MULTIMARGIN SELECTION IN LENIENCY DESIGNS

Lonjezo Sithole*

Abstract

Leniency designs exploit quasi-random assignment of cases to decision makers with different propensities to assign treatment. When decision makers weight unobserved case characteristics differently, a higher overall treatment rate can mask a lower probability of treating certain case types, with consequences for the interpretation of the IV estimand. I develop a model with finitely many latent regimes, each representing a case type with its own potential outcome distributions and decision-maker-specific treatment propensities. When each decision maker's treatment assignment within each regime and covariate cell preserves the marginal distribution of each potential outcome, observed outcome distributions admit a finite mixture representation. I show that pairwise Wald estimands weight regime-specific average treatment effects by each regime's relative contribution to the first stage. Because these weights can be negative, the IV estimand can be negative even when every regime's average treatment effect is positive. I provide conditions for identifying the number of regimes, regime-specific average treatment effects, and IV weights using either parametric restrictions on outcome distributions or auxiliary pretreatment measurements. I also characterize when the weights are nonnegative and when they are constant across decision-maker pairs, and develop pointwise asymptotically valid bootstrap tests of both restrictions for a fixed number of regimes. In an application to Miami-Dade County bail decisions, I find substantial variation across latent case types in the effects of pretrial release on sentence duration.

JEL codes: C21, C26, C38, K42.
Keywords: judge leniency, instrumental variables, monotonicity, mixture models, marginal treatment effects.

*First version: April 12, 2026. This version: November 5, 2026.
Department of Economics, University of Michigan, lsithole@umich.edu. I thank Florian Gunsilius, Kaspar Wuthrich, Eva Janssens, Andreas Hagemann, Charlie Brown, as well as participants in the Econometrics Workshop and the Econometrics Seminar at the University of Michigan, the Celebrating James MacKinnon's 75th Birthday Conference at Aarhus University, the 2026 Bristol Econometric Study Group, and the 41st Canadian Econometric Study Group for their valuable feedback and suggestions. The usual disclaimer applies.

1 Introduction

Leniency designs have become an increasingly important part of the toolkit for researchers in applied economics and adjacent fields. These empirical designs exploit exogenous variation induced by quasi-random assignment of cases to decision makers with systematically different propensities to assign treatment. In recent years, researchers have used them to estimate causal effects of pretrial detention on employment and earnings (Dobbie, Goldin and Yang, 2018) and of first patent grants on sales growth among startups (Farré-Mensa, Hegde and Ljungqvist, 2020), among other applications. Chyn, Frandsen and Leslie (2025) and Goldsmith-Pinkham, Hull and Kolesár (2026) provide surveys of the literature along with econometric guidance to practitioners.

Viewed through the lens of the latent index model of Heckman and Vytlacil (2005), cases are ordered along a single margin of latent resistance to treatment and each decision maker applies a different threshold on this margin. Under instrument exogeneity, this latent index representation is equivalent to monotonicity (Vytlacil, 2002): holding observed characteristics fixed, a more lenient decision maker treats every case treated by a stricter one. Cases, however, can vary along several unobserved dimensions, and decision makers may disagree about how much weight to place on each. In the pretrial bail setting, for example, some defendants with the same charges and criminal histories may have more precarious livelihoods than others, which could affect the employment and earnings consequences of detention. Judges may learn about these circumstances but differ in how much weight they place on them when deciding whom to release.

To fix ideas, consider two equally common latent case types and three judges with the release propensities in Table 1. Judge A releases more defendants overall than B, but fewer defendants of type 2. The judges' ordering reverses across types, violating monotonicity. Notice that judges B and C have identical overall release rates but they treat the two types differently.[1]

Table 1: Release Propensities in the Running Example

	Type 1	Type 2	Overall release rate
Judge A	0.80	0.30	0.55
Judge B	0.55	0.45	0.50
Judge C	0.30	0.70	0.50

For this illustration, let Y indicate pretrial misconduct while released, which cannot occur under detention, so $Y(0) = 0$. Suppose judges are randomly assigned and the mean effects

[1] Chyn, Frandsen and Leslie (2025, Appendix A) provide a stylized example of multidimensional judge behavior with different defendant types. Here, the types are latent, and the model provides conditions for identifying their outcome distributions and treatment propensities.

are $\tau_1 = 0.15$ and $\tau_2 = 0.35$. Within each type, the average potential outcome under release is the same among defendants released by each judge as in the type's full population. The pairwise Wald ratio for judges A and B is

$$\psi(A, B) = \frac{0.5(0.25)(0.15) + 0.5(-0.15)(0.35)}{0.55 - 0.50} = 2.5(0.15) - 1.5(0.35) = -0.15. \tag{1.1}$$

The negative weight on type 2 reverses the sign of the pairwise IV estimand despite both types having positive treatment effects. For judges B and C, the Wald ratio is not even defined: the aggregate release rates are equal, even though the average outcomes for defendants assigned to B and those assigned to C are different.

The sign reversal in the example strips the IV estimand of a weakly causal interpretation: the estimand is negative although all individual treatment effects are nonnegative (Blandhol et al., 2026). But even when monotonicity holds, and the estimand admits a weakly causal interpretation, the standard IV estimand can still average away useful variation across latent case types. A nonnegatively weighted average, for example, can be small because large positive and negative effects offset,[2] or because types with large effects receive little weight. Disentangling which of these two explanations applies in any particular setting necessarily requires picking apart the type-specific average treatment effects and their weights.

To this end, I model unobserved case heterogeneity using finitely many latent regimes. Each regime is a case type, a combination of unobserved case characteristics, with its own potential outcome distributions and treatment propensities under each decision maker. A decision maker may be more lenient than another towards cases in one regime while being stricter towards cases in another regime, so the ranking of decision makers in terms of their treatment propensity can vary across regimes. Each decision maker, therefore, has a vector of regime-specific treatment propensities: $z \mapsto (p_1(z), p_2(z), \ldots, p_{K_0}(z))$, where $p_k(z)$ represents the treatment propensity of decision maker z in regime k, instead of a single treatment propensity, $p(z)$. I establish conditions for identifying these regime-specific treatment propensities, population shares, and average treatment effects.

The main identifying restriction that I impose is that, within each regime, the cases assigned to either treatment status by any decision maker have the same potential outcome distribution as the full regime population under that treatment status. Under this restriction, decision makers can freely select whom to treat based on potential outcomes across latent regimes, but within regimes they can select on potential outcomes only in ways that preserve the regime's marginal potential outcome distribution under each treatment status. Under

[2] Under random assignment and exclusion, consider two equally common case types with average treatment effects -0.45 and 0.20. Suppose individual-level monotonicity holds and A raises treatment propensities relative to B by 0.20 and 0.45, respectively, with the additional treated cases having their type's average treatment effect. The Wald ratio averages the effects for the additional treated cases (Imbens and Angrist, 1994): $\frac{0.5(0.20)(-0.45)+0.5(0.45)(0.20)}{0.5(0.20)+0.5(0.45)} = 0$, although neither type has a zero average effect.

quasi-random assignment, observed outcome distributions in each treatment arm then form finite mixtures whose component distributions are common across decision makers and whose mixture weights vary across decision makers.

This variation in mixture weights is enough to uniquely pin down the number of regimes, provided the regime-specific outcome distributions are linearly independent, and decision makers mix these distributions in sufficiently different proportions. Recovering the regime-specific outcome distributions requires additional restrictions, such as Gaussian or negative binomial specifications, while leaving treatment propensities unrestricted across decision makers. Recovering regime-specific average treatment effects also requires pairing each regime's components under treatment and no treatment. Since the components in each treatment arm are identified only up to a permutation of their labels, components with the same label under treatment and no treatment can belong to different regimes. I use random assignment of decision makers to pair components that belong to the same regime in the two treatment arms. Random assignment requires the shares of the full population in a regime's treated and untreated components to sum to the same regime share under every decision maker. In Section 4.1, I provide conditions under which this restriction uniquely pairs the components.

I also consider a fully nonparametric alternative using three auxiliary pretreatment measurements, such as prior earnings or test scores, that are conditionally independent given the regime and observed covariates. When their distributions are sufficiently distinct across regimes, their joint distribution identifies the number of regimes, population shares, and regime-specific measurement distributions. If, within each regime, one measurement is also distributed identically across treatment arms and decision makers, its distribution among treated and untreated cases identifies each regime's share in each treatment arm and hence the regime-specific treatment propensities. Sufficient variation in these shares across decision makers then identifies the regime-specific potential outcome distributions without imposing a functional form. Section 4 formalizes these arguments.

Either route recovers each regime's average treatment effect, a causal parameter that does not depend on which decision makers are being compared. I also show that the pairwise Wald estimand is a weighted sum of these effects. The weights depend on how reassignment between the two decision makers changes treatment within each regime. Regimes whose cases contribute disproportionately more towards differences in overall leniency receive more weight than their population shares. Regimes that every decision maker treats at the same rate receive no weight, however large their population shares. For IV using all decision makers, a regime receives a negative weight when decision makers who are more lenient overall tend to be stricter toward its cases. When every regime's IV weight is nonnegative, IV is weakly causal with respect to regime-specific average treatment effects even when individual-level and average monotonicity fail. The IV estimand can still be different from the population average treatment effect because the weights it assigns to regimes do not coincide with their

population shares. When these shares are identified, reweighting these average treatment effects recovers the population average treatment effect.

The IV decomposition yields two testable restrictions with different implications for interpretation. The first, *common direction*, requires that a decision maker who is more lenient overall cannot be stricter in any regime. This condition is necessary and sufficient for nonnegative IV weights for every pair of decision makers with a nonzero first stage. It therefore characterizes when every pairwise Wald estimand is weakly causal with respect to regime-specific average treatment effects, adapting the criterion of Blandhol et al. (2026). The second restriction, *Rank-1 proportionality*, requires every decision-maker comparison to change the regime-specific treatment propensities in the same proportions. When overall leniency varies, this ensures that every pair of decision makers with a nonzero first stage places identical weights on the regimes and thus targets the same Wald estimand. I develop bootstrap tests of both restrictions and establish their pointwise asymptotic validity under regularity conditions for a given number of regimes. Neither restriction is required to identify the latent regime model.

I estimate the parametric specifications by maximum likelihood and provide a plug-in estimator using the auxiliary measurements in Section 6.2. In Monte Carlo experiments, maximum likelihood accurately recovers regime-specific average treatment effects when outcome distributions are well separated or regime composition among treated and untreated cases varies strongly across decision makers, including the design in which the IV estimand is negative despite positive average treatment effects in both regimes. In separate experiments, the Rank-1 test has near-nominal size and the common direction test is conservative; both reject in every sample under their respective alternatives.

I apply the model to data on bail decisions in Miami-Dade County, using pretrial release as the treatment and recorded sentence length as the outcome. The estimated average effects of release vary substantially across regimes. Across the fitted specifications, the regime with the largest estimated reduction in sentence length receives less IV weight than its population share.

Related Literature. This work contributes to a growing strand of literature addressing monotonicity violations in leniency designs and in the IV literature more broadly. In empirical work using the examiner IV design, Mueller-Smith (2015) allows treatment propensities to vary with *observed* defendant characteristics, so monotonicity needs to hold only within observed covariate cells. Chan, Gentzkow and Yu (2022) show that differences in diagnostic skill can generate nonmonotone decisions, which suggests that controlling for observed characteristics may not be enough to restore monotonicity. Frandsen, Lefgren and Leslie (2023) show that the IV estimand remains a positively weighted average of treatment effects under *average monotonicity* and average exclusion. Average monotonicity allows reversals between some pairs of decision makers but rules out any case whose treatment, across decision makers, is negatively

correlated with their leniency. Mogstad and Torgovitsky (2024, Section 3.4) question whether this is easier to justify than monotonicity, since which cases it rules out depends on decision makers' propensities and on how often each is assigned. Mogstad, Torgovitsky and Walters (2021) study multiple instruments under partial monotonicity, requiring monotonicity in each instrument while holding the others fixed.

My model instead takes unobserved case heterogeneity as its starting point, and identifies regime-specific average treatment effects without requiring monotonicity. The approach is in the spirit of the "forward engineering" advocated by Mogstad and Torgovitsky (2024): specify the causal parameters and a model that identifies them, then derive what a standard IV estimand averages. Recent papers test the standard assumptions of random assignment, exclusion, and monotonicity (Coulibaly et al., 2024; Mao, Sant'Anna and Song, 2026), or the weaker average monotonicity condition (Goldsmith-Pinkham, Hull and Kolesár, 2026). The model here complements these tests by identifying whether one decision maker is more lenient than another toward some regimes but stricter toward others, and by decomposing the IV estimand into regime-specific average treatment effects and their weights.

In closely related work, Hoshino and Yanagi (2022) study latent group heterogeneity in binary treatment, identifying group-specific MTE curves using group-specific continuous instruments. I instead recover regime-specific treatment propensities and regime-specific average treatment effects using discrete instrument variation that is common to all regimes. A different strand of literature on multidimensional selection considers multiple selection thresholds arising from multiple treatments, sequential treatments, or several hurdles to receiving a binary treatment (Lee and Salanié, 2018; Kamat, Norris and Pecenco, 2026). Here, the underlying problem is fundamentally how decision makers weight several unobserved case characteristics when making the same binary treatment decision. The regimes represent this unobserved heterogeneity underlying the binary treatment.

