Multi-Margin Selection in Leniency Designs

Lonjezo Sithole
Department of Economics
University of Michigan-Ann Arbor
lsithole@umich.edu

April 12, 2026

Abstract

Judge leniency designs are widely used in empirical economics, but their conventional interpretation relies on a single-index assumption: that judges differ only in overall strictness. I develop a latent-regime framework in which judge assignment can shift treatment along multiple latent margins, so decision-makers may differ in which case features they weight, not just in how strict they are overall. Under weak assumptions, the model identifies how many latent regimes are active; under stronger shape or parametric restrictions, it can point-identify the regime-specific outcome laws and mixture weights. The conventional Wald ratio then decomposes into regime-specific treatment effects with weights that depend on which judges are compared, yielding a direct test of whether a scalar leniency interpretation is adequate. In an application to felony bail decisions, the standard scalar leniency score fails the rank-1 benchmark for Black defendants but not for Non-Black defendants, implying that the same leniency variation has different economic content across groups. For the sentencing outcome, the estimated cross-race gap in the Wald ratio is driven primarily by which decision environments the instrument activates rather than by differences in what incarceration does within a given regime.

1 Introduction

A substantial literature estimates causal effects using judge or examiner leniency designs, in which cases are quasi-randomly assigned to decision-makers who differ persistently in their propensity to treat. Canonical examples exploit quasi-random assignment of cases to judges to study sentencing, bail, and pre-trial detention outcomes (e.g. Kling, 2006; Dobbie et al., 2018; Gupta et al., 2016), as well as patent examiner assignment to estimate the value of intellectual property (Farre-Mensa et al., 2020). The identifying variation in these designs comes from the fact that different decision-makers apply systematically different thresholds or standards.

The standard econometric framework for these designs is built around a *single latent margin*. A scalar leniency index orders decision-makers from strict to lenient, the instrument shifts a common threshold in a one-dimensional latent resistance space, and the objects of interest—marginal treatment effects (MTEs), marginal outcome tests (MOTs), and discrimination diagnostics—are defined with reference to this single dimension. This framework originates in the MTE/local IV tradition (Heckman and Vytlacil, 2005) and underlies influential work on measuring discrimination (Hull, 2021; Arnold et al., 2022). Its canonical formulation in the leniency-design literature hinges crucially on the Imbens and Angrist (1994) monotonicity assumption, which in this setting implies a strong restriction on the behavior of judges: a judge that is more lenient than another in one case should be more lenient than the other in every other case.

Nevertheless, monotonicity has been widely questioned in the judge leniency literature (Mueller-Smith, 2015; Norris et al., 2021; Chan et al., 2022; Frandsen et al., 2023).[1] Recent work by Sigstad (2026) documents extensive violations among judges in panel settings. Frandsen et al. (2023) relax the strict monotonicity assumption to something weaker: average monotonicity. Despite its intuitive appeal, the average monotonicity condition also has unattractive theoretical implications (see e.g. Section 3.4 in Mogstad and Torgovitsky (2024)). This paper takes a fundamentally different perspective; it takes monotonicity violations as part of a larger and more fundamental issue: the one-dimensional ordering inherent in these empirical designs may not meaningfully describe the selection mechanism. In many settings, decision-makers differ not merely in *how strict they are overall* under a common weighting of case characteristics but also in *which case features they weight*. One judge may weight charge severity most heavily; another may respond primarily to prior criminal history or to courtroom presentation. The relevant distinction is not whether a given feature is observed or unobserved, but that judges weight the available case

[1] In fact, Imbens and Angrist (1994) cite the judge example as one setting where their monotonicity condition is potentially implausible.

information differently.

When decision-makers differ in this qualitative way, there is no natural scalar that ranks them from "strict" to "lenient"; two decision-makers can have identical overall treatment rates while treating systematically different types of cases. The selection process is intrinsically multidimensional, and the vocabulary of the single-index framework (compliers, defiers, response types, and the MTE as a function of a scalar resistance index) no longer describes the relevant objects.

This paper develops a framework for leniency designs that takes multidimensional selection as its starting point. The key idea is simple: if judges differ in *what they weight*, then different judges select cases into treatment along different latent margins. I model this by positing a finite number of latent regimes $S \in \{1, \ldots, K_0\}$, each representing a distinct configuration of the observed and unobserved case features that drive the treatment decision. The framework accommodates heterogeneity in both observed and unobserved case features. In applications that already condition on rich observables, the residual problem is precisely the one the framework targets: even within a covariate cell, decision-makers may weight the remaining case information differently, and that heterogeneity is what the latent regimes are designed to capture. Different judges may be strict in one regime and lenient in another, so their regime-specific treatment rates need not move in the same direction when we compare a stricter judge to a more lenient one.

Within each regime, the model works like a standard threshold-crossing model: once the decision environment is fixed, judge variation shifts a common threshold, and outcomes depend on treatment status but not on the judge's identity. The multidimensional structure enters through the way regimes are activated. A more lenient judge might select into treatment (e.g. release) cases that have a particular configuration of case features but *fewer* cases that have another, because leniency on one dimension can coincide with strictness on another. The framework therefore makes a different trade-off from the canonical MTE model: it allows judge assignment to shift multiple latent margins rather than a common scalar threshold, but it rules out additional selection on potential outcomes within each regime.

The paper makes four main contributions:

Identification. I show what the data can and cannot reveal about the latent regime structure, proceeding in two stages. First, under weak nonparametric assumptions, the variation in outcome distributions across judges reveals how many distinct latent decision environments are active (the effective regime dimension K_0) without requiring the researcher to specify what those environments look like. The driving observation is that judge variation changes only the *mixture weights* (who is in which regime among the treated) while leaving the regime-specific outcome distributions fixed. When enough judges

shift the mixture in independent directions, the data pin down the number of latent margins along which judges' treatment decisions differ and the rank-based diagnostics developed below are identified. Second, to go beyond the regime count and recover the latent regimes themselves, I impose additional structure. For scalar outcomes, an ordered-tail condition yields full nonparametric identification of the weights and the components: sufficiently extreme tails distinguish the latent regimes even when their supports overlap. An alternative semiparametric structure, which I adopt in the empirical illustration of this framework, exploits parametric shape restrictions on the regime-specific outcome distributions, leaving the weights, which encode the selection mechanism, unrestricted. For continuous outcomes, invoking Gaussian outcome densities within each regime, with unrestricted mixing weights across judges, allows for identification of regime-specific parameters and weights up to a permutation of labels (Proposition 3.1). For binary outcomes, Appendix B gives an analogous result under a common-index logistic class.

Wald ratio decomposition. In the standard single-index framework, the Wald ratio (reduced form divided by first stage) identifies a weighted average treatment effect among compliers, with nonnegative weights. In the multi-regime framework, the Wald ratio is instead a weighted sum of regime-specific treatment effects, where the weight on each regime reflects how much a particular judge pair moves treatment *within each regime*. When different regimes move in opposite directions (one judge releases more of one type while another releases more of a different type), the weights can be negative, and the Wald ratio becomes a net contrast across offsetting flows rather than a convex average over a well-defined complier group. I formalize this in a slope-weighted decomposition (Proposition 3.2) and define a diagnostic, *NegMass*, that measures the total magnitude of the negative weight.

Testable benchmark for scalar leniency. If all judges apply the same weighting of case features and differ only in overall strictness, then the regime-specific treatment rates move proportionally across judges. This is what I call a *rank-1 incremental structure*. Under rank-1, the Wald ratio is constant across judge pairs (Proposition 4.1). Systematic variation of the Wald ratio across different judge pairs is, therefore, an empirical signature of multimargin selection. I develop a formal test for this restriction.

Discrimination decomposition. When leniency designs are used to study discrimination—for example, comparing marginal released-defendant outcomes across racial groups—the standard approach computes Wald ratios separately by group and interprets the gap as a difference in marginal risk. In the multi-regime framework, this gap further decomposes into a *within-regime* channel (do Black and Non-Black defendants in the same decision environment have different outcomes?) and a *margin-composition* channel (does the instrument activate different regimes for different groups?). The distinction has

substantive content: the first channel reflects differential treatment within comparable decision environments, whereas the second reflects differential exposure to those environments. Such a decomposition is, by design, unavailable in single-index frameworks. Because its interpretation depends on within-regime outcome stability, Appendix F also develops sensitivity bounds for both the Wald ratio and the decomposition under bounded within-regime selection.

Preview of Empirical Findings. Applying the framework to felony bail decisions in Miami-Dade County, I find a sharp racial asymmetry in the diagnostic evidence. For Black defendants, I reject both the rank-1 benchmark and the nonnegative-weight restriction. For Non-Black defendants, I do not reject either restriction. For the sentencing outcome, the estimated Black–Non-Black gap in the Wald ratio is driven primarily by margin composition rather than by within-regime outcome differences. Because the decomposition is imprecisely estimated, I interpret this evidence as directional rather than definitive, and examine its sensitivity below.

1.1 Related Literature

This paper relates to the literature on IV with heterogeneous treatment effects, judge-leniency designs, discrimination measurement, monotonicity relaxations, and finite-mixture identification.

The paper speaks directly to judge and examiner designs, in which decision-maker identity is typically scalarized into a one-dimensional leniency measure and analyzed using standard IV logic (Kling, 2006; Dobbie et al., 2018; Goldsmith-Pinkham et al., 2025; Coulibaly et al., 2024). Mueller-Smith (2015) emphasizes that these designs can feature multidimensional treatment variation and non-monotone first stages, and proposes conditioning on observables to recover monotonicity within *observable* case types. My framework can absorb both observed and unobserved case heterogeneity, but in settings that already condition on observables its empirical role is the residual problem that remains after such conditioning: even within an observable covariate cell, decision-makers may weight the remaining case features differently, so the scalar propensity score need not be a sufficient summary of judge assignment.

At a broader methodological level, the paper also fits the distinction drawn by Mogstad and Torgovitsky (2024) between reverse engineering standard IV estimands and forward engineering from an explicit selection model. The canonical MTE/LIV framework (Heckman and Vytlacil, 2005) is the leading forward-engineering benchmark under scalar threshold-crossing selection. This paper is likewise forward-engineered, but from a different primitive:

finite latent regimes with decision-maker-specific propensities. The resulting objects are regime-specific outcome distributions and slope-weighted decompositions of the Wald ratio, rather than an MTE curve indexed by a single latent margin. The paper is therefore designed for settings in which the scalar selection approximation is empirically inadequate, rather than as a nesting of the full canonical MTE model.

On discrimination measurement, Canay et al. (2024) study the logical validity of marginal outcome tests within a scalar-index Roy framework, and Arnold et al. (2022) analyze disparate impact in bail using a hierarchical MTE model with judge-specific signal precision. My contribution is different in both object and mechanism. I study cross-group differences in judge-IV Wald ratios and show that, under multidimensional selection, these gaps decompose into a within-regime channel and a margin-composition channel. That decomposition is unavailable in scalar-index models because those models posit only a single latent margin, leaving no separate composition channel to identify. The multi-margin framework is motivated by settings in which different judge comparisons change the composition of marginal cases in ways a single aggregate threshold cannot capture.

The paper is also related to work on monotonicity failures and their consequences for IV interpretation (Mogstad, Torgovitsky, and Walters, 2021; Chaisemartin, 2017; Sigstad, 2026). My question is adjacent but distinct: even when monotonicity holds in the sense characterized in my framework, namely: common-signed regime-specific first-stage changes, scalar leniency can fail because those changes need not be proportional across regimes. The distinction between sign restrictions and proportionality restrictions motivates the rank-1 diagnostic. Related work on skill or classifier heterogeneity (Chan et al., 2022; Bhuller and Sigstad, 2022; Mungan, 2023) provides economic mechanisms that can generate such departures; the present paper is agnostic about the mechanism and develops diagnostics and decompositions for the resulting multi-margin selection.

Methodologically, this work builds on two strands of the finite-mixture identification literature. One exploits observable variation in mixture weights to recover the latent structure (Henry et al., 2014; Jochmans et al., 2017), and the other provides broader identification results for finite mixtures under additional restrictions (Kasahara and Shimotsu, 2009; Kitamura and Laage, 2018). The results obtained under mild assumptions in this work are closest to the first strand; judge variation in treatment rates across regimes changes the mixture weights and identifies the dimension of the latent regime space. The stronger ordered-tail and Gaussian results are closer to the second strand; they impose additional structure to recover the latent components themselves. The paper is also connected to Bonhomme et al. (2022) and related works that use finite latent classes as tractable approximations to rich heterogeneity.

This work is especially close in spirit to Hoshino and Yanagi (2022), who study

unobserved group heterogeneity in LIV/MTE while maintaining scalar threshold-crossing selection within group. Their identification strategy relies on instruments that are specific to each latent group. By contrast, the present paper works with a common judge-IV environment and allows the instrument itself to move treatment along multiple latent margins. A different connection is to work on ordered or multivalued treatments, including Lee and Salanié (2018) on identification in multivalued treatment models; Lee and Salanié (2024) on targeting instruments with discrete multivalued treatments; Tsuda (2024) on marginal treatment effects with multivalued treatments and multidimensional unobserved heterogeneity; Heckman and Pinto (2018) on unordered monotonicity and discrete mixture representations; Kamat et al. (2023) on multidimensional threshold-crossing models with discrete instruments; and Humphries et al. (2025), who extend binary-treatment judge-IV designs to a multiple-treatment criminal-justice setting. Relative to this literature, the present framework keeps treatment binary and instead allows multidimensional heterogeneity to enter the selection mechanism itself through regime-specific responses to a common instrument.

1.2 Illustrative Example

Before presenting the general framework, a simple numerical example illustrates the key distinction between multi-margin selection and single-index leniency.[2]

Consider a bail setting with two latent *decision environments*: Regime 1 and Regime 2, representing two distinct configurations of case features that judges weight differently. Assume equal population shares $w_1 = w_2 = 0.5$ and regime-specific released-risk means $\tau_1 = 0.15$, $\tau_2 = 0.35$. Three judges $Z \in \{A, B, C\}$ make release decisions with regime-specific propensities given in Table 1.

Table 1: Regime-Specific Release Propensities: Judges Weight Factors Differently

	$S = 1$ (Regime 1)	$S = 2$ (Regime 2)	Overall $p(z)$
Judge A	0.80	0.30	0.55
Judge B	0.55	0.45	0.50
Judge C	0.30	0.70	0.50

The key observation: Judges B and C have *identical overall release rates* ($p(B) = p(C) = 0.50$) but release systematically different case types. A scalar leniency index based on $p(z)$

[2]Chyn et al. (2025, Appendix A) give a stylized grouped-defendant example of multidimensional judge behavior. The present framework replaces prespecified observed groups with latent regimes and develops the corresponding decomposition, diagnostics, and identification results.

cannot distinguish them. Judge assignment is therefore better viewed as a *vector instrument* $(p_1(z), p_2(z))$ than as a scalar shift.

This has immediate consequences for the Wald ratio $\psi(z, z') := [m(z) - m(z')]/[p(z) - p(z')]$: the (B, C) comparison is undefined because the denominator is zero even though the outcome means differ. The sections below formalize this intuition using pair-specific slope weights and a benchmark for when scalar leniency is adequate. Section 4 revisits this example once the machinery is in place.

2 Model

This section introduces the primitives and notation. I describe the latent-regime environment and distinguish observable arm-specific distributions from regime-specific distributions.

Throughout, I write all statements conditional on a fixed covariate value x when convenient; all identification results hold pointwise in x.

The theory treats Z as a generic observed instrument. In judge or examiner designs, the primitive source of quasi-random variation is assignment to a decision-maker, which I denote by J when that distinction matters. The empirical application then constructs a scalar instrument from J: specifically, Z is the leave-one-out UJIVE leniency score induced by judge assignment and residualized on court-by-time fixed effects, discretized into monotone leniency bins. Accordingly, the theoretical object Z should be read empirically as the scalar leniency instrument generated by assignment, not as literal judge identity.

2.1 Latent Regimes and Potential Outcomes

I observe i.i.d. draws of (Y, D, X, Z), where $D \in \{0, 1\}$ is treatment, $Y \in \mathcal{Y} \subseteq \mathbb{R}$ is an outcome, $X \in \mathcal{X}$ are covariates, and $Z \in \mathcal{Z}$ is an observed instrument. In leniency designs, Z may be the assigned decision-maker itself or a scalar score constructed from that assignment. Let $(Y(1), Y(0))$ denote potential outcomes and let $D(z) \in \{0, 1\}$ denote the potential treatment decision under instrument value z. Observed outcomes satisfy

$$Y \;=\; D\,Y(1) + (1 - D)\,Y(0), \qquad D = D(Z). \tag{2.1}$$

Latent regimes. Let $S \in \{1, \ldots, K_0\}$ be an unobserved *latent regime* (finite K_0), capturing qualitatively distinct case types (or decision environments) that differ in (i) how treatment responds to Z, and (ii) how outcomes differ between treated and untreated. Let V denote additional idiosyncratic heterogeneity affecting the decision. A convenient

structural representation for the decision rule is

$$D(z) = \mathbf{1}\{g_S(X, z, V) \geq 0\}, \tag{2.2}$$

where g_k denotes the regime-k decision function and g_S evaluates g_k at the realized regime $S = k$. This allows for regime-specific first stages: within regime k, the treatment propensity is $p_k(x, z) = \Pr(g_k(X, z, V) \geq 0 \mid X = x, Z = z, S = k)$.

Core assumptions.

Assumption 2.1 (Finite regimes and overlap)**.** *$S \in \{1, \ldots, K_0\}$ a.s., and for all x in the support of X,*

$$\Pr(S = k \mid X = x) =: w_k(x) > 0, \qquad k = 1, \ldots, K_0.$$

Assumption 2.2 (Regime-invariant assignment: $S \perp Z \mid X$)**.** *For all (x, z),*

$$\Pr(S = k \mid X = x, Z = z) = \Pr(S = k \mid X = x) = w_k(x). \tag{2.3}$$

Assumption 2.3 (Within-regime instrument exogeneity)**.**

$$Z \perp (Y(1), Y(0), V) \mid (X, S). \tag{2.4}$$

Assumption 2.4 (Within-regime outcome stability / no residual selection on potential outcomes)**.**

$$(Y(1), Y(0)) \perp V \mid (X, S). \tag{2.5}$$

Remark 2.1 (Relationship to MTE)**.** *This is not a generalization of the marginal treatment effects framework; it is a different model with different primitives. The MTE framework allows within-margin selection (*$\mathrm{MTE}(x, u)$ *varies continuously with the latent compliance index u) but restricts the selection process to be one-dimensional. My framework allows multidimensional selection (K_0 latent margins with non-proportional instrument responses) but restricts within-regime selection to be absent. Neither model nests the other; both impose substantive structure, and the question is which structure matches the institutional setting.*

The motivating question of this paper is what happens when decision-makers differ in *what* they weight rather than just how strictly they apply a common index. Assumption 2.4 is the modeling price for taking such a multidimensional selection mechanism seriously: it concentrates all outcome-relevant selection heterogeneity into the finite regime indicator S, making the multi-regime mixture identified, while eliminating the continuous within-regime

selection that MTE would allow. This trade-off is favorable when multidimensional selection is first-order and within-regime selection is second-order—the setting this paper targets.

The assumption is most plausible when regimes are narrow enough that within-regime risk variation is small relative to between-regime variation; that is, when the action is primarily in *which regime* a case falls into, not in fine-grained selection within regimes. It becomes mechanically weaker as K_0 increases: with more regimes, each absorbs less residual heterogeneity, making the independence condition less demanding. In the limit where regimes are singletons, it holds trivially.

Crucially, the assumption generates a *testable* restriction in principle. By Lemma 2.1, the observed conditional outcome distribution within regime k, $\Pr(Y \leq y \mid D = 1, X = x, Z = z, S = k)$, equals the regime-specific potential-outcome distribution $F_{1k}(y \mid x)$ and therefore does not vary with z. In a fully covariate-varying implementation one could test this directly from regime-specific treated outcome distributions. In the empirical application, however, the propensity-based posteriors used for the appendix comparison depend on residualized leniency bins rather than rich within-bin covariate variation, so the resulting comparison is better read as a descriptive reweighted-tercile check than as a direct test of Assumption 2.4. Appendix F therefore provides the paper's formal robustness device: it relaxes the assumption to bounded within-regime selection and gives worst-case bias bounds on the Wald ratio and on both channels of the discrimination decomposition (Propositions F.1–F.2) as a function of a scalar sensitivity parameter δ.

2.2 Observable and Latent Distributions

I use uppercase G_{dz} and F_{dk} for conditional distribution functions, and lowercase g_{dz} and f_{dk} for the corresponding densities with respect to a dominating measure μ when they exist. The observed distributions G_{dz} can vary with the instrument z; the regime-specific distributions F_{dk} cannot.

Define the regime shares and regime-specific first stages

$$w_k(x) := \Pr(S = k \mid X = x), \qquad p_k(x, z) := \Pr(D = 1 \mid X = x, Z = z, S = k). \tag{2.6}$$

Also define the observed and regime-specific conditional distribution functions

$$G_{dz}(y \mid x) := \Pr(Y \leq y \mid D = d, X = x, Z = z), \qquad d \in \{0, 1\}, \tag{2.7}$$

$$F_{dk}(y \mid x) := \Pr(Y(d) \leq y \mid X = x, S = k), \qquad d \in \{0, 1\}. \tag{2.8}$$

When these distributions admit densities with respect to μ, I denote them by $g_{dz}(y \mid x)$ and

$f_{dk}(y \mid x)$, respectively.

Lemma 2.1 (Regime-specific conditional outcome stability)**.** *Under Assumptions 2.3–2.4, for each $d \in \{0,1\}$, k, x, z, and y,*

$$\Pr(Y \le y \mid D = d, X = x, Z = z, S = k) = \Pr(Y(d) \le y \mid X = x, S = k) = F_{dk}(y \mid x). \tag{2.9}$$

Proof. Fix (d, k, x, z, y). By the consistency relation in (2.1), $Y = Y(d)$ on the event $\{D = d\}$, hence

$$\Pr(Y \le y \mid D = d, X = x, Z = z, S = k) = \Pr(Y(d) \le y \mid D = d, X = x, Z = z, S = k).$$

Now note $D = D(Z)$ and, within (X, S), the event $\{D = d\}$ is a measurable function of (Z, V). Under Assumption 2.3, $Z \perp (Y(1), Y(0), V) \mid (X, S)$, and under Assumption 2.4, $(Y(1), Y(0)) \perp V \mid (X, S)$. Therefore $(Y(1), Y(0))$ is independent of (Z, V) given (X, S). In particular, conditional on (X, S), $Y(d)$ is independent of both Z and the event $\{D(Z) = d\}$. Consequently,

$$\Pr(Y(d) \le y \mid D = d, X = x, Z = z, S = k) = \Pr(Y(d) \le y \mid X = x, S = k) = F_{dk}(y \mid x).$$

□

Interpretation. The regime-specific components in the observed data are the potential-outcome distributions F_{dk}, yielding a direct causal interpretation for regime contrasts without invoking a continuous-index MTE.

Identification roadmap. The identification argument works arm by arm and proceeds in four steps:

(i) Within each regime, the instrument shifts treatment probabilities but not potential-outcome distributions (Lemma 2.1). This links the observed selected-outcome density to regime-specific potential-outcome laws. (ii) The observed arm-d outcome density g_{dz} can therefore be written as a K_0-component mixture whose components are instrument-invariant and whose weights vary with z (Lemma 3.1). (iii) Under component separation and sufficient weight variation across instrument values (Assumptions 3.1–3.2), the number of active regimes K_0 equals the dimension of the mixture span (Theorem 3.1). This delivers the rank-based diagnostics used below. At this stage the data determine the span and the positions of observed mixtures within it, but not the individual component densities: multiple decompositions of the same span are observationally equivalent. (iv)

Point identification of the component densities and weights requires further structure that rules out rotations of the latent basis. Theorem A.1 provides a nonparametric route via ordered tail dominance; Proposition 3.1 gives the finite-Gaussian result used in the empirical application. When both arms are outcome-informative, the two arm-specific mixture systems share common regime shares w_k, and separate identification of w_k and p_k follows.

3 Identification

The identification argument connects observed arm-specific outcome distributions to regime-specific component distributions. The driving observation is that the instrument changes only the mixture weights, not the components themselves.

3.1 Mixture Representation and Identification

This subsection spells out the identification argument directly. Throughout, fix x and suppress it from the notation.

Mixture Representation Within Each Arm

For each treatment arm $d \in \{0, 1\}$, let $g_{dz}(y)$ denote the conditional density (with respect to a dominating measure) of the outcome among units with $D = d$:

$$g_{dz}(y) := f_{Y|D=d,X=x,Z=z}(y).$$

Also define the aggregate treatment propensity $p(z) := \Pr(D = 1 \mid X = x, Z = z)$.

Lemma 3.1 (Mixture representation within each treatment arm)**.** *Under Assumptions 2.2–2.4, for each $d \in \{0, 1\}$ and every z with $\Pr(D = d \mid X = x, Z = z) > 0$,*

$$g_{dz}(y) = \sum_{k=1}^{K_0} \omega_{dk}(z)\, f_{dk}(y), \tag{3.1}$$

where the posterior regime weights are

$$\omega_{1k}(z) = \frac{w_k\, p_k(z)}{p(z)}, \qquad \omega_{0k}(z) = \frac{w_k\,(1 - p_k(z))}{1 - p(z)}. \tag{3.2}$$

The components f_{dk} do not depend on z; this is the joint content of Assumptions 2.3 and 2.4 through Lemma 2.1. Without Assumption 2.4, the conditioning event $\{D(z) = d\}$ selects

on V; if $Y(d)$ depends on V given (X, S), the conditional law of $Y(d)$ among arm-d units shifts with z, and the z-invariance of the mixture components fails. Letting $\pi_k(z) := \Pr(D = 1, S = k \mid X = x, Z = z) = w_k\, p_k(z)$, the treated-arm weights satisfy $\pi_k(z) = p(z)\omega_{1k}(z)$.

Proof. Fix $d \in \{0, 1\}$. By the law of total probability and Assumption 2.2,

$$g_{dz}(y) = \sum_{k=1}^{K_0} \Pr(S = k \mid D = d, X = x, Z = z)\, f_{Y(d)|D=d,S=k,X=x,Z=z}(y).$$

By Lemma 2.1 (which uses both Assumption 2.3 and Assumption 2.4), since $D = D(Z)$ is measurable with respect to (Z, V, S) through the decision rule in (**??**), the joint independence $(Y(1), Y(0)) \perp (Z, V) \mid (X, S)$ implies $Y(d) \perp (D, Z) \mid (X, S)$. Therefore,

$$f_{Y(d)|D=d,S=k,X=x,Z=z}(y) = f_{Y(d)|X=x,S=k}(y) = f_{dk}(y \mid x),$$

i.e. the conditional density of selected arm-d outcomes within regime k equals the regime-specific potential-outcome density and does not vary with z once we condition on (X, S). Bayes' rule yields the posterior weights in (3.2). □

Identification Assumptions for the Outcome Mixtures

Assumption 3.1 (Component separation)**.** *For each $d \in \{0, 1\}$, the component densities $\{f_{dk}\}_{k=1}^{K_0}$ are linearly independent (as functions in L^1 with respect to the dominating measure, equivalently as finite signed measures).*

Lemma 3.2 (Component separation implies characteristic-function separation)**.** *Let $\{f_k\}_{k=1}^{K} \subset L^1$ and write their characteristic functions as $\varphi_k(t) := \int e^{ity} f_k(y)\, dy$.[3] If $\{f_k\}$ are linearly independent in L^1, then $\{\varphi_k\}$ are linearly independent as functions of t.*

Proof. If $\sum_{k=1}^{K} c_k \varphi_k(t) = 0$ for all t, then the Fourier transform of the L^1 function $\sum_k c_k f_k$ is identically zero. By injectivity of the Fourier transform on L^1, $\sum_k c_k f_k = 0$ almost everywhere, hence $c_k = 0$ for all k by linear independence. □

Assumption 3.2 (Instrument tilt / full-rank mixing)**.** *For at least one treatment arm $d \in \{0, 1\}$, there exist $L \geq K_0$ instrument values $z_1, \ldots, z_L$ with $\Pr(D = d \mid X = x, Z = z_\ell) > 0$ such that the $L \times K_0$ posterior-weight matrix*

$$\Omega_d := \big(\omega_{dk}(z_\ell)\big)_{\ell,k}$$

[3] If Y is not absolutely continuous, interpret f_k as the density of the component law with respect to a dominating measure; the argument applies to finite signed measures.

has full column rank K_0. For the treated arm, this is equivalent to the matrix $\Pi := (\pi_k(z_\ell))_{\ell,k}$ having rank K_0, since $\Pi = \mathrm{diag}(p(z_1), \ldots, p(z_L))\,\Omega_1$.

