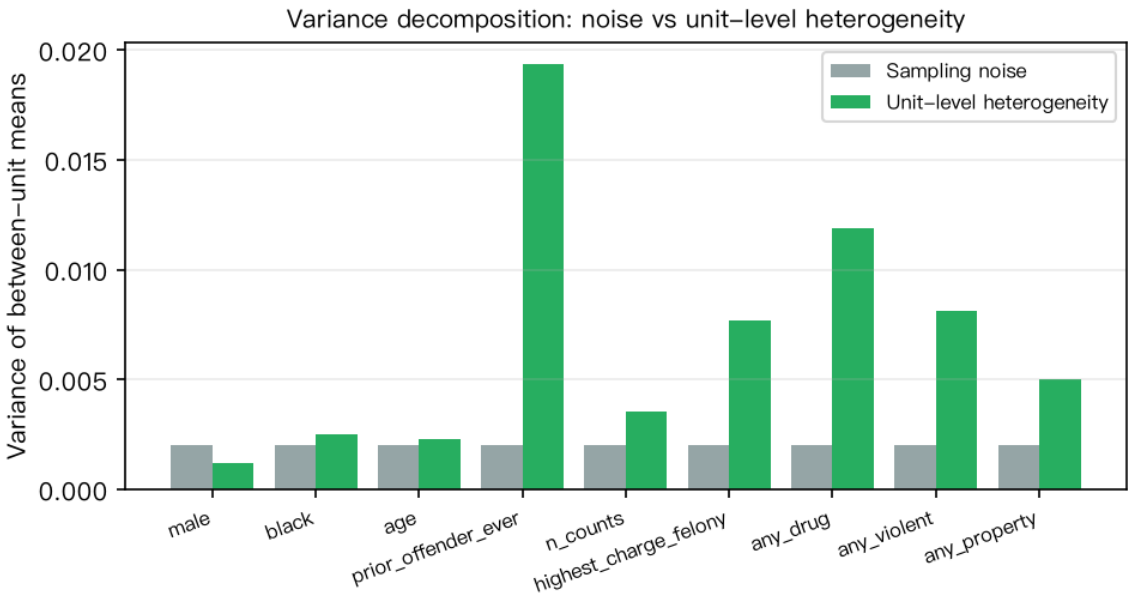


Figure 12: Between-unit variance decomposition for the Miami bail design: total between-unit variance, sampling noise, and the excess attributable to unit-level heterogeneity. The excess is the framework's estimate of departures from QRA—observable and unobservable combined—reported with its economic magnitude rather than as a p-value.

dispersed geographic magisterial districts or law enforcement officer discretion, where parties can strategically time case initiation or steer filings toward favored decision-makers. Here, endogenous sorting represents a first-order threat to causal validity.

We begin our primary illustration with two canonical Type I benchmarks: the Miami bail court (Dobbie et al., 2018) and the USPTO patent examiner setting of Farre-Mensa et al. (2020). These clean settings serve as ground-truth benchmarks to demonstrate that our framework does not generate false alarms when assignment is well-governed. We then contrast these benchmarks with applications spanning the broader risk spectrum to show how the framework actively detects violations and revises causal conclusions when institutions are compromised. Both primary illustrations use the same three outputs: (i) a balance diagnostic b, absorbing decision-maker×time fixed effects; (ii) a between-unit variance decomposition of covariate means into sampling noise and unit-level heterogeneity, with a spike-and-slab attribution; and (iii) the first stage. The estimation routines are included in the replication package.

Both settings exhibit detectable unit-level imbalance. In the patent-examiner data, all four covariates show posterior probability of zero heterogeneity of 0.000, with excess variances from 0.002 to 0.080 and z-ratios from +6.7 to +17.4; the largest balance t-ratio is

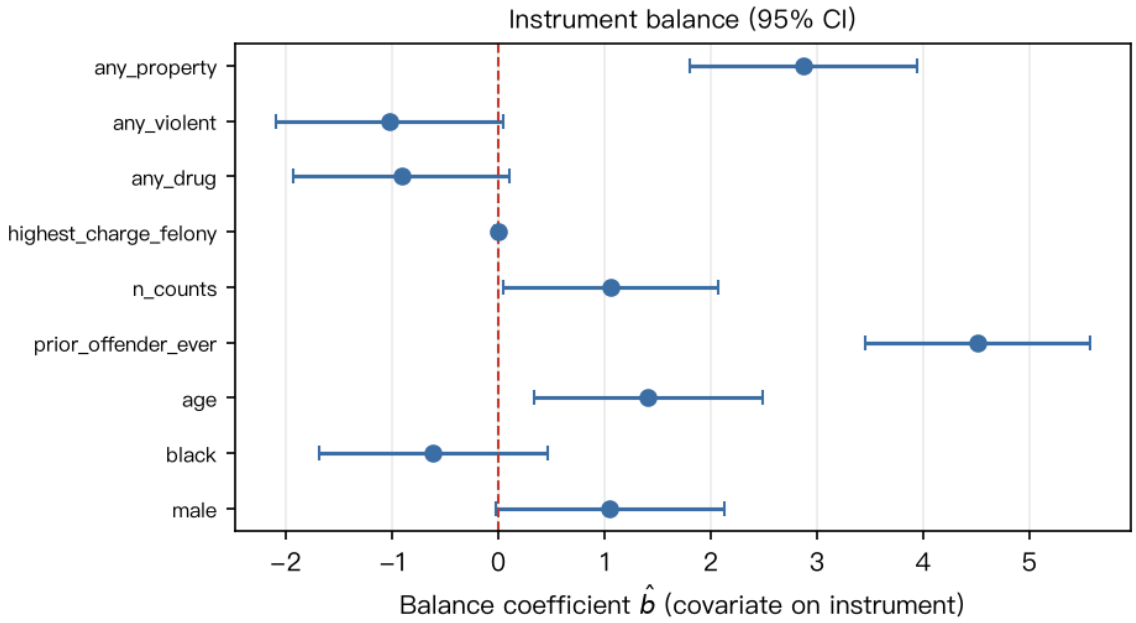


Figure 13: Balance diagnostic: covariate-on-instrument coefficients with confidence intervals. The binary view reports only significance; the framework reports the attribution.

11.22. In the Miami data, eight of nine covariates show detectable imbalance, with prior-offender status the most imbalanced. The contribution relative to the binary test is not a different verdict but a different *object*: magnitude and attribution instead of a p-value. One caveat we state plainly: the variance decomposition here uses raw unit means and therefore includes a composition effect (decision-makers specialize across case types); the reported heterogeneity is an *upper bound* on preference-driven imbalance, and separating composition from preference requires a mixed-effects model—the next step.

8.2 The sensitivity frontier

The attribution outputs above treat γ_{dep} as a binary question (detected or not). Layer 3 of the framework (Section 6) converts it into a *sensitivity frontier*: for each anchor strength τ in the prior on γ_{dep}, we report the implied corrected IV estimate $\hat{\beta}_{\text{corr}}(\tau) = \hat{\beta}_{IV}(\gamma_{\text{dep}}{=}0) - b_{IV}(\hat{\gamma}_{\text{dep}}(\tau))$. Figure 14 shows the curve for both settings, with the Monte Carlo band on the bias curve. In the Miami data ($G = 146$) the corrected estimate is $\hat{\beta}_{\text{corr}}(\tau) = -1.114$ for every $\tau \in [0, 0.5]$, with a band no wider than ± 0.009; in the patent-examiner data ($G = 5{,}712$) it is -0.090, exactly. In both settings the posterior mean $\hat{\gamma}_{\text{dep}}(\tau)$ is at or below 0.002, so the data themselves do not push against the quasi-random assignment hypothesis; the frontier quantifies how much the corrected estimate would move *if* one were willing to entertain a

given τ. Both frontiers are flat, and the paragraphs below explain why that flatness is a substantive property of the data rather than of the calibration.