My identification arguments build on the literature on finite mixtures, especially results on mixtures with varying weights (Henry, Kitamura and Salanié, 2014), and on identification of latent variable models from conditionally independent measurements (Kruskal, 1977; Hu, 2008; Hu and Schennach, 2008; Allman, Matias and Rhodes, 2009; Bonhomme, Jochmans and Robin, 2016). The causal setting presents two challenges for finite mixtures: matching the components identified among treated and untreated cases to the same regime, and converting mixture weights into treatment propensities.

Section 2 introduces the model, Section 3 gives the Wald and IV decompositions, Section 4 establishes identification, and Section 5 relates treatment propensities to IV weights. Section 6 develops estimation and inference. Sections 7 and 8 present simulation evidence and the application. Proofs, extensions, and additional simulation results are in the appendices.

2 The Model

2.1 Latent Regimes and Potential Outcomes

Let $Y \in \mathcal{Y} \subseteq \mathbb{R}$ denote the observed outcome, $D \in \{0,1\}$ treatment status, $X \in \mathcal{X}$ observed covariates, and $Z \in \mathcal{Z}$ an observed instrument with finite support. To streamline exposition in this section, I focus on the case where Z is the decision maker identity. In Section 6, I also consider instruments formed by grouping decision makers according to an estimated leniency measure.

The sample consists of i.i.d. observations of (Y, D, X, Z). 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 variables satisfy:

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

For results involving expectations, I assume $\mathbb{E}[|Y(d)|] < \infty$ for $d \in \{0,1\}$ and work at covariate values where both conditional first moments are finite.

Latent regimes. I define the *latent regime* as an unobserved variable, $S \in \{1, \ldots, K_0\}$, that bundles unobserved case characteristics simultaneously affecting potential outcome distributions and potential treatment decisions under each decision maker. Cases sharing these characteristics belong to the same regime, though cases within a regime may still differ in their realized outcomes, Y, and in the residual factors influencing the decision maker's treatment choice, V. However, I require that a case belongs to the same regime under every possible decision maker assignment: the same S appears in the potential treatment decisions, $D(z)$ (see equation (2.2) below). Assumption 2.2 further rules out dependence of the regime membership of a case on the decision maker assigned to it, conditional on X.

Regime-specific treatment probabilities are unrestricted across judges, allowing judges to differ in how they respond to unobserved case attributes. The latent regimes pertain to cases because I want to identify how average treatment effects differ across unobserved case types and how judges' treatment decisions weight those effects in the IV estimand.

The treatment rule under decision maker z admits the following general nonseparable representation

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

where g_S is the treatment decision function for regime S; and V represents residual unobserved heterogeneity in treatment decisions after conditioning on X and regime membership S.

Unlike the latent index model, the theory below does not impose restrictions on V: it may be a scalar or a vector. Both the treatment function and the distribution of V may also vary

with k. The treatment function is not required to be separable in assignment and unobserved characteristics. It can therefore admit violations of individual-level monotonicity, for which a common scalar threshold representation does not exist (Vytlacil, 2002).

Integrating $D(z)$ over V yields the regime-specific treatment propensities

$$p_k(x, z) = \Pr\{g_k(X, z, V) \geq 0 \mid X = x, Z = z, S = k\},$$

where $p_k(x, z)$ is the treatment propensity for regime k under instrument value z.

The average treatment effect within regime k is

$$\tau_k(x) := \mathbb{E}\Big[Y(1) - Y(0) \mid X = x, S = k\Big]. \tag{2.3}$$

The following are the core assumptions that I will invoke in the results below. Throughout, I condition on the same covariates X.[3]

Assumption 2.1 (Finitely many regimes with positive population shares)**.** $S \in \{1, \ldots, K_0\}$ *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 (Independence of regime membership and assignment conditional on covariates)**.** *For all* (x, z)*,*

$$\Pr(S = k \mid X = x, Z = z) = \Pr(S = k \mid X = x) = w_k(x). \tag{2.4}$$

Assumption 2.3 (Instrument exogeneity)**.**

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

Assumption 2.4 (Preservation of marginal potential outcome distributions)**.** *For almost every* x*, every instrument value* z *in the support of* Z *given* $X = x$*, and* $d \in \{0, 1\}$*,*

$$D(z) \perp Y(d) \mid S,\ X = x.$$

Assumption 2.1 specifies finitely many regimes, each with a strictly positive population share $w_k(x)$ at every covariate value in $\mathcal{X}$. Assumptions 2.2 and 2.3 follow immediately from quasi-random assignment and are therefore mild in this setting.

[3]An application may restrict regime shares, treatment propensities, and outcome distributions to depend on subsets $X_w, X_D, X_Y \subseteq X$. Such a specification assumes, for example, that conditioning the regime share on all of X is equivalent to conditioning on X_w. The identification analysis simply sets $X_w = X_D = X_Y = X$ without loss of generality.

The more substantive restriction in this setting is Assumption 2.4. Since $D(z)$ is binary, the assumption is equivalent to requiring that the cases a decision maker assigns to treatment status d have the same distribution of $Y(d)$ as all the cases in that regime. So, whenever $\Pr(D(z) = d \mid X = x, S = k) > 0$,

$$\Pr(Y(d) \leq y \mid D(z) = d, X = x, S = k) = \Pr(Y(d) \leq y \mid X = x, S = k) \qquad \text{for all } y. \quad (2.6)$$

Assumption 2.4 requires each $D(z)$ to be independent of each potential outcome, $Y(d)$, within a regime and covariate cell. Across regimes, decision makers can select whom to treat based on potential outcomes, because regimes differ in both their potential outcome distributions and their treatment propensities. Within a regime, however, they can select on potential outcomes only in ways that leave the distribution of $Y(1)$ among treated cases and of $Y(0)$ among untreated cases the same as in the regime as a whole. In the bail setting, for example, the assumption fails if, within a regime and conditional on observed characteristics, released defendants have a lower probability of misconduct than the regime population would have if everyone were released.

A stronger sufficient condition makes potential outcomes *jointly* independent of all unobserved factors associated with treatment within a regime:

$$(Y(1), Y(0)) \perp V \mid (X, S).$$

Under this condition, $D(z)$ is independent of $(Y(1), Y(0))$ given (X, S), so decision makers cannot select on potential outcomes within a regime at all. My results do not require this condition. Assumption 2.4 restricts the joint distribution of $D(z)$ with $Y(1)$ and with $Y(0)$, but does not require $D(z)$ to be independent of the pair $(Y(1), Y(0))$.

Assumption 2.4 may also hold only approximately. Appendix G starts from continuously varying unobserved case types, partitions them into finitely many regimes, and gives conditions under which selection changes the potential outcome distributions within each regime by only a small amount. The sensitivity analysis in Section G.2 then bounds the resulting errors in the Wald decomposition when, within a regime and conditional on observed covariates, the mean of $Y(d)$ among cases assigned treatment status d by a decision maker differs from its mean in the full regime population.

Remark 2.1. In the MTE model of Heckman and Vytlacil (2005), all decision makers rank cases the same way, by a single latent resistance to treatment, and each sets its own threshold on that ranking. Treatment effects can vary freely along this ranking. The finite regime model does not impose a common ranking. A decision maker can be more lenient than another toward one regime and stricter toward another, as judge A is relative to judge B in Table 1. It requires instead that, within each regime, the cases a decision maker treats have the regime's distribution of $Y(1)$ and the cases left untreated have the regime's distribution of

$Y(0)$. Adding latent regimes to the latent index model, with each decision maker applying one threshold to all regimes, implies common direction but not Rank-1.

2.2 Mixture Representation of Observed Outcome Distributions

Under instrument exogeneity, the treatment propensities defined above satisfy $p_k(x,z) = \Pr(D=1 \mid X=x, Z=z, S=k)$. Define the observed and regime-specific conditional distribution functions

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

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

When these distributions admit Radon–Nikodym derivatives with respect to a dominating measure μ, I denote the derivatives by $g_{dz}(y \mid x)$ and $f_{dk}(y \mid x)$, respectively.

Lemma 2.1 (Observed and potential outcome distributions within regimes)**.** *Under Assumptions 2.3–2.4, for almost every x and every (d,z,k) with $\Pr(D=d, Z=z, S=k \mid X=x) > 0$, the following equality holds for every y:*

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

Lemma 2.2 below shows that, under Assumptions 2.1-2.4 above, the observed outcome distribution for cases in each treatment arm handled by a decision maker, z, admits representation as a finite mixture over the regime-specific potential outcome distributions.

Fix a covariate value x at which the assumptions hold. For the mixture representation, write $g_{dz}(y) = g_{dz}(y \mid x)$, $w_k = w_k(x)$, and $p_k(z) = p_k(x,z)$, and let $p(z) := \Pr(D=1 \mid X=x, Z=z)$.

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

$$g_{dz}(y) = \sum_{k=1}^{K_0} \omega_{dk}(z)\, f_{dk}(y \mid x), \tag{2.10}$$

where $\omega_{dk}(z)$ is the conditional probability of regime k given $D=d$, $Z=z$, and $X=x$. These mixture weights satisfy

$$\omega_{1k}(z) = \frac{w_k\, p_k(z)}{p(z)}, \qquad \omega_{0k}(z) = \frac{w_k\,(1-p_k(z))}{1-p(z)}. \tag{2.11}$$

The components $f_{dk}(\cdot \mid x)$ do not depend on z.

Define $\pi_k(z) := \Pr(D = 1, S = k \mid X = x, Z = z) = w_k\, p_k(z)$. This is the joint probability of treatment and regime membership under a decision maker, z. Regime k's share among treated cases is $\omega_{1k}(z) = \pi_k(z)/p(z)$.

3 Wald and IV Estimands

Under the model, the Wald ratio for a pair of decision makers with different overall treatment propensities is a weighted sum of regime-specific average treatment effects. Each regime's weight is its relative contribution to the aggregate first stage. If the decision maker who is more lenient overall is stricter than the other toward regime k, that regime receives a negative weight in the pair's Wald ratio. The Wald ratio is weakly causal with respect to regime-specific average treatment effects if and only if all these weights are nonnegative. The population average treatment effect instead uses population regime shares as weights, and differences between the two sets of weights account for any gap between the estimands. I derive this decomposition below, extend it to the IV estimand that aggregates these pairwise Wald ratios across decision makers, and compare it with the population average treatment effect.

Let

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

denote the mean outcome and treatment propensity under decision maker z, conditional on $X = x$. Write

$$\mu_{0k}(x) := \mathbb{E}[Y(0) \mid X = x, S = k]$$

for the mean potential outcome under no treatment in regime k.

Fix a covariate value x satisfying $\mathbb{E}[|Y(d)| \mid X = x] < \infty$ for $d \in \{0, 1\}$. Under Assumptions 2.1–2.4, cases in regime k assigned to decision maker z have average outcome $\mu_{0k}(x) + p_k(x, z)\tau_k(x)$. Averaging over regimes yields

$$\begin{aligned} m(x, z) &= \sum_{k=1}^{K_0} w_k(x)\mu_{0k}(x) + \sum_{k=1}^{K_0} w_k(x)p_k(x, z)\tau_k(x), \\ p(x, z) &= \sum_{k=1}^{K_0} w_k(x)p_k(x, z). \end{aligned} \tag{3.1}$$

Decision makers face the same distribution of case types and potential outcomes. Therefore, under Assumption 2.4, differences in their average outcomes reflect differences in whom they treat. Each regime's contribution depends on how many more or fewer of its cases receive treatment and the average treatment effect within that regime. The following proposition shows how the Wald ratio weights these treatment effects.

Proposition 3.1 (Pairwise Wald decomposition)**.** *Under Assumptions 2.1–2.4, fix x with $\mathbb{E}[|Y(d)| \mid X = x] < \infty$ for $d \in \{0, 1\}$ and instrument values (z, z') in the conditional support of Z given $X = x$ such that $p(x, z) \neq p(x, z')$. Then*

$$\psi(x; z, z') := \frac{m(x, z) - m(x, z')}{p(x, z) - p(x, z')} = \sum_{k=1}^{K_0} \omega_k^*(x; z, z')\, \tau_k(x), \tag{3.2}$$

where

$$\begin{aligned} \omega_k^*(x; z, z') &:= \frac{\pi_k(x, z) - \pi_k(x, z')}{p(x, z) - p(x, z')} \\ &= \frac{w_k(x)\{p_k(x, z) - p_k(x, z')\}}{\sum_{j=1}^{K_0} w_j(x)\{p_j(x, z) - p_j(x, z')\}}. \end{aligned} \tag{3.3}$$

The weights sum to one. They are all nonnegative if and only if the decision maker who is more lenient overall is at least as lenient toward every regime.