Assumption 3.1 is a standard "distinctness" condition: it rules out redundant regimes that generate identical outcome distributions within a treatment arm. Assumption 3.2 requires that the instrument changes the *composition* of individuals in at least one arm in sufficiently many independent directions. When both arms satisfy the tilt condition, each arm independently identifies the mixture, and the common regime shares w_k enable separate recovery of w_k and p_k.

Identification of K_0 Under Weak Assumptions

Theorem 3.1 (Identification of K_0 and the mixture span from outcome mixtures)**.** *Fix x. Suppose Lemma 3.1 holds and Assumptions 3.1–3.2 are satisfied for treatment arm d.*

(i) The number of regimes K_0 is identified as the dimension of the linear span of the family $\{g_{dz} : z \in \mathcal{Z}, \Pr(D = d \mid X = x, Z = z) > 0\}$.

(ii) The family $\{g_{dz}\}_{z\in\mathcal{Z}}$ identifies a K_0-dimensional mixture span. Without additional restrictions, the latent basis $\{f_{dk}\}_{k=1}^{K_0}$ and the posterior weights $\{\omega_{dk}(z)\}_{k,z}$ are only partially identified: any alternative basis $q_d = Af_d$ together with transformed weights $\tilde{\omega}_d(z) = \omega_d(z)A^{-1}$ yields the same observables whenever the transformed components and weights remain admissible.

This nonparametric result under weak assumptions is sufficient for the span-dimension interpretation of K_0 and for the rank-based diagnostics developed below. Point identification of the latent basis requires stronger structure. Appendix A gives one nonparametric identification route under ordered tail dominance, a tail-shape restriction compatible with overlapping supports, and the empirical continuous-outcome analysis relies on a finite-Gaussian component class, where classical Gaussian-mixture results identify the component parameters and thereby the unrestricted judge-specific weights.

Theorem 3.1 isolates the part of the identification argument that survives under weak assumptions. Economically, the exclusion restriction places the instrument in the weights but not in the component laws: changing z changes the composition of latent regimes in arm d but does not change the regime-specific outcome law once the regime is fixed. That separation is what makes judge variation informative; it yields multiple reweightings of the same latent outcome laws rather than a collection of judge-specific outcome distributions. The full-rank tilt condition then requires that the instrument changes composition in sufficiently many distinct directions: different instrument values do not merely scale a single strictness margin

up and down but generate genuinely different mixtures of the same latent regimes. This suffices to reveal the mixture dimension K_0.

What full-rank tilt does *not* buy by itself is uniqueness of the latent basis. The instrument generates several different mixtures of the same underlying regime-specific outcome laws, but observing that family of mixtures does not by itself pin down the pure components: different admissible decompositions into components and weights can generate the same observable family. The component laws and posterior weights are therefore set identified, with the identified set consisting of all admissible decompositions that leave the observable family unchanged (Henry et al., 2014). Appendix A gives the proof and makes the geometry explicit.

One way to collapse that indeterminacy without going fully parametric is to restrict the relative tail behavior of the component laws. Appendix A formalizes this as a nonparametric *bridge result*: within the class of ordered-tail decompositions, lower-index regimes dominate sufficiently extreme lower tails while higher-index regimes dominate sufficiently extreme upper tails, ruling out non-permutation re-expressions of the latent basis even under overlapping and possibly full supports.

The empirical continuous-outcome analysis takes a more structured route by restricting the component class directly to a finite Gaussian family while leaving the judge-specific mixing weights unrestricted across instrument values. The next result states the corresponding arm-specific identification theorem used below.

Proposition 3.1 (Identification for single-Gaussian arm-specific mixtures)**.** *Fix an arm $d \in \{0,1\}$ and suppose that, for some $K \geq 1$, the arm-d conditional outcome law satisfies*

$$g_{dz}(y) = \sum_{k=1}^{K} \omega_{dk}(z)\, \phi(y; \mu_{dk}, \sigma_{dk}^2), \qquad y \in \mathbb{R},$$

where $\phi(\,\cdot\,; \mu, \sigma^2)$ denotes the Gaussian density with mean μ and variance σ^2, the weights satisfy $\omega_{dk}(z) \geq 0$ and $\sum_{k=1}^{K} \omega_{dk}(z) = 1$ for each z, and the parameter pairs $\{(\mu_{dk}, \sigma_{dk}^2)\}_{k=1}^{K}$ are pairwise distinct. Assume that there exists at least one instrument value $\bar{z}$ such that

$$\omega_{dk}(\bar{z}) > 0 \qquad \text{for all } k = 1, \ldots, K.$$

Then the Gaussian parameters $\{(\mu_{dk}, \sigma_{dk}^2)\}_{k=1}^{K}$ and the weights $\{\omega_{dk}(z)\}_{k=1}^{K}$ are identified from the family $\{g_{dz}\}_{z \in \mathcal{Z}}$ up to a common permutation of the regime labels.

If the same conclusion holds for both arms $d = 0, 1$, and the cross-arm pairing is uniquely determined (generically, outside knife-edge cases with identical weighted increment profiles),[4]

[4]Cross-arm pairing can fail when two regimes share the same weighted increment profile across instrument values, i.e. when $w_i\{p_i(z) - p_i(z')\} = w_j\{p_j(z) - p_j(z')\}$ for some $i \neq j$ and all z, z'. In that case, constancy of

then

$$w_k = \omega_{1k}(z)\, p(z) + \omega_{0k}(z)\, [1 - p(z)]$$

is identified, and therefore

$$\pi_k(z) = p(z)\omega_{1k}(z), \qquad p_k(z) = \frac{\pi_k(z)}{w_k}$$

are identified as well. This caveat concerns the separate recovery of w_k and $p_k(z)$ from the two-arm mixture representation. It does not affect the rank-1 diagnostic (Section 4.5), which tests the rank of the fitted propensity-schedule matrix without invoking the cross-arm pairing.

Proof. See Appendix A. □

Remark 3.1 (Why the Gaussian case is useful)**.** *Proposition 3.1 is the identification route used in the continuous empirical application. The logic has two steps. First, at any instrument value $\bar{z}$ with all K components active, classical finite-Gaussian mixture identification recovers the component parameters up to permutation. Second, because the same Gaussian components reappear at every other instrument value, the remaining arm-specific mixtures recover the corresponding judge-specific weights. The restriction is therefore on the component class, not on the mixing weights, and it allows fully overlapping supports. Gaussians are not the only class for which fixed-family identification can work, but they are the cleanest continuous-outcome specification used in the paper.*

For completeness, Appendix B records an analogous identification result for binary outcomes under a fixed-basis common-index logistic class. That extension shows that the same fixed-class identification logic is not specific to Gaussian outcomes, but it is not used in the current empirical implementation; see also Follmann and Lambert (1991) for related logistic-mixture identifiability results under discrete-design conditions.

Specialization: the bail convention. In settings where the untreated outcome is degenerate (for example, bail designs where misconduct is observed only upon release), it is convenient to adopt the *bail convention* $Y := DY(1)$, so that f_{0k} carries no substantive content and the $D = 0$ arm provides no regime-resolving outcome information. Under this convention the treated arm alone still identifies the informative mixture span and therefore the rank-based diagnostics. Recovering regime-specific treated-outcome means $\tau_k(x)$ and

w_k does not uniquely determine the pairing, because a false label matching generates the same unconditional shares. The rank-1 benchmark contains one knife-edge special case of this degeneracy: if $p_k(z) = a_k + b_k\lambda(z)$ and $w_i b_i = w_j b_j$, then the two regimes are observationally indistinguishable for this cross-arm alignment step. Identification of the individual w_k and $p_k(z)$ then requires an additional alignment condition, such as an ordering restriction on component means.

the selection blocks $\pi_k(z)$, however, requires stronger structure on the treated-arm mixture; without it, the treated-arm decomposition is only partially identified.

Remark 3.2 (What is (and is not) identified under the bail convention)**.** *Under the bail convention $Y = DY(1)$, the data identify $\{\pi_k(z)\}_{k,z}$ but do* not *separately identify $\{w_k\}$ and $\{p_k(z)\}$ without additional structure, since $\pi_k(z) = w_k p_k(z)$ admits scale–composition factorizations. This non-separability is distinct from the basis-indeterminacy problem discussed above: even after an additional restriction pins down the latent basis, the treated arm alone identifies only the products $\pi_k(z) = w_k p_k(z)$. If a nondegenerate untreated outcome $Y(0)$ is observed among $D = 0$, the latent basis is identified in both arms, and the cross-arm pairing is uniquely determined, then $\{w_k\}$ and $\{p_k(z)\}$ become separately identifiable by combining the treated and untreated mixture weights.*

This non-separability arises because the first-stage mixture constraint $p(z) = \sum_k \pi_k(z)$ is an identity: it holds for any *factorization $\pi_k(z) = w_k p_k(z)$ with $\sum_k w_k = 1$ and $0 \le p_k(z) \le 1$. The simplex constraint $\sum_k w_k = 1$ is a single equation in K_0 unknowns, which is insufficient to pin down the factorization without a second system of equations (e.g., from an untreated-outcome mixture). Three routes to separate identification are: (i) non-degenerate $Y(0)$ among $D = 0$ (provides the second system of mixture weights needed once the latent basis is identified in both arms); (ii) auxiliary covariates with known conditional regime probabilities; (iii) normalizing restrictions, e.g., fixing $p_k(z_0)$ for $K_0 - 1$ regimes at a reference instrument value, which together with the constraint $\sum_k w_k = 1$ pins down all shares.*

A Practical Rank Diagnostic for K_0

When the treated arm is the informative arm (as under the bail convention and in many release-type applications), the span-dimension characterization in Theorem 3.1(i) suggests a rank diagnostic based on moments. Let $\psi_1, \ldots, \psi_Q$ be bounded basis functions and define the $L \times Q$ matrix

$$M_Q := \Big(\mathbb{E}\big[\psi_j(Y) \mid D = 1, X = x, Z = z_\ell\big] \Big)_{\ell,j}.$$

Under (3.1), $M_Q = \Omega_1 T_Q$ where $(T_Q)_{k,j} := \int \psi_j(y)\, f_{1k}(y)\, dy$. If Ω_1 has rank K_0 (Assumption 3.2) and T_Q has rank K_0 for some $Q \ge K_0$ (a generic "moment separation" condition), then $\operatorname{rank}(M_Q) = K_0$. This provides a direct connection between K_0 and the number of non-negligible singular values of empirical analogues of M_Q, in the spirit of rank-based component identification (cf. Kasahara and Shimotsu, 2009). If the untreated arm is the informative one, an analogous construction replaces $D = 1$ and f_{1k} with $D = 0$ and f_{0k}.

3.2 Posterior Weights

Define the observed first stage $p(x,z) = \Pr(D = 1 \mid X = x, Z = z)$. By the law of total probability and Assumption 2.2,

$$p(x,z) = \sum_{k=1}^{K_0} w_k(x)\, p_k(x,z), \tag{3.3}$$

which is identified from the data.

Define posterior regime weights within treated and untreated:

$$\begin{aligned} \omega_{1k}(x,z) &:= \Pr(S = k \mid D = 1, X = x, Z = z), \\ \omega_{0k}(x,z) &:= \Pr(S = k \mid D = 0, X = x, Z = z). \end{aligned} \tag{3.4}$$

Corollary 3.1 (Posterior weights)**.**

$$\begin{aligned} \omega_{1k}(x,z) &= \frac{w_k(x) p_k(x,z)}{p(x,z)}, \\ \omega_{0k}(x,z) &= \frac{w_k(x)\big(1 - p_k(x,z)\big)}{1 - p(x,z)}. \end{aligned} \tag{3.5}$$

Corollary 3.2 (Recovering w_k and p_k from (ω_{dk}))**.**

$$w_k(x) = p(x,z)\omega_{1k}(x,z) + \big(1 - p(x,z)\big)\omega_{0k}(x,z), \tag{3.6}$$

$$p_k(x,z) = \frac{p(x,z)\omega_{1k}(x,z)}{w_k(x)}. \tag{3.7}$$

Equation (3.6) *implies a testable overidentifying restriction: the right-hand side must be constant in* z*.*

3.3 Regime-Specific Causal Effects

Whenever a stronger restriction pins down the regime-specific potential-outcome distributions F_{dk}—for example, ordered tail dominance or a correctly specified parametric submodel—any functional of those distributions becomes identified. In particular, the regime-specific average treatment effect

$$\tau_k(x) := \mathbb{E}\big[Y(1) - Y(0) \mid X = x, S = k\big] \tag{3.8}$$

is identified, as are regime-specific quantile treatment effects and distributional distances (e.g., $W_1(F_{1k}, F_{0k})$).

3.4 Wald Ratios as Regime-Weighted Contrasts

Define

$$m(x,z) := \mathbb{E}[Y \mid X = x, Z = z], \qquad p(x,z) := \mathbb{E}[D \mid X = x, Z = z].$$

Assume $z \mapsto p(x,z)$ and $z \mapsto m(x,z)$ are differentiable at the point of interest with $\partial_z p(x,z) \neq 0$.

Proposition 3.2 (Slope-weighted regime decomposition)**.** *Under Assumptions 2.1–2.4,*

$$\psi(x,z) := \frac{\partial_z m(x,z)}{\partial_z p(x,z)} = \sum_{k=1}^{K_0} \omega_k^*(x,z)\, \tau_k(x), \tag{3.9}$$

where

$$\omega_k^*(x,z) := \frac{w_k(x)\, \partial_z p_k(x,z)}{\sum_{j=1}^{K_0} w_j(x)\, \partial_z p_j(x,z)}. \tag{3.10}$$

If $\partial_z p_k(x,z)$ have a common sign across k, then $\{\omega_k^*(x,z)\}$ are convex weights; otherwise weights may be negative, reflecting monotonicity failures in leniency designs.

Proof. See Appendix A. □

Remark 3.3 (Discrete instruments)**.** *If Z is discrete (e.g., decision-maker identity), replace derivatives by differences:*

$$\psi(x; z, z') := \frac{m(x,z) - m(x,z')}{p(x,z) - p(x,z')} = \sum_{k=1}^{K_0} \omega_k^*(x; z, z')\, \tau_k(x),$$

where

$$\omega_k^*(x; z, z') := \frac{w_k(x)\big(p_k(x,z) - p_k(x,z')\big)}{\sum_{j=1}^{K_0} w_j(x)\big(p_j(x,z) - p_j(x,z')\big)}.$$

4 Application to Leniency Designs

This section translates the framework into the institutional language of leniency designs, giving the latent regimes an economic interpretation.

4.1 Institutional Setting

An observation i is a case assigned quasi-randomly to a decision-maker J_i within design strata collected in X_i. The empirical instrument $Z_i \in \{1, \ldots, L\}$ is a scalar function of that assignment, for example, a leave-one-out leniency score or a discretized leniency bin, where

L denotes the number of ordered leniency bins. Let $D_i \in \{0,1\}$ denote treatment and let $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$ as in (2.1). In bail applications, where the untreated outcome is degenerate, I use the bail convention $Y_i := D_i Y_i(1)$.

Let X_i collect observed case characteristics used to define quasi-random assignment (charge, court, hearing time, etc.). The observable first stage and reduced form are

$$p(x,z) := \Pr(D = 1 \mid X = x, Z = z), \qquad m(x,z) := \mathbb{E}[Y \mid X = x, Z = z].$$

4.2 Decision Environments and Regime Structure

I posit a finite latent variable $S_i \in \{1, \ldots, K_0\}$ indexing *decision environments* ("regimes"): configurations of case information and signals that decision-makers process differently. The regimes are not defined by observables alone; empirically they are recovered through differential first-stage response patterns across decision-makers.

Within regime k, define the decision-maker–specific treatment propensity

$$p_k(x,z) := \Pr(D = 1 \mid X = x, Z = z, S = k),$$

and the regime weight

$$w_k(x) := \Pr(S = k \mid X = x), \qquad \sum_{k=1}^{K_0} w_k(x) = 1.$$

Then the observed first stage is a varying-weights mixture:

$$p(x,z) = \sum_{k=1}^{K_0} w_k(x) p_k(x,z).$$

The reduced form decomposes as

$$m(x,z) = \sum_{k=1}^{K_0} w_k(x) \big[\mu_{0k}(x) + p_k(x,z)\, \tau_k(x)\big],$$

where $\mu_{0k}(x) := \mathbb{E}[Y(0) \mid X = x, S = k]$ and $\tau_k(x) := \mathbb{E}[Y(1) - Y(0) \mid X = x, S = k]$ are regime-specific untreated means and treatment effects. Under the bail convention ($Y(0) \equiv 0$), $\mu_{0k} = 0$ and $\tau_k(x) = \mathbb{E}[Y(1) \mid X = x, S = k]$ reduces to the regime-specific treated-outcome mean. Thus the instrument shifts outcomes by changing who is treated through the regime-specific propensities p_k, while the regime-specific treatment effects $\tau_k(x)$ are primitives of the latent decision environment.

4.3 Interpretation of the Identifying Assumptions

In leniency designs, the identifying restrictions can be read as follows. First, assignment to the primitive decision-maker J is as good as random within the design strata summarized by X, and the scalar instrument Z inherits that quasi-randomness. Second, conditional on (X, S), the potential-outcome distribution is invariant to the instrument, so decision-makers affect outcomes only through treatment rather than through regime-specific post-treatment channels. Third, the instrument must move the regimes in sufficiently different ways for the vectors $\{(p_1(x,z), \ldots, p_{K_0}(x,z)) : z \in \mathcal{Z}\}$ to span a K_0-dimensional set, while the effective regime dimension itself is finite. Together these restrictions justify reading S as stable, instrument-invariant heterogeneity in treatment effects and $p_k(x,z)$ as decision-maker-specific weighting of that heterogeneity.

4.4 Wald Ratio Decomposition

Define the pairwise Wald ratio

$$\psi(x; z, z') := \frac{m(x,z) - m(x,z')}{p(x,z) - p(x,z')},$$

whenever $p(x,z) \neq p(x,z')$. Proposition 3.2 shows that, under the mixture model,

$$\psi(x; z, z') = \sum_{k=1}^{K_0} \omega_k^*(x; z, z')\, \tau_k(x), \qquad \omega_k^*(x; z, z') := \frac{w_k(x)\big(p_k(x,z) - p_k(x,z')\big)}{\sum_{j=1}^{K_0} w_j(x)\big(p_j(x,z) - p_j(x,z')\big)}.$$

The slope weights $\omega_k^*(x; z, z')$ depend on the *vector* first stage, describing which regimes are moved by the instrument comparison (z, z') and with what intensity.

Worked Example: Full Tilt

Return to the two-regime bail example from Section 1.2, with $w_1 = w_2 = 0.5$, $\tau_1 = 0.15$, $\tau_2 = 0.35$, and regime-specific propensities as in Table 1. The posterior regime weights among the released follow from Bayes' rule, $\omega_{1k}(z) = w_k p_k(z)/p(z)$:

	ω_{11} (Regime 1)	ω_{12} (Regime 2)
Judge A releases	72.7%	27.3%
Judge B releases	55.0%	45.0%
Judge C releases	30.0%	70.0%

Figure 1 plots the regime-specific release propensities. The crossing pattern—Judge A is high on Regime 1 and low on Regime 2, while Judge C exhibits the opposite—is the graphical signature of multi-margin selection.

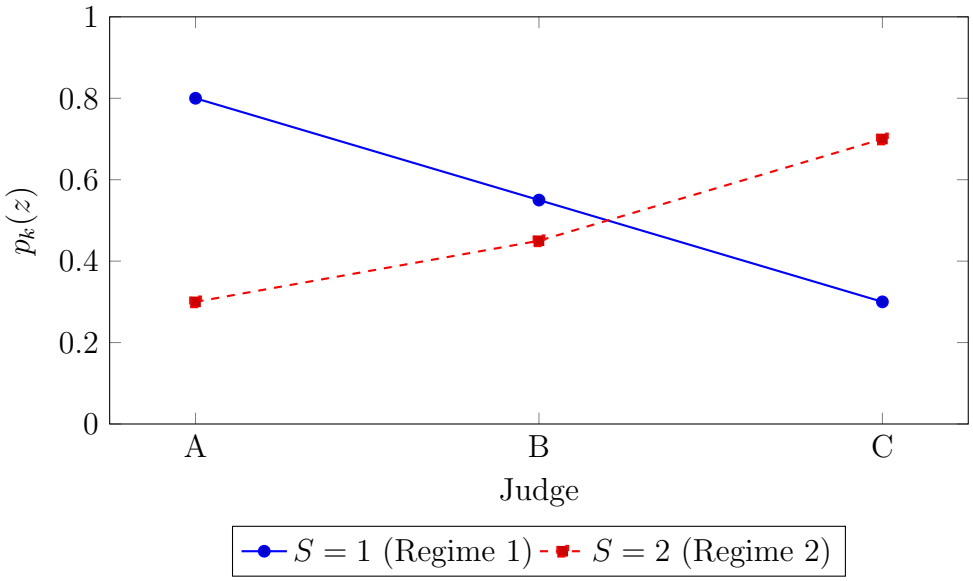

Figure 1: Regime-specific release propensities across judges (Section 1.2). The crossing pattern indicates that judges reweight regimes in opposite directions.

The overall release rates remain $p(A) = 0.55$ and $p(B) = p(C) = 0.50$, reproducing the scalar indistinguishability emphasized in Section 1.2. The posterior weights also determine expected misconduct among released defendants:

$$\mathbb{E}[Y \mid D = 1, Z = A] = 0.205, \qquad \mathbb{E}[Y \mid D = 1, Z = B] = 0.240,$$
$$\mathbb{E}[Y \mid D = 1, Z = C] = 0.290.$$

Under the bail convention, the unconditional outcome means are therefore

$$m(A) = 0.55 \times 0.205 = 0.1125,$$
$$m(B) = 0.50 \times 0.240 = 0.1200,$$
$$m(C) = 0.50 \times 0.290 = 0.1450.$$

Now apply the slope-weight decomposition. Under the bail convention, $m(z) = \sum_k w_k p_k(z) \tau_k$, and the pairwise Wald ratio decomposes as in Proposition 3.2. For the (A, C) comparison,

$$\omega_k^*(A, C) = \frac{w_k \big(p_k(A) - p_k(C)\big)}{\sum_{j=1}^{2} w_j \big(p_j(A) - p_j(C)\big)}.$$

The regime-specific first-stage numerators are

$$\begin{aligned} \text{Regime 1:} \quad & 0.5(0.80 - 0.30) = 0.25, \\ \text{Regime 2:} \quad & 0.5(0.30 - 0.70) = -0.20, \end{aligned}$$

so the denominator is only 0.05. The regime-specific first-stage changes therefore have *opposite signs*, and the small net first stage produces extreme slope weights:

$$\omega_1^*(A, C) = 5.0, \qquad \omega_2^*(A, C) = -4.0, \qquad \psi(A, C) = 5.0(0.15) - 4.0(0.35) = -0.65.$$

The negative Wald ratio does not mean "negative risk"; it reflects offsetting flows across regimes. The (A, B) comparison yields $\omega_1^*(A, B) = 2.5$, $\omega_2^*(A, B) = -1.5$, $\psi(A, B) = -0.15$—different weights and a different Wald ratio, revealing that the "marginal population" depends on which judges are compared. And because $p(B) = p(C)$ while the outcome means differ,

$$\psi(B, C) = \frac{m(B) - m(C)}{p(B) - p(C)} = \frac{0.1200 - 0.1450}{0}$$

is undefined in scalar form. The (A, C) comparison therefore combines large regime-specific movements, $\Delta p_1 = 0.50$ and $\Delta p_2 = -0.40$, into a weak net first stage $\Delta p = 0.05$. That instability is itself informative: it is the signature of offsetting margins and the reason the slope weights become extreme. These are the signatures the NegMass diagnostic and rank-1 test are designed to detect.

4.5 The Rank-1 Benchmark

A common abstraction in judge IV work is that judges differ only by a one-dimensional leniency index. In my framework, this corresponds to a *rank-1 incremental structure*:

$$p_k(x, z) = a_k(x) + b_k(x)\lambda(z), \tag{4.1}$$

for some scalar $\lambda(z)$ and regime-specific slopes $b_k(x)$. Under this restriction, the regime-specific changes $p_k(x, z) - p_k(x, z')$ are proportional across k, and Proposition 4.1 implies that $\psi(x; z, z')$ is constant across instrument pairs (up to sampling error). This gives a direct diagnostic: systematic variation of ψ across (z, z') is evidence against a single-index interpretation of leniency. The rank-1 benchmark is best understood as a local linearization of the more general scalar-index restriction $p_k(x, z) = H_k(x, \lambda(z))$; the resulting diagnostic therefore tests the adequacy of this rank-1 approximation on the observed support, rather than the existence of a scalar index itself.

Proposition 4.1 (Rank-1 incremental structure implies pairwise Wald ratio constancy)**.** *Suppose the regime-specific propensities satisfy* (4.1) *for some scalar index* $\lambda(z)$ *and functions* $\{a_k(x), b_k(x)\}_{k \le K_0}$ *with* $\sum_{j=1}^{K_0} w_j(x) b_j(x) \neq 0$*. Then for any* (z, z') *with* $\lambda(z) \neq \lambda(z')$*,*

$$\omega_k^*(x; z, z') = \bar{\omega}_k(x) := \frac{w_k(x) b_k(x)}{\sum_{j=1}^{K_0} w_j(x) b_j(x)} \qquad \textit{and} \qquad \psi(x; z, z') = \sum_{k=1}^{K_0} \bar{\omega}_k(x) \tau_k(x).$$

In particular, $\psi(x; z, z')$ *is constant across instrument pairs. If moreover* $b_k(x)$ *share a common sign across* k*, then* $\omega_k^*(x; z, z') \geq 0$ *and hence* $\mathrm{NegMass}(x; z, z') = 0$ *for all pairs.*

Proof. Under (4.1), $p_k(x, z) - p_k(x, z') = b_k(x)\{\lambda(z) - \lambda(z')\}$ for each k. Plugging into the slope-weight formula in Proposition 3.2 yields the stated simplification for $\omega_k^*(x; z, z')$. The expression for $\psi(x; z, z')$ follows by substitution into (3.9). □

Worked Example: Rank-1

Return to the two-regime setup from Section 1.2, but now suppose all judges apply the same implicit risk-scoring rule with different release thresholds: $p_k(z) = a_k + b\,\lambda(z)$ with common slope $b = 0.6$, intercepts $a_1 = 0.1$, $a_2 = 0.3$, and leniency $\lambda(A) = 1$, $\lambda(B) = 0.5$, $\lambda(C) = 0$. This gives $p_1(A) = 0.70$, $p_2(A) = 0.90$, etc. Because $p_k(z) - p_k(z') = b(\lambda(z) - \lambda(z'))$ is proportional across k, the slope weights collapse to $\omega_k^* = w_k$ for every judge pair and the Wald ratio is constant: $\psi(A, C) = \psi(A, B) = \psi(B, C) = w_1\tau_1 + w_2\tau_2 = 0.25$. This is the operational content of the rank-1 benchmark: when it holds, the Wald ratio does not depend on which judges are compared.

Remark 4.1 (Monotonicity versus rank-1 structure)**.** *Response-type monotonicity and rank-1 incremental structure are restrictions on* different objects *and are logically independent. Monotonicity is an* individual-level *restriction: for decision-makers ordered by leniency* $z \succ z'$*, it requires* $D_i(z) \geq D_i(z')$ *for every* i *(nested treatment sets, hence the sign of first-stage changes). Rank-1 is a* population-level geometric *restriction: the vector* $(p_1(z), \ldots, p_{K_0}(z))$ *must move in a one-dimensional way across instrument values (proportional shifts, hence constancy of the Wald ratio). Monotonicity rules out offsetting margins (no negative slope weights), but does not imply rank-1.*

To see this, retain the two-regime setup with $w_1 = w_2 = 0.5$ *and choose monotone but non-proportional propensities:* $p_1 = (0.80, 0.60, 0.40)$ *and* $p_2 = (0.90, 0.65, 0.20)$ *across judges* $A \succ B \succ C$*. Both schedules decrease with strictness, so monotonicity holds and all slope weights are nonnegative. But the "tilt ratios" differ:* $(p_2(A) - p_2(B))/(p_1(A) - p_1(B)) = 1.25 \neq (p_2(B) - p_2(C))/(p_1(B) - p_1(C)) = 2.25$*. The Wald ratio varies across pairs (*$\psi(A, B) = 0.261$*,* $\psi(B, C) = 0.288$*,* $\psi(A, C) = 0.277$*) even though monotonicity is satisfied. What varies is*

not the sign pattern but the relative intensity *with which each comparison moves the latent margins.*

4.6 Negative Slope Weights

When the regime-specific first-stage changes $p_k(x,z) - p_k(x,z')$ do not share a common sign, slope weights can be negative and the Wald ratio becomes a *net contrast* across offsetting regime flows rather than an average over a well-defined complier group.