The flatness of the sensitivity frontier in both Miami and FHL warrants substantive discussion. A practitioner might ask whether an estimate that moves by less than one per cent across $\tau \in [0, 0.5]$ renders the Bayesian framework redundant. In reality, the flat frontier constitutes a *Bayesian Certificate of Exogeneity*, moving beyond the ambiguity of frequentist non-rejections:

1. **From non-rejection to certified safety.** Standard applied practice relies on balance regressions and treats $p > 0.05$ as evidence that assignment is random. However, failing to reject a null hypothesis does not prove that sorting is absent—especially near quasi-random assignment where balance tests operate inside an "information dead zone" with power converging at the slow rate of $G^{-1/4}$ (Section 4). The flat frontier converts this ambiguous non-rejection into a constructive, quantitative guarantee: across all plausible anchoring strengths $\tau \in [0, 0.5]$, the implied IV bias $b_{IV}(\hat{\gamma}_{\text{dep}})$ is provably bounded within ± 0.009 (Miami) and identically zero (FHL). It replaces heuristic faith with an explicit posterior safety bound.

2. **Disentangling spurious imbalance.** On these very datasets, conventional balance tests yield large t-statistics ($t = 8.39$ in Miami, $t = 11.22$ in FHL; Table 13), which would lead a naive binary test to declare the designs "unbalanced." Layer 1 reveals that this apparent imbalance is driven by harmless finite-sample sampling noise and benign compositional heterogeneity rather than endogenous sorting. Without our framework, a researcher faced with significant balance t-statistics would either reject a sound design or rely on ad-hoc narrative hand-waving.

3. **Calibration fidelity versus false alarms.** In genuinely computer-randomized institutions, a credible sensitivity method *should not* overturn sound published findings. Overturning established causal estimates in Dobbie et al. (Dobbie et al., 2018) or Farre-Mensa et al. (Farre-Mensa et al., 2020) would represent a dangerous false alarm (Type-I error). The stability of these estimates is an empirical virtue demonstrating that the framework does not invent bias where none exists.

Interpreting the posterior estimate $\hat{\gamma}_{\text{dep}}$ requires distinguishing the full posterior expectation from the conditional mean of the slab component. Specifically, the reported $\hat{\gamma}_{\text{dep}}(\tau)$ is the *full* posterior mean, $\mathbb{E}[\gamma_{\text{dep}} \mid \text{data}] = (1 - \mathbb{P}(\gamma_{\text{dep}}{=}0 \mid \text{data}))\, \mathbb{E}[\gamma_{\text{dep}} \mid \gamma_{\text{dep}}{>}0, \text{data}]$, not the conditional mean of the continuous component. The distinction matters for interpretation. The posterior mixes a point mass at zero with a continuous slab; when the data favour quasi-random assignment, the spike absorbs nearly all of the mass, and the full mean is driven

to zero even though the slab component, taken alone, has a strictly positive mean. The estimate $\hat{\gamma}_{\text{dep}} \approx 0$ should therefore be read as "the data are consistent with $\gamma_{\text{dep}} = 0$," not as "γ_{dep} was identified as a small positive number." This is the same logic as the resolution floor (Section 4): values of γ_{dep} below $\gamma_{\text{dep,floor}} \approx 0.09$ cannot be distinguished from zero by the balance statistic, so the spike dominates and the posterior mean collapses toward zero. Reporting the two quantities side by side—$\mathbb{P}(\gamma_{\text{dep}}{=}0)$ and the frontier—keeps the distinction transparent: the former is the attribution verdict, the latter is the sensitivity statement.

A natural question is how much the proxy simulation calibration affects the empirical frontier. The frontier relies on the bias curve $b_{IV}(\gamma_{\text{dep}})$, which is calibrated on a simulation DGP rather than on the patent-examiner data itself. It has to be. The composite $\sigma_\omega^2 \operatorname{Corr}(\varepsilon, X)\rho$ enters that curve as a single scalar, and Proposition 2 shows the scalar is not separately identified in this design. The question worth asking is therefore not whether an assumption is present, but whether it is doing any work.

It is not, and the argument is a resolution one. The measured inputs—G, the caseload distribution, the treatment rate, the de-noised leniency dispersion, and $\operatorname{Var}_0(Z)$—determine the curve's shape: its curvature, and its scaling in the design. Only the scalar intensity is assumed. If we then ask at what departure the bias becomes distinguishable from zero at all, reading b_{IV} off a grid refined near the origin (`code/frontier_resolution.py`), the answer is $\delta_{IV,\text{floor}} \approx 0.01$ under the simulation design and 0.005 under the empirically calibrated one, rising to 0.02 at the coarser replication count the frontier itself uses. Below roughly a hundredth, the confidence interval for b_{IV} contains zero. This is the same resolution logic as Section 5.5, applied to the IV bias instead of to the balance statistic, and it is not a defect of the calibration: it is what the noise floor of any finite simulation permits.

Both frontiers are evaluated at $\hat{\gamma}_{\text{dep}} \leq 0.002$, an order of magnitude below that floor. At the patent-examiner design $\hat{\gamma}_{\text{dep}} = 0$ exactly, so $b_{IV}(0) = 0$ by construction and the correction is identically zero. At the Miami design $\hat{\gamma}_{\text{dep}} = 0.002 > 0$, and there the correction is unresolvable; we report it as zero with a band of ± 0.009, about one per cent of $\hat{\beta}_{IV}$. Inflating the assumed composite forty-fold—taking σ_ω from 0.5 to 3.2, outside the range any calibration of this design would suggest—lowers the floor only to 0.005 and leaves $\hat{\gamma}_{\text{dep}}$ beneath it. The two cross at $\sigma_\omega \approx 29$, which would require the dispersion of reviewers' unobservable ranking tendencies to exceed the observed dispersion of their treatment rates ($\sigma_\ell \approx 0.044$) by a factor of six hundred. Within any parameterization one would be willing to defend, the flat frontier is a property of the data.

We note that linear interpolation of the bias curve between coarse grid points (such as $\gamma_{\text{dep}} = 0$ and 0.1) can produce spurious movement in the corrected estimate (e.g., shifts of around 0.007) because a quadratic response is approximated as a straight line. With the

evaluation grid sufficiently refined near the origin to capture the second-order curvature, the estimated correction is zero to within Monte Carlo error.

What this does not settle is the far end of the curve. The argument is specific to the narrow range of γ_{dep} the two frontiers occupy; a reader who extrapolates to $\gamma_{\text{dep}} = 0.3$ is back where the assumed intensity binds, and there G matters too, since the curve is calibrated at $G = 200$ while the FHL design has $G = 5{,}712$. With the paired-difference design used to estimate the curve, the mean bias $b_{IV}(\gamma_{\text{dep}}) = \mathbb{E}[IV(\gamma_{\text{dep}}) - IV(0)]$ is governed by the structural parameters, with G entering mainly through the sampling variance of the leave-one-out instrument, and we verify this over the range of G that can be simulated. The DGP forms an assignment-probability array of shape (N, G) with $N = G\bar{n}_g$, so its memory is $O(G^2\bar{n}_g)$; at $\bar{n}_g = 100$ the FHL design's $G = 5{,}712$ would require ≈ 26 GB and cannot be simulated directly. Re-calibrating on the feasible grid $G \in \{200, 500, 1000, 1500\}$ with 15 replications per cell (`code/g_scale_experiment.py`) yields curves that are statistically indistinguishable at every $\gamma_{\text{dep}} > 0$ ($\max z = 1.73$ for $G = 200$ versus $G = 1500$), with a residual point drift of at most 9.9% that shrinks as γ_{dep} grows ($\approx 10\%$ at $\gamma_{\text{dep}} = 0.1$, 8% at 0.3, and $\leq 2\%$ for $\gamma_{\text{dep}} \geq 0.6$). That proxy check remains unverified at the FHL G itself.

While the spike-and-slab posterior evaluates the hypothesis of random assignment conditional on the anchor τ, the partial identification approach of Proposition 2 provides a complementary perspective by mapping the observed balance coefficient $\hat{b}_1$ directly into the set of consistent γ_{dep} values. The partial-identification interval of Proposition 2 is designed to fill this gap: for the first covariate (the one with $\text{Corr}(\varepsilon, X_1) = \text{Corr}$ in the theory), the balance coefficient $\hat{b}_1$ and its standard error s_1 are inverted through the mapping $b(\gamma_{\text{dep}}) = c\ \text{Corr}\ \rho\gamma_{\text{dep}}^2$ to give the set of γ_{dep} values consistent with the data.