In the running example, judge A is more lenient than B toward regime 1, but stricter toward regime 2. The two regimes have the same population share, so the additional releases in regime 1 are largely offset by fewer releases in regime 2. Only five percentage points separate the judges' overall release rates. The weights are therefore 2.5 and -1.5, and the two positive regime-specific average treatment effects produce the negative Wald ratio in (1.1).[4]

The same reasoning applies to an IV estimand that uses all decision makers. Let $q(z \mid x)$ be the probability that a case with $X = x$ is assigned to decision maker z, or another distribution over decision makers fixed in advance, and let $\zeta(Z)$ be the scalar instrument used in the IV estimand. Write Cov_q for the covariance across decision makers under $q(\cdot \mid x)$. Since $\sum_{k=1}^{K_0} w_k(x)\mu_{0k}(x)$ is constant across decision makers, the result below follows by taking the covariance of each expression in (3.1) with $\zeta(Z)$.

Proposition 3.2 (Decomposition of the IV estimand with a scalar instrument)**.** *Suppose the conditions of Proposition 3.1 hold and*

$$c(x) := \mathrm{Cov}_q\{p(x, Z), \zeta(Z)\} \neq 0.$$

Then

$$\beta^{q,\zeta}(x) := \frac{\mathrm{Cov}_q\{m(x, Z), \zeta(Z)\}}{c(x)} = \sum_{k=1}^{K_0} \Omega_k^{q,\zeta}(x)\tau_k(x), \tag{3.4}$$

[4]With a continuously distributed instrument, if $m(x, z)$ and each $\pi_k(x, z)$ are differentiable, the same argument gives $\partial_z m(x, z)/\partial_z p(x, z) = \sum_k \{\partial_z \pi_k(x, z)/\partial_z p(x, z)\}\tau_k(x)$ wherever $\partial_z p(x, z) \neq 0$. The main analysis maintains finite instrument support.

where

$$\Omega_k^{q,\zeta}(x) = \frac{\mathrm{Cov}_q\{\pi_k(x,Z),\zeta(Z)\}}{c(x)} = \frac{w_k(x)\,\mathrm{Cov}_q\{p_k(x,Z),\zeta(Z)\}}{c(x)}, \qquad \sum_{k=1}^{K_0}\Omega_k^{q,\zeta}(x) = 1.$$

If the treatment rates $p(x,z)$ are distinct across instrument values, or if $\zeta(z) = \zeta(z')$ whenever $p(x,z) = p(x,z')$, then $\beta^{q,\zeta}(x)$ is a weighted sum of the Wald ratios for pairs with different treatment rates. The weights on these pairs sum to one and are proportional to

$$q(z \mid x)q(z' \mid x)\{p(x,z) - p(x,z')\}\{\zeta(z) - \zeta(z')\}.$$

If $\zeta(z) = p(x,z)$, the weights on pairs are nonnegative and proportional to squared first-stage differences.

With $\zeta(z) = p(x,z)$, the IV estimand is a nonnegatively weighted average of pairwise Wald ratios. A regime's IV weight is the same average of its weights in those Wald ratios, and those weights can be negative. The regime's IV weight is negative exactly when decision makers who are more lenient overall tend to be stricter toward cases in that regime:

$$\mathrm{Cov}_q\{p_k(x,Z),p(x,Z)\} < 0.$$

When every regime's IV weight is nonnegative, IV is weakly causal with respect to regime-specific average treatment effects even when individual-level and average monotonicity fail. Section 5.1 gives the formal definition and the comparison with average monotonicity.

Nonnegative IV weights can still give some regimes more weight, and others less, than their population shares. The population average treatment effect conditional on $X = x$ is

$$\mathrm{ATE}(x) := \mathbb{E}[Y(1) - Y(0) \mid X = x] = \sum_{k=1}^{K_0} w_k(x)\tau_k(x).$$

It follows that

$$\beta^{q,\zeta}(x) - \mathrm{ATE}(x) = \sum_{k=1}^{K_0}\{\Omega_k^{q,\zeta}(x) - w_k(x)\}\tau_k(x).$$

With $\zeta(z) = p(x,z)$, $\Omega_k^{q,p}(x) = w_k(x)s_k(x)$, where

$$s_k(x) := \frac{\mathrm{Cov}_q\{p_k(x,Z),p(x,Z)\}}{\mathrm{Var}_q\{p(x,Z)\}}$$

is the slope from regressing regime k's treatment propensity on the overall treatment propensity across decision makers, with an intercept and assignment weights q. A regime's treatment effect receives zero IV weight when every decision maker treats the same fraction of its cases: there is no first stage within the regime. Regimes with $s_k(x) > 1$ receive more than

their population shares. Because $\sum_k w_k(x)s_k(x) = 1$, the difference above is the covariance between $s_k(x)$ and $\tau_k(x)$ across regimes, weighted by their population shares. IV exceeds the population average treatment effect when the regimes whose treatment responds most to leniency have the largest effects. For the pairwise Wald ratio, the relevant slope is $\{p_k(x,z) - p_k(x,z')\}/\{p(x,z) - p(x,z')\}$.

Using the identification arguments in Section 4, I can recover the average treatment effects within each regime and probe the sign of its IV weight.

4 Identification

By Lemma 2.2, for each decision maker, the distribution of outcomes among treated cases is a weighted average of the same regime-specific outcome distributions. The same holds for untreated cases. The identification problem in this section is to recover the *unobserved* regime-specific outcome distributions and their mixture weights from the *observed* outcome distributions. Section 4.1 first shows how we can uniquely pin down the number of regimes using variation in decision maker assignment when the underlying regimes are sufficiently distinct. Next, it shows how we can identify the regime-specific potential outcome distributions and their mixture weights when the distributions belong to some parametric families. Section 4.2 dispenses with such restrictions, but at the price of an additional data requirement: three pretreatment measurements that are conditionally independent given the regimes.

4.1 Identification from Outcome Distributions

Because every observed outcome distribution in a treatment arm is a weighted average of the same K_0 regime-specific outcome distributions, all observed distributions lie in their linear span. Theorem 4.1 shows that if the regime-specific outcome distributions are linearly independent and decision makers mix them in sufficiently different proportions in that arm, the observed outcome distributions span this entire K_0-dimensional space, identifying K_0. The formal conditions are as follows:

Assumption 4.1 (Linear independence)**.** *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).*

Assumption 4.2 (Full-rank variation in mixture weights)**.** *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 $K_0 \times L$ mixture weight matrix*

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

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

These two assumptions rule out different reasons why observed outcome distributions might suggest fewer regimes than actually exist. The linear independence assumption requires that no regime-specific outcome distribution can be written as a linear combination of the others. If, for example, one of three regimes has a potential outcome distribution that is a linear combination of those in the other two regimes, the observed outcome distributions in that treatment arm can span at most two dimensions, regardless of how decision makers select cases across regimes. Theorem 4.1 below adapts the argument of Henry, Kitamura and Salanié (2014) to this setting:

Theorem 4.1 (Identification of the number of regimes and the linear span of component distributions)**.** *Fix x. Under Assumptions 2.2–2.4 and Assumptions 4.1–4.2 for treatment arm d:*

(i) The number K_0 of regimes coincides with 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 the K_0-dimensional linear span of the component distributions.

Part (*ii*) shows why these conditions are not enough to pin down the regime-specific outcome distributions themselves and their weights; different combinations of weights and regime-specific outcome distributions admit observationally equivalent finite mixtures. To see this, define an alternative basis $q_d = A f_d$ and introduce new weights $\tilde{\omega}_d(z) = A^{-\top} \omega_d(z)$. Provided the new components and weights are admissible, they generate an observationally equivalent distribution, g_{dz}.

Gaussian, censored normal, and negative binomial specifications provide one route to identifying the regime-specific outcome distributions. For each treatment status, choose a decision maker under whom every regime has a positive mixture weight. Under the conditions below, the chosen specification identifies the regime-specific outcome distributions from the observed outcomes for that treatment status, up to the labeling of regimes. We can then identify each of the mixture weights for the remaining decision makers from the observed outcome distribution, exploiting the varying-weight mixture representation in Lemma 2.2. Under misspecification, the likelihood targets only an approximation within the chosen family.

Suppose that each regime outcome distribution is Gaussian. An adaptation of standard finite mixture results recovers the component parameters and weights up to a permutation of the regime labels (Teicher, 1963; Yakowitz and Spragins, 1968).

Proposition 4.1 (Identification under Gaussian potential outcome distributions)**.** *Suppose each potential outcome density f_{dk} is Gaussian, so that*

$$g_{dz}(y) = \sum_{k=1}^{K_0} \omega_{dk}(z)\, \phi(y; \mu_{dk}, \sigma_{dk}^2),$$

where the weights $\omega_{dk}(z) \geq 0$ sum to one. Assume the parameter pairs $\{(\mu_{dk}, \sigma_{dk}^2)\}_{k=1}^{K_0}$ are distinct and that all weights are positive at some instrument value $\bar{z}$. Then the Gaussian parameters and the weights $\{\omega_{dk}(z)\}_{k=1}^{K_0}$ are identified up to a common permutation of the regime labels.

If both arms satisfy these conditions and there is a unique permutation σ of the labels for the distributions under no treatment making

$$z \mapsto p(z)\omega_{1k}(z) + \{1 - p(z)\}\omega_{0,\sigma(k)}(z)$$

constant in z for every k, then w_k, $\pi_k(z) = p(z)\omega_{1k}(z)$, and $p_k(z) = \pi_k(z)/w_k$ are identified.

The first part of Proposition 4.1 identifies the outcome distributions and mixture weights separately in each treatment arm up to label permutation. Recovering regime-specific average treatment effects requires matching the treated and untreated components that belong to the same regime. The second part of the proposition therefore requires a *unique cross-arm pairing* under which, for each pair, the shares of the whole population in the treated and untreated components sum to a constant regime share across decision makers, a condition I characterize below.

A regime's outcome distribution may itself be a mixture of Gaussian distributions, called subcomponents, that share the regime's treatment propensities under each decision maker. Under the conditions of Theorem C.1 in Appendix C, variation across decision makers identifies which subcomponents belong to the same regime. Additional Gaussian components can therefore better capture the regime's outcome distributions without incorrectly introducing additional regimes.

Corollary 4.1 (Censored normal mixtures)**.** *Under Assumptions 2.1–2.4, fix x and known finite limits $\ell < u$, common across regimes and instrument values. Let $C_{\ell,u}(t) := \min\{u, \max\{\ell, t\}\}$. Suppose, for treatment status d, that F_{dk} is the distribution of $C_{\ell,u}(T_{dk})$, where $T_{dk} \sim N(\nu_{dk}, \sigma_{dk}^2)$ and $\sigma_{dk} > 0$. If the parameter pairs are distinct and $\omega_{dk}(\bar{z}) > 0$ for every k at some instrument value $\bar{z}$, the observed outcome distributions conditional on $D = d$, $X = x$, and $Z = z$ identify K_0, the normal parameters, and the weights $\omega_{dk}(z)$ at every z with $\Pr(D = d \mid X = x, Z = z) > 0$, up to one common permutation.*

For regimes containing several normal subcomponents, the conclusions of Theorem C.1 continue to hold after applying $C_{\ell,u}$, under the conditions spelled out for that theorem. In either

case, if the regime-specific potential outcome distributions and mixture weights are identified separately in both treatment arms and satisfy the unique pairing condition in Proposition 4.1, the regime shares, treatment propensities, and treatment effects on the censored outcome are identified. The corollary establishes identification within the censored-normal family used to approximate sentence-length distributions in Section 8. The identification argument uses the continuous density on (ℓ, u), between the lower censoring limit ℓ and the upper censoring limit u. Within each regime, the treatment effect is the difference between mean censored outcomes under treatment and no treatment, not the difference between the underlying normal means. The proof can be found in Appendix A.

A negative binomial specification provides an analogous identification argument for count outcomes. Write its mass function as

$$b(y; r, q) = \frac{\Gamma(y + r)}{\Gamma(r)\Gamma(y + 1)}(1 - q)^r q^y, \qquad y \in \mathbb{N}_0, \quad r > 0, \quad q \in (0, 1). \tag{4.1}$$

The mean dispersion parameterization is $r = 1/\alpha$ and $q = \mu/(r + \mu)$, where $\mu > 0$ and $\alpha > 0$.

Theorem 4.2 (Identification with negative binomial outcome distributions)**.** *Fix a treatment status d and suppose*

$$g_{dz}(y) = \sum_{k=1}^{K_0} \omega_{dk}(z) b(y; r_{dk}, q_{dk}).$$

Assume the pairs $\{(r_{dk}, q_{dk})\}_{k=1}^{K_0}$ are distinct and there is an instrument value $\bar{z}$ at which $\omega_{dk}(\bar{z}) > 0$ for every k. Then K_0, the component parameters, and the weights $\{\omega_{dk}(z)\}_{k,z}$ are identified up to one permutation. If both arms satisfy these conditions, then under Assumption 2.2 and the unique pairing condition in Proposition 4.1, w_k, $\pi_k(z)$, and $p_k(z)$ are also identified.