Definition 4.1 (Negative weight mass)**.** *For a given* $(x; z, z')$ *with* $p(x,z) \neq p(x,z')$*, define*

$$\text{NegMass}(x; z, z') := \sum_{k=1}^{K_0} \left|\omega_k^*(x; z, z')\right| \mathbf{1}\{\omega_k^*(x; z, z') < 0\}.$$

Under response-type monotonicity (nested treatment sets along a decision-maker ordering), $\omega_k^*(x; z, z') \geq 0$ and hence $\text{NegMass}(x; z, z') = 0$. Positive values diagnose *offsetting margins*: some regimes increase treatment while others decrease treatment for the same instrument comparison. Large values often coincide with "weak net first stage" (small $p(x,z) - p(x,z')$ due to cancellation), producing unstable slope weights even when regime-specific movements are large. In the terminology of Mogstad and Torgovitsky (2024), negative slope weights imply that the pairwise Wald ratio is no longer *weakly causal*; when all slope weights are nonnegative it remains a comparison-specific weakly causal average even if rank-1 fails.

5 Estimation and Inference

This section describes the estimands, inference procedures, and implementation. Formal likelihoods and consistency results are deferred to the appendices.

5.1 Overview

The empirical sentencing analysis uses a finite-dimensional two-arm Gaussian MLE with regime-specific propensity schedules and single-Gaussian outcome densities in each arm. The formal bail-convention estimator is the corresponding sieve MLE that targets the treated-arm densities f_{1k} and the selection-block weights $\pi_k(x,z) = \Pr(D = 1, S = k \mid X = x, Z = z)$ when only the treated arm is outcome-informative. In both cases, the fitted conditional law of $(Y, D) \mid (X, Z)$ delivers the observable building blocks: Wald ratios are direct plug-in functionals, while signed slope

weights, NegMass, and the cross-group decomposition additionally use the recovered component representation.

Operationally, estimation proceeds by EM on the observed-data likelihood of $(Y, D) \mid (X, Z)$: the E-step computes posterior regime responsibilities from each observation's contribution to the fitted joint law, and the M-step updates the selection block and the outcome block using those responsibilities as weights. The structural interpretation of the sentencing estimates relies on Proposition 3.1 together with the two-arm parametric specification. Appendix D collects the formal likelihoods and consistency statements; Appendices E and I give the bail-convention and two-arm proofs. The latter is not a corollary of the former: it covers the fixed-dimensional two-arm Gaussian model actually estimated in the data rather than the bail-convention sieve problem.

5.2 Targets for Estimation and Inference

The regime-specific treatment effect is estimated as

$$\widehat{\tau}_k(x) = \int y \, \widehat{f}_{1k}(y \mid x) \, dy - \int y \, \widehat{f}_{0k}(y \mid x) \, dy.$$

Under the bail convention ($\mu_{0k}(x) = 0$), this reduces to the regime-specific treated-outcome mean $\widehat{\tau}_k(x) = \int y \, \widehat{f}_{1k}(y \mid x) \, dy$. The key applied estimands are computed by plug-in from $\{\widehat{\pi}_k, \widehat{\tau}_k\}$. Under the bail convention, the population unconditional mean $m(x, z) := \mathbb{E}[Y \mid X = x, Z = z]$ admits the decomposition $m(x, z) = \sum_{k=1}^{K_0} \pi_k(x, z) \, \tau_k(x)$. The corresponding plug-in estimator is

$$\widehat{m}(x, z) := \sum_{k=1}^{K} \widehat{\pi}_k(x, z) \, \widehat{\tau}_k(x). \tag{5.1}$$

Note that $\widehat{m}(x, z)$ is *not* the mean outcome among the treated; that object is $\widehat{m}(x, z)/\widehat{p}(x, z) = \sum_k \widehat{\omega}_{1k}(x, z) \, \widehat{\tau}_k(x)$, where $\widehat{\omega}_{1k} := \widehat{\pi}_k/\widehat{p}$ are the posterior regime weights among the treated.

Wald ratio. For an instrument pair (z, z') with $|\widehat{p}(x, z) - \widehat{p}(x, z')| > \delta_n$ (trimmed to avoid small-denominator instability):

$$\widehat{\psi}(x; z, z') = \frac{\widehat{m}(x, z) - \widehat{m}(x, z')}{\widehat{p}(x, z) - \widehat{p}(x, z')}. \tag{5.2}$$

Slope-weight decomposition.

$$\widehat{\psi}(x; z, z') = \sum_{k=1}^{K} \widehat{\omega}_k^*(x; z, z')\, \widehat{\tau}_k(x), \qquad \widehat{\omega}_k^*(x; z, z') := \frac{\widehat{\pi}_k(x, z) - \widehat{\pi}_k(x, z')}{\widehat{p}(x, z) - \widehat{p}(x, z')}. \tag{5.3}$$

NegMass. $\widehat{\mathrm{NegMass}}(x; z, z') := -\sum_{k:\, \widehat{\omega}_k^* < 0} \widehat{\omega}_k^*(x; z, z')$.

Discrimination decomposition. For two demographic groups r, r', the MOT gap decomposes into *within-regime* and *margin-composition* components:

$$\widehat{\psi}^r - \widehat{\psi}^{r'} = \underbrace{\sum_k \bar{\omega}_k \big(\widehat{\tau}_k^r - \widehat{\tau}_k^{r'}\big)}_{\text{within-regime}} + \underbrace{\sum_k (\widehat{\omega}_k^{*r} - \widehat{\omega}_k^{*r'})\, \bar{\tau}_k}_{\text{margin-composition}}, \tag{5.4}$$

where $\bar{\omega}_k := \frac{1}{2}(\widehat{\omega}_k^{*r} + \widehat{\omega}_k^{*r'})$ and $\bar{\tau}_k := \frac{1}{2}(\widehat{\tau}_k^r + \widehat{\tau}_k^{r'})$ are symmetric midpoint references, ensuring the decomposition is exact. By Theorem D.1(iii), each of these objects is consistent.

Cross-group regime commensurability. The decomposition (5.4) requires regime labels to be aligned across group-specific fits. In the empirical application I therefore use a common K, align regimes by minimizing joint distance in $(\widehat{\tau}_k, \widehat{p}_k(z))$, and assess sensitivity through the stability checks in Section 5.4.2. When two candidate matchings are nearly tied, the resulting decomposition can be locally sensitive to the alignment rule; re-solving the alignment problem inside each bootstrap replicate propagates part of that uncertainty, but it does not eliminate the underlying commensurability issue. A fully pooled estimator would make regimes automatically comparable across groups, but only at the cost of imposing common within-regime outcome laws across groups. That restriction would shut down the very within-regime channel that the decomposition is designed to measure.

Under the general two-arm convention, the corresponding plug-in for the conditional mean is

$$\widehat{m}(x, z) = \sum_{k=1}^{K} \widehat{w}_k(x) \Big[\widehat{\mu}_{0k}(x) + \widehat{p}_k(x, z)\, \widehat{\tau}_k(x)\Big].$$

Equation (5.1) is the bail-convention specialization used when $Y(0) \equiv 0$, in which case $\pi_k(x, z) = w_k(x) p_k(x, z)$ and the mean simplifies accordingly.

5.3 Inference

Bootstrap. I use a nonparametric bootstrap that resamples cases (Y_i, D_i, X_i, Z_i) within court-by-time strata and refits the model in each replicate. As in other finite-mixture

problems, standard bootstrap arguments are cleanest when $K = K_0$ and the true active weights are interior; when $K > K_0$, boundary issues can arise (Chen and Li, 2009). In practice, I therefore report inference conditional on a selected $\widehat{K}$ and use the bootstrap to propagate the resulting estimation uncertainty.

Formal tests. I formalize two tests that have direct economic content in leniency designs. Both use a *parametric bootstrap under the null* to handle the boundary issues inherent in mixture models.

Test A. NegMass / offsetting margins. The null is $H_0 : \omega_k^*(x; z, z') \geq 0$ for all k and valid pairs (z, z'). I aggregate the estimated negative weight mass across pairs using $T_n^{\max} = \max_{(z,z')\in\mathcal{P}} \widehat{\text{NegMass}}(x; z, z')$ and $T_n^{\text{CvM}} = \sum_{(z,z')\in\mathcal{P}} \widehat{\text{NegMass}}(x; z, z')^2$. Critical values are obtained by parametric bootstrap under the null, implemented by projecting the fitted model onto the nonnegative-weight cone and refitting in each bootstrap draw.

Test B. Rank-1 / Wald ratio constancy. The null is that the incremental selection structure is rank-1, so the Wald ratio is constant across instrument pairs. I use the minimum-distance statistic

$$T_n^{\text{rank}} = \min_{c\in\mathbb{R}} \sum_{(z,z')\in\mathcal{P}} \left(\widehat{\psi}(x; z, z') - c\right)^2 \widehat{\sigma}_{zz'}^{-2}, \tag{5.5}$$

where the sum is over trimmed instrument pairs and $\widehat{\sigma}_{zz'}^2$ is the bootstrap variance of $\widehat{\psi}(x; z, z')$. Critical values come from a parametric bootstrap under the rank-1 null: I first fit the unrestricted K-regime model, project the estimated propensity schedules onto the rank-1 manifold, simulate bootstrap samples from that projected null, and then re-estimate the unrestricted model in each replicate.

Test A detects offsetting margins. Test B detects failures of proportional regime movement across instrument pairs. Rejection of rank-1 is informative about the inadequacy of scalar leniency, but does not by itself determine K_0.

5.4 Implementation

In the empirical application, the primitive quasi-randomized object is judge assignment J, while the instrument in the fitted model is the scalar UJIVE leniency measure induced by J and discretized into monotone bins. The fitted propensity schedules $\widehat{p}_k(z)$ should therefore be read as regime-specific treatment propensities indexed by leniency bins rather than by literal

judge labels. Algorithmic details are deferred to Appendix G, and the resulting robustness exercises are reported in Appendix H.

5.4.1 Diagnostics

I complement the formal rank-1 and NegMass tests with three practical checks: singular-value scree plots for moment and CDF operators, re-estimation at nearby values of K, and descriptive reweighted-tercile outcome comparisons across judge-leniency groups. These checks are meant to assess whether the fitted regime structure and resulting decompositions are adequate for the quantities of interest.

5.4.2 Selecting the Number of Regimes

For selecting K, I use cross-validated held-out log-likelihood as the primary device, inspect rank evidence as an exploratory lower bound, and treat functional stability as the deciding robustness check when nearby values of K have similar predictive performance. In the empirical application I evaluate $K \in \{1, 2, 3\}$: $K = 1$ is the scalar benchmark, $K = 2$ is the smallest genuinely multi-regime alternative, and $K = 3$ is the next richer specification against which functional stability can be assessed. The substantive criterion is whether the main objects—Wald-ratio profiles, signed-weight decompositions, and discrimination decompositions—have stabilized. The appendix consistency results are conditional on a fixed working value of K, so this selection rule should be read as a pragmatic model-adequacy device rather than as a theorem-backed recovery procedure for K_0. Appendix G records the full procedure.

5.5 Computation

The estimator is computed by multi-start EM routines. Under the bail-convention sieve model, the E-step updates posterior regime responsibilities, the selection-block M-step updates the sieve coefficients for $\pi_k(x, z)$, and the outcome-block M-step updates the KW mixing measures. Under the empirical two-arm implementation, the same architecture is specialized to weighted Gaussian MLEs for $\log(1 + \text{days\_prison})$: if r_{ik} denotes the posterior responsibility of regime k for observation i, then the M-step updates $\widehat{p}_k(z_j)$ by weighted treatment shares within judge-leniency bin z_j and updates $(\widehat{\mu}_{dk}, \widehat{\sigma}^2_{dk})$ by weighted Gaussian means and variances within arm d. I retain the run with the highest attained log-likelihood and monitor both likelihood improvement and parameter stability.

Because regimes are unlabeled, reported components are ordered by the magnitude of $\widehat{\tau}_k$ (or equivalently by arm means when more transparent). For cross-group decompositions,

labels are re-aligned by minimizing joint standardized distance in $(\widehat{\tau}_k, \widehat{p}_k(z))$, and the same rule is applied within each bootstrap replicate. Under the continuous-outcome Gaussian specification, the finite-Gaussian class identifies the regime basis up to label permutation. Under the bail convention, by contrast, only $\widehat{\pi}_k(x,z) = \widehat{w}_k(x)\widehat{p}_k(x,z)$ is identified without an informative untreated arm, so the reported $\widehat{w}_k$ and $\widehat{p}_k(z)$ should be read as a working factorization of the identified selection block. Appendix G records the full initialization, convergence, alignment, and numerical-stabilization details.

6 Simulation Evidence

This section evaluates the finite-sample performance of the latent-regime framework, both for structural recovery and for the formal diagnostics introduced in Section 4. The simulations stress distinct economic mechanisms: one-dimensional leniency shifts, multi-margin monotone tilts, and true monotonicity violations from crossing propensity schedules. All designs use the bail convention, the primary application setting and the case in which regime identification relies on a single treatment arm; the two-arm implementation has separate consistency guarantees (Appendix I) and is not separately simulated.

6.1 Design and Targets

All designs use the latent-regime structure in Section 2.1. I consider four DGPs:

For the $K_0 = 2$ DGPs, I set $(\tau_1, \tau_2) = (0.15, 0.35)$ and $(w_1, w_2) = (0.5, 0.5)$ so differences in diagnostics are driven by selection geometry rather than outcome-level shifts. The regime-specific first stages are constructed from deterministic functional forms to ensure exact population properties: the Rank-1 DGP uses affine schedules $p_k(z) = a_k + b_k\lambda(z)$ with $b_1/b_2 = 3/2$, guaranteeing proportional increments; the Mono-tilt uses $p_1(z) \propto \sqrt{\lambda(z)}$ and $p_2(z) \propto \lambda(z)^3$, which are both increasing in λ but with different curvatures; the Full-tilt uses $p_1(z) = 0.20 + 0.60\lambda(z)$ and $p_2(z) = 0.30 + 0.50\sin(\pi\lambda(z))$, so the first schedule increases while the second rises and then falls, creating crossing propensity schedules and negative slope weights. The baseline configuration is $n = 5000$ and $J = 20$ judges. I run:

- Part A: 50 replications for K_0 selection (held-out likelihood and SVD diagnostics),
- Parts B–C: 200 replications for parameter recovery and diagnostic separation,
- Part D: 100 replications on a grid $n \in \{2000, 5000, 10000\}$ and $J \in \{10, 20, 50\}$,
- Part E: 100 replications for bootstrap size/power with $B = 199$ resamples.

Table 2: Simulation DGPs and theoretical signatures

DGP	K_0	Regime-specific first stage $p_k(z)$	Monotonicity	Rank-1	Theoretical implication
Single	1	One latent margin, linear in judge index	Yes	Yes	Wald ratio constant; diagnostics should indicate one regime
Rank-1	2	Proportional shifts across regimes	Yes	Yes	Wald ratio constant at the population level; both formal tests should have size near α
Mono-tilt	2	Monotone but non-proportional shifts	Yes	No	No negative slope weights, but pairwise Wald ratio varies across judge comparisons
Full-tilt	2	Crossing/hump-shaped schedule in one regime	No	No	Signed slope weights and large monotonicity violations

6.2 Diagnostic Targets

Selection dimension and recovery (Parts A–B). Part A asks whether finite samples recover the true number of regimes and whether moment/CDF-grid rank diagnostics are informative. Part B evaluates the EM estimator's ability to recover regime means τ_k, mixing weights w_k, and judge-regime propensities $p_k(z)$.

Model diagnostics (Part C). I compare three diagnostic families across DGPs:

1. singular-value ratios from moment and CDF-grid matrices,

2. projection-residual rank-1 distance,

3. model-based NegMass and Wald ratio range summaries.

The target pattern is: small rank-1 distance in Rank-1, moderate in Mono-tilt, and large in Full-tilt; and large NegMass only in Full-tilt. This provides a preliminary check on diagnostic separation, but relies on ad-hoc thresholds. The formal tests in Part E address this limitation.

Formal tests (Part E). I implement two parametric bootstrap tests:

1. **NegMass test** (monotonicity null): H_0 imposes nonnegative slope weights for all valid instrument pairs, as defined in Section 5.3. The test statistics are $T_n^{\max}$ and T_n^{CvM} (the

max and Cramér–von Mises aggregates of $\widehat{\text{NegMass}}(x; z, z')$ across pairs); critical values are calibrated by a contact-set bootstrap that projects the estimated model onto the nonnegative-slope-weight cone (Romano et al., 2014). To construct the contact set, judges are ordered by fitted overall leniency and the projection enforces nonnegative regime-specific propensity increments between adjacent bins, which is the implemented form of the monotonicity null.

2. **Rank-1 test** (single-index null): H_0 is rank-1 incremental structure, tested with a projection-distance statistic and bootstrap calibration under the projected null.

The need for formal inference is apparent from the Part C diagnostics themselves: a NegMass threshold of 0.05 produces a 10.5% false positive rate in the Rank-1 DGP, while the formal bootstrap test controls size at 1% (Table 5).

6.3 Results

(i) Regime selection and recovery are strong. Table 3 shows held-out K_0 recovery rates around 0.82–0.88. Conditional on $K_0 = 2$, biases for τ_k are below 0.001 in absolute value in all DGPs, with RMSE below 0.003, confirming that the EM estimator recovers regime-specific treatment effects accurately. The propensity surface $\widehat{p}_k(z)$ has RMSE around 0.05–0.06, small relative to the scale of regime-specific propensity variation. (Under the bail convention, only the products $\pi_k(z) = w_k p_k(z)$ are identified (Remark 3.2); the reported $\widehat{p}_k(z)$ values reflect the EM algorithm's working factorization $\pi_k = w_k p_k$ after normalization $\sum_k w_k = 1$. The RMSE figures should be understood as targeting this normalized factorization, which is the object the EM algorithm produces; the structurally identified objects $\pi_k(z)$ and τ_k are the ones with formal consistency guarantees.)

Table 3: Parts A–B: finite-sample recovery of regime dimension and parameters

DGP	True K_0	$\Pr(\widehat{K} = K_0)$	Bias($\widehat{\tau}_1$)	Bias($\widehat{\tau}_2$)	RMSE($\widehat{p}_k(z)$)
Single	1	0.820	–	–	–
Rank-1	2	0.880	0.0001	-0.0001	0.0563
Mono-tilt	2	0.820	-0.0003	0.0001	0.0543
Full-tilt	2	0.880	0.0002	0.0000	0.0599

(ii) Diagnostics separate mechanisms as theory predicts. Table 4 confirms the predicted separation pattern. Rank-1 distance separates all three DGPs: the 90th percentile for Rank-1 (0.294) falls below the 10th percentile for Full-tilt (0.607), with Mono-tilt intermediate. NegMass provides a clean monotonicity diagnostic: it is near zero

for both Rank-1 and Mono-tilt (which satisfy monotonicity by construction) but large for Full-tilt (population value 0.874, estimated mean 0.464). The Wald ratio range also separates correctly, with Full-tilt showing nearly four times the variation of the monotone DGPs, driven by the signed-weight structure of Proposition 3.2.

Note that the Rank-1 DGP has population Wald ratio range of exactly zero (the Wald ratio is identical for every judge pair when increments are proportional), but the estimated Wald ratio range is 0.545 due to finite-sample noise in $\widehat{p}_k(z)$. This sampling variability in diagnostics, present even when the population object is well-behaved, motivates the formal bootstrap tests in Part E.

Table 4: Part C: diagnostic separation across DGPs (mean, sd in parentheses)

Diagnostic	Rank-1	Mono-tilt	Full-tilt
SV ratio (moments)	0.0425 (0.006)	0.1247 (0.010)	0.0540 (0.006)
SV ratio (CDF-grid)	0.0640 (0.009)	0.0724 (0.007)	0.0829 (0.006)
Rank-1 distance	0.2447 (0.040)	0.3242 (0.036)	0.6439 (0.027)
Model NegMass mean	0.0299 (0.015)	0.0239 (0.011)	0.4640 (0.079)
Model Wald ratio range	0.5450 (0.155)	0.4993 (0.138)	1.9592 (0.332)

(iii) Scaling in (n, J) connects to the consistency theory. Figure 2 shows propensity-surface RMSE decreases with n at approximately the $n^{-1/2}$ rate (left panel), consistent with parametric convergence of the regime-specific parameters established in Section 5. Holding n fixed, RMSE increases sublinearly in J (right panel): the propensity surface has $K_0 \times J$ free parameters, so the effective sample per judge-regime cell is $w_k \cdot n/J$, and doubling J with fixed n halves this cell count; the resulting variance scales proportionally to J, implying RMSE grows approximately as $\sqrt{J}$. This scaling is practically relevant because empirical judge designs vary substantially in J (from $\sim$10 to $\sim$100), and the simulation confirms that the framework remains informative at n/J ratios typical of applied settings.

(iv) Formal tests have excellent power and controlled size. Table 5 and Figure 3 summarize rejection rates. The two tests are complementary in the intended way: the Rank-1 test detects non-proportional selection (rejecting in both Mono-tilt and Full-tilt with power 1.000), while the NegMass test isolates genuine monotonicity violations (rejecting only in Full-tilt with power 0.980). Together, they distinguish between multi-margin selection that preserves the sign structure of the IV estimand (Mono-tilt) and multi-margin selection that generates offsetting flows and signed weights (Full-tilt), precisely the diagnostic hierarchy developed in Section 4.

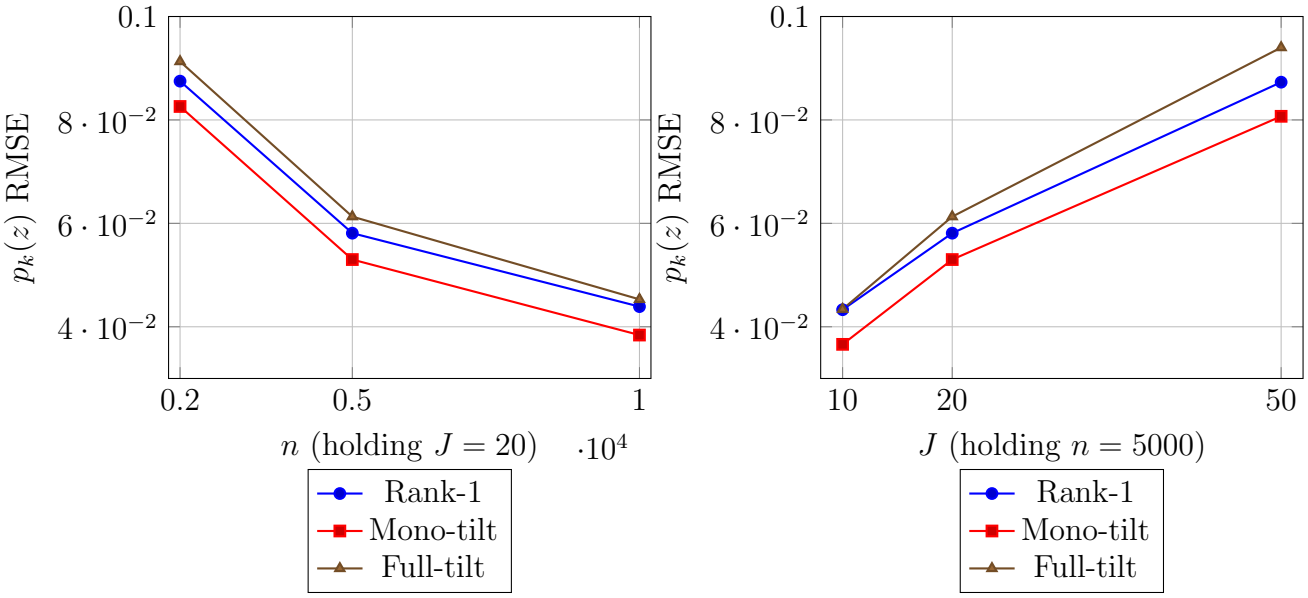

Figure 2: Part D: scaling of propensity-surface recovery. Left: RMSE falls with n at approximately the parametric rate. Right: RMSE grows with J as the effective cell size $w_k n/J$ shrinks.

Table 5: Part E: bootstrap size and power (100 replications, $B = 199$, $\alpha = 0.05$)

Test	DGP	Null status	Rejection rate	Mean statistic	Mean p-value
NegMass	Rank-1	Size	0.010	1.8469	0.655
NegMass	Mono-tilt	Size	0.010	2.1318	0.615
NegMass	Full-tilt	Power	0.980	7.5293	0.006
Rank-1	Rank-1	Size	0.070	0.2432	0.488
Rank-1	Mono-tilt	Power	1.000	0.3192	0.006
Rank-1	Full-tilt	Power	1.000	0.6399	0.005

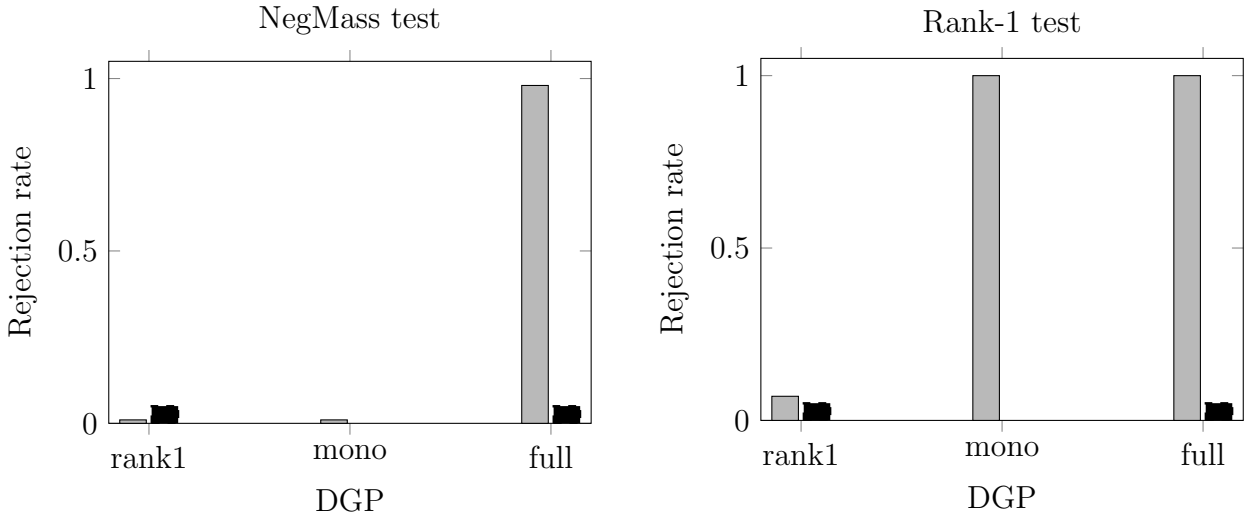

Figure 3: Part E rejection rates. Dashed line is nominal size $\alpha = 0.05$.

Interpretation for the theory. The simulation evidence supports the paper's central claim: when judges differentially reweight latent margins, the one-dimensional benchmark fails in predictable and detectable ways, and the proposed diagnostics recover this structure with high power.

The main caveat is finite-sample test calibration. The NegMass test is conservative under the null (about 1% rejection at nominal 5%), with median p-values around 0.65–0.73. This is a well-documented property of moment inequality tests at the boundary of the null: when most monotonicity constraints are slack, the contact-set bootstrap generates a null distribution that is too dispersed relative to the least favorable configuration.[5] For the Rank-1 test, the null rejection rate is 7% at the nominal 5% level, and the Monte Carlo standard error of the rejection rate is $\sqrt{0.05 \times 0.95/100} \approx 0.022$. In these designs, the test therefore appears close to nominal size. In applied work, I recommend reporting p-values directly alongside diagnostic statistics rather than relying solely on fixed-α rejection decisions.

The empirical section now turns to the two-arm Gaussian specification used for the continuous sentencing outcome. The simulation evidence above remains relevant because it validates the regime-selection logic, the diagnostic hierarchy, and the finite-sample behavior of the latent-regime machinery in the harder bail-convention case; the empirical implementation then adds an informative untreated arm and relies on the separate two-arm consistency result in Appendix I. The scaling exercise also includes the empirically relevant $J = 10$ case, which matches the bin count used in the application.