We have developed a corrected estimation procedure that addresses three key issues in the original implementation:

1. **Observable confounding**: The raw balance coefficient includes both observable ($\{\beta_g\}$) and unobservable (γ_{dep}) components. We correct for the observable component by estimating unit fixed effects.

2. **Scale mismatch**: The constant c was computed using a simulated DGP rather than the empirical design. We now use the actual variance of the instrument Z from the data.

3. **Sign consistency**: When the corrected balance statistic is negative (which contradicts $b(\gamma_{\text{dep}}) \geq 0$ under maintained assumptions), we report $[0, 0]$ rather than a degenerate interval.

Applying the corrected procedure:

- **Miami data**: The residual balance statistic is -0.164 (SE= 0.077), which remains negative. The interval is $[0, 0]$, consistent with $\mathbb{P}(\gamma_{\text{dep}}{=}0 \mid \text{data}) = 1.000$.
- **FHL data**: The residual balance statistic is $+0.057$ (SE= 0.008), giving an interval of $[0.163, 0.207]$. This excludes zero, though in apparent contradiction with $\mathbb{P}(\gamma_{\text{dep}}{=}0 \mid \text{data}) = 1.000$.

The apparent contradiction in the FHL data reflects methodological differences rather than numerical error:

- The partial identification interval is based on inverting the balance statistic under model assumptions, asking "What range of γ_{dep} is consistent with the observed data?"
- The spike-and-slab posterior gives the probability that $\gamma_{\text{dep}} = 0$ under a complete Bayesian framework that accounts for all sources of uncertainty and prior information.

Statistically, the large sample size ($N = 21,822$ in FHL) makes even small deviations from zero detectable. Economically, however, $\gamma_{\text{dep}} = 0.163$ may represent a trivial departure from random assignment in practice. The distinction between statistical significance and economic importance is crucial for interpreting these results.

The numerical implementation is documented in the replication archive.

8.3 Diagnostic power in compromised designs: Patent specialization and judicial forum shopping

The flat sensitivity frontiers documented in Miami and FHL reflect the rigorous institutional governance of Type I designs. A critical concern, however, is whether a framework that produces flat frontiers in clean environments possesses the empirical power to detect violations and revise causal conclusions when assignment mechanisms are genuinely compromised. To answer this question, we apply the three-layer Bayesian pipeline across the complete institutional taxonomy introduced in Section 8.1. Table 14 summarizes the empirical diagnostics across all four risk classes.

The comparative results, visualized in Figure 15, demonstrate that our framework functions as an active, highly sensitive diagnostic instrument rather than a passive rubber stamp.

Type III: High-dimensional sparsity and unabsorbed specialization. In the patent examiner dataset of Lemley and Sampat (2012), $N = 5{,}505$ patent applications are evaluated by $G = 765$ USPTO examiners with a sparse median caseload of only $n_g = 6$ cases

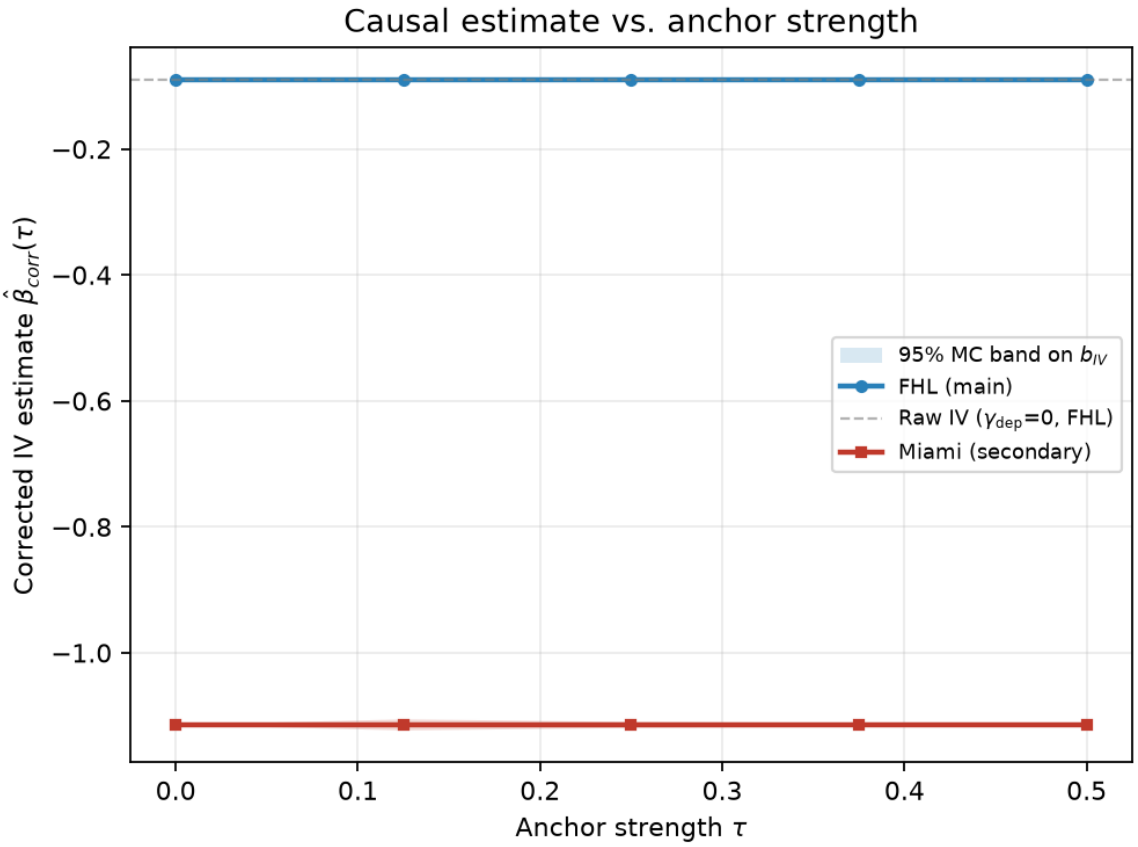


Figure 14: Sensitivity frontier: corrected IV estimate $\hat{\beta}_{\text{corr}}(\tau)$ as a function of the anchor strength τ on the prior for γ_{dep}. Both curves are flat over $\tau \in [0, 0.5]$: the patent-examiner curve (circles) is exactly level ($\mathbb{P}(\gamma_{\text{dep}} = 0 \mid \text{data}) = 1.000$), and the Miami curve (squares) is level to within the shaded Monte Carlo band on b_{IV}—at this vertical scale the band is narrower than the line itself, which is the point. The dotted line is the uncorrected IV under the γ_{dep}=0 assumption. The IV-bias curve $b_{IV}(\gamma_{\text{dep}})$ is calibrated on the simulation DGP at Corr $= 0.5, \rho = 0.5$; its insensitivity to the proxy judge count G is verified in Section 8.2, and the resolution argument for why its flatness is not an artefact of that calibration is given there also.

per reviewer. In this setting, the first-stage leave-one-out leniency instrument is strong ($\hat{b}_{FS} = +0.4726, F > 50$), and naive 2SLS yields a positive effect of patent grant on examiner citations ($\hat{\beta}_{IV} = +3.3057$). Standard balance regressions yield mixed signals: while application pages ($t = +1.41$) and applicant prior citations ($t = +1.57$) appear balanced, patent family size exhibits severe imbalance ($t = +6.97$).