Matching potential outcome distributions to the same regime. I return to the unique cross-arm pairing condition introduced in Proposition 4.1. This condition can be characterized using changes in the identified joint probabilities of regime membership and treatment status across decision makers. The characterization applies to both the Gaussian and negative binomial specifications. For two components paired under the label k, their shares of the entire population must sum to the regime's population share:

$$w_k(x) = p(x, z)\omega_{1k}(x, z) + \{1 - p(x, z)\}\omega_{0k}(x, z). \tag{4.2}$$

Under Assumption 2.2, this sum must be constant across decision makers. There is, however, nothing inherent to this condition that rules out more than one pairing satisfying it.

Assignment can change how many cases in regime k receive treatment while leaving its population share fixed. A valid pairing must therefore make the treated and untreated

components' shares of the full population sum to that regime share under every assignment. Taking z_0 as a reference assignment, this requires:

$$\begin{aligned}\pi_k(x,z) - \pi_k(x,z_0) = &\{1 - p(x,z_0)\}\omega_{0k}(x,z_0) \\ &- \{1 - p(x,z)\}\omega_{0k}(x,z).\end{aligned}$$

Provided a valid pairing exists, it is unique if and only if, for every $k \neq \ell$, the identified probabilities $\pi_k(x,z)$ and $\pi_\ell(x,z)$ have different changes from the reference assignment z_0 for at least one z:

$$\pi_k(x,z) - \pi_k(x,z_0) \neq \pi_\ell(x,z) - \pi_\ell(x,z_0).$$

If instead these changes are equal for every z for some $k \neq \ell$, exchanging the potential outcome distributions under no treatment assigned to regimes k and ℓ, together with their mixture weights, leaves both sums in (4.2) constant across decision makers. Assumption 2.2 cannot therefore tell the two pairings apart. Proposition B.2 in Appendix B.2 provides this characterization. Once the pairing is identified, the constant sum in (4.2) identifies each regime's population share, and its treatment propensity follows from

$$p_k(x,z) = \frac{p(x,z)\omega_{1k}(x,z)}{w_k(x)}. \tag{4.3}$$

The auxiliary measurement conditions in Section 4.2 identify the regimes before recovering their outcome distributions. Both treatment arms therefore use the same regime labels, without recourse to a pairing condition. Under either approach, identifying both potential outcome distributions, F_{1k} and F_{0k}, for regime k identifies its average treatment effect. Once the regime-specific outcome distributions are pinned down, we can also recover other features of the potential outcome distributions, such as regime-specific quantile treatment effects.

Recovering regime-specific average treatment effects does not always require identification of the full potential outcome distributions and their mixture weights for both treatment statuses. Provided we can identify the joint probabilities $\pi_k(x,z)$, we only need sufficiently rich variation in these probabilities across decision makers to identify these effects from differences in observed average outcomes across decision makers. Proposition B.3 in Appendix B.3 requires full row rank in these probability changes. For this result, the distributional equality in Assumption 2.4 can be weakened to equality of means. Under the full distributional restriction in Assumption 2.4, the same rank condition also identifies F_{0k} without specifying a parametric family for these distributions. This alternative argument, however, does not generally identify the population regime shares or the population average treatment effect.

4.2 Identification Using Pretreatment Measurements

To avoid parametric functional-form restrictions on the outcome distributions, I show how to identify the model nonparametrically using three pretreatment measurements—such as prior outcomes or information from administrative records—that proxy for the latent regimes. Write these measurements as $R = (R^{(1)}, R^{(2)}, R^{(3)})$, each possibly vector-valued, with regime-specific distributions

$$F_k^{(r)}(\cdot) = \Pr\Big(R^{(r)} \in \cdot \mid S = k, X = x\Big), \qquad r = 1, 2, 3.$$

Under the conditions below, identification proceeds in three steps. First, the joint distribution of the measurements identifies the number of regimes, their population shares, and each regime's measurement distributions. Second, a measurement whose distribution within each regime does not depend on treatment status identifies each regime's proportions among treated and untreated cases under each decision maker, and hence the regime-specific treatment propensities. Third, sufficient variation in these proportions across decision makers identifies the regime-specific potential outcome distributions. Notice that none of these steps requires a parametric family for the outcome distributions.

Step 1: The number of regimes, population shares, and measurement distributions.

Assumption 4.3 (Local independence)**.** *Conditional on (X, S), the measurements $R^{(1)}$, $R^{(2)}$, $R^{(3)}$ are mutually independent.*

This assumption allows the measurements to be related but, analogous to standard arguments in finite mixtures with multiple measurements, it restricts the dependence to occur only via the latent regimes, S, and observed characteristics, X.

Theorem 4.3 (Identification of the number of regimes, their shares and measurement distributions)**.** *Let Assumptions 2.1 and 4.3 hold. After reordering the measurements, suppose the distributions $\{F_k^{(r)}\}_{k=1}^{K_0}$ are linearly independent for $r = 1, 2$ and pairwise distinct for $r = 3$ (Assumption D.1). Then the distribution of R given $X = x$ identifies K_0 as the smallest number of components with strictly positive weights in a mixture of product distributions. Among representations with K_0 components, the regime shares $w_k(x)$ and the regime-specific measurement distributions $\{F_k^{(r)}\}$ are identified, up to one common permutation of the regime labels.*

The theorem uses only the measurements; neither outcomes nor sufficient variation in the instrument is needed.[5] The uniqueness argument builds on Kruskal (1977) and Allman, Matias and Rhodes (2009). For the constructive proof, I adapt the matrix diagonalization argument of Bonhomme, Jochmans and Robin (2016) to the joint distribution of the pretreatment

[5] Replacing a component by identical copies whose weights sum to its original weight leaves the distribution of R unchanged. The theorem therefore identifies the smallest number of components.

measurements. The matrix diagonalization below is closely related to an argument of Hu (2008) and is a finite-dimensional counterpart of the operator decomposition in Hu and Schennach (2008).

Fix x. All expectations below are conditional on $X = x$, which I suppress from the notation. For finite vectors $a(\cdot)$ and $b(\cdot)$ of bounded measurable functions of the first two measurements, Assumption 4.3 gives

$$\mathbb{E}\Big[a(R^{(1)})b(R^{(2)})'\Big] = \sum_{k=1}^{K_0} w_k \, \mathbb{E}[a(R^{(1)}) \mid S = k] \, \mathbb{E}[b(R^{(2)}) \mid S = k]'.$$

Each summand has rank at most one, so the matrix has rank at most K_0. Similarly, the rank of this matrix cannot exceed the number of product components in any representation of the distribution of R. Because the regime shares are positive and the regime-specific distributions of each of the first two measurements are linearly independent, some choice of K_0 functions in each vector attains rank K_0 (Lemma D.1). The largest attainable rank therefore identifies K_0, and no representation has fewer components. Without these linear independence conditions, the rank provides a lower bound on the number of components; see Kasahara and Shimotsu (2014).

Choose K_0 functions of each of the first two measurements and write

$$a = (a_1, \ldots, a_{K_0})', \qquad b = (b_1, \ldots, b_{K_0})'.$$

Lemma D.1 permits the first function in each vector to be constant one, $a_1 \equiv b_1 \equiv 1$, while choosing the remaining functions so that the matrix of cross-moments has rank K_0.

For a bounded measurable function c of the third measurement, Assumption 4.3 delivers

$$\begin{aligned} M(c) &:= \mathbb{E}\Big[a(R^{(1)})b(R^{(2)})'c(R^{(3)})\Big] \\ &= A \, \operatorname{diag}\{w_k \bar{c}_k\}_{k=1}^{K_0} \, B', \end{aligned}$$

where

$$A_{jk} = \mathbb{E}[a_j(R^{(1)}) \mid S = k], \qquad B_{jk} = \mathbb{E}[b_j(R^{(2)}) \mid S = k],$$

and $\bar{c}_k = \mathbb{E}[c(R^{(3)}) \mid S = k]$. The chosen functions make A and B nonsingular. Since the regime shares are positive, $M(1)$ is invertible, so

$$M(c)M(1)^{-1} = A \, \operatorname{diag}\{\bar{c}_k\}_{k=1}^{K_0} \, A^{-1}.$$

The observable matrix on the left hand side has eigenvalues $\bar{c}_k$, with the columns of A as eigenvectors. Distinct eigenvalues identify these columns up to scale and order. Pairwise distinct distributions of the third measurement guarantee a bounded function c whose conditional expectation is unique to each regime (Lemma D.2). Since $a_1 \equiv 1$, the first entry

of every column of A equals one, fixing its scale. The order amounts to an arbitrary labeling of the regimes. With A identified, the shares solve

$$\mathbb{E}[a(R^{(1)})] = Aw,$$

and $M(1) = A\,\mathrm{diag}(w)B'$ then identifies B. If we use indicators of measurable sets as the bounded functions, we identify each regime's probabilities of those sets by linear inversion, and hence the distributions of its measurements. The proof is in Appendix D.

The third measurement needs only to provide distinct eigenvalues, whereas the first two must have linearly independent regime-specific distributions. A measurement with fewer than K_0 possible values cannot have K_0 linearly independent distributions, but it can have pairwise distinct ones. When $K_0 = 3$, for example, a binary third measurement suffices if its success probabilities are pairwise distinct across regimes, whereas each of the first two measurements must have at least three possible values and linearly independent regime-specific distributions.

Step 2: Mixture weights and treatment propensities. Theorem 4.3 in Step 1 provided conditions for identifying population regime shares and the regime-specific measurement distributions. I now turn to regime-specific treatment propensities, for which we also need the mixture weights of Lemma 2.2: each regime's proportion among cases receiving treatment status d under decision maker z,

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

We further invoke the following assumptions:

Assumption 4.4 (Conditional independence of measurements and assignment)**.**

$$R \perp Z \mid (X, S).$$

Assumption 4.5 (Conditional independence of a measurement and treatment)**.** *There is a measurement $Q := R^{(r^*)}$, designated ex ante, such that*

$$Q \perp D \mid (Z, X, S).$$

These assumptions jointly imply $Q \perp (D, Z) \mid (X, S)$: within each regime and conditional on observed covariates, Q has the same distribution among treated cases, untreated cases, and under every decision maker. Random assignment satisfying $Z \perp (S, R) \mid X$ justifies Assumption 4.4. Assumption 4.5, however, imposes an additional requirement: conditioning on treatment preserves the distribution of Q. For example, suppose disability examiners observe an applicant's average earnings before application, while the researcher has the annual earnings records. Include the average in X and use the annual records as Q. Assumption 4.5

holds if, conditional on (Z, X, S), the records are independent of the additional information the examiner uses to award benefits. The assumption can fail if the examiner uses medical evidence correlated with past earnings even after conditioning on (Z, X, S).

Proposition 4.2 (Identification of regime-specific treatment propensities and potential outcome distributions)**.** *Fix x. Suppose the conditions of Theorem 4.3 and Assumptions 2.2, 4.4, and 4.5 hold, and the regime-specific distributions $F_1^Q, \dots, F_{K_0}^Q$ are linearly independent.*

(i) *For every treatment status and assignment with positive conditional probability, the mixture weights $\omega_{dk}(x, z)$ are identified. The joint probabilities $\pi_k(x, z)$ and regime-specific treatment propensities $p_k(x, z)$ are also identified, with regime labels common across treatment arms and decision maker assignments.*

(ii) *Suppose further that consistency and Assumptions 2.3 and 2.4 hold. For a treatment status d, choose instrument values $z_1, \dots, z_L$ such that*

$$\Pr(D = d \mid Z = z_\ell, X = x) > 0, \qquad \ell = 1, \dots, L,$$

and define the $K_0 \times L$ matrix

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

If $\operatorname{rank}\{\Omega_d(x)\} = K_0$, the regime-specific potential outcome distributions $F_{dk}(\cdot \mid x)$ are identified. If this condition holds for both treatment arms, the two potential outcome distributions within each regime are identified under the same labeling. Their means, when finite, identify the regime-specific average treatment effects.

The first part uses the measurement distributions identified in Step 1. Suppress x, and choose bounded functions $a_1, \dots, a_{K_0}$ such that the matrix

$$A_{jk}^Q = \mathbb{E}[a_j(Q) \mid S = k, X = x]$$

is nonsingular. Linear independence of distributions F_k^Q ensures such functions exist. For each observed treatment arm and assignment, define

$$m_d^Q(z) = \big(\mathbb{E}[a_j(Q) \mid D = d, Z = z, X = x]\big)_{j=1}^{K_0}.$$

Since conditioning on treatment and assignment leaves each regime's distribution of Q unchanged,

$$m_d^Q(z) = A^Q \omega_d(z), \qquad \omega_d(z) = (A^Q)^{-1} m_d^Q(z).$$

Both A^Q and $m_d^Q(z)$ are identified, so this equation recovers the unknown mixture weights. For assignments with $p(z) > 0$,

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

The second equality uses Assumption 2.2. When $p(z) = 0$, both probabilities are zero.

This step requires neither outcome data nor variation in treatment propensities across decision makers. Note that this step also identifies the IV weights for pairs of decision makers with a nonzero first stage.

Because the argument uses only the conditional means of finitely many functions of Q, Assumption 4.5 can be weakened to require that treated and untreated cases have the same means of these functions within each regime and covariate cell under every decision maker (Appendix D.4).