7 Empirical Application: Bail Decisions in Miami-Dade

I apply the latent-regime framework to felony bail decisions in Miami-Dade County using the administrative data assembled by Chyn et al. (2025). The setting is well suited to the framework: cases are quasi-randomly assigned to judges, judges exhibit persistent differences in release behavior, and institutional accounts suggest that judges may differ in *which* case features they weight—charge severity, criminal history, or courtroom demeanor—rather than only in overall strictness. The primitive assignment variable is judge identity J; the instrument Z is the scalar leave-one-out leniency score induced by J after residualizing on court-by-time fixed effects and discretizing into monotone bins. Accordingly, the empirical question is whether judge assignment operates through a

[5] The conservatism implies that power against weaker alternatives, e.g., smaller crossing magnitudes, would be reduced. Tighter calibration via subsampling or hybrid bootstrap methods (Andrews and Barwick, 2012) is a direction for future work.

single-index leniency score. Rejections below should therefore be read as failures of that scalar-leniency benchmark, not as direct nonparametric statements about literal judge identity. I focus on the continuous sentencing outcome $\log(1 + \text{days\_prison})$, for which the finite-Gaussian submodel identifies the latent basis up to permutation.

7.1 Data, institutional setting, and design choices

Treatment is pretrial release ($D_i = \mathbf{1}\{\text{bail met ever}\}$). The outcome is $\log(1 + \text{days\_prison})$, estimated using the two-arm Gaussian specification so that both treated and untreated arm means are modeled directly. Judge leniency is constructed using UJIVE (Goldsmith-Pinkham et al., 2025), then discretized into $L = 10$ ordered bins with first-stage-monotone bin ordering. Inference uses contact-set bootstrap tests with $B = 399$ draws.

Table 6: Empirical Sample Summary

Outcome	Sample	n	Release rate	$\mathbb{E}[Y \mid D = 1]$	$\mathbb{E}[Y \mid D = 0]$
$\log(1 + \text{days\_prison})$	Full	94,210	0.469	0.875	1.242
$\log(1 + \text{days\_prison})$	Black	49,381	0.450	1.002	1.318
$\log(1 + \text{days\_prison})$	Non-Black	44,829	0.491	0.746	1.153

Notes: Release rate is $\Pr(D = 1)$ with treatment $D = \mathbf{1}\{\text{bail\_met\_ever}\}$. Outcome is $\log(1 + \text{days\_prison})$.

7.2 Regime Structure and Model Selection

Table 7 reports cross-validated held-out log-likelihood by K, and Table 8 reports fitted regime weights and arm means at the selected $\hat{K}$.

Table 7: Cross-Validated Log-Likelihood by Regime Count

CV-LL ($K = 1$)	CV-LL ($K = 2$)	CV-LL ($K = 3$)	$\hat{K}$
-2.8486	1.5607	1.6268	3

Table 8: Estimated Regime Structure for $\log(1 + \text{days\_prison})$

Sample	$\hat{K}$	$\hat{w}$	$\hat{\mu}_1$	$\hat{\mu}_0$
Full	3	[0.053, 0.755, 0.191]	[1.134, 0.000, 4.723]	[3.086, 0.000, 5.298]
Black	3	[0.723, 0.068, 0.209]	[0.000, 1.148, 4.675]	[0.000, 3.086, 5.266]
Non-Black	3	[0.791, 0.046, 0.163]	[0.000, 1.102, 4.785]	[0.000, 3.176, 5.541]

Notes: Under the general convention, μ_1 and μ_0 are regime-specific means for treated and control arms.

Cross-validation selects $\hat{K} = 3$ in the full sample and in both race-stratified samples. The improvement from $K = 1$ to $K = 2$ is large, with a further modest gain at $K = 3$. The magnitude of the jump reflects improvement in the outcome-density channel, not merely the selection channel: at $K = 1$, a single Gaussian must fit the entire outcome distribution, whereas at $K \geq 2$ regime-specific Gaussians with smaller variances fit the treated and untreated outcome tails much better.

The estimated regime structure is economically interpretable. In the full sample, the dominant regime ($\hat{w}_2 = 0.755$) has essentially zero treatment effect in both arms, corresponding to defendants with negligible prison exposure regardless of release status. The remaining regimes separate a small high-effect group ($\hat{w}_1 = 0.053$, $\hat{\tau}_1 = -1.95$) from a larger moderate-effect group ($\hat{w}_3 = 0.191$, $\hat{\tau}_3 = -0.58$). Across racial groups, the regime-specific treatment effects are similar, but the regime weights and propensity schedules differ in ways that matter for the Wald-ratio decomposition below.

7.3 Diagnostic Evidence

Table 9 reports the formal diagnostic tests for the sentencing outcome.

Table 9: Empirical Diagnostic Tests for log(1 + days_prison)

Sample	n	$\hat{K}$	NegMass max		NegMass CvM		Rank-1	
			T	p	T	p	T	p
Full	94,210	3	5.439	0.0175**	85.157	0.0075***	0.571	0.0175**
Black	49,381	3	3.829	0.0250**	26.675	0.0200**	0.505	0.0050***
Non-Black	44,829	3	2.797	0.2225	19.261	0.1900	0.472	0.3275

Notes: Contact-set bootstrap with $B = 399$ and $J = 10$ leniency bins.

The full sample rejects both rank-1 and NegMass at conventional levels, indicating that the leniency score induced by judge assignment does not behave like a one-dimensional shift and that some leniency-bin comparisons induce offsetting regime flows. The race-stratified results reveal a sharp asymmetry. For Black defendants, both rank-1 and NegMass reject: the leniency score fails the scalar-leniency benchmark and behaves as if multiple latent margins with nontrivial cancellation are being activated. For Non-Black defendants, neither test rejects, so the standard single-index interpretation is a reasonable approximation for that score in the subsample.

This asymmetry is central for interpretation. Among Black defendants, different leniency-bin comparisons activate different regime mixtures and can even place negative net weight on some regimes. Among Non-Black defendants, the fitted regime schedules are much closer to proportional shifts, so the Wald ratio behaves more like a conventional complier-weighted

average.

7.4 What the Wald Ratio Aggregates

The diagnostic tests establish *that* the single-index model fails; the decomposition shows *how*. Table 10 reports, for the strict-versus-lenient comparison $(1, 10)$ in each sample: the scalar Wald estimand $\hat{\psi}$, regime-specific treatment effects $\hat{\tau}_k$, signed slope weights $\hat{\omega}_k^*$, regime contributions $\hat{\omega}_k^* \hat{\tau}_k$, and the largest signed-weight distortion across valid pairs.

Table 10: Strict-vs-Lenient Wald Decomposition by Regime for $\log(1 + \text{days\_prison})$

Sample	Pair	$\hat{\psi}$	$\hat{\tau}$	$\hat{\omega}^*$	$\hat{\omega}^* \odot \hat{\tau}$	$\max_{(z,z')}\{\sum_k \lvert\hat{\omega}_k^*\rvert - 1\}$
Full	$(1, 10)$	-0.098	[-1.95, +0.00, -0.58]	[+0.02, +0.87, +0.12]	[-0.03, +0.00, -0.07]	+0.425
Black	$(1, 10)$	-0.230	[+0.00, -1.94, -0.59]	[+0.79, +0.08, +0.14]	[+0.00, -0.15, -0.08]	+0.392
Non-Black	$(1, 10)$	-0.047	[+0.00, -2.07, -0.76]	[+0.95, +0.01, +0.04]	[+0.00, -0.02, -0.03]	+0.385

Notes: Pair is the largest first-stage gap among valid judge-bin comparisons ($|\Delta\hat{p}| \geq 0.03$). $\hat{\psi}$ is the model-implied Wald ratio, $\hat{\tau}$ are regime treatment effects, and $\hat{\omega}^*$ are signed slope weights. The final column reports the largest signed-weight distortion observed across valid pairs.

Two patterns are worth noting. First, in the full sample the dominant zero-effect regime absorbs most of the slope weight, so the scalar Wald ratio is small in magnitude not because release has no effect on sentencing, but because judge variation primarily moves defendants in a regime with little prison exposure. Second, the Black–Non-Black difference in the sentencing Wald ratio is driven much more by weights than by regime-specific treatment effects. The high-effect regime is similar across groups ($\hat{\tau} \approx -2$), but Black defendants place meaningfully more marginal weight on the non-zero-effect regimes than Non-Black defendants do.

7.5 Implications for discrimination measurement

The race asymmetry in the diagnostics has a direct consequence for outcome-test style comparisons. Under the standard single-index framework, the Black–Non-Black gap in Wald ratios is interpreted as a like-for-like difference in marginal released risk. In the latent-regime framework, that gap decomposes into a within-regime channel and a margin-composition channel:

$$\hat{\psi}^{\text{Black}} - \hat{\psi}^{\text{Non-Black}} = \underbrace{\sum_k \bar{\omega}_k \left(\hat{\tau}_k^{\text{Black}} - \hat{\tau}_k^{\text{Non-Black}}\right)}_{\text{within-regime}} + \underbrace{\sum_k \left(\hat{\omega}_k^{*\text{Black}} - \hat{\omega}_k^{*\text{Non-Black}}\right) \bar{\tau}_k}_{\text{margin-composition}},$$

where $\bar{\omega}_k$ and $\bar{\tau}_k$ are symmetric midpoint references. To ensure comparability, both race groups are estimated at the common full-sample value $K = 3$ and regime labels are aligned by joint distance in treatment effects and propensity schedules, as described in Section 5.2. The group-specific Wald ratios reported below are therefore computed from this common-K, cross-group aligned refit rather than copied from Table 10; they need not coincide exactly with the standalone values even when the common K equals a group's selected regime count.

Table 11: Black–Non-Black MOT Gap Decomposition for log(1 + days_prison)

Pair	$\hat{\psi}^B$	$\hat{\psi}^{NB}$	Gap	Within	Composition	B_{valid}
$(1, 10)$	-0.230	-0.011	-0.219	-0.002	-0.217	399
	(0.215)	(5.541)	(5.552)	(5.201)	(1.819)	

Notes: Decomposition uses equation (5.4) on the strict-vs-lenient pair $(1, 10)$. Bootstrap standard errors (parentheses) are from a parametric bootstrap with re-fitting under the general convention. Both race groups are estimated with a common regime count $K = 3$ to maintain channel comparability. Because the decomposition is computed from a separate common-K refit with cross-group label alignment, the reported group-specific Wald ratios need not coincide exactly with the standalone values in Table 10, even when the common K equals a group's selected regime count. The bootstrap distribution is heavy-tailed because some replicates generate near-zero effective first stages within non-trivial regimes, so the percentile intervals reported in the text are more informative than Wald-type intervals.

This is the identified decomposition in the paper: under the Gaussian submodel, the latent basis is identified up to permutation. The point estimates attribute essentially the entire cross-race gap to margin composition (-0.217 out of a total gap of -0.219), with a negligible within-regime component. The bootstrap distribution is heavy-tailed because some replicates generate near-zero effective first stages within the non-trivial regimes, so the confidence intervals remain wide. I therefore interpret the sentencing decomposition as directional evidence that composition, rather than within-regime outcome differences, is carrying the gap.

Economic interpretation. The decomposition implies that the racial difference in marginal sentencing outcomes is not primarily coming from judges treating comparable Black and Non-Black defendants differently within the same decision environment. Instead, the judicial instrument activates different latent margins for the two groups. In that sense, the standard scalar outcome test conflates who is marginal with what happens within a given regime.

7.6 Robustness checks

Two appendix-based checks matter for interpretation. First, Appendix H.1 reports the sentencing decomposition at $K = 1, 2, 3$. The main takeaway is that $K = 1$ is clearly

inadequate: by construction it assigns the entire gap to the within-regime channel and shuts down the composition channel. That is the strongest scalar-leniency benchmark, and it is inconsistent with the within-group rank-1 evidence once the model allows multiple regimes. Once $K \geq 2$, the composition channel remains dominant, and the $K = 2$ and $K = 3$ decompositions are close.

Second, Appendix H.2 reports descriptive reweighted-tercile outcome comparisons for the sentencing outcome. Because the propensity-based posteriors in this implementation depend only on residualized leniency bins, these comparisons are not direct tests of within-regime outcome stability. Their value is descriptive: after regime-specific reweighting, the race-stratified tercile distributions do not display large residual discrepancies. I therefore do not treat them as validating Assumption 2.4, and instead rely on the sensitivity analysis in Appendix F to quantify the remaining scope for within-regime selection.

That sensitivity analysis is the formal robustness check for the decomposition. For the sentencing outcome at the anchor pair $(1, 10)$, a nontrivial benchmark of $\delta = 0.02$ implies at most 0.04 worst-case bias in the within-regime channel, versus an estimated composition component of 0.217; under the race-neutral benchmark, the within-regime channel is unaffected and the composition component shifts by at most 0.007. So the main empirical message does not rest on the tercile comparison: it survives moderate within-regime selection and is overturned only under substantially larger race-differential selection.

7.7 Discussion

The empirical evidence supports the paper's core claim in the following narrower sense: in this application, the leniency score induced by judge assignment behaves like a vector instrument that reweights multiple latent margins, and single-index interpretations miss economically relevant heterogeneity. The rank-1 restriction fails precisely in the subsample where the standard interpretation would be most policy-relevant, and the Gaussian decomposition shows that the Black–Non-Black gap in the Wald ratio is driven primarily by margin composition rather than within-regime outcome differences. The confidence intervals are wide, but the formal sensitivity analysis shows that overturning this composition-dominant interpretation would require much larger race-differential within-regime selection than the moderate benchmarks considered in Appendix F. The same leniency-score variation therefore means different things for different groups because it activates different latent margins.

8 Conclusion

This paper develops an econometric framework for leniency designs in which decision-makers differ along multiple latent margins, varying in *what* they weight and not only in how strictly they decide. Relative to the canonical MTE model, it allows multidimensional selection across latent regimes at the cost of ruling out within-regime selection on potential outcomes. The framework's estimable objects are a slope-weighted decomposition of the Wald ratio into regime-specific treatment effects, a decomposition of cross-group disparities into within-regime and margin-composition channels, and diagnostics for when multi-margin structure is empirically relevant.

In an application to felony bail decisions in Miami-Dade County, the scalar leniency score fails the rank-1 benchmark for Black defendants but not for Non-Black defendants, and the cross-race gap in the Wald ratio is driven primarily by which decision environments the instrument activates rather than by within-regime outcome differences. The finding is robust across regime counts $K \geq 2$ and survives moderate within-regime selection under the formal sensitivity analysis.

The framework turns the standard single-index interpretation from an untested maintained assumption into a testable empirical restriction, and is most useful when the data reject a scalar account of selection. Extensions to continuous instruments, multi-valued treatments, or dynamic decision environments are natural next steps.

A Proofs

Proof of Theorem 3.1

The proof records the weak nonparametric result stated in the main text. I state the argument for a generic arm d satisfying the assumptions; the same logic applies to each arm independently.

Step (i): K_0 is the span dimension. By (3.1), each g_{dz} lies in $V := \text{span}\{f_{dk}\}_{k=1}^{K_0}$; Assumption 3.1 gives $\dim V = K_0$, so the span dimension is at most K_0. Conversely, choose $z_1, \ldots, z_L$ as in Assumption 3.2, where $L \geq K_0$ and $\Omega_d := [\omega_{dk}(z_\ell)]_{\ell,k}$ has rank K_0. Since $\text{rank}(\Omega_d) = K_0$, there exist K_0 rows, say those indexed by $z_{\ell_1}, \ldots, z_{\ell_{K_0}}$, forming a nonsingular $K_0 \times K_0$ submatrix Ω_d^*. Restricting attention to these K_0 instrument values: if $\sum_{j=1}^{K_0} c_j g_{dz_{\ell_j}} \equiv 0$, then (3.1) implies

$$\sum_{k=1}^{K_0} \Big(\sum_j c_j \omega_{dk}(z_{\ell_j}) \Big) f_{dk} \equiv 0,$$

hence $\sum_j c_j \omega_{dk}(z_{\ell_j}) = 0$ for each k by Assumption 3.1. In matrix form, $c^\top \Omega_d^* = 0$ where $c \in \mathbb{R}^{K_0}$ and Ω_d^* is $K_0 \times K_0$ nonsingular, so $c = 0$. Therefore $\{g_{dz_{\ell_j}}\}_{j=1}^{K_0}$ are linearly independent and the span dimension equals K_0.

Step (ii): What remains unidentified under the weak assumptions. The same observable family $\{g_{dz}\}_{z \in \mathcal{Z}}$ can admit multiple K_0-component representations when the component basis is allowed to rotate within the identified span. Concretely, if $q = (q_1, \ldots, q_{K_0})^\top$ is any basis of the same span as $(f_{d1}, \ldots, f_{dK_0})^\top$ and A is an invertible matrix satisfying $q = A f_d$, then the corresponding coefficients can be transformed as $\tilde{\omega}_d(z) = \omega_d(z) A^{-1}$. Whenever the transformed coefficients remain admissible weights and the transformed basis elements remain admissible component laws, the observable mixture family is unchanged. Assumptions 3.1–3.2 rule out collapse of the span dimension, but they do not by themselves rule out such change-of-basis indeterminacy.

Step (iii): Stronger routes to identifying the latent basis. To recover regime-specific component laws, posterior weights, and causal contrasts, one must add further structure that rules out non-permutation changes of basis. Theorem A.1 gives one such route under nonparametric tail-shape restrictions that replace exact support exclusions with asymptotic tail ordering. The empirical continuous-outcome specification uses a finite-Gaussian component class, for which classical Gaussian-mixture results identify the component parameters and hence the unrestricted arm-specific weights.

The computational procedure is separate from the identification argument. The WFR-

based Kiefer–Wolfowitz NPMLE provides a flexible way to fit the maintained model class, but its fitted components receive a structural interpretation only under the identification conditions imposed above.

Proof of Proposition 3.2

Using Lemma 2.1, within regime k,

$$\begin{aligned}\mathbb{E}[Y \mid X = x, Z = z, S = k] &= \mathbb{E}[Y(0) \mid X = x, S = k] + p_k(x, z)\, \mathbb{E}[Y(1) - Y(0) \mid X = x, S = k] \\ &= \mu_{0k}(x) + p_k(x, z)\, \tau_k(x),\end{aligned}$$

where $\mu_{0k}(x) := \mathbb{E}[Y(0) \mid X = x, S = k]$. Taking expectations over S with weights $w_k(x)$ yields

$$m(x, z) = \sum_{k=1}^{K_0} w_k(x)\mu_{0k}(x) + \sum_{k=1}^{K_0} w_k(x)p_k(x, z)\tau_k(x).$$

Differentiate with respect to z to obtain

$$\partial_z m(x, z) = \sum_{k=1}^{K_0} w_k(x)\tau_k(x)\partial_z p_k(x, z).$$

Similarly, $p(x, z) = \sum_{k=1}^{K_0} w_k(x)p_k(x, z)$ implies

$$\partial_z p(x, z) = \sum_{k=1}^{K_0} w_k(x)\partial_z p_k(x, z).$$

Dividing yields (3.9)–(3.10).

Proof of Proposition 3.1

The argument is pointwise in x: fix a covariate value x and suppress it from the notation throughout.

The proof has two steps.

Step 1: Distinct Gaussian densities are linearly independent. Suppress the arm index d and write

$$\phi_k(y) := \phi(y; \mu_k, \sigma_k^2), \qquad k = 1, \ldots, K,$$

with pairwise distinct parameter pairs $\{(\mu_k, \sigma_k^2)\}_{k=1}^K$. Suppose

$$\sum_{k=1}^{K} c_k \phi_k(y) \equiv 0, \qquad y \in \mathbb{R},$$

for some real coefficients $c_1, \ldots, c_K$.

Taking moment generating functions yields

$$\sum_{k=1}^{K} c_k \exp\bigl(\mu_k t + \tfrac{1}{2}\sigma_k^2 t^2\bigr) \equiv 0, \qquad t \in \mathbb{R}.$$

This step is specific to Gaussian components: it uses the existence of the moment generating function for all $t \in \mathbb{R}$ and therefore does not automatically extend to heavy-tailed component classes. Order the indices lexicographically by (σ_k^2, μ_k): first by largest variance, and among ties by largest mean. Let m denote the unique maximal index. Divide the identity by

$$\exp\bigl(\mu_m t + \tfrac{1}{2}\sigma_m^2 t^2\bigr).$$

This gives

$$c_m + \sum_{k \neq m} c_k \exp\bigl((\mu_k - \mu_m)t + \tfrac{1}{2}(\sigma_k^2 - \sigma_m^2)t^2\bigr) \equiv 0.$$

For each $k \neq m$, either $\sigma_k^2 < \sigma_m^2$, in which case the exponent has a strictly negative quadratic coefficient, or $\sigma_k^2 = \sigma_m^2$ and $\mu_k < \mu_m$, in which case the quadratic coefficient is zero and the linear coefficient is strictly negative. Hence every term in the sum vanishes as $t \to +\infty$. Taking the limit therefore yields

$$c_m = 0.$$

Removing the mth term and repeating the same argument iteratively shows that $c_k = 0$ for all k. Thus the Gaussian densities $\{\phi(y; \mu_k, \sigma_k^2)\}_{k=1}^K$ are linearly independent.

Step 2: Recover the component parameters and weights. At the instrument value $\bar{z}$,

$$g_{d\bar{z}}(y) = \sum_{k=1}^{K} \omega_{dk}(\bar{z})\, \phi(y; \mu_{dk}, \sigma_{dk}^2),$$

with all weights strictly positive by assumption. By the classical identifiability of finite Gaussian mixtures (Teicher, 1963; Yakowitz and Spragins, 1968), the mixing measure

$$\{(\omega_{dk}(\bar{z}), \mu_{dk}, \sigma_{dk}^2)\}_{k=1}^K$$

is uniquely determined from $g_{d\bar{z}}$ up to permutation. In particular, the set of Gaussian parameter pairs $\{(\mu_{dk}, \sigma_{dk}^2)\}_{k=1}^K$ is identified up to permutation.

Now fix any other instrument value z. Since the Gaussian parameters are already identified and common across instrument values, the representation

$$g_{dz}(y) = \sum_{k=1}^{K} \omega_{dk}(z)\, \phi(y; \mu_{dk}, \sigma_{dk}^2)$$

is a linear combination of known, linearly independent functions. By Step 1, its coefficients are uniquely determined. Hence the weights $\{\omega_{dk}(z)\}_{k=1}^K$ are identified, again up to the same common permutation of labels established at $\bar{z}$.

This proves identification of the arm-specific Gaussian parameters and weights. If the same conclusion holds for both arms and the cross-arm pairing is uniquely determined, then

$$w_k = \omega_{1k}(z)\, p(z) + \omega_{0k}(z)\, [1 - p(z)]$$

follows from the two-arm accounting identity, and

$$\pi_k(z) = p(z)\omega_{1k}(z), \qquad p_k(z) = \frac{\pi_k(z)}{w_k}$$

follow immediately.

Nonparametric Identification Under Ordered Tail Dominance

Theorem A.1 (Nonparametric identification under ordered tail dominance)**.** *Fix x and an arm $d \in \{0, 1\}$. Suppose the observed family admits the minimal K_0-component representation*

$$g_{dz}(y) = \sum_{k=1}^{K_0} \omega_{dk}(z)\, f_{dk}(y), \qquad z \in \mathcal{Z},$$

where each f_{dk} is a probability density on $(\mathcal{Y}, \mu)$, $\omega_{dk}(z) \geq 0$, and $\sum_{k=1}^{K_0} \omega_{dk}(z) = 1$ for every z. Let

$$\underline{y} := \inf \mathcal{Y}, \qquad \bar{y} := \sup \mathcal{Y},$$

and define the lower-tail and upper-tail masses

$$L_{dk}(t) := \int_{(-\infty, t] \cap \mathcal{Y}} f_{dk}(y)\, d\mu(y), \qquad U_{dk}(t) := \int_{[t, \infty) \cap \mathcal{Y}} f_{dk}(y)\, d\mu(y).$$

Assume:

(P1) ***Minimality.*** *The observed family* $\{g_{dz}\}_{z\in\mathcal{Z}}$ *has span dimension* K_0.

(P2) ***Ordered tail dominance of the true components.*** *After a relabeling of the true components, for every* $j > k$,

$$\frac{L_{dj}(t)}{L_{dk}(t)} \to 0 \qquad \text{as } t \downarrow \underline{y},$$

and

$$\frac{U_{dk}(t)}{U_{dj}(t)} \to 0 \qquad \text{as } t \uparrow \bar{y}.$$

Moreover, for each k, $L_{dk}(t) > 0$ *for all* t *sufficiently close to* $\underline{y}$, *and* $U_{dk}(t) > 0$ *for all* t *sufficiently close to* $\bar{y}$.

(P3) ***Maintained tail-order class restriction.*** *Any alternative* K_0*-component representation of the same observed family by nonnegative densities with nonnegative mixing weights summing to one also satisfies* (P2) *after a relabeling of its component indices.*

Then the component densities $\{f_{dk}\}_{k=1}^{K_0}$ *and the weights* $\{\omega_{dk}(z)\}_{k=1}^{K_0}$ *are identified from the observed family* $\{g_{dz}\}_{z\in\mathcal{Z}}$ *up to a common permutation of component labels.*

Proof. Suppress the arm index d for notational simplicity. Let

$$g_z(y) = \sum_{k=1}^{K_0} \omega_k(z) f_k(y)$$

be the given representation, and let

$$g_z(y) = \sum_{k=1}^{K_0} \tilde{\omega}_k(z) \tilde{f}_k(y)$$

be any alternative representation satisfying Condition (P3) of Theorem A.1. After relabeling the alternative components, we may assume that both $(f_1, \ldots, f_{K_0})$ and $(\tilde{f}_1, \ldots, \tilde{f}_{K_0})$ satisfy Condition (P2) in the same order.

Because the observed family has span dimension K_0, the true components $\{f_k\}_{k=1}^{K_0}$ must be linearly independent. Likewise, if the alternative components were linearly dependent, then their span would have dimension strictly less than K_0, contradicting the fact that the observed family $\{g_z\}_{z\in\mathcal{Z}}$ lies in that span and has span dimension K_0. Hence both tuples are linearly independent bases of the same K_0-dimensional subspace of $L^1(\mu)$. Therefore there

exists an invertible $K_0 \times K_0$ matrix A such that

$$\tilde{f} = Af, \qquad \tilde{\omega}(z) = \omega(z)A^{-1},$$

where $f = (f_1, \ldots, f_{K_0})^\top$ and $\tilde{f} = (\tilde{f}_1, \ldots, \tilde{f}_{K_0})^\top$.