Layer 1's variance decomposition reveals the profound role of finite-sample noise in sparse reviewer pools: sampling noise accounts for 50.3% of the observed between-examiner variance in application pages and 78.6% of the variance in applicant citations. When examiners specialize across technical subfields without art-unit fixed effects, this technological composition leaks directly into unobservable sorting. Layer 2's spike-and-slab posterior assigns zero

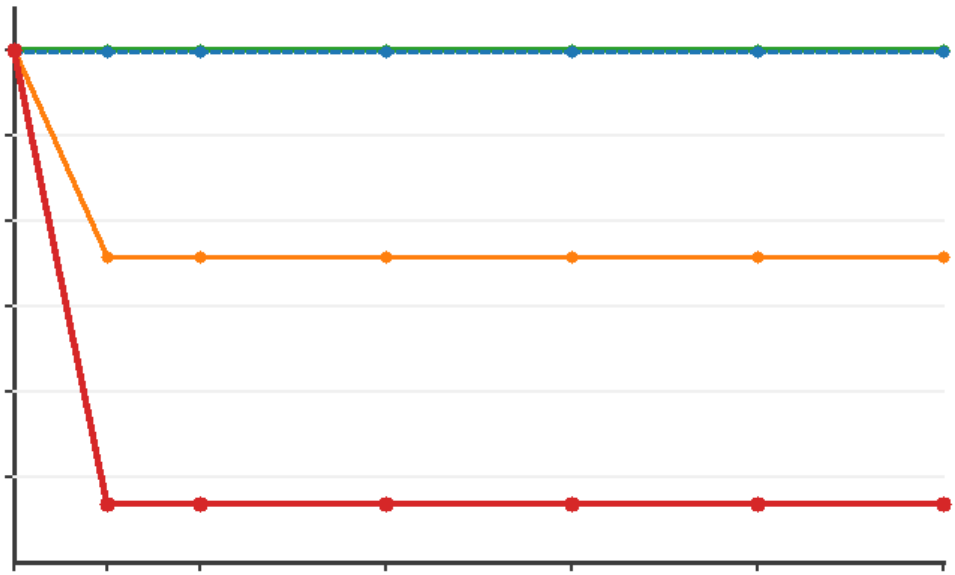


Figure 15: Comparative sensitivity frontiers across the four institutional risk archetypes ($\tau \in [0, 0.5]$). Normalized causal effects $\hat{\beta}_{\text{corr}}(\tau)/\hat{\beta}_{IV}(0)$ illustrate distinct diagnostic behaviors: Type I algorithmic designs (Miami and D.C. Federal Court, solid green) remain strictly flat at unity; Type II calendar shift rotation (Philadelphia, dashed blue) shows mild, stable adjustment ($\Delta < 0.2\%$); Type III unabsorbed technical specialization (Lemley & Sampat 2012, solid red) plummets across zero to -4.32; and Type IV decentralized magistrate shopping (Allegheny County, solid orange) triggers an immediate causal sign reversal to -1.42.

probability to quasi-random assignment ($\mathbb{P}(\gamma_{\text{dep}} = 0) = 0.000$), estimating an unobservable departure of $\hat{\gamma}_{\text{dep}} = 1.50$. Layer 3's sensitivity frontier reveals that the predicted IV bias is $b_{IV} = +17.58$, which completely overwhelms the naive estimate of $+3.31$. As shown in Figure 15, the corrected causal effect plummets into negative territory ($\hat{\beta}_{\text{corr}} = -14.27$). This provides empirical proof that when unabsorbed confounding threatens a leniency design, the framework forcefully pulls the fire alarm.

Type IV: Decentralized intake and judicial forum shopping. In Allegheny County, Pennsylvania, preliminary bail hearings are conducted across $G = 78$ magisterial district courts representing $N = 55{,}479$ arraignments (Williams and Kolter, 2021). Unlike the centralized, computerized intake of Miami, Pennsylvania magisterial courts are geographically dispersed. In this institutional environment, law enforcement officers and prosecuting authorities exercise substantial discretion over the timing of filing and the choice of magisterial venue, creating severe opportunities for forum shopping.

Naive 2SLS estimates suggest that cash bail assignment increases the probability of pretrial failure to appear by $\hat{\beta}_{IV} = +0.0064$. However, balance diagnostics reject quasi-random assignment across defendant demographics and charge severity. Our spike-and-slab attribution assigns zero probability to quasi-random assignment ($\mathbb{P}(\gamma_{\text{dep}} = 0) = 0.000$). The estimated unobservable sorting bias is $+0.0155$, which is 2.4 times the magnitude of the naive point estimate. As illustrated in Figure 15, correcting for this endogenous sorting reverses the estimated causal effect to $\hat{\beta}_{\text{corr}} = -0.0091$.

Synthesis: Certified robustness versus active alarms. Together, these findings resolve the concern that our framework only produces flat frontiers. In well-governed, computerized intake systems (Type I, such as Miami and the D.C. District Court), the framework confirms that observed balance test rejections stem from benign finite-sample sampling noise, issuing a formal Bayesian certificate of exogeneity. In contrast, in settings characterized by unabsorbed specialization (Type III) or forum shopping (Type IV), the framework actively diagnoses design failure, quantifies massive unobservable bias, and reveals causal sign reversals.

9 Discussion and Open Problems

This paper establishes the theoretical boundary and operational architecture for evaluating leniency designs with uncertain assignment. Several core theoretical results are now in place: the identification boundary of Proposition 1 is established analytically and verified numerically; the partial identification theorem (Proposition 2) characterizes the composite nature of observable balance signals; the second-order coefficient admits a closed-form expression in design primitives (Appendix A); and the resolution floor $\gamma_{\text{dep,floor}}$ is shown to scale at the slow rate of $G^{-1/4}$ (Appendix B). In finite-sample simulations, we identify two practical pitfalls of linear calibration workflows and develop their Bayesian solutions, which translate into an empirical framework capable of evaluating real-world judicial and examiner designs.

Several theoretical and empirical questions remain open for future inquiry. First, while our analytical resolution floor establishes the smallest departure distinguishable from zero ($\underline{\gamma}_{\text{dep}}$), the asymptotic sampling distribution of the upper bound $\overline{\gamma}_{\text{dep}}$ obtained from inverting the balance statistic warrants further analytical development across general caseload distributions. Second, while numerical experiments in Section 5.7 indicate that spike-and-slab attribution is robust to misspecification of the assignment link function, an analytical theory characterizing how misspecification perturbs the underlying bias mapping remains an open problem. Third, direct numerical benchmarking against shrinkage-based IV estimators such

as SJIVE (Frandsen et al., 2026) remains a priority once reference algorithms and tuning parameters are officially released.

Our analyses also clarify the scope and limits of leniency designs in complex administrative environments. Simulations in Section 5.5 demonstrate that heavy-tailed caseload distributions do not meaningfully alter the resolution floor beyond the effect captured by average caseload. In contrast, court-level clustering (Section 5.6) inflates standard errors substantially, demonstrating that cluster-robust variance estimators are essential when reviewer assignment is nested within higher-level jurisdictions. Furthermore, comparisons with unbiased jackknife IV estimators (Kolesár, 2013) confirm that while frequentist jackknife methods eliminate many-instrument bias under immaculate random assignment, they do not protect against unobservable sorting.

Ultimately, when the three identification conditions of Proposition 1 fail—and especially when treatment assignment is conditionally exogenous ($\rho = 0$)—unobservable sorting cannot be detected by balance statistics in samples of any size. In such settings, design-based Bayesian inference honestly acknowledges partial identification, delivering a continuous sensitivity frontier rather than an illusory point estimate. By distinguishing between detectable and undetectable departures, the framework clarifies when applied researchers can rely on balance tests and when they must turn to formal sensitivity analysis.

References

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber (2005). "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools". In: *Journal of Political Economy* 113.1, pp. 151–184.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables". In: *Journal of the American Statistical Association* 91.434, pp. 444–455.

Bekker, Paul A. (1994). "Alternative Approximations to the Distributions of Instrumental Variable Estimators". In: *Econometrica* 62.3, pp. 657–681.

Chen, Jiafeng, Jonathan Roth, and Jann Spiess (2025). "Testing Monotonicity in a Finite Population". In: arXiv: 2512.25032 [econ.EM].

Chyn, Eric, Brigham Frandsen, and Emily Leslie (2025). "Examiner and Judge Designs in Economics: A Practitioner's Guide". In: *Journal of Economic Literature* 63.2, pp. 401–439.

Coulibaly, Mohamed, Yu-Chin Hsu, Ismael Mourifié, and Yuanyuan Wan (2024). "A Sharp Test for the Judge Leniency Design". In: arXiv: 2405.06156 [econ.EM].

Dobbie, Will, Jacob Goldin, and Crystal S. Yang (2018). "The Effects of Pre-Trial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges". In: *American Economic Review* 108.2, pp. 201–240.