Step 3: Regime-specific potential outcome distributions. By consistency, instrument exogeneity, and Assumption 2.4, the outcome distribution of cases with treatment status d under decision maker z is a weighted average of the regime-specific distributions of $Y(d)$, with the weights identified in Step 2:

$$\Pr(Y \leq y \mid D = d, Z = z, X = x) = \sum_{k=1}^{K_0} \omega_{dk}(x, z) F_{dk}(y \mid x).$$

Thus, after Step 2, the potential outcome distributions are the remaining unknowns in these equations.

For the instrument values in part (ii) of Proposition 4.2, collect the observed conditional outcome CDFs in the vector $\mathbf{G}_d(y)$, and the regime-specific CDFs of $Y(d)$ in $\mathbf{F}_d(y)$. Then

$$\mathbf{G}_d(y) = \Omega_d' \mathbf{F}_d(y).$$

When the mixture weight matrix, Ω_d, is full row rank, we can back out the regime-specific potential outcome distributions:

$$\mathbf{F}_d(y) = (\Omega_d \Omega_d')^{-1} \Omega_d \mathbf{G}_d(y).$$

Proofs of both parts of the proposition are relegated to Appendix D.

Full row rank requires sufficiently rich variation in treatment decisions across decision makers: for each treatment status, no regime's vector of mixture weights can be written as a linear combination of the vectors for the other regimes. This requires at least K_0 instrument values at which that treatment status has positive conditional probability. The argument

requires neither linear independence nor a parametric specification of the potential outcome distributions; these distributions may coincide across regimes.

Because the regime labels in both treatment arms are from Step 1, F_{1k} and F_{0k} from this step belong to the same regime. Therefore, we do not need the unique pairing condition. The difference between their average outcomes is regime k's average treatment effect, and weighting these effects by the regime shares from Step 1 yields the average treatment effect at $X = x$.

To use the nonparametric route in an application, a researcher needs measurements for the full sample; measurements that vary within covariate cells; and arguments for why local independence and Assumption 4.5 hold in the application. Random assignment of decision makers justifies neither assumption. Appendix D discusses practical considerations when choosing measurements and handling missing observations.

IV weights and population shares. The joint probabilities $\pi_k(x, z) = w_k(x)p_k(x, z)$ identify the IV weights without requiring separate identification of the regime shares and treatment propensities. Once the regime-specific average treatment effects are identified, recovering the population average treatment effect generally also requires the shares. For example, when the outcome is defined to be zero without treatment, Proposition B.1 shows that the shares satisfy

$$w_k(x) \geq \max_z \pi_k(x, z), \qquad \sum_k w_k(x) = 1.$$

Each regime must be at least as large as the largest group of its cases treated by any decision maker. Any population mass left after satisfying these lower bounds can be allocated across regimes, potentially changing the population average treatment effect. The shares can be recovered from the regime-specific outcome distributions and mixture weights identified for both treatment statuses, provided the pairing is unique, or from pretreatment measurements under the conditions in Section 4.2. Appendix F extends the comparison between IV and the population average treatment effect to averages over covariate values.

5 Treatment Propensities and IV Weights

Nonnegative weights in every pairwise Wald ratio make each ratio weakly causal with respect to regime-specific average treatment effects, while constant weights make the Wald ratios for different pairs average those effects in the same proportions. These properties require different restrictions on treatment propensities: common direction and Rank-1 proportionality. I first define weak causality and relate it to individual-level and average monotonicity. Then I characterize these two restrictions.

5.1 Weak Causality and Monotonicity

The weak causality criterion of Blandhol et al. (2026) pertains to whether an estimand can have the wrong sign when the average treatment effects in all groups have the same sign. Their groups are defined by potential treatment choices. The following definition adapts this criterion to average treatment effects within regimes.

Definition 5.1 (Weak causality with respect to regime-specific average treatment effects)**.** Fix x, the regime shares, and the regime-specific treatment propensities. An estimand is *weakly causal with respect to regime-specific average treatment effects* if, for every vector of regime-specific potential outcome means, it is nonnegative whenever $\tau_k(x) \geq 0$ for every k and nonpositive whenever $\tau_k(x) \leq 0$ for every k.

With x, the regime shares and treatment propensities fixed, the Wald and IV estimands are linear in the regime-specific potential outcome means. For such estimands, Definition 5.1 requires both *level independence* and nonnegative treatment effect weights, paralleling the characterization in Blandhol et al. (2026). Level independence means that shifting both potential outcome means in a regime by the same amount leaves the estimand unchanged. To see why Definition 5.1 requires this property, suppose treatment has no average effect in any regime, so $\tau_k(x) = 0$ for every k. The estimand must then be both nonnegative and nonpositive, and hence zero, whatever the values of $\mu_{0k}(x)$. When a linear estimand is written in terms of $\mu_{0k}(x)$ and $\tau_k(x)$, the coefficients on $\mu_{0k}(x)$ must therefore vanish, leaving dependence only on the treatment effects.

Under the model, level independence holds because decision maker assignment leaves the regime composition unchanged. Average potential outcomes under no treatment therefore add the same amount to every decision maker's average outcome, so differences in these averages across regimes do not contaminate IV. The remaining requirement is that the weights on regime-specific average treatment effects be nonnegative.

Corollary 5.1 (IV weights and weak causality)**.** *Under Assumptions 2.1–2.4, fix a pair (z, z') with $p(x, z) \neq p(x, z')$. The conditional Wald estimand $\psi(x; z, z')$ has the following properties:*

(i) (Level independence.) *It depends on the regime-specific potential outcome means only through their differences $\tau_k(x)$.*

(ii) (Sign preservation.) *It is weakly causal with respect to regime-specific average treatment effects if and only if*

$$\omega_k^*(x; z, z') \geq 0 \quad \text{for every } k.$$

Suppose $\omega_k^*(x; z, z') < 0$ for some k. If the regime-specific average treatment effects are otherwise unrestricted, choosing $\tau_k(x) > 0$ and $\tau_j(x) = 0$ for $j \neq k$ gives $\psi(x; z, z') < 0$.

The same argument applies to the IV estimand in Proposition 3.2. It is weakly causal with respect to regime-specific average treatment effects if and only if $\Omega_k^{q,\zeta}(x) \geq 0$ for every

k. Whether a negative weight reverses the sign depends on the treatment effects. If average treatment effects are equal across regimes, the IV estimand equals that common effect because its weights sum to one, even if some are negative.

To compare Definition 5.1 with the criterion of Blandhol et al. (2026), consider a pair with $p(x,z) > p(x,z')$. If the mean potential outcomes of compliers, defiers, always-takers, and never-takers can take any values, the pairwise Wald ratio satisfies the criterion of Blandhol et al. (2026) if and only if there are no defiers:

$$\Pr\{D(z) < D(z') \mid X = x\} = 0.$$

This condition implies nonnegative weights on regime-specific average treatment effects. Definition 5.1 allows defiers within a regime, provided no regime's treatment propensity falls. The two criteria coincide for this Wald estimand when $D(z)$ and $D(z')$ are deterministic conditional on (X, S). Test A in Section 6.4 tests the sign restrictions for conditional pairwise Wald ratios. For residualized 2SLS, the original criterion also involves the *rich covariate* conditions of Blandhol et al. (2026).

Average monotonicity provides a sufficient condition for nonnegative weights when the overall treatment propensity is used as the instrument. Take $q(z \mid x) = \Pr(Z = z \mid X = x)$, set $\zeta(z) = p(x,z)$, and suppose $\mathrm{Var}_q\{p(x,Z)\} > 0$. For case i, define

$$a_i(x) := \mathrm{Cov}_q\{D_i(Z), p(x,Z)\}.$$

The covariance is taken across decision makers for a fixed case. Under the assignment assumptions, averaging within regime k gives

$$\mathbb{E}[a_i(x) \mid X = x, S_i = k] = \mathrm{Cov}_q\{p_k(x,Z), p(x,Z)\}.$$

Consequently,

$$\Omega_k^{q,p}(x) = \frac{w_k(x)\mathbb{E}[a_i(x) \mid X = x, S_i = k]}{\mathrm{Var}_q\{p(x,Z)\}}. \tag{5.1}$$

This identity uses no restriction on potential outcomes. The conditional version of average monotonicity in Frandsen, Lefgren and Leslie (2023) requires $a_i(x) \geq 0$ almost surely. Always-takers and never-takers satisfy this condition trivially. For any other case, the decision makers who would treat it must be at least as lenient on average as those who would not, weighting by assignment probabilities. Every regime then receives a nonnegative IV weight. The converse fails: nonnegative regime-specific IV weights allow individual cases to violate this condition when other cases in the same regime offset their negative contributions.

Average monotonicity therefore has a testable implication: every regime's weight in this IV estimand is nonnegative, for the assignment probabilities q and the instrument $p(x,Z)$. Because the weights are identified under the model, a significantly negative estimate is evidence

against average monotonicity or against the model's identifying assumptions. Negative weights for individual pairs of decision makers are compatible with average monotonicity, which allows reversals between some pairs.

5.2 Common Direction and Negative Weights

I measure negative weighting by summing the absolute values of the negative weights on regime-specific average treatment effects. The sum, denoted $\text{NegMass}(x; z, z')$ below, is zero exactly when all weights for that pair are nonnegative.

Definition 5.2 (Negative weight mass)**.** For a given $(x; z, z')$ with $p(x, z) \neq p(x, z')$, define

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

Let $\Delta_k(x; z, z') := p_k(x, z) - p_k(x, z')$ and $\Delta(x; z, z') := \sum_j w_j(x)\Delta_j(x; z, z')$. For an instrument pair with nonzero aggregate first stage, nonnegative IV weights require the *common direction* condition:

$$\Delta_k(x; z, z')\Delta(x; z, z') \geq 0 \quad \text{for all } k \text{ with } w_k(x) > 0.$$

Because $\omega_k^*(x; z, z') = w_k(x)\Delta_k(x; z, z')/\Delta(x; z, z')$, common direction is equivalent to nonnegative weights for every regime with $w_k(x) > 0$, and hence to $\text{NegMass}(x; z, z') = 0$. Conditional on $X = x$, individual-level monotonicity excludes defiers, whereas common direction requires only that compliers be at least as prevalent as defiers within every regime. Common direction therefore suffices for the Wald ratio to be weakly causal with respect to regime-specific average treatment effects, without individual-level monotonicity or Rank-1.

With fixed IV weights, overstating effects in positively weighted regimes and understating them in negatively weighted regimes both raise the weighted average. If each error is at most δ in absolute value, the error in the weighted average is bounded by $\delta\{1 + 2\text{NegMass}(x; z, z')\}$ (Proposition G.2). The joint probabilities $\pi_k(x, z)$ identify NegMass, allowing this bound to be evaluated under the identification conditions in Section 4. If only the linear span of the regime-specific outcome distributions is identified, different mixtures consistent with the observed distributions can imply different values of NegMass.

5.3 The Rank-1 Benchmark

When the overall treatment propensity varies across decision makers, each regime receives the same weight in every pairwise Wald ratio with a nonzero first stage only if differences in treat-

ment propensities occur in fixed proportions across regimes. I represent this proportionality by the following affine specification:

$$p_k(x,z) = a_k(x) + b_k(x)\lambda(z), \tag{5.2}$$

for some scalar $\lambda(z)$ and regime-specific slopes $b_k(x)$. Each regime makes the same relative contribution to the first stage for every pair of decision makers with a nonzero first stage. Proposition 5.1 expresses this restriction in terms of the centered matrix of regime-specific treatment propensities and the weights for the pairwise Wald ratio.

Proposition 5.1 (Characterization of Rank-1 proportionality)**.** *Fix x and instrument values $z_1, \ldots, z_L$. Let*

$$P_x := \big(p_k(x,z_\ell)\big)_{k,\ell}, \qquad C_L := I_L - L^{-1}\mathbf{1}_L\mathbf{1}_L',$$

and suppose $p(x,z)$ is not constant on this support. The following statements are equivalent:

(i) *The regime-specific treatment propensities satisfy (5.2) for some $a_k(x)$, $b_k(x)$, and $\lambda(z)$.*

(ii) $\operatorname{rank}(P_x C_L) = 1$.

(iii) *There are constants $\alpha_k(x)$ and $\beta_k(x)$ such that*

$$p_k(x,z) = \alpha_k(x) + \beta_k(x)p(x,z) \qquad \text{for every } k, z.$$

(iv) *There are constants $\bar{\omega}_k(x)$ such that $\omega_k^*(x;z,z') = \bar{\omega}_k(x)$ for every pair with $p(x,z) \neq p(x,z')$, and $p_k(x,z) = p_k(x,z')$ for every k whenever $p(x,z) = p(x,z')$.*

Under these equivalent conditions,

$$\bar{\omega}_k(x) = w_k(x)\beta_k(x) \qquad \text{and} \qquad \psi(x;z,z') = \sum_{k=1}^{K_0} \bar{\omega}_k(x)\tau_k(x).$$

Thus the Wald ratio is constant across instrument pairs with a nonzero aggregate first stage. If the $\beta_k(x)$ are nonnegative, then $\operatorname{NegMass}(x;z,z') = 0$ *for every pair.*

In Table 1, judges B and C release half of all defendants overall but 55 and 30 percent of type 1 defendants, respectively. They violate Rank-1, which requires equal treatment propensities within every regime whenever overall treatment propensities are equal.