Fix a row index r. Write

$$\tilde{f}_r = \sum_{j=1}^{K_0} A_{rj} f_j.$$

Let

$$m(r) := \min\{j : A_{rj} \neq 0\}, \qquad M(r) := \max\{j : A_{rj} \neq 0\}.$$

Also define

$$\tilde{L}_r(t) := \int_{(-\infty,t]\cap\mathcal{Y}} \tilde{f}_r(y)\, d\mu(y), \qquad \tilde{U}_r(t) := \int_{[t,\infty)\cap\mathcal{Y}} \tilde{f}_r(y)\, d\mu(y).$$

Then

$$\tilde{L}_r(t) = \sum_{j=1}^{K_0} A_{rj} L_j(t), \qquad \tilde{U}_r(t) = \sum_{j=1}^{K_0} A_{rj} U_j(t).$$

Because $L_j(t)/L_{m(r)}(t) \to 0$ for every $j > m(r)$ as $t \downarrow \underline{y}$, we have

$$\frac{\tilde{L}_r(t)}{L_{m(r)}(t)} = A_{r,m(r)} + \sum_{j>m(r)} A_{rj} \frac{L_j(t)}{L_{m(r)}(t)} \longrightarrow A_{r,m(r)}.$$

Since $\tilde{L}_r(t) \geq 0$ for all t and $L_{m(r)}(t) > 0$ for all t sufficiently close to $\underline{y}$ by Condition (P2), it follows that $A_{r,m(r)} \geq 0$. By definition of $m(r)$, this coefficient is nonzero, so in fact

$$A_{r,m(r)} > 0.$$

Thus

$$\tilde{L}_r(t) = A_{r,m(r)} L_{m(r)}(t)\{1 + o(1)\} \qquad \text{as } t \downarrow \underline{y}.$$

Similarly, because $U_j(t)/U_{M(r)}(t) \to 0$ for every $j < M(r)$ as $t \uparrow \bar{y}$, we have

$$\frac{\tilde{U}_r(t)}{U_{M(r)}(t)} = A_{r,M(r)} + \sum_{j<M(r)} A_{rj} \frac{U_j(t)}{U_{M(r)}(t)} \longrightarrow A_{r,M(r)}.$$

Since $\tilde{U}_r(t) \geq 0$ and $U_{M(r)}(t) > 0$ for all t sufficiently close to $\bar{y}$, it follows that

$$A_{r,M(r)} > 0.$$

Hence

$$\tilde{U}_r(t) = A_{r,M(r)} U_{M(r)}(t)\{1 + o(1)\} \qquad \text{as } t \uparrow \bar{y}.$$

We next show that the indices $m(r)$ are strictly increasing in r. Fix $r < s$. Because the alternative components satisfy Condition (P2),

$$\frac{\tilde{L}_s(t)}{\tilde{L}_r(t)} \to 0 \qquad \text{as } t \downarrow \underline{y}.$$

But the lower-tail expansions imply

$$\frac{\tilde{L}_s(t)}{\tilde{L}_r(t)} = \frac{A_{s,m(s)} + o(1)}{A_{r,m(r)} + o(1)} \cdot \frac{L_{m(s)}(t)}{L_{m(r)}(t)}.$$

If $m(s) = m(r)$, the ratio converges to the strictly positive constant $A_{s,m(s)}/A_{r,m(r)}$, contradiction. If $m(s) < m(r)$, then by Condition (P2),

$$\frac{L_{m(r)}(t)}{L_{m(s)}(t)} \to 0,$$

so

$$\frac{L_{m(s)}(t)}{L_{m(r)}(t)} \to \infty,$$

again a contradiction. Therefore

$$m(r) < m(s).$$

Thus $m(1), \ldots, m(K_0)$ is a strictly increasing sequence of integers in $\{1, \ldots, K_0\}$, so necessarily

$$m(r) = r \qquad \text{for each } r = 1, \ldots, K_0.$$

An analogous argument using upper tails shows that the indices $M(r)$ are also strictly increasing. Indeed, for $r < s$, Condition (P2) gives

$$\frac{\tilde{U}_r(t)}{\tilde{U}_s(t)} \to 0 \qquad \text{as } t \uparrow \bar{y},$$

while the upper-tail expansions imply

$$\frac{\tilde{U}_r(t)}{\tilde{U}_s(t)} = \frac{A_{r,M(r)} + o(1)}{A_{s,M(s)} + o(1)} \cdot \frac{U_{M(r)}(t)}{U_{M(s)}(t)}.$$

If $M(r) = M(s)$, the ratio converges to the strictly positive constant $A_{r,M(r)}/A_{s,M(s)}$, contra-

diction. If $M(r) > M(s)$, then by Condition (P2),

$$\frac{U_{M(s)}(t)}{U_{M(r)}(t)} \to 0,$$

so

$$\frac{U_{M(r)}(t)}{U_{M(s)}(t)} \to \infty,$$

again a contradiction. Therefore

$$M(r) < M(s).$$

Thus

$$M(r) = r \qquad \text{for each } r = 1, \ldots, K_0.$$

Since for each row r we have $m(r) = M(r) = r$, row r of A has exactly one nonzero entry, namely in column r. Hence A is diagonal. Finally, because both f_k and $\tilde{f}_k$ are probability densities,

$$\sum_{j=1}^{K_0} A_{rj} = 1 \qquad \text{for each } r.$$

Since A is diagonal, this implies $A_{rr} = 1$ for every r, so $A = I$ in the chosen ordering. Undoing the relabeling of the alternative components shows that the original change-of-basis matrix is a permutation matrix.

Therefore the component densities and weights are uniquely determined up to a common permutation of the component labels. □

B Binary-Outcome Extension: Common-Index Logistic Mixtures

This appendix records an analogous identification result for binary outcomes under a fixed-basis common-index logistic class. The result is included as an extension rather than part of the active empirical path. Its role is to show that the fixed-class identification logic is not specific to Gaussian outcomes. The result is related to Follmann and Lambert (1991), who study identifiability of finite mixtures of logistic regressions under discrete-design conditions. The proof below uses the common-index structure together with interval support of the index, so the route is different from theirs.

Proposition B.1 (Point identification for common-index logistic mixtures)**.** *Fix an arm $d \in \{0, 1\}$ and a sieve level m. Let $\psi_m(x)$ denote a fixed m-dimensional basis in the covariates,*

and let

$$u_d(x) := \beta_d^\top \psi_m(x)$$

be a common index whose support contains a nondegenerate interval. Suppose that, for a binary outcome $Y \in \{0,1\}$*,*

$$\Pr(Y = 1 \mid D = d, X = x, Z = z) = \sum_{k=1}^{K} \omega_{dk}(z)\, \Lambda(\gamma_{dk} + u_d(x))\,,$$

where $\Lambda(u) = e^u/(1+e^u)$*, the intercepts* $\gamma_{d1}, \ldots, \gamma_{dK}$ *are pairwise distinct, and the weights satisfy* $\omega_{dk}(z) \geq 0$ *and* $\sum_{k=1}^{K} \omega_{dk}(z) = 1$ *for every* z*. Assume there exist instrument values* $z_1, \ldots, z_K$ *such that the* $K \times K$ *weight matrix*

$$\Omega_d := \big(\omega_{dk}(z_\ell)\big)_{\ell,k}$$

is nonsingular.

Then the component success functions

$$q_{dk}(x) := \Lambda(\gamma_{dk} + u_d(x))$$

and the weights $\{\omega_{dk}(z)\}_{k=1}^{K}$ *are identified from the family of binary regression functions*

$$x \mapsto \Pr(Y = 1 \mid D = d, X = x, Z = z)$$

up to a common permutation of component labels.

Proof. Write

$$g_{dz}(x) := \Pr(Y = 1 \mid D = d, X = x, Z = z) = \sum_{k=1}^{K} \omega_{dk}(z)\, q_{dk}(x).$$

Step 1: a common-index logistic component cannot be a nontrivial linear combination of distinct members of the same class. Consider the admissible component class

$$\mathcal{A}_{m,d}^{\log} := \{x \mapsto \Lambda(\gamma + u_d(x)) : \gamma \in \mathbb{R}\}$$

and let $q_k(x) = \Lambda(\gamma_k + u_d(x))$ for pairwise distinct intercepts $\gamma_1, \ldots, \gamma_r$, with $r \leq K$.

Suppose

$$h(x) = \Lambda(\gamma + u_d(x))$$

lies in $\operatorname{span}\{q_1, \dots, q_r\}$. Then there exist coefficients $a_1, \dots, a_r$ such that, writing $u = u_d(x)$,

$$\Lambda(\gamma + u) = \sum_{k=1}^{r} a_k \Lambda(\gamma_k + u)$$

for all u in a nondegenerate interval by the support assumption on the common index.

Set $t = e^u$, $c = e^{-\gamma}$, and $c_k = e^{-\gamma_k}$. Then on an open interval of $t > 0$,

$$\frac{t}{t + c} = \sum_{k=1}^{r} a_k \frac{t}{t + c_k}.$$

Since both sides are rational functions of t, equality on an open interval implies equality for all admissible t. Dividing by t and rearranging,

$$\frac{1}{t + c} = \sum_{k=1}^{r} a_k \frac{1}{t + c_k}.$$

By uniqueness of partial-fraction decompositions with simple poles, this is possible only if $c = c_j$ for some j and $a_j = 1$ while $a_k = 0$ for all $k \neq j$. Hence $h = q_j$.

Step 2: distinct members of $\mathcal{A}^{\log}_{m,d}$ are linearly independent. If

$$\sum_{k=1}^{r} b_k q_k(x) = 0$$

on a nondegenerate interval of the common index, then the same change of variables gives

$$\sum_{k=1}^{r} \frac{b_k}{t + c_k} = 0$$

for all admissible $t > 0$. Uniqueness of the partial-fraction decomposition therefore forces $b_k = 0$ for every k.

Step 3: any alternative logistic decomposition must use the same component functions up to relabeling. Now consider any alternative representation of the same regression-function family by K functions in $\mathcal{A}^{\log}_{m,d}$:

$$g_{dz}(x) = \sum_{k=1}^{K} \tilde{\omega}_{dk}(z)\, \tilde{q}_{dk}(x).$$

By Step 2, $\{q_{dk}\}_{k=1}^{K}$ is linearly independent. Since Ω_d is nonsingular, the observable family at $z_1, \dots, z_K$ spans the same K-dimensional space as $\{q_{dk}\}_{k=1}^{K}$. The same observable functions

$g_{dz_1}, \ldots, g_{dz_K}$ also lie in $\text{span}\{\tilde{q}_{d1}, \ldots, \tilde{q}_{dK}\}$, so that span must have dimension at least K. Because it is generated by only K functions, it has dimension exactly K, and therefore the alternative functions $\{\tilde{q}_{dk}\}_{k=1}^{K}$ are linearly independent as well. In particular, each $\tilde{q}_{dk}$ lies in $\text{span}\{q_{d1}, \ldots, q_{dK}\}$. Step 1 then implies that each $\tilde{q}_{dk}$ must equal one of the q_{dj}. Because the alternative functions are linearly independent, no two of them can equal the same q_{dj}, so this correspondence is a permutation. Thus the component success functions are identified up to a common relabeling.

Step 4: the weights are the unique coordinates in the identified basis. Once the component functions are pinned down, each observable regression function g_{dz} has a unique coefficient vector in the identified basis $\{q_{d1}, \ldots, q_{dK}\}$ because that basis is linearly independent. Those coefficients are exactly $\{\omega_{dk}(z)\}_{k=1}^{K}$. Therefore the weights are identified up to the same permutation. □

Remark B.1 (Scope of the logistic extension)**.** *Proposition B.1 is a fixed-basis result. The basis $\psi_m(x)$ may be enriched with m, but at each fixed sieve level the admissible component class remains the fixed common-index family*

$$x \mapsto \Lambda(\gamma + u_d(x)),$$

so latent regimes differ only through intercept shifts in a common index. This does not cover generic logistic mixtures with regime-specific slope vectors or unrestricted nonlinear components. The result is included to document a binary-outcome identification route under a stronger structured class, not to claim identification of unrestricted binary varying-weight mixtures beyond the span-based results above.

C Role of Within-Regime Outcome Stability

This appendix collects the precise role of Assumption 2.4 (within-regime outcome stability, $(Y(1), Y(0)) \perp V \mid (X, S)$) in the identification argument.

What holds under Assumptions 2.1–2.3 alone

Under the first three assumptions:

- The joint law $\mathcal{L}(Y, D \mid X = x, Z = z)$ is a K_0-component finite mixture.
- The regime weights $w_k(x)$ do not depend on z.
- The regime-specific first stages $p_k(x, z) := \Pr(D = 1 \mid X = x, Z = z, S = k)$ remain defined as latent treatment probabilities within regime.

However, the conditional outcome law among treated units in regime k,

$$f_{Y|D=1,X=x,Z=z,S=k}(y),$$

can depend on z. This is because different z values induce different selection sets $\{g_k(x,z,V) \geq 0\}$, and if $Y(1)$ depends on V given (X,S), the selected distribution of $Y(1)$ shifts with z. Consequently, the mixture (3.1) fails: both the components *and* the weights vary with z, and the varying-weight mixture identification framework does not apply.

Assumption 2.4 imposes $(Y(1),Y(0)) \perp V \mid (X,S)$: conditional on regime membership, potential outcomes are independent of the selection heterogeneity V. This ensures

$$\mathcal{L}(Y(1) \mid D(z)=1, X=x, S=k) = \mathcal{L}(Y(1) \mid X=x, S=k) = F_{1k}(y \mid x),$$

which is z-invariant. The mixture (3.1) then holds with z-invariant components, and both the span-based result and the stronger identification arguments become available.

Economic interpretation

An intuitive way to read Assumption 2.4 is through analogy to the panel-data literature on discrete heterogeneity or latent groups. In those models, a finite latent type absorbs the persistent, outcome-relevant unobserved heterogeneity, and the remaining disturbance is treated as idiosyncratic. The present framework uses the same logic, but in a stronger way. The latent regime S is not only a device for parsimoniously approximating rich heterogeneity; it also organizes the causal structure of the model. Conditional on (X,S), the remaining heterogeneity V may still affect the treatment decision, but it is restricted not to carry additional information about potential outcomes. Equivalently, S absorbs the outcome-relevant unobserved heterogeneity, while V is an idiosyncratic selection shifter. This is the sense in which the model concentrates the economically meaningful latent heterogeneity into a finite regime index. The payoff from that restriction is the regime-invariant mixture representation: changing z changes the mixture weights through selection, but does not alter the regime-specific potential-outcome laws themselves.

As discussed in Remark 2.1, this is not a weakening or strengthening of the standard MTE framework but a different model entirely. The MTE framework allows within-margin selection (MTE(x,u) varies continuously with u) but restricts the selection process to be one-dimensional. My framework allows multidimensional selection (K_0 latent margins with non-proportional instrument responses) but restricts within-regime selection to be absent. The trade-off is favorable when the institutional question concerns *what* decision-makers weight (multidimensional selection is first-order) rather than fine-grained ranking within a

single dimension (within-regime selection is first-order). The plausibility of the trade-off depends on the richness of the regime structure: larger K_0 makes Assumption 2.4 more tenable because more of the selection heterogeneity is absorbed into regime membership.

D Further Estimation Theory

D.1 Likelihood and Parameterization

I observe i.i.d. data $\{(Y_i, D_i, X_i, Z_i)\}_{i=1}^n$, where $D_i \in \{0,1\}$ is the treatment decision, $X_i \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ (compact) are covariates, and $Z_i \in \mathcal{Z} = \{1, \ldots, J\}$ is a finite instrument (for example, a discretized leniency bin induced by decision-maker assignment). Throughout, $K \geq K_0$ is a working upper bound on the number of latent regimes; I discuss selecting K in Section 5.4.2.

Two-arm extension. When untreated outcomes are non-degenerate and informative, a natural two-arm extension augments the bail-convention model with a second arm-specific mixture:

$$h(y, d \mid x, z) := \begin{cases} \displaystyle\sum_{k=1}^{K} \pi_k(x,z)\, f_{1k}(y \mid x), & d = 1, \\ \displaystyle\sum_{k=1}^{K} \big[w_k(x) - \pi_k(x,z)\big]\, f_{0k}(y \mid x), & d = 0, \end{cases} \tag{D.1}$$

where $\pi_k(x,z) = \Pr(D = 1, S = k \mid X = x, Z = z)$ and the untreated-arm weights are $w_k(x) - \pi_k(x,z) = \Pr(D = 0, S = k \mid X = x, Z = z)$. Because the two arm-specific mixtures share common regime shares $w_k(x)$, this extension can in principle recover w_k and p_k separately in addition to the arm-specific outcome densities. I use this architecture empirically for outcomes with non-degenerate untreated outcomes; Appendix I states a finite-dimensional parametric MLE for this setting and establishes its consistency.

Specialization: the bail convention. When the untreated outcome is degenerate—as in bail designs where misconduct is observed only upon release—the $D = 0$ arm collapses to $f_{0k}(0 \mid x) = 1$ for all k, providing no regime-resolving outcome information. The identified parameter reduces to $\theta := \big\{\pi_k(\cdot,\cdot),\ f_{1k}(\cdot \mid \cdot)\big\}_{k=1}^K$, the separate factorization of π_k into w_k and p_k is no longer identified (Remark 3.2), and the likelihood simplifies to

$$h(y, d \mid x, z;\, \theta) = \begin{cases} \displaystyle\sum_{k=1}^{K} \pi_k(x,z)\, f_{1k}(y \mid x), & d = 1, \\ 1 - p(x,z), & d = 0,\ y = 0. \end{cases} \tag{D.2}$$

The treated-arm sum is the only term that carries regime-specific outcome information; the untreated-arm sum contributes only to estimation of the aggregate first stage $p(x, z)$.

Scope of the formal theory. The sieve framework above is stated at full generality: it allows the within-regime outcome densities f_{dk} to be represented flexibly through compactly supported kernel mixtures. Formal consistency results are proved separately for each convention: Theorem D.1 and Appendix E cover the bail-convention likelihood (D.2); Theorem I.1 and Appendix I cover the two-arm likelihood (D.1). At the identification level, Proposition 3.1 is already broader than a fully parametric model: it restricts the component outcome laws to a finite Gaussian family but leaves the arm-specific mixing weights unrestricted across instrument values. In the empirical application, I use a *parametric specialization* of that broader identified class in which each regime-arm density is a single Gaussian ($G_{dk} = \delta_\eta$, i.e., a point mass in the mixing-distribution space) for $\log(1 + \text{days\_prison})$ and the selection block is represented with a fixed softmax basis. This parametric restriction corresponds to setting G_{dk} to a point mass in $\mathcal{P}(\mathcal{H})$—a special case of the sieve estimator, not a different estimator. For the continuous-outcome Gaussian specification, Theorem I.1 directly covers the implemented parametric MLE under correct specification together with the identification condition in Assumption I.1. Identification is still stated conditional on covariates X, but in the empirical application the judge leniency instrument is first residualized on court-by-time fixed effects via UJIVE, so the identifying variation is already conditioned on the standard design strata. Conditional on this residualized instrument, the implemented estimator uses the covariate-restricted special case of the two-arm model: regime shares and arm-specific kernel parameters are taken to be constant within the estimation sample, while regime-specific treatment propensities vary only by judge-leniency bin. Under misspecification—if the true within-regime density is not well approximated by the chosen single-Gaussian family—the estimator converges to the best restricted KL projection of the true conditional law within the chosen parametric class. Whether this pseudo-true approximation is adequate for the estimated quantities of interest (π_k, τ_k, slope weights) is an empirical question. The functional stability diagnostics in Section 5.4.1 address K-regime adequacy: if the estimated quantities stabilize as K varies, the regime-count specification is not driving the conclusions. Kernel-shape adequacy is harder to diagnose directly; in the covariate-restricted implementation, the within-regime stability exercise is best read as a descriptive reweighted-tercile comparison rather than a direct test of the model's z-invariance restriction. Formal robustness to violations of within-regime stability is handled instead by the bounded-selection sensitivity analysis in Appendix F. The general KW–NPMLE framework is presented because it establishes that the identification and consistency arguments do not depend on the single-kernel restriction,

and because richer within-regime density approximations may be needed in applications where outcome distributions are multimodal or heavily skewed.

The sentencing application in the paper is estimated under the general convention using the two-arm EM architecture described in Section 5.5.

D.2 Sieve Parameterization

Selection block. I parameterize $\pi_k(x,z)$ directly via a softmax sieve over $K+1$ categories (K treated-in-regime-k, plus untreated):

$$\pi_k(x,z;\gamma) = \frac{\exp\big(\gamma_k^\top b_n(x,z)\big)}{1+\sum_{\ell=1}^{K}\exp\big(\gamma_\ell^\top b_n(x,z)\big)}, \qquad 1-p(x,z;\gamma) = \frac{1}{1+\sum_{\ell=1}^{K}\exp\big(\gamma_\ell^\top b_n(x,z)\big)}, \tag{D.3}$$

where $b_n(x,z) \in \mathbb{R}^{q_n}$ is a sieve basis whose dimension q_n grows slowly with n. In judge designs with J judges, $b_n(x,z)$ can include judge indicators interacted with a low-dimensional basis in x (e.g., polynomials, B-splines), or a scalar leave-one-out leniency index and its interactions with x. The softmax parameterization targets π_k directly, so the non-identifiability of the $w_k \times p_k$ factorization under bail never arises.

Outcome densities (KW–NPMLE). For each regime k and each treatment arm $d \in \{0,1\}$, I model the outcome density as a kernel mixture:

$$f_{dk}(y \mid x;\, G_{dk}) = \int \varphi_d\big(y \mid x,\eta\big)\, dG_{dk}(\eta), \tag{D.4}$$

where $\varphi_d(\cdot \mid x,\eta)$ is a known kernel family and G_{dk} is a probability measure on a *compact* kernel parameter set $\mathcal{H}$. For continuous outcomes, I use Gaussian location-scale kernels: $\varphi_d(y \mid x,\eta) = \varphi(y;\mu(x,\eta),\sigma^2(\eta))$ with

$$\mathcal{H}_{\text{Gauss}} := \big\{\eta : \|\eta_1\| \le M,\ \eta_2 \in [\underline{\sigma},\bar{\sigma}]\big\}, \qquad 0<\underline{\sigma}<\bar{\sigma}<\infty, \quad M<\infty. \tag{D.5}$$

The lower bound $\underline{\sigma} > 0$ is essential: without it the Gaussian-mixture likelihood is unbounded (a component can spike at a single observation), and the MLE does not exist in the classical sense. I write $\mathcal{H}$ for this compact Gaussian kernel parameter space and let $\mathcal{P}(\mathcal{H})$ denote the space of Borel probability measures on $\mathcal{H}$, equipped with the weak topology. By Prokhorov's theorem, $\mathcal{P}(\mathcal{H})$ is compact.

As in standard Kiefer–Wolfowitz (KW) theory, the NPMLE over $\mathcal{P}(\mathcal{H})$ can be taken to be a discrete (finite-atomic) measure, so $\widehat{f}_{dk}$ is automatically a finite kernel mixture whose number of atoms adapts to the data. Under the bail convention, only the treated-arm

densities $\{f_{1k}\}$ are estimated; the untreated-arm densities $\{f_{0k}\}$ are degenerate and carry no parameters. For outcomes with non-degenerate $Y(0)$, both arms are estimated, adding K additional KW–NPMLE blocks.

Formal parameter space (bail convention). For the formal estimator and theorem below, the sieve parameter space is

$$\Theta_n := \Gamma_n \times \mathcal{P}(\mathcal{H})^K, \qquad \Gamma_n := \big\{\gamma = (\gamma_1, \ldots, \gamma_K) : \|\gamma_k\|_\infty \le R_n\big\}, \tag{D.6}$$

where $R_n \to \infty$ slowly (the sieve grows with sample size) and $\mathcal{P}(\mathcal{H})^K$ contains the K treated-arm mixing distributions (with $\mathcal{H} = \mathcal{H}_{\text{Gauss}}$ for continuous outcomes; the bail-convention appendix writes this as Θ_0). For the two-arm extension used with non-degenerate untreated outcomes, Appendix I defines a finite-dimensional parametric parameter space with a fixed softmax basis, regime-specific propensities, and single-kernel outcome densities per regime-arm, and proves consistency directly. The empirical implementation in Section 7 uses this parametric specification.

D.3 The Sieve MLE

I define the formal estimator as the global maximizer of the bail-convention sample log-likelihood over the sieve:

$$\widehat{\theta}_n \in \arg\max_{\theta \in \Theta_n} \ell_n(\theta). \tag{D.7}$$

The estimator is defined as the unpenalized sieve MLE. The maximizer exists because: (i) for fixed $\gamma \in \Gamma_n$ (compact), the KW–NPMLE over $\mathcal{P}(\mathcal{H})^K$ exists and is discrete; (ii) the log-likelihood is continuous in γ on the compact set Γ_n; and (iii) the lower bound $\sigma^2 \ge \underline{\sigma}^2 > 0$ (for continuous outcomes) ensures boundedness. Practical regularization choices used to stabilize computation are discussed in Appendix G.

D.4 Consistency of the Bail-Convention Estimator

I now state the main theoretical guarantee for the bail-convention likelihood (D.2); the general-convention extension is discussed in the scope paragraph of Section 5. The proof, which follows a standard sieve-MLE argument (uniform law of large numbers, near-optimality, identification), is in Appendix E. I first summarize the regularity conditions informally, then state the precise numbered assumptions below.

The key regularity conditions are:

(C1) Correct specification. The true conditional density $h_0(y, d \mid x, z)$ belongs to the

bail-convention model family with $K_0 \le K$ active regimes. The true selection-block functions π_k^0 are smooth enough to be approximated by the sieve (D.3), and the true treated-arm outcome densities f_{1k}^0 lie in the closure of the Gaussian-mixture class (D.4).

(C2) Identification. The parameter $\theta_0 = \{\pi_k^0, f_{1k}^0\}_{k \le K_0}$ is unique (up to label permutation) within the maintained model class. This is stronger than the weak span result in Theorem 3.1 and should be understood as requiring an additional restriction of the kind developed elsewhere in the paper (for example, ordered tail dominance, mutual singularity, or a correctly specified finite-Gaussian component class, which identifies the outcome block and thereby the unrestricted mixing weights).

(C3) Sieve approximation. There exists a sequence $\tilde{\theta}_n \in \Theta_n$ that approximates the truth: $\sup_{x,z} |\tilde{\pi}_k(x,z) - \pi_k^0(x,z)| \to 0$ and $\sup_x d_H(\tilde{f}_{1k}(\cdot \mid x), f_{1k}^0(\cdot \mid x)) \to 0$ for each k, where d_H is the Hellinger distance.

(C4) Entropy control. The sieve complexity grows slowly enough that a uniform law of large numbers holds: $K q_n \log R_n = o(n)$.

(C5) Bounded likelihood. The aggregate treatment rate satisfies $p^0(x,z) \in [\varepsilon_0, 1-\varepsilon_0]$ for some $\varepsilon_0 > 0$, the kernel parameter space Θ_0 is compact with $\sigma^2 \ge \underline{\sigma}^2 > 0$, the outcome space $\mathcal{Y}$ is bounded, and the sieve bound R_n is chosen so that $p(x,z;\gamma)$ remains bounded away from $\{0,1\}$ uniformly over Γ_n. (In bail applications, Y is a bounded risk score and the first-stage bounds are mild.)

Define the *observational Hellinger distance*

$$\rho(\theta, \theta') := \left\{ \mathbb{E}_{P_0}\left[\int \left(\sqrt{h(y,d \mid X,Z;\theta)} - \sqrt{h(y,d \mid X,Z;\theta')} \right)^2 d\nu(y,d) \right] \right\}^{1/2},$$

where ν is the product of μ (dominating measure on $\mathcal{Y}$) and counting measure on $\{0,1\}$.

Theorem D.1 (Consistency of the sieve MLE)**.** *Under conditions* (C1)–(C5)*:*

(i) ***Density consistency.*** $\rho(\widehat{\theta}_n, \theta_0) \xrightarrow{p} 0$.

(ii) ***Parameter consistency.*** *There exist relabelings of the active fitted components and merged active-regime weights $\bar{\pi}_k^{(n)}$ such that, for each active regime $k \le K_0$,*

$$\left\| \bar{\pi}_k^{(n)} - \pi_k^0 \right\|_{L^2(P_{XZ})} \xrightarrow{p} 0, \qquad \left\| \widehat{f}_{1,\sigma_n(k)}(\cdot \mid \cdot) - f_{1k}^0(\cdot \mid \cdot) \right\|_{L^2(\mu \otimes P_X)} \xrightarrow{p} 0.$$

When $K = K_0$, the relabeling reduces to a permutation and $\bar{\pi}_k^{(n)} = \widehat{\pi}_{\sigma_n(k)}$. When $K > K_0$, $\bar{\pi}_k^{(n)}$ aggregates any surplus fitted indices whose limit components duplicate active regime k; see Remark E.2 in Appendix E for details.

(iii) ***Functional consistency.*** *For any functional $\Psi(\theta)$ that is continuous with respect to ρ, $\Psi(\widehat{\theta}_n) \xrightarrow{p} \Psi(\theta_0)$. This includes the fitted conditional law itself, the first stage $p(x, z)$, the reduced form $m(x, z)$, and the Wald ratio for instrument pairs whose first-stage gap is bounded away from zero. When $K = K_0$ (or after consistently merging duplicate components in the overfitted case), the same conclusion extends to regime-specific means $\tau_k(x) = \mathbb{E}[Y(1) \mid X = x, S = k]$, slope weights, and the NegMass statistic.*

This theorem is conditional on a fixed working value of K. It does not address recovery of the true regime count K_0 or model selection over K, which would require additional separation or eigenvalue-gap conditions to rule out local merging of distinct components.

Economic content. Part (i) says the estimated conditional distribution of (Y, D) given (X, Z) converges to the truth—the model "fits" in the sense relevant for prediction. Part (ii) says the structural objects the bail-convention estimator targets—the selection-block weights π_k that determine *who* is marginal in each regime, and the treated-arm densities f_{1k} that determine *what happens* to them—are consistently estimated (up to the inherent label ambiguity, and up to merging in the overfitted case). Part (iii) says that observables built continuously from the fitted conditional law, such as the Wald ratio, are consistently estimated. Component-level objects such as regime means, slope weights, and NegMass are also consistently estimated once the active component representation is pinned down, either because $K = K_0$ or because duplicate components are merged.