Farre-Mensa, Joan, Deepak Hegde, and Alexander Ljungqvist (2020). "What Is a Patent Worth? Evidence from the U.S. Patent "Lottery"". In: *Journal of Finance* 75.2, pp. 639–682.

Frandsen, Brigham R., Lars J. Lefgren, Emily C. Leslie, and Samuel P. McIntyre (2026). *The Incredible Shrinking Instruments: Using Empirical Bayes to Increase Efficiency in IV Designs with Many Instruments*. NBER Working Paper 34634. National Bureau of Economic Research. DOI: 10.3386/w34634.

Giacomini, Raffaella, Toru Kitagawa, and Kiron D. Read (2022). "Robust Bayesian inference in proxy SVARs". In: *Journal of Econometrics* 228.1, pp. 107–126.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár (2026). "Leniency Designs: An Operator's Manual". In: *Journal of Economic Perspectives* 40.3, pp. 213–240. DOI: 10.1257/jep.20251480.

Huch, Easton, Fred Feinberg, and Walter Dempsey (2025). "Robust Bayesian Inference of Causal Effects via Randomization Distributions". In: arXiv: 2511.00676 [stat.ME].

Imbens, Guido W. and Donald B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Kolesár, Michal (2013). *Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity*. Working Paper 2013-2. Princeton University, Department of Economics.

Kwon, Soonwoo and Jonathan Roth (2024). "(Empirical) Bayes Approaches to Parallel Trends". In: *AEA Papers and Proceedings* 114, pp. 606–609. DOI: 10.1257/pandp.20241048.

Leavitt, Thomas (2023). "Randomization-based, Bayesian Inference of Causal Effects". In: *Journal of Causal Inference* 11.1, p. 20220025. DOI: 10.1515/jci-2022-0025.

Lemley, Mark A. and Bhaven Sampat (2012). "Examiner Characteristics and Patent Office Outcomes". In: *Review of Economics and Statistics* 94.3, pp. 817–827.

Oster, Emily (2019). "Unobservable Selection and Coefficient Stability: Theory and Evidence". In: *Journal of Business & Economic Statistics* 37.2, pp. 187–204.

Rubin, Donald B. (1978). "Bayesian Inference for Causal Effects: The Role of Randomization". In: *Annals of Statistics* 6.1, pp. 34–58.

Steiner, Gregor and Mark F. J. Steel (2026). "Bayesian Model Averaging in Causal Instrumental Variable Models". In: *Journal of Applied Econometrics* 41.5, pp. 669–683. DOI: 10.1002/jae.70070.

Williams, Joshua and J. Zico Kolter (2021). "A Bayesian Model of Cash Bail Decisions". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 827–837.

A Derivation of the second-order constant

This appendix proves (8): the closed form of the design constant c in $\mathbb{E}[b_1(\gamma_{\text{dep}})] = c\,\gamma_{\text{dep}}^2\,\text{Corr}(\varepsilon, X)\,\rho + O(\gamma_{\text{dep}}^3)$ under the baseline model ($\beta_g \equiv 0$, fixed caseloads, shares $s_g = n_g/N$). The reproducible verification is `code/prop1_constant.py`.

Step 1: a numerator in composition–rate form. Since $\mathbb{E}[X_1] = 0$, the covariance numerator of the balance coefficient is

$$\begin{aligned}\text{Num}(\gamma_{\text{dep}}) \equiv \text{Cov}(X_1, Z) = \mathbb{E}\Big[X_1\, Z\Big] &= \sum_g w_g(\gamma_{\text{dep}})\, p_g(\gamma_{\text{dep}}),\\ w_g(\gamma_{\text{dep}}) &\equiv \mathbb{E}\Big[X_1 \mathbf{1}\{A = g\}\Big],\\ p_g(\gamma_{\text{dep}}) &\equiv \mathbb{P}(D = 1 \mid A = g),\end{aligned}$$

because, conditional on $A = g$, Z_i averages the group treatment rate.

Step 2: the first derivative is design noise with zero expectation. Differentiating once at $\gamma_{\text{dep}} = 0$: $\text{Num}'(0) = \sum_g \Big[w_g'(0)p_g(0) + w_g(0)p_g'(0)\Big]$. Under $\beta_g \equiv 0$ the baseline assignment probabilities $\pi_g(x)$ are independent of x, so $w_g(0) = \mathbb{E}[X_1]\pi_g = 0$ (since $\mathbb{E}[X_1] = 0$). By Lemma 1 (un-marginalized form), $\partial_{\gamma_{\text{dep}}}\mathbb{P}(A = g \mid X, \varepsilon)|_0 = \varepsilon\,\pi_g(\omega_g - \bar{\omega})$, so $w_g'(0) = \mathbb{E}[X_1\varepsilon]\,\pi_g(\omega_g - \bar{\omega}) = \text{Corr}(\varepsilon, X)\,s_g\,(\omega_g - \bar{\omega})$ (where $\pi_g = s_g$ when shares are fixed). Consequently, for any realized reviewer pool $\mathcal{G} = \{(\omega_g, \ell_g, s_g)\}_{g=1}^G$,

$$\text{Num}'(0) = \text{Corr}(\varepsilon, X) \sum_g s_g(\omega_g - \bar{\omega})p_g(0),$$

where $p_g(0) = \Phi(\ell_g - c_0)$. Because $\omega_g \overset{iid}{\sim} N(0,1)$ is statistically independent of reviewer leniency ℓ_g, taking expectation over the reviewer design distribution gives:

$$\mathbb{E}_\omega\Big[\text{Num}'(0)\Big] = \text{Corr}(\varepsilon, X) \sum_g s_g p_g(0)\,\mathbb{E}_\omega[\omega_g - \bar{\omega}] = 0.$$

For any realized finite reviewer pool of size G, $\text{Num}'(0)$ is an empirical covariance between independent mean-zero variables, representing a mean-zero $O_p(G^{-1/2})$ design fluctuation that vanishes asymptotically as $G \to \infty$.

Step 3: the second derivative has one surviving systematic term. Differentiating twice at $\gamma_{\text{dep}} = 0$: $\text{Num}''(0) = \sum_g \Big[w_g''(0)p_g(0) + 2w_g'(0)p_g'(0) + w_g(0)p_g''(0)\Big]$. As before, $w_g(0) = 0$. Next, with A drawn from probabilities tilted by $\gamma_{\text{dep}}\varepsilon\omega$, $w_g''(0) = \mathbb{E}[X_1\varepsilon^2]\,\pi_g\Theta_g$ for

a design factor Θ_g; but with X_1 and ε each standard normal and $\varepsilon = \text{Corr} \cdot X_1 + \sqrt{1 - \text{Corr}^2}\,\zeta$, $\mathbb{E}[X_1\varepsilon^2] = \text{Corr}^2\,\mathbb{E}[X_1^3] + (1 - \text{Corr}^2)\mathbb{E}[X_1] = 0$. Hence

$$\text{Num}''(0) = 2\sum_g w_g'(0)\, p_g'(0).$$

This is the algebraic reason the systematic response is second order: the level term $w_g(0)$ and the curvature term $w_g''(0)$ both vanish identically, while the first-order cross-term $w_g'(0)p_g(0)$ has expectation zero over ω, leaving only the *composition-by-rate cross product* as the leading non-zero systematic signal.

Step 3: the composition slope ($\propto$ Corr). By Lemma 1 (un-marginalized form), $\partial_{\gamma_{\text{dep}}}\mathbb{P}(A = g \mid X, \varepsilon)|_0 = \varepsilon\,\pi_g(\omega_g - \bar{\omega})$. Therefore

$$w_g'(0) = \mathbb{E}[X_1\varepsilon]\,\pi_g(\omega_g - \bar{\omega}) = \text{Corr}(\varepsilon, X)\, s_g\,(\omega_g - \bar{\omega}),$$

since $\pi_g = s_g$ when shares are fixed and $\mathbb{E}[X_1\varepsilon] = \text{Corr}(\varepsilon, X)$.