Unequal pairwise Wald ratios also rule out Rank-1, but equal ratios do not establish it. If all regimes have the same average treatment effect, every pairwise Wald ratio with a nonzero first stage equals that effect regardless of the weights.

Even with many decision makers, Rank-1 allows the observed outcome distributions within either treatment status to span at most two dimensions. With $a = (a_k(x))_k$, $b = (b_k(x))_k$, and $\lambda = (\lambda(z_\ell))_\ell$, the matrix of joint probabilities of treatment and regime membership satisfies

$$\Pi_x = \text{diag}(w)\{a\mathbf{1}_L' + b\lambda'\}, \qquad \text{rank}(\Pi_x) \leq 2.$$

The matrix of joint probabilities of no treatment and regime membership has the same form with $\mathbf{1}_{K_0} - a$ and $-b$. Normalizing the columns by positive probabilities of the relevant treatment status preserves rank. When $K_0 > 2$, the linear span of the observed outcome distributions therefore cannot identify the number of regimes under this restriction. The tests below instead use identification of the joint probabilities $\pi_k(x, z)$ through restrictions on the component distributions or auxiliary information.

Remark 5.1 (Monotonicity versus Rank-1 structure)**.** When individual-level monotonicity holds, each regime's Wald weight is its share of the pair's compliers. Rank-1 requires these shares to be the same across pairs. In the latent index model, compliers have resistance to treatment between the two decision makers' thresholds, and different threshold intervals can contain different proportions of each regime. Monotonicity therefore allows the Wald weights to vary across pairs.

Under Assumption 2.4, common direction and Rank-1 also restrict the joint distribution of treatment and observed outcomes. In Appendix E.1, I provide testable implications of these restrictions on the joint distribution, which can be analyzed without first identifying the regime-specific potential outcome distributions.

Under conditional random assignment, grouping decision makers preserves the affine specification in (5.2). A rejection of Rank-1 for grouped assignments therefore also rejects Rank-1 for individual decision makers. The same implication holds for common direction when the groups are ordered intervals of the true conditional treatment propensity (Proposition E.2 in Appendix E). Estimated leniency measures may fail to preserve that ordering, and failure to reject either restriction can reflect averaging over violations within groups. Grouping can also reduce the rank of the mixture weight matrix and hence limit identification of the number of regimes from the span of observed outcome distributions.

6 Estimation and Inference

This section furnishes the estimation and inference procedures for the model proposed above. I use a maximum likelihood procedure for estimating the parametric models. To compute the estimator, I adapt standard expectation-maximization algorithms. From the fitted model, I estimate the regime-specific average treatment effects and IV weights and test common direction and Rank-1. For nonparametric estimation of the model with pretreatment

measurements, I propose a plug-in estimator based on the constructive identification arguments in Section 4.2. The asymptotic theory for this route is deferred to future work. Both the parametric and nonparametric estimation routes hold the number of regimes fixed.

6.1 Parametric Likelihood Estimation

Hold the number of regimes fixed at K and let θ collect the parameters from the shares, $w_k(x)$; treatment propensities, $p_k(x,z)$; and the potential outcome densities, $f_{dk}(y \mid x)$ for $d \in \{0,1\}$. The joint density of treatment and the observed outcome is:

$$h(y,d \mid x,z;\theta) = \sum_{k=1}^{K} w_k(x;\theta) p_k(x,z;\theta)^d \{1 - p_k(x,z;\theta)\}^{1-d} f_{dk}(y \mid x;\theta). \tag{6.1}$$

Given observations $\{(Y_i, D_i, X_i, Z_i)\}_{i=1}^n$, the maximum likelihood estimator maximizes $\sum_{i=1}^n \log h(Y_i, D_i \mid X_i, Z_i; \theta)$. One benefit of this likelihood procedure is that the components across treatment arms are paired, since each regime has the same population share under every decision maker.[6] The pairing is identified under the unique-pairing condition in Section 4.1.

The Gaussian, censored normal, and negative binomial specifications in Section 4 admit finite-dimensional likelihood estimation, even when there are multiple components within a regime. To compute the accompanying estimators, I use variations of the standard expectation-maximization (EM) algorithm. In general, the algorithm alternates between computing posterior regime probabilities given the data, and updating the parameters, θ. In Theorem H.1, I show that the maximum likelihood estimator is consistent given sufficient regularity, up to relabeling of the regimes and any within-regime outcome components, if the model is correctly specified and there is a unique pairing of labels across treatment arms. For selection of K, I propose using held-out likelihood to compare candidate values, as in Section 7. The inference results below take K as fixed. Appendix H contains the proof and computational details.

6.2 Estimation Using Pretreatment Measurements

For the pretreatment measurement model in Section 4.2, fix the number of regimes at K and hold x fixed. Choose finite vectors of bounded test functions $\varphi^{(r)}$, with first coordinate equal to one, that preserve the rank and distinctness conditions in Section 4.2. Write

[6] If $Y(0) \equiv 0$, I can write the likelihood in terms of the joint probabilities $\pi_k(x,z) = w_k(x)p_k(x,z)$. Untreated observations still contribute through their treatment status. Appendix H.2 develops this special case and the restrictions on π_k required for constant regime shares.

$A_{jk}^{(r)} = \mathbb{E}[\varphi_j^{(r)}(R^{(r)}) \mid S = k, X = x]$. Estimate the three-way cross-moments and fit their rank-K representation:

$$\widehat{T}_{j_1 j_2 j_3} = \widehat{\mathbb{E}}\left[\prod_{r=1}^{3} \varphi_{j_r}^{(r)}(R^{(r)}) \mid X = x\right] \approx \sum_{k=1}^{K} \widehat{w}_k \prod_{r=1}^{3} \widehat{A}_{j_r k}^{(r)}.$$

Here $\widehat{\mathbb{E}}$ denotes an estimated conditional expectation. Normalize $\widehat{A}_{1k}^{(r)} = 1$ and constrain $\widehat{w}$ to the simplex $\Delta_K = \{v \geq 0 : \mathbf{1}'v = 1\}$.

For a measurement $Q = R^{(r^*)}$ satisfying Assumption 4.5, choose its test functions so that $A^Q = A^{(r^*)}$ has full column rank, and set $\widehat{A}^Q = \widehat{A}^{(r^*)}$. Constrained least squares estimates the mixture weights in each treatment arm:

$$\begin{aligned} \widehat{m}_d^Q(z) &= \widehat{\mathbb{E}}[\varphi^{(r^*)}(Q) \mid D = d, Z = z, X = x], \\ \widehat{\omega}_d(z) &= \arg\min_{v \in \Delta_K} \|\widehat{m}_d^Q(z) - \widehat{A}^Q v\|^2. \end{aligned}$$

With $\widehat{p}(z) \in [0, 1]$ estimating the overall treatment probability, calculate

$$\begin{aligned} \widehat{w}_k(z) &= \{1 - \widehat{p}(z)\}\widehat{\omega}_{0k}(z) + \widehat{p}(z)\widehat{\omega}_{1k}(z), \\ \widehat{\pi}_k(z) &= \widehat{p}(z)\widehat{\omega}_{1k}(z), \qquad \widehat{p}_k(z) = \frac{\widehat{\pi}_k(z)}{\widehat{w}_k(z)} \quad \text{when } \widehat{w}_k(z) > 0. \end{aligned}$$

The accounting identity keeps the estimated treatment propensities in $[0, 1]$. When $\widehat{p}(z) = 0$ or 1, omit the zero-weight term. The joint probabilities $\widehat{\pi}_k(z)$ give the IV weights through the formulas in Section 6.3.

The potential outcome moments within each regime,

$$\mu_{dk,h} = \mathbb{E}[h(Y(d)) \mid S = k, X = x],$$

can be estimated by linear inversion using variation in the estimated weights $\widehat{\omega}_{dk}(z)$ across assignments. For an integrable function h, collect these moments in $\boldsymbol{\mu}_{d,h} = (\mu_{dk,h})_{k=1}^K$. For assignments $z_1, \ldots, z_L$ with $\Pr(D = d \mid Z = z_\ell, X = x) > 0$, set

$$\widehat{\Omega}_d = (\widehat{\omega}_{dk}(z_\ell))_{k,\ell}, \qquad \widehat{g}_{d,h} = \left(\widehat{\mathbb{E}}[h(Y) \mid D = d, Z = z_\ell, X = x]\right)_{\ell=1}^L.$$

Under the rank condition in Proposition 4.2, calculate, when $\widehat{\Omega}_d$ has full row rank,

$$\widehat{\boldsymbol{\mu}}_{d,h} = (\widehat{\Omega}_d \widehat{\Omega}_d')^{-1} \widehat{\Omega}_d \, \widehat{g}_{d,h}.$$

With $h(y) = y$, this gives the regime-specific means and $\widehat{\tau}_k = \widehat{\mu}_{1k,h} - \widehat{\mu}_{0k,h}$, when the means are finite.

I leave Wald and bootstrap inference for smooth functionals of the measurement estimator to future work. Both require joint asymptotic results for the estimated component moments and conditional moments after label normalization, with additional rate conditions if the number of test functions grows.

6.3 Inference for Treatment Effects and IV Weights

I estimate the regime-specific average treatment effects and IV weights from the fitted model, then compare the model-implied IV estimate with the estimated population average treatment effect using the decompositions in Section 3. For pairwise Wald ratios and their weights, I restrict attention to pairs satisfying $|\widehat{p}(x,z) - \widehat{p}(x,z')| \geq \delta$ for a fixed $\delta > 0$.

For a correctly specified model with a fixed number of regimes, Propositions I.1 and I.2 justify pointwise normal and bootstrap inference for the regime-specific average treatment effects, the IV weights and the IV–ATE comparison. They require an asymptotically normal estimator for a parameter value in the interior of the parameter space; regime labels that do not switch near that true parameter value; and first stages that are bounded away from zero. Appendix I spells out all the required conditions, and handles issues of inference with estimated leniency measures and binning of the leniency measures.

6.4 Tests of Common Direction and Rank-1

Both the common direction and Rank-1 restrictions can be tested using the joint probabilities $\pi_k(x,z) = w_k(x)p_k(x,z)$. At a fixed x, collect these probabilities in $\Pi_x = (\pi_k(x,z_\ell))_{k,\ell}$ and define $C_L = I_L - L^{-1}\mathbf{1}_L\mathbf{1}_L'$. Next, order the decision makers by their overall treatment propensity. The common direction restriction is satisfied when

$$g_{k\ell}(x) := \pi_k(x,z_{\ell+1}) - \pi_k(x,z_\ell) \geq 0$$

for every regime and adjacent pair. Hence the null hypothesis is

$$H_0^N: \quad g_{k\ell}(x) \geq 0 \quad \text{for all } k = 1,\ldots,K, \quad \ell = 1,\ldots,L-1.$$

I call the test for common direction Test A. The test statistic is

$$T_n^{\max} = \max_{k,\ell}\left[-\frac{\sqrt{n}\,\widehat{g}_{k\ell}(x)}{\widehat{\sigma}_{k\ell}}\right]_+,$$

where $[a]_+ = \max\{a,0\}$, $\widehat{g}_{k\ell}(x) = \widehat{\pi}_k(x,z_{\ell+1}) - \widehat{\pi}_k(x,z_\ell)$, and $\widehat{\sigma}_{k\ell}$ consistently estimates the asymptotic standard deviation of $\sqrt{n}\{\widehat{g}_{k\ell}(x) - g_{k\ell}(x)\}$. I estimate this standard deviation by bootstrap. Test A uses the largest studentized violation of the common direction restriction.

On the other hand, the null hypothesis for the Rank-1 restriction is

$$H_0^R: \quad \operatorname{rank}(\Pi_x C_L) = 1.$$

The matrix $\Pi_x C_L$ subtracts each regime's mean joint probability across decision makers. Rank one requires the changes in these probabilities to be proportional across regimes. When the overall treatment propensity is nonconstant, each regime then has the same weight across all pairwise Wald ratios with a nonzero first stage. I call this test Test B. The accompanying statistic is

$$D_n^{\text{rank}} = \left(\frac{\sum_{j \geq 2} \sigma_j^2(\widehat{M}_x)}{\sum_{j \geq 1} \sigma_j^2(\widehat{M}_x)} \right)^{1/2},$$

where $\widehat{M}_x = \widehat{\Pi}_x C_L$ and $\sigma_j(\widehat{M}_x)$ are its singular values in decreasing order. The statistic measures the normalized distance from $\widehat{M}_x$ to its best rank-one approximation.

For Test A, I estimate which inequalities bind using $\sqrt{n}\,\hat{g}_{k\ell}/\hat{\sigma}_{k\ell} \leq \kappa_n$, where $\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$, and take the bootstrap maximum over this estimated contact set. With probability approaching one, the set contains the binding inequalities and excludes fixed strictly slack inequalities. I center the bootstrap perturbations under each null. Before testing, I pool ties in the overall treatment propensities for the bootstrap procedure used in simulating the sampling distribution for the test statistic. I keep the ordering fixed in the bootstrap; an estimated ordering must satisfy Assumption I.3. Appendix I gives the bootstrap procedures used to approximate each statistic's distribution under its null hypothesis.