What is (not) consistently estimated under each convention. Under the bail convention, only the products $\pi_k(x, z) = w_k(x)\, p_k(x, z)$ and the treated-arm densities f_{1k} are consistently estimated; w_k and p_k separately are not. Wald ratios depend only on the fitted conditional law and are therefore consistently estimated under Theorem D.1. Regime-specific treated-outcome means, slope weights, and NegMass are additionally recovered when the active component representation is pinned down (for example, when $K = K_0$ or after consistent merging of duplicates in the overfitted case). Under the two-arm extension (Appendix I), w_k, p_k, and the untreated-arm densities f_{0k} are additionally recovered, enabling estimation of regime-specific treatment effects $\tau_k(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x, S = k]$.

Remark D.1 (Misspecification and the interpretation of regimes)**.** *If the true DGP has a continuum of latent types—as in a standard MTE model with continuous unobserved heterogeneity U_D—the finite-regime model is misspecified and the MLE converges to the best K-component KL-projection of the true conditional density onto the K-regime family. In this case, the estimated regimes should be interpreted not as "true" discrete types, but as a*

low-dimensional approximation to the latent selection structure chosen to fit the observed conditional law.

This interpretation does not license a literal structural reading of the fitted components under misspecification. The sieve MLE still approximates the observed conditional distribution $h(y, d \mid x, z)$, but observables computed from the fitted law need not equal their true counterparts unless the approximation error is negligible for that target. In particular, even the model-implied first stage, reduced form, and Wald ratio should be read as plug-in objects from the fitted law, not as automatically consistent estimators of the true corresponding quantities under arbitrary misspecification. Component-level summaries such as the signed slope-weight decomposition, NegMass, and the discrimination decomposition require more: they rely on the K-regime approximation providing a stable summary of the latent selection structure, which is why the paper emphasizes regime-count stability, descriptive reweighted-tercile checks, and the bounded-selection sensitivity analysis. What matters for empirical conclusions is whether the K-regime approximation is empirically adequate *for the quantities of interest being reported.*

The diagnostics in Section 5.4.1 are designed precisely for this purpose. If Wald ratio profiles, slope-weight decompositions, and discrimination estimates stabilize across $K = 1, 2, 3, \ldots$, the K-regime summary is adequate for the questions being asked, regardless of whether the true heterogeneity is discrete or continuous. If these objects change substantially as K increases, the researcher should increase K or report results for multiple values to assess robustness.

E Proof of Consistency (Theorem D.1)

This appendix provides the formal proof. I state the argument for the bail-convention likelihood (D.2), which uses only the treated-arm mixture; Appendix I provides the analogous result for the two-arm likelihood (D.1) using a finite-dimensional parametric model. The argument proceeds in five steps: (1) uniform law of large numbers over the sieve, (2) sieve approximation bias vanishes, (3) criterion convergence, (4) identification-based parameter recovery, and (5) functional consistency for the quantities of interest.

Remark E.1 (Two-arm extension)**.** *Appendix I provides a self-contained consistency theorem for the finite-dimensional two-arm parametric MLE used in the empirical application. That result uses a fixed compact parameter space (softmax selection block with fixed basis dimension, plus single-kernel outcome densities), so the proof reduces to a standard Wald M-estimator argument. The bail-convention proof below handles the more general infinite-dimensional case*

(Gaussian-mixture sieves with growing atom counts).

Formal assumptions

Assumption E.1 (Correct specification)**.** *The true conditional density $h_0(y, d \mid x, z)$ is representable by the model family with $K_0 \leq K$ active regimes. Specifically, there exist true structural objects $\{\pi_k^0(x, z), f_{1k}^0(\cdot \mid x)\}_{k \leq K_0}$ such that $h_0(y, d \mid x, z)$ has the mixture form* (D.2) *with these components, and:*

(i) The true selection-block functions $\pi_k^0(x, z)$ lie in the closure of $\{\pi_k(\cdot; \gamma) : \gamma \in \bigcup_n \Gamma_n\}$ under the supremum norm (i.e., are approximable by the sieve).

(ii) The true treated-outcome densities $f_{1k}^0(\cdot \mid x)$ lie in the Hellinger closure of $\{f_{1k}(\cdot \mid x; G) : G \in \mathcal{P}(\Theta_0)\}$, uniformly in x (i.e., are approximable by the Gaussian-mixture class).

The true parameter θ_0 may therefore be a limit of the sieve parameterization rather than an element of any finite sieve space Θ_n; Assumption E.3 ensures these approximations are achieved at suitable rates.

Condition (ii) holds whenever each true density $f_{1k}^0(\cdot \mid x)$ lies in the Hellinger closure of the Gaussian-mixture class generated by $\mathcal{P}(\Theta_0)$, uniformly in x. In particular, it holds under correct specification of a finite Gaussian-mixture outcome block with parameters in Θ_0. More general approximation results require the kernel set to be chosen rich enough, for example by allowing smaller variances and wider support bounds.

Assumption E.2 (Identification)**.** *If $h(\cdot; \theta) = h_0(\cdot)$ P_0-a.s., then there exists a permutation σ of $\{1, \ldots, K_0\}$ such that $\pi_{\sigma(k)}(\cdot; \gamma) = \pi_k^0(\cdot)$ and $f_{1,\sigma(k)}(\cdot; G_{\sigma(k)}) = f_{1k}^0(\cdot)$ for all $k \leq K_0$, and $\pi_k(\cdot; \gamma) = 0$ for $k > K_0$.*

This condition is imposed directly for the sieve consistency argument. In applications of interest it can be justified by stronger restrictions such as ordered tail dominance or a correctly specified parametric submodel.

Assumption E.3 (Sieve approximation)**.** *For each n, there exists $\tilde{\theta}_n = (\tilde{\gamma}_n, \tilde{G}_1^{(n)}, \ldots, \tilde{G}_K^{(n)}) \in \Theta_n$ such that:*

(i) $\sup_{(x,z) \in \mathcal{X} \times \mathcal{Z}} |\pi_k(x, z; \tilde{\gamma}_n) - \pi_k^0(x, z)| \to 0$ for each $k \leq K_0$.

(ii) $\sup_{x \in \mathcal{X}} d_H\big(f_{1k}(\cdot \mid x; \tilde{G}_k^{(n)}), f_{1k}^0(\cdot \mid x)\big) \to 0$ for each $k \leq K_0$.

(iii) $\pi_k(x, z; \tilde{\gamma}_n) \to 0$ uniformly for $k > K_0$.

Since $\mathcal{Z}$ is finite, condition (i) requires the sieve to approximate only the x-dependence of π_k^0. For s-smooth π_k^0 with B-spline sieves of dimension $s_n = q_n/J$ (e.g., tensor-product splines when $X \in \mathbb{R}^{d_x}$), the worst-case approximation error is $O(s_n^{-s/d_x})$, so any $q_n \to \infty$ suffices for approximation. Condition (ii) holds with $\tilde{G}_k^{(n)} = G_k^0$ whenever $G_k^0 \in \mathcal{P}(\Theta_0)$, which requires the compact kernel parameter space to be chosen wide enough to contain the truth.

Assumption E.4 (Entropy control)**.** *The sieve class* $\mathcal{F}_n := \{\log h(\cdot;\theta) : \theta \in \Theta_n\}$ *satisfies*

$$\int_0^1 \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}_n, L^2(P_0))}\, d\varepsilon = o(\sqrt{n}),$$

where $N_{[]}(\varepsilon, \mathcal{F}_n, L^2(P_0))$ *is the* ε*-bracketing number.*

If the sieve basis $b_n(x,z)$ is uniformly bounded, the softmax selection block is Lipschitz in its Kq_n coefficients on Γ_n, so it contributes parametric bracketing entropy of order $Kq_n \log(R_n/\varepsilon)$. For the KW block, standard compact-mixture entropy bounds over $\mathcal{P}(\Theta_0)$ with Θ_0 compact and $\sigma^2 \geq \underline{\sigma}^2$ give bracketing entropy of order $O((\log(1/\varepsilon))^{d_\eta+1})$ per regime. Assumption E.5 keeps $h(y, d \mid x, z; \theta)$ uniformly bounded away from zero and infinity on the sieve, so passing from the parameter blocks to the log-likelihood class $\mathcal{F}_n$ preserves these entropy orders up to constants. The KW contribution is n-free and contributes only an $O(1)$ term to the bracketing integral, so a sufficient condition for Assumption E.4 is $Kq_n \log R_n = o(n)$.

Assumption E.5 (Bounded likelihood ratios)**.** *(i) There exists* $\varepsilon_0 > 0$ *such that* $p^0(x,z) \in [\varepsilon_0, 1-\varepsilon_0]$ *for all* $(x,z) \in \mathcal{X} \times \mathcal{Z}$*.*

(ii) The kernel parameter space Θ_0 *is compact with* $\sigma^2 \geq \underline{\sigma}^2 > 0$*.*

(iii) The outcome space $\mathcal{Y}$ *is bounded:* $\mathcal{Y} \subseteq [-C_Y, C_Y]$ *for some* $C_Y < \infty$*.*

(iv) The sieve parameter space Γ_n *is restricted so that* $p(x,z;\gamma) := \sum_k \pi_k(x,z;\gamma) \in [\varepsilon_0/2,\, 1-\varepsilon_0/2]$ *for all* $(x,z) \in \mathcal{X} \times \mathcal{Z}$ *and* $\gamma \in \Gamma_n$*.*

Condition (iii) is natural for the bail application, where Y is a misconduct indicator or bounded risk score. For unbounded $\mathcal{Y}$, the same argument can instead be based on an integrable envelope for the log-likelihood. In the Gaussian case with compact mean and variance bounds, a finite second moment of Y is enough to control $|\log h(Y, D \mid X, Z; \theta)|$ uniformly over the sieve by a quadratic envelope. Condition (iv) restricts the sieve parameter space, not just the truth. With a softmax parameterization and growing R_n, the unrestricted coefficient box would eventually include candidate parameters whose aggregate propensity $p(x,z;\gamma)$ is arbitrarily close to 0 or 1, which would break the uniform envelope argument for

the log-likelihood. This restriction is asymptotically harmless for approximation: by part (i), the true aggregate first stage satisfies $p^0(x,z) \in [\varepsilon_0, 1-\varepsilon_0]$, so any sieve sequence $\tilde{\gamma}_n$ that uniformly approximates the true selection block will eventually satisfy the same propensity bound.

Step 1: Uniform law of large numbers

Define the population criterion $M(\theta) := \mathbb{E}_{P_0}[\log h(Y, D \mid X, Z; \theta)]$. By the Kullback–Leibler inequality,

$$M(\theta) = M(\theta_0) - \mathrm{KL}(h_0 \,\|\, h_\theta) \leq M(\theta_0), \tag{E.1}$$

with equality iff $h(\cdot;\theta) = h_0(\cdot)$ P_0-a.s.

Lemma E.1 (ULLN)**.** *Under Assumptions E.4–E.5,* $\sup_{\theta\in\Theta_n} |\ell_n(\theta) - M(\theta)| = o_p(1)$.

Proof. I verify a constant envelope. **Arm** $d=0$**:** $|\log(1-p(x,z;\gamma))| \leq |\log(\varepsilon_0/2)|$ by Assumption E.5(iv). **Arm** $d=1$**:** Since $\mathcal{Y} \subseteq [-C_Y, C_Y]$ (Assumption E.5(iii)) and $\sigma^2 \geq \underline{\sigma}^2$ (Assumption E.5(ii)), each Gaussian kernel satisfies $\varphi(y;\mu,\sigma^2) \in [\underline{\varphi}, \bar{\varphi}]$ for $y \in \mathcal{Y}$, $\eta \in \Theta_0$, and $x \in \mathcal{X}$, where $\underline{\varphi} > 0$ depends on C_Y, M, $\bar{\sigma}$ and $\bar{\varphi} = (\sqrt{2\pi}\,\underline{\sigma})^{-1}$. Hence $f_{1k} \in [\underline{\varphi}, \bar{\varphi}]$ on $\mathcal{Y}$, and $\sum_k \pi_k f_{1k} \geq (\varepsilon_0/2)\underline{\varphi} > 0$. The log-likelihood is bounded: $|\log h| \leq B := \max\{|\log(\varepsilon_0/2)|, |\log((\varepsilon_0/2)\underline{\varphi})|, \log\bar{\varphi}\} < \infty$. The maximal inequality for bracketed empirical processes gives:

$$\mathbb{E}_0\left[\sup_{\theta\in\Theta_n} |\ell_n(\theta) - M(\theta)|\right] \leq \frac{C\int_0^1 \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}_n, L^2(P_0))}\, d\varepsilon}{\sqrt{n}} = o(1)$$

by Assumption E.4. □

Step 2: Sieve approximation bias

Lemma E.2 (Approximation bias vanishes)**.** *Under Assumptions E.1, E.3, and E.5,* $M(\tilde{\theta}_n) \to M(\theta_0)$.

Proof. By (E.1), $M(\theta_0) - M(\tilde{\theta}_n) = \mathrm{KL}(h_0 \,\|\, h_{\tilde{\theta}_n})$. It suffices to show $d_H(h_0, h_{\tilde{\theta}_n}) \to 0$.

The $d=0$ arm. $h(0,0 \mid x,z;\tilde{\theta}_n) = 1 - \tilde{p}_n(x,z)$ with $\tilde{p}_n \to p^0$ uniformly by Assumption E.3(i). The Hellinger contribution from this component vanishes.

The $d=1$ arm. I use a mixture Hellinger bound. Write $g_n = \sum_k \tilde{\pi}_k^{(n)} f_{1k}^{(n)}$ and $g_0 = \sum_k \pi_k^0 f_{1k}^0$, where for $k > K_0$ I adopt the standard overfitted-mixture padding convention $\pi_k^0 \equiv 0$ (the choice of f_{1k}^0 for $k > K_0$ is immaterial since these terms are multiplied by zero).

By the elementary inequality $(\sqrt{a} - \sqrt{b})^2 \leq |a - b|$ for $a, b \geq 0$:

$$d_H^2(g_n, g_0) = \int (\sqrt{g_n} - \sqrt{g_0})^2 \, d\mu \leq \int |g_n - g_0| \, d\mu = \|g_n - g_0\|_{L^1(\mu)}.$$

Adding and subtracting only the active-regime terms $\sum_{k \leq K_0} \tilde{\pi}_k^{(n)} f_{1k}^0$:

$$\begin{aligned} \|g_n - g_0\|_{L^1} &\leq \textstyle\sum_{k \leq K_0} \tilde{\pi}_k^{(n)} \|f_{1k}^{(n)} - f_{1k}^0\|_{L^1} \;+\; \sum_{k > K_0} \tilde{\pi}_k^{(n)} \|f_{1k}^{(n)}\|_{L^1} \;+\; \sum_{k \leq K_0} |\tilde{\pi}_k^{(n)} - \pi_k^0| \, \|f_{1k}^0\|_{L^1} \\ &\leq \textstyle\sum_{k \leq K_0} \|f_{1k}^{(n)} - f_{1k}^0\|_{L^1} + \sum_k |\tilde{\pi}_k^{(n)} - \pi_k^0|, \end{aligned}$$

where I used $\tilde{\pi}_k \leq 1$, $\|f_{1k}^{(n)}\|_{L^1} = \|f_{1k}^0\|_{L^1} = 1$, and $\pi_k^0 \equiv 0$ for $k > K_0$. The first sum vanishes by Assumption E.3(ii) (Hellinger convergence implies L^1); the second by Assumption E.3(i) and (iii). Hence $d_H(h_{\tilde{\theta}_n}, h_0) \to 0$. Since $d_H \to 0$ and the likelihood ratios are bounded (Assumption E.5), KL $\to 0$ by dominated convergence of $\log(h_0 / h_{\tilde{\theta}_n})$. □

Step 3: Criterion convergence

Since $\widehat{\theta}_n$ maximizes ℓ_n over Θ_n and $\tilde{\theta}_n \in \Theta_n$, $\ell_n(\widehat{\theta}_n) \geq \ell_n(\tilde{\theta}_n)$. Combining with the ULLN:

$$M(\widehat{\theta}_n) \geq \ell_n(\tilde{\theta}_n) + o_p(1) = M(\tilde{\theta}_n) + o_p(1) = M(\theta_0) + o_p(1).$$

Since $M(\widehat{\theta}_n) \leq M(\theta_0)$ always, $\mathrm{KL}(h_0 \,\|\, h_{\widehat{\theta}_n}) \xrightarrow{p} 0$. By Pinsker's inequality and $d_H^2 \leq d_{\mathrm{TV}}$:

$$\rho(\widehat{\theta}_n, \theta_0) \xrightarrow{p} 0,$$

establishing Theorem D.1(i). □

Step 4: From density to parameter consistency

Lemma E.3 (Identification implies parameter convergence)**.** *Let $\{\theta_n\}$ satisfy $\rho(\theta_n, \theta_0) \to 0$. Then there exist representative indices $\sigma_n(k) \in \{1, \ldots, K\}$ and merged active-regime weights $\bar{\pi}_k^{(n)}$ such that, for each $k \leq K_0$:*

$$\|\bar{\pi}_k^{(n)} - \pi_k^0\|_{L^2(P_{XZ})} \to 0, \qquad \|f_{1,\sigma_n(k)}^{(n)} - f_{1k}^0\|_{L^2(\mu \otimes P_X)} \to 0.$$

When $K = K_0$, the representatives can be chosen as a permutation and $\bar{\pi}_k^{(n)} = \pi_{\sigma_n(k)}^{(n)}$.

Proof. The argument proceeds in four stages: extract a good subsequence via compactness, identify the limiting mixture weights, upgrade to L^2 convergence, and then promote the result to the full sequence by contradiction.

First, extract a "good" subsequence. Since $\rho(\theta_n, \theta_0) \to 0$ implies $\|h(\cdot;\theta_n) - h_0(\cdot)\|_{L^1(P_0)} \to 0$, extract a subsequence (still indexed by n) along which

$$h(y, d \mid x, z; \theta_n) \to h_0(y, d \mid x, z) \qquad \text{for } (P_0 \otimes \nu)\text{-a.e. } (y, d, x, z). \tag{E.2}$$

This is possible by the standard $L^1 \to$ a.e. subsequence lemma.

KW component: By Prokhorov's theorem ($\mathcal{P}(\Theta_0)$ is compact), pass to a further subsequence with $G_k^{(n)} \Rightarrow G_k^*$ weakly for each $k = 1, \ldots, K$. Since $\eta \mapsto \varphi(y; \mu(x,\eta), \sigma^2(\eta))$ is bounded and continuous on Θ_0 (using $\sigma^2 \geq \underline{\sigma}^2 > 0$), the portmanteau theorem gives

$$f_{1k}(y \mid x; G_k^{(n)}) \to f_{1k}(y \mid x; G_k^*) \qquad \text{for every } (y, x). \tag{E.3}$$

Second, propagate to pointwise a.e. convergence of the relevant weights. For each instrument value $z_j \in \mathcal{Z}$ (finitely many), the $d = 0$ component of (E.2) gives:

$$1 - \textstyle\sum_k \pi_k(x, z_j; \gamma_n) \;\to\; 1 - p^0(x, z_j) \qquad P_X\text{-a.e.}$$

The $d = 1$ component gives: for $(\mu \otimes P_X)$-a.e. (y, x),

$$\textstyle\sum_{k=1}^{K} \pi_k(x, z_j; \gamma_n)\, f_{1k}(y \mid x; G_k^{(n)}) \;\to\; \sum_{k=1}^{K_0} \pi_k^0(x, z_j)\, f_{1k}^0(y \mid x).$$

By (E.3) and $\pi_k \in [0,1]$, replacing $f_{1k}^{(n)}$ by f_{1k}^* introduces an error bounded by $\sum_k |f_{1k}^{(n)}(y \mid x) - f_{1k}^*(y \mid x)| \to 0$ for every (y, x). So for $(\mu \otimes P_X)$-a.e. (y, x) and each z_j:

$$\textstyle\sum_{k=1}^{K} \pi_k(x, z_j; \gamma_n)\, f_{1k}(y \mid x; G_k^*) \;\to\; \sum_{k=1}^{K_0} \pi_k^0(x, z_j)\, f_{1k}^0(y \mid x). \tag{E.4}$$

Now invoke the maintained identification assumption for the sieve problem. The convergence (E.4) shows $\sum_k \pi_k^{(n)}(x, z_j) f_{1k}^*(\cdot \mid x) \to \sum_k \pi_k^0(x, z_j) f_{1k}^0(\cdot \mid x)$ in $L^1(\mu)$ for a.e. x and each z_j. Since $\pi_k^{(n)} \in [0,1]$, any cluster point of the weight vector $(\pi_1^{(n)}(x, z_j), \ldots, \pi_K^{(n)}(x, z_j))$ defines a mixture representation of the true density using components $\{f_{1k}^*\}$. Assumption E.2 implies that, within the sieve model, any representation of the true limit density is unique up to label permutation and null surplus components. Hence the limit mixing measure must equal the true mixing measure up to label permutation: for each distinct active component f_{1k}^0, the total weight aggregated across all indices ℓ with $f_{1\ell}^* = f_{1k}^0$ converges to $\pi_k^0(x, z_j)$.

After relabeling the distinct active limit components if necessary, define the duplicate sets

$$I_k := \{\ell \in \{1, \ldots, K\} : f_{1\ell}^* = f_{1k}^0\}, \qquad k = 1, \ldots, K_0,$$

and the corresponding *merged* active-regime weights

$$\bar{\pi}_k^{(n)}(x, z_j) := \sum_{\ell \in I_k} \pi_\ell(x, z_j; \gamma_n), \qquad k = 1, \ldots, K_0.$$

By the uniqueness argument, $\bar{\pi}_k^{(n)}(x, z_j) \to \pi_k^0(x, z_j)$ for P_X-a.e. x and each z_j. Any surplus component whose limit density $f_{1\ell}^*$ is not among $\{f_{1k}^0\}_{k \le K_0}$ must receive vanishing aggregate weight, since otherwise the limit mixture would contain a component absent from the true K_0-component representation.

It remains to show that $\bar{\pi}_k^{(n)}$ can be identified with a single index. When $K = K_0$, the merging is trivially a relabeling. When $K > K_0$, duplicate components may split weight; in this case, the parameter consistency statement below holds for the merged weights $\bar{\pi}_k^{(n)}$ (which are the structurally meaningful objects) rather than for individual labeled indices. Empirically, this means that component-level objects should be interpreted through the merged active-regime representation whenever the selected working value of K exceeds the active regime count.

The merged weights satisfy the same $K_0 \times K_0$ linear system used below. Fix x in the full-P_X-measure set where the convergence holds for a.e. y. Integrate both sides against K_0 test functions $\psi_1, \ldots, \psi_{K_0} \in L^\infty(\mu)$ chosen so that the matrix $A(x) := (\int \psi_m f_{1k}^0 \, d\mu)_{m,k}$ is nonsingular (such functions exist because $\{f_{1k}^0(\cdot \mid x)\}_{k \le K_0}$ are linearly independent by Assumption 3.1; for instance, the indicator functions of the approximate support regions from Step 4 of identification, or the moment functions $\psi_m(y) = y^{m-1}$ when the component means are distinct). After merging duplicate active components and dropping the vanishing nonmatching surplus terms, the system gives

$$\underbrace{\left(\int \psi_m f_{1k}^0 \, d\mu\right)_{m,k}}_{=: A(x), \text{ nonsingular}} \begin{pmatrix} \bar{\pi}_1^{(n)}(x, z_j) \\ \vdots \\ \bar{\pi}_{K_0}^{(n)}(x, z_j) \end{pmatrix} \to A(x) \begin{pmatrix} \pi_1^0(x, z_j) \\ \vdots \\ \pi_{K_0}^0(x, z_j) \end{pmatrix},$$

where any surplus component not assigned to one of the sets I_k has vanishing aggregate weight, and any duplicate active component has already been absorbed into $\bar{\pi}_k^{(n)}$. Since $A(x)$ is nonsingular:

$$\bar{\pi}_k^{(n)}(x, z_j) \to \pi_k^0(x, z_j) \qquad P_X\text{-a.e., for each } k \le K_0 \text{ and each } z_j. \tag{E.5}$$

Third, upgrade to L^2 convergence by dominated convergence. Since $|\bar{\pi}_k^{(n)}(x, z) - \pi_k^0(x, z)|^2 \le 1$ (bounded) and converges to 0 for P_{XZ}-a.e. (x, z) by (E.5) (and $\mathcal{Z}$ is finite), the Dominated Convergence Theorem gives $\|\bar{\pi}_k^{(n)} - \pi_k^0\|_{L^2(P_{XZ})} \to 0$. For each

$k \le K_0$, choose any representative index $\sigma_n(k) \in I_k$. Then $f^{(n)}_{1,\sigma_n(k)} \to f^0_{1k}$ in $L^2(\mu \otimes P_X)$ by (E.3) and the constant bound $f_{1k} \le (\sqrt{2\pi}\,\underline{\sigma})^{-1}$ on $\mathcal{Y}$ (bounded). When $K = K_0$ (no surplus components), the representatives form a permutation and $\bar{\pi}^{(n)}_k = \pi^{(n)}_{\sigma_n(k)}$.

Finally, promote the argument from a subsequence to the full sequence by contradiction. The above holds along a subsequence. Suppose the full sequence fails: then there exist $\varepsilon > 0$ and $\{n'\} \subseteq \{n\}$ such that no further subsequence admits representative indices and merged active-regime weights with

$$\max_{k \le K_0} \Big[\|\bar{\pi}^{(n')}_k - \pi^0_k\|_{L^2} + \|f^{(n')}_{1,\sigma_{n'}(k)} - f^0_{1k}\|_{L^2} \Big] \le \varepsilon.$$

But $\rho(\theta_{n'}, \theta_0) \to 0$, so the first three stages applied to $\{n'\}$ yield a sub-subsequence together with exactly such representatives and merged weights, contradicting the assumption.

This establishes Theorem D.1(ii). □ □

Remark E.2 (Parameter consistency in the overfitted case $K > K_0$)**.** *When $K > K_0$, the unpenalized MLE can split an active component's weight across multiple indices (e.g., assigning positive weight to two indices k and k' with $f^*_{1k} = f^*_{1k'}$). In this case, individual labeled weights $\pi^{(n)}_k$ need not converge to zero for surplus indices; only the* merged *weights $\bar{\pi}^{(n)}_k$ (aggregated across all indices sharing the same limit component) converge to π^0_k. Theorem D.1(ii) should therefore be read as follows: there exist label permutations σ_n and a merging map such that the merged active-regime weights converge in $L^2(P_{XZ})$ to π^0_k and the corresponding component densities converge in $L^2(\mu \otimes P_X)$ to f^0_{1k}. When $K = K_0$, merging is trivially a relabeling and the componentwise statement holds as written. Functional consistency for observables that depend only on the fitted conditional density $h(\cdot; \widehat{\theta}_n)$, such as the Wald ratio, is unaffected by duplication. By contrast, component-level functionals such as slope weights and NegMass should be interpreted using the merged active-regime representation whenever the fitted working value of K exceeds the active regime count.*

Step 5: Functional consistency

Theorem D.1(iii) follows from the continuous mapping theorem. I verify continuity for each target functional under the L^2 topology on (π_k, f_{1k}).

Regime-specific means. $\tau_k(x) = \int y\, f_{1k}(y \mid x)\, dy$. Since $\mathcal{Y} \subseteq [-C_Y, C_Y]$ (Assumption E.5(iii)) and $\mu(\mathcal{Y}) < \infty$:

$$\|\widehat{\tau}_k - \tau^0_k\|^2_{L^2(P_X)} = \int \Big| \int y\big(\widehat{f}_{1k}(y \mid x) - f^0_{1k}(y \mid x)\big)\, d\mu(y) \Big|^2 dP_X(x) \le C_Y^2 \int \Big(\int |\widehat{f}_{1k} - f^0_{1k}|\, d\mu \Big)^2 dP_X.$$

By Cauchy–Schwarz, $(\int |\widehat{f}_{1k} - f^0_{1k}|\, d\mu)^2 \le \mu(\mathcal{Y}) \int (\widehat{f}_{1k} - f^0_{1k})^2\, d\mu$. Hence

$$\|\widehat{\tau}_k - \tau^0_k\|^2_{L^2(P_X)} \le C_Y^2\, \mu(\mathcal{Y})\, \|\widehat{f}_{1k} - f^0_{1k}\|^2_{L^2(\mu \otimes P_X)} \to 0,$$

using the $L^2(\mu \otimes P_X)$ convergence established in Step 4.