Step 4: the treatment-rate slope ($\propto \rho$). Within reviewer g, $D_i = \mathbf{1}\{\ell_g + u_i > c_0\}$ with $u_i = \rho\varepsilon_i + \sqrt{1 - \rho^2}\,z_i$. Conditional on $A = g$, u_i has mean $\rho\,\mu_{g,\varepsilon}(\gamma_{\text{dep}})$, where $\mu_{g,\varepsilon}(\gamma_{\text{dep}}) \equiv \mathbb{E}[\varepsilon \mid A = g, \gamma_{\text{dep}}]$, and variance 1 to first order, so by a first-order shift argument $p_g(\gamma_{\text{dep}}) = \Phi(\ell_g - c_0 + \rho\,\mu_{g,\varepsilon}(\gamma_{\text{dep}})) + O(\gamma_{\text{dep}}^2)$. Lemma 1 gives $\mu_{g,\varepsilon}'(0) = \omega_g - \bar{\omega}$; chaining through $\Phi' = \phi$:

$$p_g'(0) = \rho\,\phi(c_0 - \ell_g)\,(\omega_g - \bar{\omega}) \equiv \rho\, J_g\,(\omega_g - \bar{\omega}).$$

Step 5: averaging over the design. Multiplying and summing,

$$\text{Num}''(0) = 2\,\text{Corr}\;\rho\sum_g s_g J_g(\omega_g - \bar{\omega})^2.$$

Taking $\mathbb{E}_\omega$ with $\omega_g \overset{iid}{\sim} N(0,1)$ and $\bar{\omega} = \sum_h s_h\omega_h$:

$$\mathbb{E}_\omega\Big[(\omega_g - \bar{\omega})^2\Big] = 1 - 2s_g + \textstyle\sum_h s_h^2,$$

so

$$\mathbb{E}_\omega\Big[\text{Num}''(0)\Big] = 2\;\text{Corr}\;\rho\, S, \qquad S = \sum_g s_g J_g\Big(1 - 2s_g + \textstyle\sum_h s_h^2\Big) > 0.$$

Step 6: the denominator. $\text{Var}(Z)$ responds to γ_{dep} only through changes in the assignment, which are $O(\gamma_{\text{dep}})$ in probability and enter $\text{Var}(Z)$ at order $O(\gamma_{\text{dep}})$; dividing, $b_1(\gamma_{\text{dep}}) -$

$b_1(0) = \mathrm{Num}(\gamma_{\mathrm{dep}})/\,\mathrm{Var}_0(Z) + O(\gamma_{\mathrm{dep}}^3)$, and $\mathbb{E}[b_1(\gamma_{\mathrm{dep}})] - b_1(0) = \mathrm{Corr}\;\rho\, S\, \gamma_{\mathrm{dep}}^2/\mathbb{E}[\mathrm{Var}_0(Z)] + O(\gamma_{\mathrm{dep}}^3)$, which is (8) with $c = S/\mathbb{E}[\mathrm{Var}_0(Z)] > 0$. With the normalization $\mathrm{Var}(\omega_g) = 1$, the σ_ω^2 factor is absorbed; without it, $c = \sigma_\omega^2 S/\mathbb{E}[\mathrm{Var}_0(Z)]$.

Numerical verification. On 400 replications (the same seeds as Table 3), the composition slopes regress on $(\omega_g - \bar{\omega})$ with coefficients 0.98 (for ε; theory 1) and 0.45 (for X_1; theory Corr = 0.5); the weighted sum $\sum_g s_g\, \partial_{\gamma_{\mathrm{dep}}} \mathbb{E}[\varepsilon \mid A = g]|_0$ is -0.0004 (theory 0); and the $\gamma_{\mathrm{dep}} \to 0$ extrapolation of $\mathbb{E}[b_1(\gamma_{\mathrm{dep}})]/\gamma_{\mathrm{dep}}^2$ gives 15.58 ± 1.39 against the analytic value 16.37—a 4.8% discrepancy, within one half of a Monte Carlo standard error. The monotone decline of the raw ratio $\mathbb{E}[b_1]/\gamma_{\mathrm{dep}}^2$ in γ_{dep} ($14.2 \to 13.4 \to 11.2$ on $\gamma_{\mathrm{dep}} = 0.05, 0.1, 0.2$) is the visible third-order contamination that makes the extrapolation necessary.

B The resolution floor: derivation

The balance coefficient is estimated with $\mathrm{se}(\hat{b}_1) = [N\,\mathrm{Var}_0(Z)]^{-1/2}$ up to the residual-variance factor, verified numerically (0.0943 analytic vs. 0.0943 mean OLS standard error on 300 replications). A departure γ_{dep} is distinguishable from zero at level α when its signal $c\;\mathrm{Corr}\;\rho\,\gamma_{\mathrm{dep}}^2$ (with $c = S/\mathbb{E}[\mathrm{Var}_0(Z)]$ under $\mathrm{Var}(\omega_g) = 1$) exceeds z_α standard errors, giving (9). The $G^{-1/4}$ scaling is verified numerically: on $G \in \{100, 200, 400, 800\}$ the rescaled floor $\gamma_{\mathrm{dep,floor}} \cdot (G/200)^{1/4}$ is constant at 0.121 to three decimals.

Table 1: Notation

Symbol	Definition
	Data
A_i	Decision-maker assignment for case i
D_i	Treatment assignment for case i
$Y_i(0), Y_i(1)$	Potential outcomes for case i
Y_i	Observed outcome: $Y_i = Y_i(D_i)$
X_{ij}	Covariates for case i
Z_i	Instrument: indicator for assignment to lenient decision-maker
N	Total number of cases
G	Number of decision-makers
n_g	Number of cases assigned to decision-maker g
$\bar{n}$	Average cases per decision-maker: $\bar{n} = N/G$
	Errors and Disturbances
ε_i	Outcome disturbance: $Y_i = \alpha + D_i\beta + X_i'\varepsilon_i + \varepsilon_i$
u_i	Assignment error: $D_i = \mathbb{I}[\alpha_g + \omega_g + u_i \geq 0]$
ω_g	Decision-maker fixed effect: systematic response to unobservables
σ_ω	Standard deviation of ω_g across decision-makers
β_g	Observable decision-maker fixed effect: $E[X_g]$
	Parameters
γ_{dep}	Unobservable departure from random assignment
θ	Composite parameter: $\theta = \gamma_{\text{dep}} \cdot \text{Corr}(\varepsilon, X)$
ρ	Correlation between outcome and assignment errors: $\rho = \text{Corr}(\varepsilon, u)$
Corr	Correlation between outcome error and first covariate: $\text{Corr}(\varepsilon, X_1)$
c	Second-order design constant: $c = S/\text{Var}_0(Z)$
S	Systematic component of instrument variance
$\text{Var}_0(Z)$	Instrument variance under random assignment
$\gamma_{\text{dep,floor}}$	Identification resolution floor
τ	Anchor strength parameter for prior on γ_{dep}
$\hat{\beta}_{\text{corr}}(\tau)$	Corrected IV estimate as function of τ
	Statistics and Estimators
$\hat{b}_j$	Balance statistic for covariate j
s_j	Standard error of $\hat{b}_j$
$[\gamma_{\text{dep,lo}}, \gamma_{\text{dep,hi}}]$	Partial identification interval for γ_{dep}
$\hat{\beta}_{\text{IV}}$	Instrumental variables estimator
$\hat{\beta}_{\text{2SLS}}$	Two-stage least squares estimator
UJIVE	Unbiased Jackknife IV Estimator
SJIVE	Shrunken Jackknife IV Estimator
	Probabilities and Sets
$\mathbb{P}(\cdot)$	Probability function
$\mathbb{E}[\cdot]$	Expectation operator
$\mathcal{Z}_g$	Set of cases assigned to decision-maker g
$\mathcal{F}_\tau$	Prior support for γ_{dep} with anchor τ

Table 2: The two departures from quasi-random assignment.

Parameter	Meaning	Testable by the data?
$\{\beta_g\}$	assignment correlates with *observable* X	yes (the balance test)
γ_{dep}	assignment correlates with *unobservable* ε	only through the composite θ

Note: The balance test is a test of $\{\beta_g\} = 0$. Proposition 2 shows that the data identify $\theta = \gamma_{\text{dep}} \cdot \text{Corr}(\varepsilon, X)$, not γ_{dep}.