Given the conditions of Theorem I.1, Test B has correct asymptotic size: it rejects a true null with limiting probability α. For Test A, this probability is at most α, with equality when at least one inequality binds. These are pointwise results for a correctly specified model with a fixed number of regimes. They require distinct components, positive regime shares and a stable instrument order. When the order is estimated, distinct overall treatment propensities must be separated, and exact ties are pooled.

These tests have different implications for the pairwise Wald weights: common direction rules out negative Wald weights, whereas Rank-1 requires each regime's weight to be the same for every pair of decision makers with a nonzero first stage. Either restriction can hold without the other. To test the null that at least one restriction holds, I reject only when both tests reject; by the intersection–union principle (Berger, 1982), this rule has pointwise asymptotic size at most α.

7 Simulation Evidence

The first set of simulations examines whether the estimator recovers regime-specific average treatment effects and their IV weights when outcomes under both treatment statuses help

distinguish the regimes. I vary the overlap between the outcome distributions and the differences in regime composition among treated and untreated cases across judges. These simulations include cases where both average treatment effects are positive but the IV estimand is negative. I use a separate set of simulations to examine the tests of common direction and Rank-1.

I simulate samples of 2,500 and 10,000 cases, assigning each case to one of ten judges with equal probability and independently of its regime. The two regimes have Gaussian potential outcome distributions and average treatment effects $(\tau_1, \tau_2) = (0.2, 1.4)$. I choose the means so that the regime with the higher mean under treatment has the lower mean without treatment. Matching the distributions by their mean ranks would therefore pair different regimes and give the wrong average treatment effects. I estimate the joint model by maximum likelihood with $K = 2$ and use the true overall treatment propensity as the scalar instrument to estimate the IV weights.

I combine three patterns of treatment propensities with three degrees of overlap between the regime-specific outcome distributions. When treatment propensities increase proportionally across judges in both regimes, common direction and Rank-1 hold. When treatment propensities increase in both regimes but the increases are not proportional across judges, only common direction holds. When judges who are more lenient toward one regime are stricter toward the other, both restrictions fail. For each pattern, I consider outcome distributions that are well separated, overlap more under both treatment statuses, or overlap more only without treatment. All designs satisfy the Gaussian identification conditions. Appendix J gives the data-generating process and estimation details. Table 2 reports estimation error and how often the estimated means recover the correct ordering across treatment statuses.

The estimator accurately recovers the positive average treatment effects when judges who are more lenient toward one regime are stricter toward the other. The IV weights in these simulations are $(2.585, -1.585)$, and the IV estimand is -1.703. Across both sample sizes and all three specifications of the outcome distributions, the estimated means have the correct ordering across treatment statuses in all 3,000 fits. Even when the distributions overlap under both treatment statuses, a sample of 10,000 cases gives RMSEs of 0.034 and 0.052 for average treatment effects of 0.2 and 1.4. The IV weights are estimated less precisely, with an RMSE of 0.254 in this case, reflecting the offsetting contributions of the two regimes to the aggregate first stage.

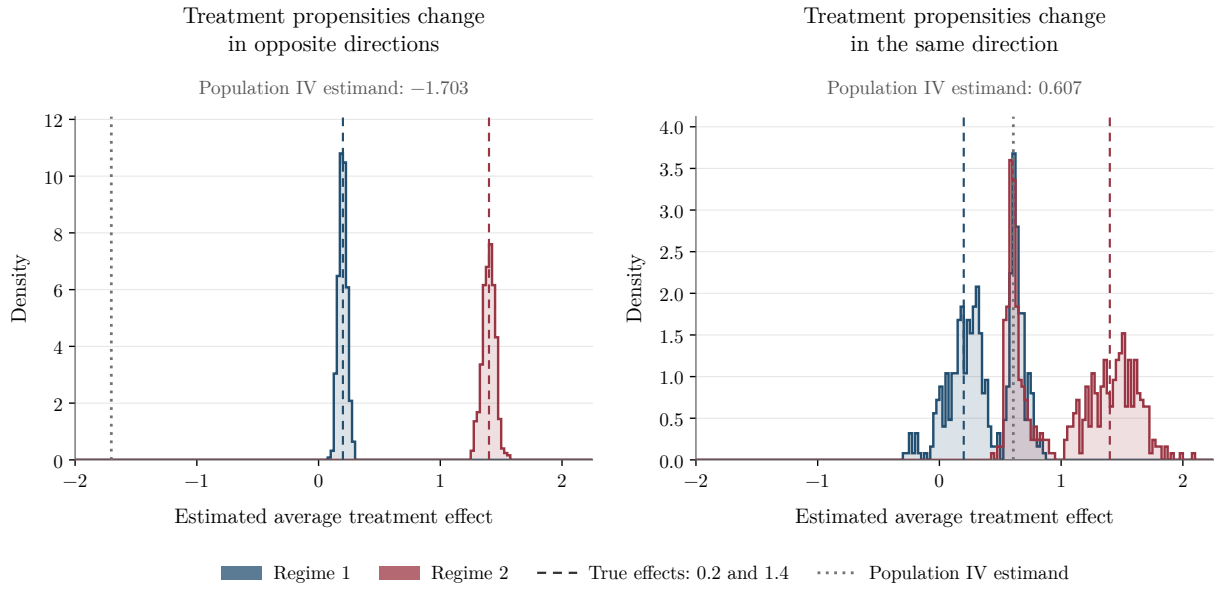
With separated outcome distributions and $n = 10{,}000$, I recover the correct ordering of the means in every replication under all three patterns of treatment propensities, with RMSEs for the average treatment effects ranging from 0.008 to 0.015. When the distributions overlap more, estimation depends more on how regime composition among treated and untreated cases differs across judges. When treatment propensities increase in both regimes but not proportionally across judges, and component distributions overlap under both treatment statuses, the estimated means have the correct ordering in only 58.8 percent of samples at

Table 2: Recovery of treatment effects and IV weights using both treatment arms

Treatment propensities across judges	Overlap between component distributions	Correct ordering (%)	RMSE $\widehat{\tau}_1$	RMSE $\widehat{\tau}_2$	RMSE $\widehat{\Omega}_2$	RMSE $\widehat{\beta}_{\text{IV}}$
$n = 2{,}500$						
Increase proportionally in both regimes	Separated	100.0	0.020	0.029	0.039	0.042
	Both statuses	64.2	0.263	0.609	0.298	0.069
	No treatment	72.6	0.205	0.544	0.055	0.077
Increase in both regimes, but not proportionally	Separated	98.8	0.063	0.079	0.037	0.036
	Both statuses	53.4	0.390	0.608	0.294	0.055
	No treatment	53.4	0.297	0.543	0.046	0.056
Increase in one regime and decrease in the other	Separated	100.0	0.016	0.022	0.477	0.568
	Both statuses	100.0	0.066	0.104	0.558	0.646
	No treatment	100.0	0.086	0.124	0.570	0.680
$n = 10{,}000$						
Increase proportionally in both regimes	Separated	100.0	0.009	0.013	0.019	0.022
	Both statuses	81.4	0.173	0.427	0.204	0.034
	No treatment	89.0	0.121	0.354	0.026	0.038
Increase in both regimes, but not proportionally	Separated	100.0	0.010	0.015	0.016	0.017
	Both statuses	58.8	0.315	0.534	0.222	0.025
	No treatment	59.0	0.260	0.491	0.023	0.028
Increase in one regime and decrease in the other	Separated	100.0	0.008	0.011	0.214	0.257
	Both statuses	100.0	0.034	0.052	0.254	0.296
	No treatment	100.0	0.053	0.079	0.232	0.288

Notes: Each row includes all 500 replications with $K = 2$ fixed and ten equally likely judges. The true average treatment effects are $(\tau_1, \tau_2) = (0.2, 1.4)$. Labels are ordered once by the estimated mean potential outcomes under treatment. Correct ordering means $\widehat{\mu}_{01} > \widehat{\mu}_{02}$, recovering the reversed ordering across treatment arms. This measure does not distinguish incorrect pairing from estimation error in the means. RMSEs include every replication. Outcome standard deviations (σ_0, σ_1) are $(0.2, 0.2)$ for separated distributions, $(0.4, 0.4)$ for overlap in both arms, and $(0.6, 0.2)$ for overlap under no treatment. The population IV weight Ω_2 is 0.204 when treatment propensities increase proportionally in both regimes, 0.340 when they increase in both regimes but not proportionally, and -1.585 when they increase in one regime and decrease in the other. Monte Carlo standard errors of the percentages are at most 2.24 percentage points.

Figure 1: Recovery of average treatment effects



10,000 observations; 500 replications; overlapping outcome distributions in both treatment arms

Notes: Histograms show the estimated regime-specific average treatment effects over all 500 replications with $K = 2$. Colored dashed lines mark the true average treatment effects; gray dotted lines mark the population IV estimand under each simulation design. In the right panel, treatment propensities increase in both regimes, but the increases are not proportional across judges. The panels use the same bins and horizontal limits, with different vertical scales.

$n = 10{,}000$. The RMSEs for the average treatment effects are 0.315 and 0.534, even though the IV RMSE is only 0.025. When judges who are more lenient toward one regime are stricter toward the other, regime composition among treated and untreated cases differs more across judges, and I continue to estimate the average treatment effects accurately with the same outcome distributions (Figure 1). The aggregate IV estimate can be accurate even when the average treatment effects within regimes are poorly estimated.

With $n = 10{,}000$, held-out log likelihood selects the correct number of regimes in 87–89 percent of samples when outcome distributions are well separated and in 64–89 percent when they overlap under both treatment statuses. Appendix Table 7 reports how often each number of regimes is selected and the estimation errors for the population average treatment effect and IV estimand after selection. Appendix Tables 5 and 6 report estimation errors for the potential outcome means, regime shares, and treatment propensities.

Table 3 reports rejection rates from separate simulations with $Y(0) = 0$. Only Test B rejects when treatment propensities increase in both regimes but not proportionally across judges, whereas both tests reject when some judges are more lenient toward one regime and stricter toward the other. When both restrictions hold, Test B rejects in 4.01 and 3.26 percent of samples. When one regime's treatment propensity is constant, Test A rejects in 1.25

percent of samples at the five percent nominal level. Its estimated contact set includes some slack inequalities, which can raise the critical value. Appendix J.3 describes the simulations, bootstrap procedures, and Monte Carlo uncertainty.

Table 3: Rejection rates at the five percent level

Treatment propensities across judges	Test A	Test B	Both reject
Increase proportionally in both regimes	0.00	4.01	0.00
Increase in both regimes, but not proportionally	0.00	100.00	0.00
Increase in one regime and decrease in the other for some judge pairs	100.00	100.00	100.00
One regime's treatment propensity constant	1.25	3.26	0.50

Notes: Entries are percentages from 399 replications per design, with $n = 5{,}000$, ten judges, fixed $K = 2$, and 1999 bootstrap draws per test. These experiments use Gaussian outcomes among treated cases and $Y(0) = 0$. Test A tests common direction; Test B tests Rank-1. Both nulls hold in the first and fourth rows. Only common direction holds in the second row, and neither restriction holds in the third. Appendix Table 8 reports counts and Monte Carlo intervals.

8 Empirical Application: Bail Decisions and Sentencing

In this section, I provide an empirical illustration of some of the theoretical results we have developed in this paper. The illustration is based on the parametric identification route in Section 4.1, for which asymptotic guarantees have been fully worked out. Using the Miami-Dade County bail data assembled by Chyn, Frandsen and Leslie (2025), I estimate the regime-specific average effects of pretrial release on the recorded sentence length and the IV weights on these effects. In the model with two regimes and three outcome components, release reduces mean recorded sentence length by about 192 days for a fifth of cases and by about one day for the remainder. The regime with the larger reduction receives less IV weight than its population share in the point estimates, although this difference is imprecisely estimated.

8.1 Data and Specification

Treatment is pretrial release, measured by whether bail is ever met, and the outcome is the maximum recorded sentence length across the included charges per case, capped at 3,650 days. I use recorded sentence length, which may not coincide with the time the defendant actually serves. The data contain felony and misdemeanor cases from 2006 to 2014, which are then split at the defendant level into a training sample of 28,034 and an inference sample of 65,764 cases, each containing 164 judges. I use the training sample to construct the instrument and

to estimate models whose predictions are evaluated on the inference sample. All reported treatment effects, population shares and IV weights are evaluated on the inference sample, including those from models estimated on the training sample.

Using the leave-one-out UJIVE construction of Goldsmith-Pinkham, Hull and Kolesár (2026), I construct the instrument—a measure of a judge's leniency, henceforth "leniency measure"—from judge indicators along with additive court and time controls in the training sample. Estimating release probabilities judge by judge would require 164 treatment intercepts within every regime, which would result in very noisy estimates. I therefore pool cases across judges into ten groups of the estimated leniency measure when estimating these release probabilities. Judges within a group may still have different release probabilities for the same regime and observed characteristics, so this pooling averages out some of the judge-level variation. In Section 8.3, I verify whether the estimated treatment effects and IV weights change when I use 20 or 40 groups. On the inference sample, the estimated coefficients and boundaries for the groups of the leniency measure are held fixed. Adjusting for court and time controls, however, can still move cases across the groups of the leniency measure. Therefore, the groups depend on the controls, and the IV weights reported below apply to the instrument constructed from these groups.