Wald ratio. $\psi(x; z, z') = [\sum_k \pi_k(x, z)\tau_k(x) - \sum_k \pi_k(x, z')\tau_k(x)]/[p(x, z) - p(x, z')]$ is a continuous function of (π_k, τ_k) when $|p^0(x, z) - p^0(x, z')| > \delta > 0$. Since $\pi_k \to \pi^0_k$ and $\tau_k \to \tau^0_k$ in L^2, the numerator converges in L^1 (by Hölder), and the denominator is bounded from zero. The continuous mapping theorem therefore gives $L^1(P_X)$ consistency; $L^2(P_X)$ consistency then follows from the bounded-$\mathcal{Y}$ assumption.

Slope weights. $\omega^*_k(x; z, z') = [\pi_k(x, z) - \pi_k(x, z')]/[p(x, z) - p(x, z')]$ is continuous in π_k when the first-stage gap $|p^0(x, z) - p^0(x, z')| > \delta > 0$ and the component representation is pinned down. In the overfitted case $K > K_0$, the same statement applies after replacing π_k by the merged active-regime weights.

NegMass and discrimination decomposition. NegMass is a Lipschitz function of the componentwise slope weights, since it sums the negative parts of ω^*_k. Therefore it inherits consistency whenever the slope weights themselves are consistently estimated, i.e. when $K = K_0$ or after merging duplicate components. The discrimination decomposition in the main text combines two separately estimated group-specific decompositions and therefore requires, in addition, stable cross-group regime alignment; Theorem D.1(iii) supplies consistency of the within-group building blocks, while the commensurability issue is handled separately in the discussion around (5.4). Informally, the overall convergence rate is governed by the slower of the sieve approximation rate for the selection block and the finite-mixture rate for the outcome block. □

Bootstrap validity

The nonparametric (case) bootstrap is valid for smooth functionals of θ_0 under the following conditions:

(a) $K = K_0$ (correct number of regimes selected), so that θ_0 lies in the *interior* of the parameter space;

(b) all active selection-block weights satisfy $\inf_{x,z} \pi^0_k(x, z) > 0$ for $k \le K_0$;

(c) the functional Ψ is sufficiently smooth at θ_0 (for example, Hadamard differentiable), as is the case for regime means, Wald ratios, and slope weights when the relevant first-stage gaps are bounded away from zero.

Under these conditions, bootstrap validity is the regular interior case for smooth functionals of an asymptotically regular estimator. A full bootstrap theorem for the present sieve mixture estimator would require additional work to verify the relevant local expansion and regime-label stability conditions, so I treat this paragraph as a scope statement rather than a proved theorem. In the empirical application, inference is conducted conditional on a fixed working value of K, and the cleanest justification applies when the selected specification is effectively regular and interior. I likewise do not invoke subsampling as a generic fallback: in mixture settings with possible boundary weights, label instability, and data-driven choice of K, subsampling is not automatically valid without its own separate verification. The practical inference strategy in the paper therefore stays with the regular fixed-K implementation and the null-based parametric bootstrap for the formal tests.

When $K > K_0$, some true weights $\pi_k^0 \equiv 0$, placing θ_0 on the boundary of Θ_n. The naive bootstrap can produce non-standard or inconsistent confidence intervals in this setting (a well-known issue in mixture models). Two strategies address this:

(i) **Condition on $\widehat{K}$.** First select K using the procedure in Section 5.4.2, then bootstrap conditional on $K = \widehat{K}$. This is the practical route used in the paper. Its cleanest justification is the case in which the selected working value matches the active regime count, so the fitted parameter lies effectively in the interior and conditions (a)–(c) apply.

(ii) **Null-based bootstrap for the formal tests.** For the specific hypothesis tests in Section 5.3, simulate from the restricted model under H_0. For the rank-1 test, this is a smooth equality-restricted null and the projected model is the natural bootstrap DGP. For the NegMass test, by contrast, the null is an inequality/contact-set restriction, so the projection may lie on the boundary of the restricted parameter space; the paper therefore uses a contact-set bootstrap rather than claiming that boundary issues disappear. In both cases the goal is to approximate the null distribution of the test statistic conditional on the restricted fit.

I recommend strategy (i) for general confidence intervals and (ii) for the two formal tests.

F Sensitivity to Within-Regime Selection

Assumption 2.4 requires that potential outcomes are independent of the within-regime compliance index V conditional on (X, S). This appendix develops a sensitivity analysis that relaxes this assumption to *bounded* within-regime selection, following the approach outlined in the discussion of Assumption 2.4.

Bounded within-regime selection

Assumption F.1 (Bounded within-regime selection)**.** *For each regime $k \leq K_0$ and $x \in \mathcal{X}$, there exists $\delta_k(x) \geq 0$ such that*

$$\sup_{v,v' \in \text{supp}(V|X=x,S=k)} \left|\mathbb{E}[Y(1) \mid X = x, S = k, V = v] - \mathbb{E}[Y(1) \mid X = x, S = k, V = v']\right| \leq \delta_k(x).$$

The parameter $\delta_k(x)$ bounds how much the mean potential outcome can vary with the within-regime compliance index V. When $\delta_k(x) = 0$, Assumption 2.4 holds exactly and the results in the main text apply. When $\delta_k(x) > 0$, the regime-specific selected-outcome mean $\tau_k^{\text{rel}}(x, z) := \mathbb{E}[Y(1) \mid X = x, S = k, D = 1, Z = z]$ can differ from the regime-specific potential-outcome mean $\tau_k(x) := \mathbb{E}[Y(1) \mid X = x, S = k]$, because decision-makers who treat more cases within regime k may be selecting on unobserved risk.

Lemma F.1 (Bias bound on regime-specific means)**.** *Under Assumption F.1, for each regime k, covariate value x, and instrument value z with $\Pr(D = 1 \mid X = x, S = k, Z = z) > 0$:*

$$\left|\tau_k^{\text{rel}}(x, z) - \tau_k(x)\right| \leq \delta_k(x).$$

Proof. Write $\tau_k^{\text{rel}}(x, z) = \mathbb{E}[\mathbb{E}[Y(1) \mid X = x, S = k, V] \mid D = 1, X = x, S = k, Z = z]$. This is a weighted average of $\mathbb{E}[Y(1) \mid X = x, S = k, V = v]$ over the distribution of V among the treated in regime k. The unconditional mean $\tau_k(x)$ is a weighted average of the same function over the unconditional distribution of V given $(X = x, S = k)$. Both are averages of a function whose range has diameter at most $\delta_k(x)$. Any two averages of a function with range $\delta_k(x)$ differ by at most $\delta_k(x)$. □

Implications for the Wald Ratio and Its Decomposition

The estimation procedure in Section D.3 targets the treated-outcome densities $f_{1k}(\cdot \mid x)$ and selection-block weights $\pi_k(x, z)$. Under bounded within-regime selection, the estimated regime-specific means $\widehat{\tau}_k(x)$ are consistent for $\tau_k^{\text{rel}}(x, z)$ (averaged over the instrument distribution), not for $\tau_k(x)$. The following proposition bounds the resulting bias in the Wald ratio and its decomposition.

Proposition F.1 (Sensitivity of the Wald ratio)**.** *Under bounded within-regime selection with parameter $\delta_k(x)$, the bias in the Wald ratio satisfies*

$$\left|\psi^{\text{rel}}(x; z, z') - \psi(x; z, z')\right| \leq \sum_{k=1}^{K_0} |\omega_k^*(x; z, z')|\, \delta_k(x),$$

where $\psi(x; z, z') = \sum_k \omega_k^* \tau_k$ *is the Wald ratio under exact within-regime stability, and* $\psi^{\text{rel}}(x; z, z')$ *is the probability limit of the plug-in Wald ratio under bounded within-regime selection. Specifically, the plug-in estimator targets instrument-averaged selected means* $\bar{\tau}_k^{\text{rel}}(x) := \mathbb{E}_Z[\tau_k^{\text{rel}}(x, Z) \mid X = x]$ *(where the expectation averages over the instrument distribution weighted by within-regime treatment shares), and* $\psi^{\text{rel}}(x; z, z') := \sum_k \omega_k^*(x; z, z')\, \bar{\tau}_k^{\text{rel}}(x)$.

Proof. By linearity:

$$\left|\psi^{\text{rel}} - \psi\right| = \left|\sum_k \omega_k^* \left(\bar{\tau}_k^{\text{rel}} - \tau_k\right)\right| \leq \sum_k |\omega_k^*|\, |\bar{\tau}_k^{\text{rel}} - \tau_k| \leq \sum_k |\omega_k^*|\, \delta_k(x),$$

where the last inequality uses Lemma F.1: since $\bar{\tau}_k^{\text{rel}}(x)$ is a weighted average of $\tau_k^{\text{rel}}(x, z)$ over z, and each $\tau_k^{\text{rel}}(x, z)$ lies within $\delta_k(x)$ of $\tau_k(x)$, the averaged quantity $\bar{\tau}_k^{\text{rel}}(x)$ also lies within $\delta_k(x)$ of $\tau_k(x)$. □

The bound has a transparent structure: the sensitivity of the Wald ratio is the sum of absolute slope weights times the within-regime selection parameter. This connects directly to the NegMass diagnostic: when slope weights are large in absolute value (as happens with offsetting margins or weak net first stages), the Wald ratio is fragile to even modest within-regime selection. Conversely, when all slope weights are positive and moderate (as under approximate single-index leniency), the bound is tight.

Proposition F.2 (Sensitivity of the discrimination decomposition)**.** *Under bounded within-regime selection with group-specific biases* $b_k^r := \tau_k^{r,\text{rel}} - \tau_k^r$ *satisfying* $|b_k^r| \leq \delta$ *for each regime* k *and group* $r \in \{B, NB\}$*, the bias in each channel of the decomposition* (5.4) *satisfies:*

(i) ***Within-regime channel.*** $\left|\text{bias}_W\right| = \left|\sum_k \bar{\omega}_k \left(b_k^B - b_k^{NB}\right)\right| \leq 2\delta \sum_k |\bar{\omega}_k| =: M_W \cdot \delta.$

(ii) ***Margin-composition channel.*** $\left|\text{bias}_C\right| = \left|\sum_k (\omega_k^{*B} - \omega_k^{*NB}) \cdot \tfrac{1}{2}(b_k^B + b_k^{NB})\right| \leq \delta \sum_k |\omega_k^{*B} - \omega_k^{*NB}| =: M_C \cdot \delta.$

Proof. Part (i): The within-regime channel under biased regime means is $\sum_k \bar{\omega}_k[(\tau_k^B + b_k^B) - (\tau_k^{NB} + b_k^{NB})] = \sum_k \bar{\omega}_k(\tau_k^B - \tau_k^{NB}) + \sum_k \bar{\omega}_k(b_k^B - b_k^{NB})$. By the triangle inequality and $|b_k^B - b_k^{NB}| \leq 2\delta$, the bound follows. Part (ii): The composition channel under biased means is $\sum_k (\omega_k^{*B} - \omega_k^{*NB}) \cdot \frac{1}{2}[(\tau_k^B + b_k^B) + (\tau_k^{NB} + b_k^{NB})] = \sum_k (\omega_k^{*B} - \omega_k^{*NB})\bar{\tau}_k + \sum_k (\omega_k^{*B} - \omega_k^{*NB}) \cdot \frac{1}{2}(b_k^B + b_k^{NB})$. By $|\frac{1}{2}(b_k^B + b_k^{NB})| \leq \delta$, the bound follows. □

Remark F.1 (Two benchmark scenarios)**.** *The bounds in Proposition F.2 are not simultaneously tight. Two polar cases clarify which channel is fragile:*

1. ***Race-neutral within-regime selection*** *($b_k^B = b_k^{NB}$ for all k): the within-regime channel has* zero bias *regardless of δ (since $b_k^B - b_k^{NB} = 0$), while the composition channel absorbs all bias, bounded by $M_C \cdot \delta$. Under this scenario, the finding that margin composition dominates is* strengthened *by violations of Assumption 2.4.*

2. ***Worst-case race-differential selection*** *($b_k^B = -b_k^{NB}$ for all k): the composition channel has zero bias (since $b_k^B + b_k^{NB} = 0$), and the within-regime channel faces its maximal bias $M_W \cdot \delta$. This is the scenario most damaging to the headline finding.*

The race-neutral benchmark is arguably the natural default under the model's regime structure: S absorbs outcome-relevant heterogeneity, so within-regime selection operates through the residual index V, and the baseline model does not assign race-specific roles to that residual index once one conditions on the latent decision environment. Race-differential within-regime selection would require the selected-mean bias term b_k^r to vary systematically by race within the same latent decision environment*, a substantively stronger departure from the model.*

Calibration

The parameter $\delta_k(x)$ cannot generally be identified from the data without additional structure, because it concerns variation with an unobserved index. I recommend treating $\delta_k = \delta$ (common across regimes) as a sensitivity parameter and reporting results for a grid of values (e.g., $\delta \in \{0, 0.01, 0.02, 0.05\}$ in risk-probability units).

Two sources of calibration information are available:

(a) **Observed cross-instrument variation within regimes (necessary lower bound).** From the inequality $|\tau_k^{\text{rel}}(x,z) - \tau_k^{\text{rel}}(x,z')| \leq 2\delta_k(x)$, any observed variation in estimated regime-specific selected means across instrument values implies $\delta_k(x) \geq \frac{1}{2} \max_{z,z'} |\widehat{\tau}_k^{\text{rel}}(x,z) - \widehat{\tau}_k^{\text{rel}}(x,z')|$. *Implementation note:* since the fitted mixture model imposes z-invariant regime-specific outcome densities $f_{1k}(y \mid x)$ by construction, $\widehat{\tau}_k^{\text{rel}}(x,z)$ must be computed outside the model. To obtain regime-specific means that vary with z, the posterior regime responsibilities must depend on individual-level covariates X_i, not solely on the discrete instrument value. Specifically, define individual-level responsibilities $\widehat{r}_{ik} := \widehat{w}_k(X_i)\widehat{p}_k(X_i, Z_i)/\widehat{p}(X_i, Z_i)$ using the fitted model's covariate-dependent regime shares and propensities, and compute $\widehat{\tau}_k^{\text{rel}}(x,z) := \sum_{i:D_i=1,Z_i=z} \widehat{r}_{ik}\, Y_i \big/ \sum_{i:D_i=1,Z_i=z} \widehat{r}_{ik}$. Because $\widehat{r}_{ik}$ varies across individuals within each instrument bin (through X_i), these weighted means are not mechanically constant in k, and cross-z variation within a given k reflects differential selection into treatment within regimes. (When the empirical implementation

conditions on residualized leniency bins without additional X_i variation, the responsibilities $\widehat{r}_{ik}$ are constant within z and this calibration is uninformative.)

(b) **Descriptive reweighted-tercile comparison.** The appendix comparison in Section 5.4.1 reweights treated outcomes by propensity-based posteriors and compares leniency-tercile distributions within each fitted regime. In the covariate-restricted implementation this is not a direct comparison of $\widehat{f}_{1k}(y \mid x)$ across judge groups, but it can still reveal whether large residual outcome differences remain after regime-specific reweighting.

Empirical Calibration of Sensitivity Bounds

D.1 Wald-ratio-level sensitivity (Proposition F.1)

To calibrate Proposition F.1, I compute the signed-weight multiplier

$$M(z, z') := \sum_k |\hat{\omega}_k^*(z, z')|$$

for all judge-bin pairs with $|\Delta\hat{p}(z, z')| \geq 0.03$. Under a common sensitivity parameter δ, the pairwise Wald ratio bias obeys $|\psi^{\text{rel}} - \psi| \leq M(z, z')\delta$.

Table 12: Empirical Sensitivity Multipliers and Bias Envelopes

			Multiplier M			Bias bound $Q_{95} \cdot \delta$		
Outcome	Sample	Pairs	Q_{50}	Q_{95}	Max	$\delta = .01$	$\delta = .02$	$\delta = .05$
log(1 + days_prison)	Full	23	1.000	1.107	1.425	0.011	0.022	0.055
log(1 + days_prison)	Black	24	1.000	1.107	1.392	0.011	0.022	0.055
log(1 + days_prison)	Non-Black	19	1.000	1.224	1.385	0.012	0.024	0.061

Notes: $M(z, z') := \sum_k |\hat{\omega}_k^*(z, z')|$ over valid pairs ($|\Delta\hat{p}| \geq 0.03$). Under common within-regime sensitivity parameter δ, Proposition F.1 implies $|\psi^{\text{rel}} - \psi| \leq M\delta$.

Table 12 reports the calibration for log(1 + days_prison) across all three samples.

For log(1+days_prison), the median multiplier is exactly 1.000 in all three samples and the 95$^{\text{th}}$ percentile is close to one ($Q_{95} \approx 1.1$–1.2), meaning the typical judge-pair comparison has well-behaved weights even though NegMass rejects in the full-sample and Black subsamples. More generally, the multiplier $M = \sum_k |\omega_k^*|$ exceeds one whenever some slope weights are negative, so the NegMass diagnostic and the sensitivity analysis reinforce each other: cells with larger offsetting margins are also the cells where the Wald ratio is most fragile to within-regime selection violations.

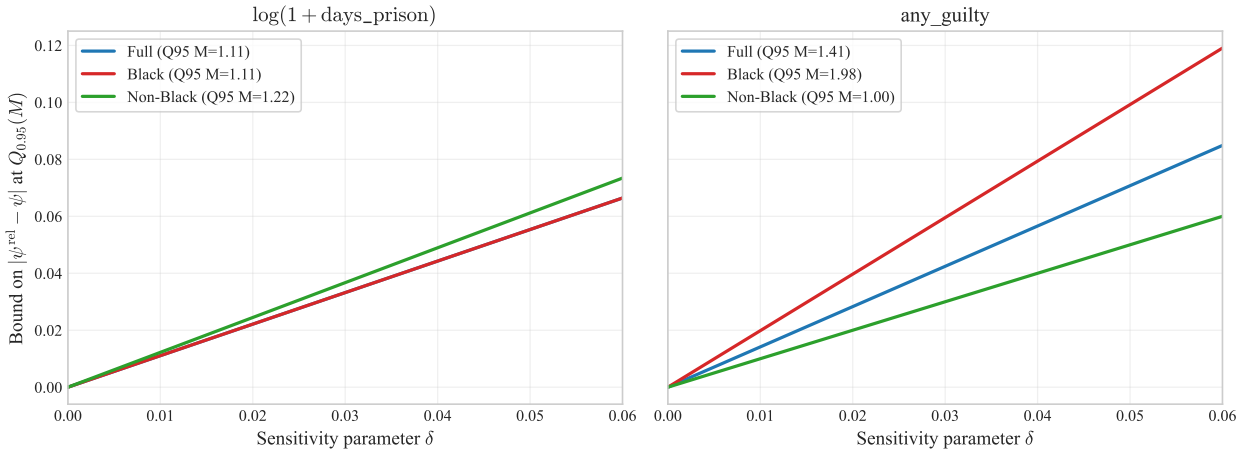

Figure 4: Sensitivity envelopes for the Wald ratio based on the 95$^{\text{th}}$ percentile of $M(z, z')$ for each outcome-sample cell. Each curve plots the maximum bias $M_{95} \cdot \delta$ as a function of the within-regime selection parameter δ.

D.2 Decomposition-specific sensitivity (Proposition F.2)

The Wald-ratio-level sensitivity analysis (Proposition F.1) targets the bias in a single-group Wald ratio. But the headline finding—that margin composition drives 98.5% of the racial MOT gap—concerns the *decomposition* of $\hat{\psi}^B - \hat{\psi}^{NB}$ into within-regime and margin-composition channels. These two channels have distinct sensitivity structures.

Under bounded within-regime selection (Assumption F.1) with $|b_k^r| \leq \delta$, write $\hat{\tau}_k^r = \tau_k^r + b_k^r$. The bias in each channel decomposes as:

$$\text{Within-regime bias:} \quad \sum_k \bar{\omega}_k (b_k^B - b_k^{NB}), \tag{F.1}$$

$$\text{Margin-composition bias:} \quad \sum_k (\hat{\omega}_k^{*B} - \hat{\omega}_k^{*NB}) \cdot \tfrac{1}{2}(b_k^B + b_k^{NB}). \tag{F.2}$$

The within-regime channel is sensitive to *differential* within-regime selection across races ($b_k^B - b_k^{NB}$), while the margin-composition channel is sensitive to *average* selection ($b_k^B + b_k^{NB}$). This asymmetry motivates two benchmark scenarios:

1. **Race-neutral within-regime selection** ($b_k^B = b_k^{NB}$ for all k): the within-regime channel has *zero bias* regardless of δ, and the composition channel bias is bounded by $M_C \cdot \delta$ where $M_C := \sum_k |\hat{\omega}_k^{*B} - \hat{\omega}_k^{*NB}|$. Under this scenario, the finding that margin composition dominates is *strengthened* by violations of Assumption 2.4, since the within-regime channel remains uncontaminated while the composition channel absorbs whatever bias exists.

2. **Worst-case race-differential selection** ($b_k^B = -b_k^{NB}$ for all k): the composition channel has zero bias, and the within-regime channel bias is bounded by $M_W \cdot \delta$ where $M_W := 2\sum_k |\bar{\omega}_k|$. This is the scenario most damaging to my headline finding: it maximally inflates the within-regime channel while leaving the composition channel unaffected.

Table 13: Decomposition-Specific Sensitivity Multipliers (Proposition F.2)

			Within worst case			Comp. race-neutral		
Outcome	M_W	M_C	$\delta = .01$	$\delta = .02$	$\delta = .05$	$\delta = .01$	$\delta = .02$	$\delta = .05$
log(1 + days_prison) (point est: $\|\widehat{\text{within}}\| = 0.0018$)	2.000	0.348	0.0200	0.0400	0.1000	0.0035	0.0070	0.0174

Notes: $M_W := 2\sum_k |\bar{\omega}_k|$ is the worst-case (race-differential) multiplier for the within-regime channel; $M_C := \sum_k |\hat{\omega}_k^{*B} - \hat{\omega}_k^{*NB}|$ is the race-neutral multiplier for the margin-composition channel. Under race-neutral within-regime selection ($b_k^B = b_k^{NB}$), the within-regime channel has zero bias regardless of δ, and the composition channel bias is bounded by $M_C\delta$. Under worst-case race-differential selection ($b_k^B = -b_k^{NB}$), the within-regime channel bias is bounded by $M_W\delta$ and the composition channel has zero bias. Pair is $(1, 10)$.

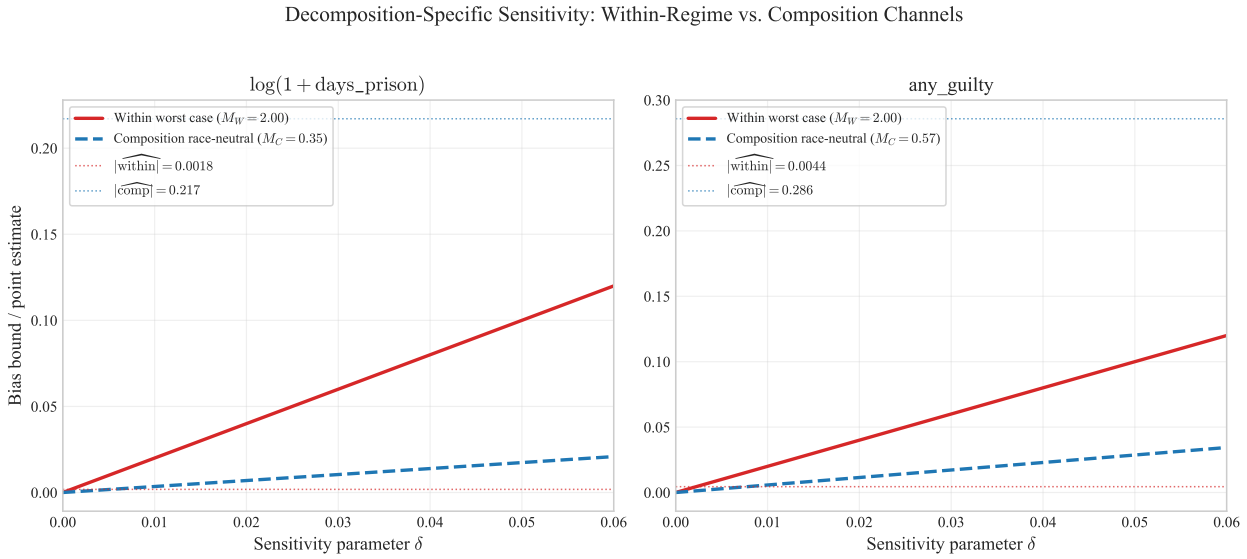

Figure 5: Decomposition-specific sensitivity envelopes. Solid red: within-regime worst-case bias envelope ($M_W \cdot \delta$). Dashed blue: composition race-neutral bias envelope ($M_C \cdot \delta$). Dotted horizontal lines: point estimates $|\widehat{\text{within}}|$ and $|\widehat{\text{comp}}|$. For the within-regime worst-case to overturn the finding that composition dominates, δ must be large enough that $M_W \cdot \delta$ exceeds $|\widehat{\text{comp}}|$—a condition that requires substantial race-differential within-regime selection.

For the sentencing outcome at the anchor pair $(1, 10)$, a nontrivial benchmark of $\delta = 0.02$ implies at most $M_W \times 0.02$ worst-case bias in the within-regime channel, which remains well below the margin-composition point estimate of $|\widehat{\text{comp}}| = 0.217$. Under the race-neutral

benchmark, the within-regime channel has zero bias regardless of δ, and the composition channel is contaminated only by $M_C \cdot \delta$, which shifts the level of the composition estimate without affecting the qualitative dominance of composition over within-regime effects.

G Further Practical Implementation Details

G.1 Regularization and numerical stabilization

The formal results analyze the unpenalized estimator. In computation, however, modest numerical regularization is useful. For the sieve-based selection block, small ℓ_1 penalties on the coefficient vector and soft group penalties on redundant regimes help prevent explosive coefficients when the empirical first stage is nearly degenerate. These penalties are used only to stabilize optimization; the estimand itself remains the unpenalized likelihood-based solution. In the empirical implementation, reported specifications are retained only after checking that weakening or removing the regularization does not materially change the fitted objects.

G.2 Diagnostic checks and model adequacy

I use three families of practical checks.

First, I inspect singular-value scree plots for empirical moment and CDF operators. These are not formal estimators of K_0, but they reveal whether the selected specification is obviously too coarse relative to the data. In particular, they are useful for ruling out a spurious $K = 1$ interpretation when several singular values remain energetic.

Second, I examine functional stability across nearby values of K. The main quantities of interest in the paper are the Wald-ratio profiles, signed-weight decompositions, and cross-group discrimination decompositions. If those objects stabilize once $K \geq 2$, then the discrete regime summary is adequate for the question of interest even if the true heterogeneity is continuous.

Third, I inspect descriptive reweighted-tercile outcome comparisons. The model implies that regime-specific outcome laws are invariant to the instrument once one conditions on the latent regime. In the covariate-restricted empirical implementation, however, the propensity-based posteriors depend only on the leniency bin, so the resulting Kolmogorov–Smirnov and QQ diagnostics do not recover pure regime-specific outcome laws. Instead they ask whether large treated-outcome differences remain across judge-leniency terciles after regime-specific reweighting. I use them as descriptive model-adequacy checks, not as direct tests of Assumption 2.4.

G.3 Choosing the number of regimes

The working dimension K is the number of latent decision environments in the mixture. It is distinct from the number of support points used by the KW subproblem, which adapts automatically within each regime.

The empirical workflow combines three inputs:

1. **Rank evidence.** The identification argument implies that K_0 is the dimension of the span of the informative arm-specific laws. Empirically, I approximate this with singular values of moment and CDF operators evaluated at fixed covariate strata.

2. **Cross-validated predictive fit.** I estimate the model for $K = 1, 2, 3$ and select the value that maximizes held-out log-likelihood for the joint model $(Y, D \mid Z)$.

3. **Functional stability.** When cross-validation is flat across nearby values of K, I require the decomposition objects of economic interest to stabilize before treating the fitted regime structure as adequate.

This procedure is deliberately conservative. Rank evidence is useful as a lower bound, but it can overstate K_0 if one lets covariate heterogeneity enter too flexibly. Cross-validation guards against that problem, and functional stability ensures that the selected model is not merely predictively adequate but also economically interpretable.