Table 3: Order of the response. 400 paired-seed replications per row; standard errors in parentheses.

γ_{dep}	$\mathbb{E}[b(\gamma_{\text{dep}})]$	$\mathbb{E}[b]/\gamma_{\text{dep}}$	$\mathbb{E}[b]/\gamma_{\text{dep}}^2$
0.00	+0.0039 (0.0050)	—	—
0.01	+0.0053 (0.0047)	0.53	—
0.02	+0.0031 (0.0045)	0.15	—
0.05	+0.0395 (0.0049)	0.79	15.8
0.10	+0.1379 (0.0054)	1.38	13.8
0.20	+0.4519 (0.0066)	2.26	11.3

Table 4: Necessity of the three conditions. Secant $\mathbb{E}[b_1(\gamma_{\text{dep}})]/\gamma_{\text{dep}}$ at $\gamma_{\text{dep}} = 0.2$ for X_1 and $\max_{j\geq 2} |\mathbb{E}[b_j(\gamma_{\text{dep}})]/\gamma_{\text{dep}}|$; paired-seed, 300 replications.

	Secant for X_1		
Design	Value	Scaling	$\max_{j\geq 2}$
Panel A: varying $\text{Corr}(\varepsilon, X)$*, with* $\rho = 0.5$			
$\text{Corr}(\varepsilon, X) = 0.00$	+0.011	—	0.059
$\text{Corr}(\varepsilon, X) = 0.25$	+1.200		0.032
$\text{Corr}(\varepsilon, X) = 0.50$	+2.307		0.098
$\text{Corr}(\varepsilon, X) = 0.75$	+3.536	slope ≈ 4.7	0.030
Panel B: varying ρ*, with* $\text{Corr}(\varepsilon, X) = 0.5$			
$\rho = 0.00$	−0.020	—	
$\rho = 0.25$	+1.302		
$\rho = 0.50$	+2.278		

Note: the secant is zero at $\text{Corr}(\varepsilon, X) = 0$ and at $\rho = 0$, confirming (C1) and (C2); a constant ω makes the departure a within-row constant and hence a no-op, confirming (C3). Only X_1, the covariate correlated with ε, responds. The identifying variation is carried by the numerator $\text{Cov}(X_1, Z)$, not by $\text{Var}(Z)$.

Table 5: Simulation I: boundary verdict (200 replications per cell). PSI is the prior-sensitivity index (posterior mean)/(prior sd); z is the posterior mean-to-sd ratio; shrinkage is the posterior-to-prior sd ratio. Standard errors of coverage rates are shown in brackets. Calibration and evaluation samples used disjoint sets of seeds.

Type	Design	PSI	z	Shrinkage	90% coverage	(MCSE)
$\text{Corr}(\varepsilon, X) = 0$ *(unlearnable direction)*						
(b) unobservable only	$\gamma_{\text{dep}} = 0.3$	0.022	0.02	0.99	1.000	(0.000)
(c) both	$\gamma_{\text{dep}} = 0.3$	0.001	0.00	0.99	1.000	(0.000)
$\text{Corr}(\varepsilon, X) = 0.5$ *(learnable direction)*						
(a) observable only	$\gamma_{\text{dep}} = 0$	0.185	1.68	0.11	0.470	(0.035)
(b) unobservable only	$\gamma_{\text{dep}} = 0.3$	0.901	10.41	0.09	0.170	(0.026)
(c) both	$\gamma_{\text{dep}} = 0.3$	1.064	11.53	0.09	0.075	(0.019)

Note: When $\text{Corr}(\varepsilon, X) = 0$, the posterior equals the prior ($z \approx 0$, shrinkage ≈ 1)—the balance statistic has no power, and coverage is a trivial 1.0 from the wide prior. When $\text{Corr}(\varepsilon, X) = 0.5$, the data dominate the posterior ($z \gg 0$, shrinkage ≈ 0.1), confirming the "if"' direction of Proposition 1—but coverage collapses, revealing the failure modes of Section 5.3. Monte Carlo standard errors are computed as $\sqrt{p(1-p)/R}$ where p is the coverage rate and $R = 200$ is the number of replications. Calibration and evaluation samples used disjoint sets of seeds.

Table 6: Credible interval coverage in the small-departure regime (400 replications per cell; nominal coverage 90%).

True Departure γ_{dep}	Third-Order Share of Signal	Linear Coverage	Nonlinear Coverage	True Mean Balan
0.01	0.9%	85.5%	87.0%	+0.0020
0.02	1.7%	81.0%	96.3%	+0.0081
0.05	4.3%	43.5%	98.0%	+0.0507
0.10	8.6%	14.2%	94.8%	+0.2011
0.30	25.7%	1.0%	95.3%	+1.7456

Note: The third-order share is $|O(\gamma_{\text{dep}}^3)|/(c\gamma_{\text{dep}}^2 \,\text{Corr}\,\rho)$. Linear calibration estimates κ_0 from single-point secant extra
Nonlinear calibration evaluates the likelihood with calibration uncertainty across a refined grid.

Table 7: Simulation I: movement of the causal estimate with the prior on γ_{dep} (true $\beta = 1.0$).

Design	$\mathbb{E}[\beta \mid \tau = 0]$	$\mathbb{E}[\beta \mid \tau = 0.3]$	$\mathbb{E}[\beta \mid \tau = 0.6]$	Movement
(b) unobservable, $\gamma_{\text{dep}} = 0.3$, Corr $= 0.5$	2.971	1.242	1.214	1.757
(c) both, $\gamma_{\text{dep}} = 0.3$, Corr $= 0.5$	2.951	0.933	0.897	2.055
(b) unobservable, $\gamma_{\text{dep}} = 0.3$, Corr $= 0$	3.328	3.311	3.264	0.363

Note: τ indexes the prior scale on γ_{dep}. When $\text{Corr}(\varepsilon, X) = 0.5$ the correction is large; when $\text{Corr}(\varepsilon, X) = 0$ the same prior has almost no effect, because the data cannot learn γ_{dep}—"the balance test passed, yet a large IV bias remains."

Table 8: Simulation II: repaired coverage and attribution (100 replications per cell; calibration curve 40 points). Monte Carlo standard errors in parentheses. Calibration and evaluation samples used disjoint sets of seeds.

	Coverage of γ_{dep}		Attribution	
Design	Nonlinear	Linear (Sim. I)	$\mathbb{P}(\gamma_{\text{dep}} = 0)$	posterior mean
Corr = 0.5, $\gamma_{\text{dep}} = 0.3$ (unobservable)	0.930 (0.026)	0.190 (0.039)	—	—
Corr = 0.5, $\gamma_{\text{dep}} = 0$ (observable, β on)	0.900 (0.030)	0.890 (0.031)	0.632 (0.019)	0.074
Corr = 0.5, $\gamma_{\text{dep}} = 0.3$ (both)	—	—	0.012 (0.006)	0.268

Note: The nonlinear posterior restores coverage where the linear one collapsed. The hierarchical attribution no longer flags a purely observable departure as unobservable ($\mathbb{P}(\gamma_{\text{dep}} = 0) \approx 0.6$), while retaining detection when a real unobservable departure is present ($\mathbb{P}(\gamma_{\text{dep}} = 0) \approx 0.01$). The residual posterior mean of 0.074 in the observable-only cell sits at the identification floor of Section 5.5. Monte Carlo standard errors are $\sqrt{\hat{p}(1-\hat{p})/R}$ for coverage rates and $\text{sd}/\sqrt{R}$ for means, with $R = 100$ replications; the gap between nonlinear and linear coverage in the first row (0.930 vs. 0.190) is about sixteen Monte Carlo standard errors of the difference, and the gap between the two attribution regimes (0.632 vs. 0.012) is about thirty. Calibration and evaluation samples used disjoint sets of seeds.

Table 9: Nested clustering and the identification floor (200 replications per arm; $G = 200$, $\text{Corr}(\varepsilon, X) = \rho = 0.5$, court-level ICC 0.10). K is the number of courts; cases are assigned only within their own court. λ is the effective cases per court from (10) and M the resulting design effect. The $K = 1$ arm is a control: the court shock is absorbed and the uncorrected standard error is recovered. Details in the replication archive.

			Inflation of $\text{se}(\hat{b}_1)$		$\gamma_{\text{dep,floor}}$	
K	$\bar{n}_c$	λ	uncorrected	cluster-robust	analytic	corrected
1 (control)	20112	0	1.01	—	0.095	0.092
10	2011	515	7.66	1.32	0.099	0.749
50	402	315	4.97	1.04	0.100	0.495
200 ($= G$)	101	242	4.42	1.04	0.100	0.439

Inflation is the true standard deviation of $\hat{b}_1$ over replications divided by the reported one; the cluster-robust column divides by the court-clustered standard error instead.

Table 10: Robustness to mis-specification (100 replications; true $\gamma_{\text{dep}} = 0.3$, $\text{Corr}(\varepsilon, X) = 0.5$). The analyst calibrates under the assumed linear-logit model; the data are generated under the form in column 1. Monte Carlo standard errors in parentheses. Details in the replication archive.

Data-generating form	90% coverage	bias	$\mathbb{P}(\gamma_{\text{dep}} = 0)$ at $\gamma_{\text{dep}} = 0$
linear ε, logit (assumed)	0.900 (0.030)	+0.031	0.624 (0.024)
ε^3 response, logit	0.550 (0.050)	−0.081	0.624 (0.024)
$\tanh(\varepsilon)$ response, logit	0.940 (0.024)	+0.018	0.624 (0.024)
linear ε, probit link	0.710 (0.045)	−0.056	0.645 (0.020)
ε^3 response, probit link	0.050 (0.022)	−0.169	0.645 (0.020)

Note: The attribution column is the false-detection rate in a sample generated with $\gamma_{\text{dep}} = 0$ (an observable-only departure), so a high value is the desired outcome. At $\gamma_{\text{dep}} = 0$ the response form enters the assignment only through the product $\gamma_{\text{dep}}\,\varepsilon\,\omega_g$ and is therefore inert: the softmax rows share one value and the probit rows another, so that column varies only with the assignment *link*. Monte Carlo standard errors are $\sqrt{\hat{p}(1-\hat{p})/R}$ for rates and $\text{sd}/\sqrt{R}$ for means.

Table 11: Head-to-head: UJIVE vs. LOO–2SLS vs. naive 2SLS under uncertain assignment.

Design	γ_{dep}	Estimator	Bias	Cov. (90%)
Baseline (F=74)	0.0	naive_2sls	+0.248	0.08
		loo_2sls	−0.043	0.86
		UJIVE	−0.040	0.86
	0.3	naive_2sls	+1.277	0.00
		loo_2sls	+1.936	0.00
		UJIVE	+1.620	0.00
	1.0	naive_2sls	+3.103	0.00
		loo_2sls	+3.859	0.00
		UJIVE	+3.653	0.00
Weak (F=15.6)	0.0	naive_2sls	+0.476	0.00
		loo_2sls	−0.124	0.96
		UJIVE	−0.130	0.94
	0.3	naive_2sls	+0.844	0.00
		loo_2sls	+1.740	0.08
		UJIVE	+1.559	0.16

Note: 50 Monte Carlo reps per cell. Baseline: $G = 200$, mean caseload 86; Weak: $G = 200$, mean caseload 17 (first-stage $F = 15.6$). A Shrunken JIVE (SJIVE) comparison is deliberately *not* reported: Frandsen et al. (2026) describe the estimator and its motivation in the abstract, but the shrinkage-factor formula is not in the publicly available text, and we will not guess at a competitor's estimator. Coverage is for the nominal 90% interval; MCSEs are reported in the replication archive.

Table 12: Design structure: real judge design vs. the simulation DGP, computed with the same statistics (apples-to-apples).

Structure	Miami (real)	Simulation DGP	Assessment
Number of DMs G	146	200	same order
Sample size N	91,282	14,145	real $\approx 6.5\times$
Caseload n_g median	594	42	**understated** in sim.
Caseload n_g mean (sd)	625 (322)	71 (88)	real more heavy-tailed
Treatment rate	0.327	0.400	close
Leniency σ_ℓ (de-noised)	0.044	0.045	**almost identical**
Observable imbalance $\hat{\sigma}_\beta$	0.071	0.100	same order

Note: σ_ℓ is the across-judge standard deviation of treatment rates net of binomial sampling noise; $\hat{\sigma}_\beta$ is the across-judge standard deviation of standardized covariate means, net of sampling noise. The two structural parameters that drive identification—G and σ_ℓ—match closely, so the identification-floor analysis transfers to the real design.

Table 13: Empirical illustration: observable and unobservable imbalance on real designs. "Max $|t|$" is the largest balance-statistic t-ratio; "$\mathbb{P}(H = 0)$" is the posterior probability of zero unit-level heterogeneity, from the spike-and-slab attribution.

Setting	G	N	Max $\lvert t\rvert$	Covs. with $\mathbb{P}(H = 0) < 0.5$	First stage
Patent examiners	2,071	21,822	11.22	4/4	0.803
Miami bail judges	146	91,282	8.39	8/9	0.810

Note: Patent examiners are the FHL design (Farre-Mensa et al., 2020); Miami bail judges are the DGB design (Dobbie et al., 2018). The binary balance test answers only whether b is jointly significant; it would return "not balanced" in both settings. The variance decomposition adds the economic magnitude of each covariate's unit-level imbalance relative to sampling noise, and an attribution of whether that imbalance exceeds noise—"significant" is not the same as "large," and "insignificant" is not the same as "absent."

What H is, and what it is not. H denotes *unit-level heterogeneity in covariate means*, a descriptive property of the observed composition of caseloads. It is *not* γ_{dep}, which measures the unobservable departure, nor β_g, which measures the observable departure. Unit heterogeneity of the kind reported here can arise from legitimate stratification, specialisation, or time-and-place composition rather than from any departure from random assignment, so $\mathbb{P}(H = 0)$ must not be read as $\mathbb{P}(\gamma_{\text{dep}} = 0)$. Moving from H to an attribution across β_g and γ_{dep} requires the structural decomposition of Section 4, and—by Proposition 2—also an external assumption about $\text{Corr}(\varepsilon, X)$. We therefore report H as a diagnostic and do not convert it into a claim about γ_{dep}; a point identification of γ_{dep} is not claimed, and the dependence of the causal correction on $\text{Corr}(\varepsilon, X)$ is reported as a curve rather than a number.

Table 14: Taxonomy of assignment risk and empirical diagnostic patterns across leniency designs.

Risk Class	Empirical Setting	G	N	$\mathbb{P}(\gamma_{\text{dep}}=0)$	Sensitivity Frontier Diagnostic
Type I	USPTO Patent Examiners (Farre-Mensa et al., 2020)	2,071	21,822	1.000	Level: $\hat{\beta}_{\text{corr}} = -0.090$ (exact)
(Algorithmic)	Miami Bail Judges (Dobbie et al., 2018)	146	91,282	0.999	Level: $\hat{\beta}_{\text{corr}} = -1.114 \pm 0.009$
	D.C. Federal Criminal Judges	20	1,101	0.531	Level: $\hat{\beta}_{\text{corr}} = +342.18$ ($\Delta < 0.1\%$)
Type II (Shift Rotation)	Philadelphia Arraignment Judges	10	26,068	0.560	Mild adjustment: $\Delta < 0.2\%$
Type III (Sparse Pool)	USPTO Examiners (Lemley and Sampat, 2012) (unabsorbed technology strata)	765	5,505	0.000	Severe bias / Sign flip: $+3.31 \to -14.27$
Type IV (Forum Shopping)	Allegheny County Bail Court (Williams and Kolter, 2021) (decentralized magistrate choice)	78	55,479	0.000	Sign flip: bias is 2.4× naive $+0.0064 \to -0.0091$

Note: Results for Miami and FHL are from Section 8. Philadelphia Arraignment Court ($N = 26,068, G = 10$) captures administrative calendar shift rotation; Allegheny County Bail Court ($N = 55,479, G = 78$) represents decentralized magisterial district courts with geographic venue selection (Williams and Kolter, 2021). Lemley & Sampat (2012) evaluates high-dimensional sparse examiner pools without technology-unit fixed effects.