G.4 Regime labeling and alignment

All mixture components are identified only up to permutation. Within a single fit, I label regimes by their estimated treatment-effect magnitude $\widehat{\tau}_k$, or by the pair $(\widehat{\mu}_{1k}, \widehat{\mu}_{0k})$ when that ordering is more transparent. Across group-specific fits or bootstrap replicates, labels are re-aligned by minimum-distance matching on the joint object $(\widehat{\tau}_k, \widehat{p}_k(z))$, implemented as a standardized distance that uses both treatment effects and propensity schedules. In the discrimination decomposition, each bootstrap replicate re-estimates both group models and then re-solves the alignment problem so that label uncertainty is propagated into the reported confidence intervals rather than suppressed ex post.

G.5 EM–KW–WFR algorithm

The bail-convention sieve estimator combines an EM outer loop with a KW subproblem for the regime-specific outcome laws. Let r_{ik} denote the posterior responsibility of regime k for observation i. The generic E-step takes the form

$$r_{ik} \propto w_k(X_i)\, p_k(X_i, Z_i)^{D_i} \big(1 - p_k(X_i, Z_i)\big)^{1-D_i} f_{D_i k}(Y_i \mid X_i),$$

with normalization across k. In the bail convention, f_{0k} is degenerate and only the treated arm contributes an outcome density term; in the general convention both arms contribute.

The empirical implementation specializes the M-step to the two-arm Gaussian model used in the paper:

$$\widehat{w}_k = \frac{1}{n}\sum_i r_{ik}, \qquad \widehat{p}_k(z_j) = \frac{\sum_i r_{ik} D_i \mathbf{1}\{Z_i = z_j\}}{\sum_i r_{ik}\mathbf{1}\{Z_i = z_j\}}, \qquad \widehat{\mu}_{dk} = \frac{\sum_{i:D_i=d} r_{ik} Y_i}{\sum_{i:D_i=d} r_{ik}}, \qquad \widehat{\sigma}^2_{dk} = \frac{\sum_{i:D_i=d} r_{ik}\big(Y_i - \widehat{\mu}_{dk}$$

Under the bail-convention Gaussian sieve, the selection block is updated by weighted multinomial logistic regression and the outcome block by a weighted KW–NPMLE, which I compute by Wasserstein–Fisher–Rao particle descent.

Algorithm G.1 (Implementation summary for the empirical estimator)**.**

1. *Choose K and initialize the regime-specific parameters from multiple random starts.*
2. *Iterate until convergence:*
   (a) ***E-step:*** *compute posterior responsibilities r_{ik}.*
   (b) ***Selection-block M-step:*** *update the regime shares and propensity schedules using weighted responsibilities (empirical implementation) or weighted multinomial logistic regression plus the corresponding propensity update (sieve implementation).*
   (c) ***Outcome-block M-step:*** *update regime-specific Gaussian parameters by weighted MLE; under the bail-convention sieve, update the KW mixing distribution by WFR particle descent.*
3. *Stop when the relative log-likelihood change falls below the numerical tolerance and parameters have stabilized.*
4. *Keep the run with the highest attained log-likelihood.*

For the empirical runs reported in the paper, baseline fits use 20 random starts, a maximum of 500 EM iterations, and the relative stopping rule

$$|\ell^{(t)} - \ell^{(t-1)}| < 10^{-7}\max\{1, |\ell^{(t-1)}|\}.$$

Cross-validation uses lighter budgets (5 starts and 150 iterations per fold), while bootstrap refits use moderate budgets to control runtime without changing the selection procedure.

G.6 Inference and bootstrap implementation

Inference is based on parametric bootstrap procedures tailored to the null hypotheses of interest.

For the NegMass tests, I first estimate the unrestricted model, then project it onto the monotone null. The empirical implementation uses a contact-set projection for the inequality-constrained null so that only near-binding constraints are imposed, following the logic of moment-inequality procedures. Each bootstrap sample is generated from that projected null, the unrestricted model is re-estimated, and the bootstrap test statistic is compared with the observed one. The rank-1 test follows the same structure, except that the null projection is onto the rank-1 restriction. In the empirical application I use $B = 399$ bootstrap draws.

The optimization budgets are deliberately asymmetric. The observed unrestricted fit uses a larger budget (8 starts and 250 iterations inside the bootstrap wrappers) than each bootstrap refit (5 starts and 150 iterations). This is a standard computational compromise: the observed statistic should be estimated as accurately as possible, whereas the bootstrap distribution requires many refits and benefits more from stable repetition than from maximal optimization in each draw.

The discrimination decomposition uses a separate bootstrap routine because the object is cross-group. Each draw re-estimates both group-specific models at the common comparison value of K, realigns labels by weighted schedule matching, recomputes the within-regime and composition channels, and then forms percentile intervals. This is the appropriate place to handle label switching, because the decomposition is not permutation-invariant once one compares regime-specific objects across groups.

G.7 Common Computational Issues

The main computational failure modes are familiar from finite-mixture work.

Label switching. If two regimes are weakly separated, different random starts may return equivalent solutions with permuted labels. This is harmless for permutation-invariant diagnostics, but it matters for regime-by-regime reporting. I therefore align labels before tabulation and treat frequent relabeling across bootstrap draws as evidence that the selected K may be too rich for the available sample.

Vanishing regimes. When a regime receives negligible posterior mass, the corresponding weighted MLE can become numerically unstable. In practice I clip regime weights away from zero during iteration, monitor whether a regime repeatedly collapses across starts, and interpret persistent collapse as evidence that a smaller K is preferable.

Weak net first stages. Signed weights can be unstable when large regime-specific first-stage shifts offset in the aggregate. This is economically meaningful rather than a software bug, but it implies that some pairwise Wald ratios are poorly behaved. I therefore report valid-pair thresholds, signed-weight profiles, and the NegMass distortion measure rather than relying on a single scalar ratio.

Diagnostic plots. The practical toolkit includes singular-value scree plots, profile plots of $\widehat{p}_k(z)$, signed-weight profiles, and QQ plots for the continuous reweighted-tercile outcome comparison. These diagnostics are not substitutes for theory, but they make it much easier to see when a fitted model is doing something pathological.

H Further Details on the Empirical Application

H.1 Functional Stability of the Sentencing Decomposition

This appendix reports the sentencing decomposition at $K = 1, 2, 3$, delivering on the functional-stability diagnostic summarized in Section 7.6. The purpose is to show how the conclusions depend on the selected regime count and to document that $K = 2$ provides a transparent robustness check for the baseline $K = 3$ specification.

Table 14 reports the plug-in decomposition at each K for the anchor pair $(1, 10)$.

Table 14: Functional Stability of the Sentencing Decomposition Across K

K	$\hat{\psi}^B$	$\hat{\psi}^{NB}$	Gap	Within	Comp.	abs	gap
1	−0.316	−0.406	+0.090	+0.0904	+0.0000	—†	
2	−0.296	−0.053	−0.243	−0.0386	−0.2043	84.1%	84.1%
3	−0.230	−0.011	−0.219	−0.0018	−0.2171	99.2%	99.2%

Notes: Plug-in decomposition at the strict-vs-lenient pair $(1, 10)$ for each $K \in \{1, 2, 3\}$. Both race groups are estimated at the common K. †At $K = 1$, the composition channel is zero by construction. "abs" is $|\text{Comp}|/(|\text{Within}| + |\text{Comp}|)$; "gap" is Comp/Gap.

Three features of the results are worth emphasizing. First, $K = 1$ is clearly inadequate: it mechanically sets the composition channel to zero and attributes the entire gap to within-regime differences. That is the strongest scalar-leniency benchmark, and it is inconsistent with the within-group rank-1 evidence once the model allows multiple regimes. Second, once the model allows at least two regimes, the composition channel dominates the decomposition. Third, the $K = 2$ and $K = 3$ decompositions are close, which supports the view that the baseline $K = 3$ specification is not manufacturing the composition result.

H.2 Descriptive Reweighted-Tercile Outcome Comparisons

Assumption 2.4 implies that, within each estimated regime, the distribution of treated outcomes should not vary with the instrument. In a fully covariate-varying implementation one could test that restriction directly from regime-specific treated outcome distributions. The covariate-restricted implementation used here does not permit that direct check. Treated observations are reweighted by propensity-based posteriors $\hat{\omega}_{1k}(z) = \hat{w}_k \hat{p}_k(z)/\hat{p}(z)$, which depend only on the selection block of the model and, in this application, only on the residualized leniency bin. Within a given bin those weights are constant across observations, so within any leniency tercile the reweighted outcome distribution is just a bin-level reweighting of the overall within-tercile outcome distribution; it does not isolate the structural regime-specific densities whose means may be far apart (as in Table 8). Table 15 therefore reports descriptive comparisons of reweighted tercile mixtures rather than direct estimates of regime-specific outcome laws. The table reports the maximum KS statistic and minimum p-value across the three pairwise tercile comparisons within each regime.

Table 15: Reweighted-Tercile Outcome Comparisons for $\log(1 + \text{days\_prison})$

Sample	Regime	$\hat{\tau}_k$	n_k^{eff}	$\text{KS}_{\max}$	$p_{\min}$	$\bar{Y}_{\text{strict}}$	$\bar{Y}_{\text{lenient}}$
Full	1	-1.952	43,895.9	0.030	0.000***	0.924	0.831
Full	2	-0.575	44,171.0	0.030	0.000***	0.926	0.834
Full	3	0.000	44,191.0	0.030	0.000***	0.925	0.832
Black	1	-1.938	22,171.5	0.008	0.981	1.007	1.006
Black	2	-0.591	22,137.0	0.007	0.995	1.013	1.009
Black	3	0.000	22,192.8	0.007	0.983	1.007	1.007
Non-Black	1	-2.075	21,825.6	0.024	0.040**	0.769	0.765
Non-Black	2	-0.756	21,922.0	0.026	0.021**	0.771	0.773
Non-Black	3	0.000	21,995.2	0.024	0.039**	0.771	0.764

Notes: Within each fitted regime k, treated observations are reweighted by propensity-based posteriors and split into judge leniency terciles. In the covariate-restricted implementation these are descriptive tercile-comparison statistics rather than direct estimates of regime-specific outcome laws. $\text{KS}_{\max}$ is the maximum weighted Kolmogorov–Smirnov statistic across the three pairwise tercile comparisons; $p_{\min}$ uses asymptotic KS calibration with effective sample sizes.

The race-stratified results, which are the relevant inputs for the discrimination decomposition, show small discrepancies by this descriptive metric. For Black defendants, no cell rejects at conventional levels and the strict-versus-lenient mean differences are tiny. Non-Black defendants show slightly larger KS statistics, but the deviations remain modest in economic magnitude. The pooled full-sample rejections indicate that the reweighted tercile mixtures differ more visibly in the combined sample.

Two caveats remain important. First, because the mixture model imposes within-regime

distributional stability by construction, any post-estimation comparison can only be informative rather than decisive. I use propensity-based posteriors precisely to avoid the circularity that would arise from outcome-conditioned hard assignments. Second, in the present implementation the comparison is joint: it reflects both adequacy of the fitted regime structure and residual tercile-level outcome differences after reweighting. I therefore do not interpret the small race-stratified KS statistics as validating Assumption 2.4; the formal robustness device remains the sensitivity analysis in Appendix F.

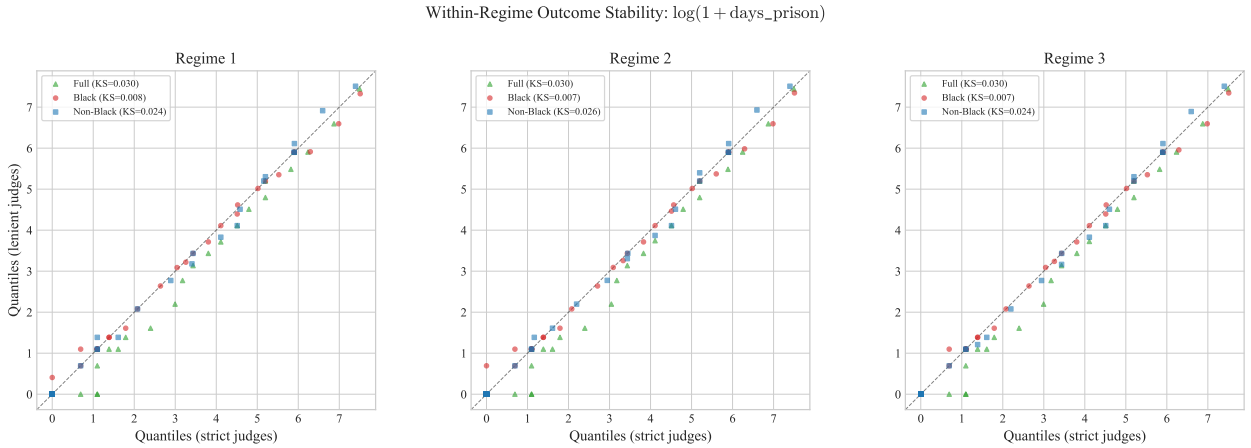

Figure 6: Reweighted-tercile QQ plots for log(1 + days_prison): strict versus lenient judge-tercile outcome quantiles, by regime and sample. Points near the 45-degree line indicate small visual differences in the reweighted tercile distributions.

I Two-Arm Parametric MLE: Consistency Under the General Convention

This appendix establishes consistency of the parametric MLE used in the empirical application under the general (two-arm) convention, in which outcomes are observed under both treatment states. The bail-convention sieve theory in Appendix E handles a richer nonparametric density class with a growing parameter space; here the goal is a self-contained result for the finite-dimensional parametric estimator that produced the empirical tables. This is not a corollary of the bail-convention theorem: Appendix E is written for the bail convention, where only the treated arm is outcome-informative and the identified selection object is the product $\pi_k = w_k p_k$, whereas the empirical estimator uses two informative outcome arms and a different finite-dimensional parameterization. In the current paper, this is the load-bearing consistency result for the continuous-outcome Gaussian specification.

The identification results in the main text are stated conditional on covariates X. In the empirical application, those covariates enter at the instrument-construction stage: the judge leniency measure is formed after residualizing on court-by-time fixed effects via UJIVE. Conditional on this residualized leniency variation, the implemented estimator uses the covariate-restricted special case of the model below: regime shares are constant within the estimation sample, regime-specific propensity schedules vary only by judge-leniency bin, and the arm-specific outcome kernels have no additional x-dependence. Because nonlinear mixture models do not satisfy a Frisch–Waugh–Lovell theorem, this scalarization remains a maintained approximation: residual dependence of the leniency score on the original design strata can in principle be absorbed by the mixture as apparent regime heterogeneity. I state the theorem below in slightly more general notation so that this implementation is nested as the intercept-only / fixed-basis specialization.

Model

Fix $K \geq 1$. For each (x, z), define the joint model for (Y, D) as

$$h(y, d \mid x, z; \theta) = \sum_{k=1}^{K} w_k(x; \alpha)\, p_k(x, z; \beta)^d \,(1 - p_k(x, z; \beta))^{1-d}\, f_{dk}(y \mid x; \eta_{dk}), \qquad d \in \{0, 1\}, \tag{I.1}$$

where $(w_1, \ldots, w_K)$ are parameterized via a softmax function with a fixed finite basis $b_w(x) \in \mathbb{R}^{q_w}$, while each regime-specific propensity is parameterized separately as a logistic index $p_k(x, z) = \Lambda(b_p(x, z)'\beta_k)$ with fixed finite basis $b_p(x, z) \in \mathbb{R}^{q_p}$ (e.g., judge-bin indicators interacted with covariates), and $f_{dk}(y \mid x; \eta_{dk})$ is a Gaussian kernel:

- For the sentencing outcome $\log(1 + \text{days\_prison})$: Gaussian with $\eta_{dk} = (\mu_{dk}, \sigma^2_{dk}) \in \mathcal{H}_{\text{Gauss}} := [-M, M] \times [\underline{\sigma}^2, \bar{\sigma}^2]$.

The full parameter vector is $\theta := (\alpha, \beta, \eta) \in \Theta$, where α collects the regime-share softmax coefficients, $\beta := \{\beta_k\}_{k=1}^K$ the regime-specific propensity-logit coefficients, and $\eta := \{\eta_{dk}\}_{d,k}$ the $2K$ kernel parameters.

Parameter space

Because the basis dimension q is fixed and the softmax coefficients are bounded (e.g., $\|\alpha\|_\infty \leq R$, $\|\beta\|_\infty \leq R$ for a chosen $R < \infty$), the full parameter space

$$\Theta := \{\alpha : \|\alpha\|_\infty \leq R\} \times \{\beta : \|\beta\|_\infty \leq R\} \times \mathcal{H}^{2K}$$

is a compact subset of a finite-dimensional Euclidean space. (Here $\mathcal{H} = \mathcal{H}_{\text{Gauss}}$.) The bound R is chosen large enough that $\theta_0 \in \Theta$ (Assumption A1 below).

Assumptions and result

Assumption I.1 (Two-arm parametric regularity)**.**

(A1) (Correct specification.) *There exists $\theta_0 \in \Theta$ such that $h(y, d \mid x, z; \theta_0)$ equals the true conditional density of (Y, D) given $(X, Z) = (x, z)$, for P_{XZ}-a.e. (x, z).*

(A2) (Compactness.) *Θ is compact (immediate from its definition as a bounded subset of Euclidean space).*

(A3) (Continuity and envelope.) *The map $\theta \mapsto \log h(y, d \mid x, z; \theta)$ is continuous on Θ for $\mu \otimes P_{XZ}$-a.e. (y, d, x, z), and there exists an integrable function $M(y, d, x, z)$ such that*

$$\sup_{\theta \in \Theta} \left| \log h(y, d \mid x, z; \theta) \right| \leq M(y, d, x, z), \qquad \mathbb{E}[M(Y, D, X, Z)] < \infty.$$

(A4) (Identification.) *The population criterion $Q(\theta) := \mathbb{E}[\log h(Y, D \mid X, Z; \theta)]$ is uniquely maximized at θ_0, up to permutation of the regime labels $\{1, \ldots, K\}$.*

Conditions (A1)–(A4) are the standard Wald conditions for consistency of an M-estimator on a fixed compact parameter space. Unlike the bail-convention sieve proof in Appendix E (which handles infinite-dimensional outcome densities on a growing parameter space, requiring sieve-approximation and entropy conditions), here the parameter space is finite-dimensional and does not grow with n, so the argument reduces to textbook extremum-estimator theory. The envelope in (A3) holds whenever the outcome space is bounded and the aggregate treatment rate is bounded away from $\{0, 1\}$, and for the Gaussian continuous-outcome specification it also holds under the compact mean/variance restriction together with a finite second moment of the true outcome distribution. For the empirical sentencing specification, Assumption I.1(A4) is supplied by Proposition 3.1 applied arm by arm under the maintained finite-Gaussian class, together with the two-arm accounting identities that recover the common regime shares and regime-specific propensity schedules once both arm-specific mixtures are pinned down. This theorem is likewise conditional on a fixed working value of K; it does not address recovery of K_0 or data-driven selection of K.

Theorem I.1 (Two-arm parametric MLE consistency)**.** *Suppose Assumption I.1 holds. Let $\widehat{\theta}_n \in \arg\max_{\theta \in \Theta} \ell_n(\theta)$, where $\ell_n(\theta) := n^{-1} \sum_{i=1}^{n} \log h(Y_i, D_i \mid X_i, Z_i; \theta)$. Then, up to label*

permutation, $\widehat{\theta}_n \to \theta_0$ *in probability. Moreover, the implied fitted conditional densities converge in* L_1*:*

$$\int \left|h(y,d \mid x,z;\widehat{\theta}_n) - h(y,d \mid x,z;\theta_0)\right| d\mu(y,d) \xrightarrow{p} 0$$

for P_{XZ}*-a.e.* (x,z)*, where* $\mu_{\mathcal{Y}}$ *is Lebesgue measure on* $\mathcal{Y}$ *and* $\mu = \mu_{\mathcal{Y}} \times \nu_D$ *with* ν_D *counting measure on* $\{0,1\}$*. Hence any continuous functional of the identified parameter* θ_0 *is consistently estimated.*

Proof. This is a direct application of the Wald consistency theorem for M-estimators on compact spaces (see, e.g., Newey and McFadden, 1994, Theorem 2.1).

Step 1: Uniform convergence. Θ is compact (A2), $\theta \mapsto \log h(\cdot;\theta)$ is continuous a.e. with integrable envelope (A3). By the uniform law of large numbers for continuous functions on compact sets (Newey and McFadden, 1994, Lemma 2.4):

$$\sup_{\theta\in\Theta} \left|\ell_n(\theta) - Q(\theta)\right| \xrightarrow{p} 0. \tag{I.2}$$

Step 2: Identification. By (A1), $\theta_0 \in \Theta$, so $\widehat{\theta}_n$ maximizes ℓ_n over a set containing θ_0. By (A4), θ_0 is the unique maximizer of Q over Θ (up to label permutation).

Step 3: Parameter convergence. Standard argmax reasoning: uniform convergence (I.2) plus unique maximization of Q at θ_0 imply $\widehat{\theta}_n \to \theta_0$ in probability, up to label permutation. (Compactness of Θ ensures every subsequence has a convergent further subsequence; continuity of Q and (I.2) force the limit to maximize Q; (A4) pins the limit at θ_0.)

Step 4: Density convergence. Parameter convergence and continuity (A3) imply $h(y,d \mid x,z;\widehat{\theta}_n) \to h(y,d \mid x,z;\theta_0)$ for a.e. (y,d). Since $\int h(y,d \mid x,z;\theta)\, d\mu(y,d) = 1$ for all θ, Scheffé's lemma gives L_1 convergence. □

References

Arnold, David et al. (2022). "Measuring Racial Discrimination in Bail Decisions". In: *American Economic Review*.

Bhuller, Manudeep and Henrik Sigstad (2022). "Errors and Monotonicity in Judicial Decision-Making". In: *Economics Letters* 215, p. 110486. DOI: 10.1016/j.econlet.2022.110486.

Bonhomme, Stéphane et al. (2022). "Discretizing Unobserved Heterogeneity". In: *Econometrica* 90.2, pp. 625–643. DOI: 10.3982/ECTA15238.

Canay, Ivan A. et al. (2024). "On the Use of Outcome Tests for Detecting Bias in Decision Making". In: *The Review of Economic Studies* 91.4, pp. 2135–2167. DOI: 10.1093/restud/rdad082.

Chaisemartin, Clément de (2017). "Tolerating Defiers in Instrumental Variables Estimation". In: *Quantitative Economics*.

Chan, David C. et al. (2022). "Selection with Variation in Diagnostic Skill: Evidence from Radiologists". In: *Quarterly Journal of Economics* 137.2, pp. 729–783. DOI: 10.1093/qje/qjab048.

Chen, Jiahua and Pengfei Li (2009). "Hypothesis Test for Normal Mixture Models: The EM Approach". In: *Annals of Statistics* 37.5A, pp. 2523–2542. DOI: 10.1214/08-AOS651.

Chyn, Eric et al. (2025). "Examiner and Judge Designs in Economics: A Practitioner's Guide". In: *Journal of Economic Literature* 63.2, pp. 401–439. DOI: 10.1257/jel.20241719.

Coulibaly, Mohamed et al. (2024). *A sharp test for the judge leniency design*. NBER Working Paper 32456. National Bureau of Economic Research. DOI: 10.3386/w32456.

Dobbie, Will et al. (2018). "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges". In: *American Economic Review*.

Farre-Mensa, Joan et al. (2020). "What Is a Patent Worth? Evidence from the U.S. Patent "Lottery"". In: *Journal of Finance* 75.2, pp. 639–682. DOI: 10.1111/jofi.12867.

Follmann, Dean and Diane Lambert (1991). "Identifiability of Finite Mixtures of Logistic Regression Models". In: *Journal of Statistical Planning and Inference* 27.3, pp. 375–381. DOI: 10.1016/0378-3758(91)90050-O.

Frandsen, Brigham et al. (Jan. 2023). "Judging Judge Fixed Effects". In: *American Economic Review* 113.1, pp. 253–277. DOI: 10.1257/aer.20201860. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20201860.

Goldsmith-Pinkham, Paul et al. (2025). *Leniency Designs: An Operator's Manual*. NBER Working Paper 34473. National Bureau of Economic Research. DOI: 10.3386/w34473.

Gupta, Arpit et al. (2016). "The Heavy Costs of High Bail: Evidence from Judge Randomization". In: *The Journal of Legal Studies* 45.2, pp. 471–505. DOI: 10.1086/688907.

Heckman, James J. and Rodrigo Pinto (2018). "Unordered Monotonicity". In: *Econometrica* 86.1, pp. 1–35. DOI: 10.3982/ECTA13777.

Heckman, James J. and Edward J. Vytlacil (2005). "Structural Equations, Treatment Effects, and Econometric Policy Evaluation". In: *Econometrica* 73.3, pp. 669–738.

Henry, Marc et al. (2014). "Partial Identification of Finite Mixtures in Econometric Models". In: *Quantitative Economics* 5.1, pp. 123–144. DOI: 10.3982/QE170.

Hoshino, Tadao and Takahide Yanagi (2022). "Estimating Marginal Treatment Effects under Unobserved Group Heterogeneity". In: *Journal of Causal Inference* 10.1, pp. 197–216. DOI: 10.1515/jci-2021-0052.

Hull, Peter (2021). *What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making*. NBER Working Paper 28503. National Bureau of Economic Research. DOI: 10.3386/w28503.

Humphries, John Eric et al. (2025). "Conviction, incarceration, and recidivism: Understanding the revolving door". In: *The Quarterly Journal of Economics* 140.4, pp. 2907–2962.

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

Jochmans, Koen et al. (2017). "Inference on Two-Component Mixtures under Tail Restrictions". In: *Econometric Theory* 33.3, pp. 610–635. DOI: 10.1017/S0266466616000098.

Kamat, Vishal et al. (2023). "Identification in Multiple Treatment Models under Discrete Variation". In: *arXiv preprint arXiv:2307.06174*. URL: https://arxiv.org/abs/2307.06174.

Kasahara, Hiroyuki and Katsumi Shimotsu (2009). "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices". In: *Econometrica*.

Kitamura, Yuichi and Louise Laage (2018). "Nonparametric analysis of finite mixtures". In: *arXiv preprint arXiv:1811.02727*.

Kling, Jeffrey R. (2006). "Incarceration Length and the Returns to Education". In: *The Quarterly Journal of Economics*.

Lee, Sokbae and Bernard Salanié (2018). "Identifying Effects of Multivalued Treatments". In: *Econometrica* 86.6, pp. 1939–1963. DOI: 10.3982/ECTA15155.

— (2024). *Treatment Effects with Targeting Instruments*. cemmap Working Paper CWP23/24. cemmap. DOI: 10.47004/wp.cem.2024.2324.

Mogstad, Magne and Alexander Torgovitsky (2024). "Instrumental Variables with Unobserved Heterogeneity in Treatment Effects". In: *Handbook of Labor Economics*. Vol. 5. Elsevier, pp. 1–114. DOI: 10.1016/bs.heslab.2024.11.003.

Mogstad, Magne, Alexander Torgovitsky, and Christopher R. Walters (2021). "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables". In: *American Economic Review* 111.11, pp. 3663–3698. DOI: 10.1257/aer.20190221.

Mueller-Smith, Michael (2015). *The Criminal and Labor Market Impacts of Incarceration*. Working Paper. Version dated August 18, 2015. University of Michigan.

Mungan, Murat C. (2023). "Endogenous Judge Decision Quality, Monotonicity, and Treatment Effects". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.4322684.

Norris, Samuel et al. (2021). "The Effects of Parental and Sibling Incarceration: Evidence from Ohio". In: *American Economic Review* 111.9, pp. 2926–2963. DOI: 10.1257/aer.20190415.

Romano, Joseph P. et al. (2014). "A Practical Two-Step Method for Testing Moment Inequalities". In: *Econometrica* 82.5, pp. 1979–2002. DOI: 10.3982/ECTA11011.

Sigstad, Henrik (Jan. 2026). "Monotonicity among Judges: Evidence from Judicial Panels and Consequences for Judge IV Designs". In: *American Economic Review* 116.1, pp. 189–208. DOI: 10.1257/aer.20231104. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20231104.

Teicher, Henry (1963). "Identifiability of Finite Mixtures". In: *Annals of Mathematical Statistics* 34.4, pp. 1265–1269. DOI: 10.1214/aoms/1177703862.

Tsuda, Toshiki (2024). *Treatment Effects with Multidimensional Unobserved Heterogeneity: Identification of the Marginal Treatment Effect*. Revised December 2024. DOI: 10.48550/arXiv.2209.11444. arXiv: 2209.11444 [econ.EM].

Yakowitz, Sidney J. and John D. Spragins (1968). "On the Identifiability of Finite Mixtures". In: *Annals of Mathematical Statistics* 39.1, pp. 209–214. DOI: 10.1214/aoms/1177698520.