For IV, each case is assigned its group's mean leniency measure from the training sample. This instrument is then demeaned within covariate cells, each containing all cases with the same values of all controls. Unless stated otherwise, the IV weights below refer to this instrument. Whenever I mention *observed* IV below, I mean the IV calculated directly from the release decisions and sentence lengths recorded in the data. I contrast this with the *model-implied* IV where each case's observed release status is replaced with its predicted probability of release and its recorded sentence is replaced with the sentence length predicted by the model in this paper. These predictions depend on the case's controls and the leniency group. This model-implied IV is the one that admits the decomposition described in Section 3.

To estimate the model, I use a censored-normal mixture for $\log(1 + \text{days})$, where days capture sentence length. Then I back out the expected sentence length in days by integrating over the fitted distributions. Treatment propensities follow a logit model, with a separate intercept for each combination of regime and leniency group and covariate coefficients common across regimes and leniency groups. These treatment propensities are not required to increase with the leniency measure in every regime. Thus, a group with a higher leniency measure can be stricter toward cases in a particular regime.

To distinguish an additional regime from greater flexibility in the distribution of sentences within a regime, I compare three specifications. Two of these specifications have, respectively, two and three regimes, with one censored-normal outcome component per regime under each treatment status. The third specification has two regimes and three outcome components. It combines two components into one regime by requiring them to share release propensities at every value of the instrument and controls. Each component's weight within this regime is also

the same under release and detention, conditional on the controls. BIC favors this two-regime model with three outcome components. Its fit on the inference sample is almost as good as that of the three-regime model, and its predictive likelihood on that sample is higher when both models are estimated on the training sample. Allowing the two components combined into one regime to have different treatment propensities, therefore, does not materially change the fit of the model. Appendix K reports the specifications, likelihood comparisons and IV formulas.

8.2 Regime-Specific Average Treatment Effects and IV Weights

The estimated average effects of release vary substantially across regimes (Table 4). In the two-regime model with three outcome components fitted to the inference sample, 80.9 percent of cases belong to a regime whose fitted sentence distribution places about 93 percent of its mass at zero under both release and detention. Mean sentence length is 1.5 days under detention and 0.2 days under release, so the estimated reduction is just 1.3 days. For the other 19.1 percent of cases, mean sentence length falls from about 594 days under detention to 402 days under release. The estimated reduction is about 192 days, with a 95 percent confidence interval of 151 to 234 days [7]. It bears emphasis here that these are latent regimes recovered by the model and not groups defined by cutoffs in observed sentence length.

More interestingly, although the regime with the larger reduction accounts for 19.1 percent of cases, it receives only 8.5 percent of the IV weight. Part of this gap reflects how IV weights covariate cells: their weights depend on the first stage as well as their population frequencies. Nonetheless, even when IV's weights across cells are used instead, the regime's population share is 17.5 percent—still about twice its IV weight. The confidence interval for the estimated difference between the IV weight and this covariate-adjusted share runs from -19.3 to 1.1 percentage points.

Most of the IV weight falls on the regime in which release reduces mean sentence length by about only a day. The model-implied IV estimate is consequently a reduction of 17 days, compared with an estimated population average reduction of 38 days. Weighting regimes within covariate cells accounts for 16.5 days—roughly four-fifths—of this 21-day gap. The remaining 4.2 days reflect how IV weights the covariate cells. The confidence interval for the difference between the model-implied IV estimate and the population average treatment effect runs from -3.2 to 44.6 days.

IV calculated directly from the data points to an increase of 169 days, but its defendant-clustered standard error is 208 days. Despite the opposite signs, the model-implied estimate is within one standard error of the observed IV estimate. Maximum likelihood fits the

[7] I compare specifications using BIC on the inference sample. The confidence intervals in the main text treat the specification as fixed. Appendix K also reports Bonferroni-adjusted intervals that account for selection among the candidate specifications and regimes.

Table 4: Average treatment effects and IV weights

Quantity	Estimate	95% confidence interval
Regime with the larger reduction		
Average treatment effect (days)	−192.48	[−233.70, −151.26]
Population share (%)	19.11	[18.78, 19.43]
IV weight (%)	8.46	[−3.92, 20.84]
IV weight minus adjusted population share (percentage points)	−9.08	[−19.29, 1.13]
Regime with the smaller reduction		
Average treatment effect (days)	−1.29	[−1.34, −1.25]
Population comparisons		
Population average treatment effect (days)	−37.82	[−45.78, −29.85]
Model-implied IV estimate (days)	−17.15	[−40.60, 6.30]
Model-implied IV estimate minus population average treatment effect (days)	20.67	[−3.21, 44.55]

Notes: Estimates are from the two-regime model with three censored-normal outcome components fitted to the inference sample of 65,764 cases. Confidence intervals use defendant-clustered standard errors and hold the specification fixed. The adjusted population share uses the same covariate weights as IV. Appendix K reports the inference procedure and additional results.

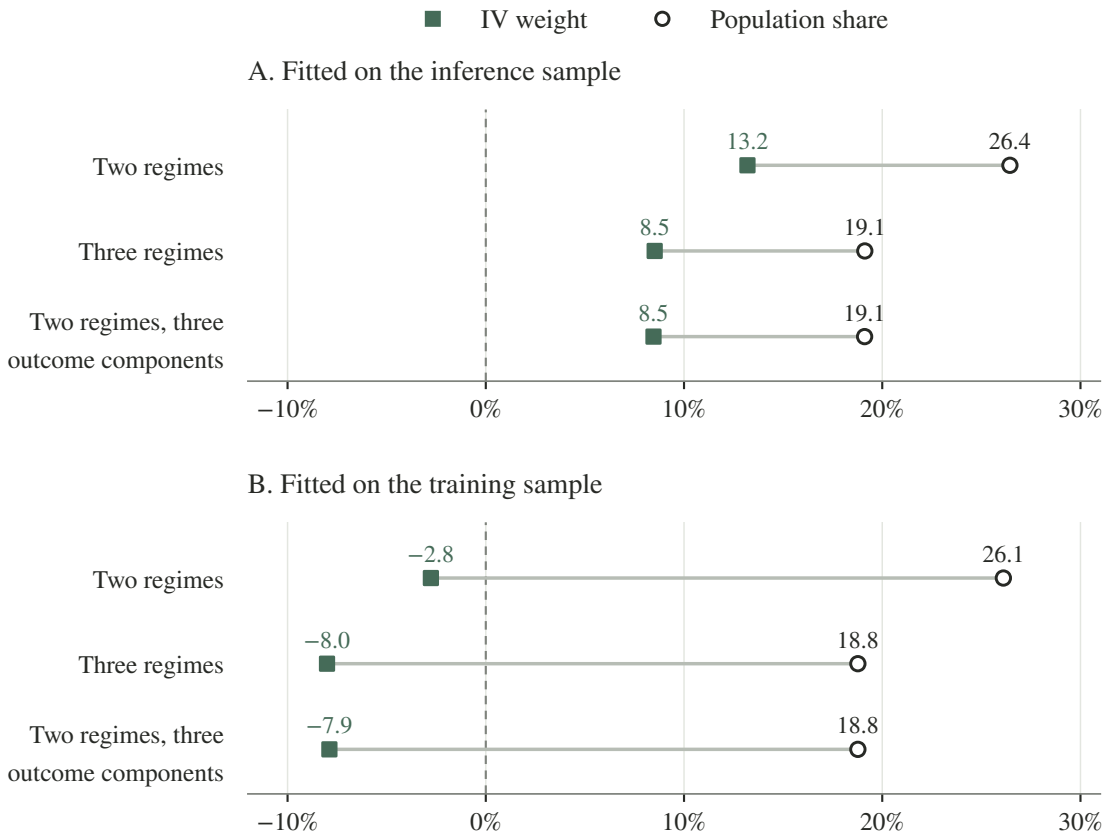
joint distribution of release and sentence length without requiring the fitted first-stage and reduced-form covariances to coincide with their sample values.

8.3 Sensitivity Analyses

The estimated reductions in sentence length are similar across the three models fitted to each sample. For the regime with the largest reduction, they range from 192 to 218 days, and in all six fits this regime receives less IV weight than its population share (Figure 2). The sign of the IV weight, however, changes between samples. When I estimate the two-regime model with three outcome components on the training sample, the regime with the larger reduction has a population share of 18.8 percent and an IV weight of −7.9 percent. Both regime-specific average treatment effects are negative, yet the model-implied IV estimate is an increase of 16 days. This is the possibility illustrated in Section 3: the IV estimand has the opposite sign from every regime-specific average treatment effect. When I estimate the model on the inference sample, this regime's estimated IV weight is 8.5 percent, with a 95 percent confidence interval of −3.9 to 20.8 percent.

The negative IV weight in the training sample fit indicates that, after accounting for the controls, groups with higher estimated leniency tend to be stricter toward cases in this regime. An incorrect ordering of judges by the estimated leniency measure can produce this pattern across groups even without reversals among individual judges. I also check whether the underweighting depends on how finely I group the same leniency measure. Using 20 or

Figure 2: Population shares and IV weights in Miami-Dade



Notes: Each row shows the regime with the largest estimated average reduction in recorded sentence length, measured in days. Both panels evaluate the estimates on the same inference sample of 65,764 cases; regime membership may differ across specifications. Lines connect estimates of the population share and IV weight, not confidence limits. The instrument is each of the ten leniency groups' mean leniency measure in the training sample, demeaned within covariate cells.

40 groups leaves the estimated population shares and treatment effects nearly unchanged. The regime with the largest estimated reduction continues to receive less IV weight than its population share, whether the instrument is the mean leniency measure in each of the original ten groups, the mean in each finer group or the continuous measure. As before, group means are calculated in the training sample and each instrument is demeaned within covariate cells. Neither finer grouping improves BIC in either estimation sample. Models estimated on the training sample also have lower predictive likelihoods on the inference sample than the model with ten groups.

These robustness checks with respect to finer groupings keep the outcome specification fixed. When I allow the two components within a regime to have different weights under release and detention, the regime with the largest estimated reduction still receives less IV weight than its population share in both estimation samples. However, BIC is higher and the model estimated on the training sample has a lower predictive likelihood on the

inference sample than the specification imposing the same component weights under release and detention.

One outcome component fits many detained cases with sentences recorded as exactly 21 days. Its estimated standard deviation is 0.0204 on the log scale, just above the minimum of 0.02 imposed in estimation. To assess sensitivity of the estimates to this bound, I raise it to 0.05 and 0.10 and re-estimate the model. The larger estimated reduction becomes about 206 and 213 days, respectively, and the regime with that reduction still receives less IV weight than its adjusted population share (Appendix K).

9 Conclusion

Standard leniency designs can obscure multidimensional unobserved case heterogeneity. Summarizing decisionmakers by their overall treatment propensity effectively treats decisionmakers with the same share of treated cases as identical, even when they systematically disagree on which types of cases to treat more often. I develop a finite regime model that allows the treatment propensities of decision makers to vary across latent types so decision makers can disagree on which types to treat more. I provide conditions for identifying these type-specific treatment propensities, the types' average treatment effects and their population shares. My identification results use either restrictions on outcome distributions or suitably chosen pretreatment measurements, without invoking monotonicity.

With overall treatment propensity as the instrument, I show that the IV estimand admits a decomposition as a weighted sum of type-specific average treatment effects, with weights that depend on each type's population share and how its treatment rate varies with overall leniency across decision makers. These weights can be negative even when the weights on the pairwise Wald estimands across decision makers are positive, so the estimand can lose its weakly causal interpretation with respect to type-specific average treatment effects. Neither individual-level monotonicity (Imbens and Angrist, 1994) nor average monotonicity (Frandsen, Lefgren and Leslie, 2023) is required for a weakly causal interpretation when these weights are nonnegative.

Even when the conditions for a weakly causal interpretation are satisfied, recovering the type-specific average treatment effects may still be instructive. In particular, a small IV estimand can aggregate large type-specific average treatment effects of opposite signs or put little weight on types with large effects. The latter possibility is borne out by the application of the model to Miami-Dade County bail data where, re-analyzing the effect of pretrial release on sentencing outcomes, I find that the point estimates of the IV weights put less weight on the type with the largest estimated reduction in recorded sentence length than its estimated population share.

References

Allman, Elizabeth S., Catherine Matias, and John A. Rhodes. 2009. "Identifiability of Parameters in Latent Structure Models with Many Observed Variables." *Annals of Statistics*, 37(6A): 3099–3132.

Andrews, Donald W. K., and Gustavo Soares. 2010. "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection." *Econometrica*, 78(1): 119–157.

Berger, Roger L. 1982. "Multiparameter Hypothesis Testing and Acceptance Sampling." *Technometrics*, 24(4): 295–300.

Bhuller, Manudeep, Gordon B. Dahl, Katrine V. Løken, and Magne Mogstad. 2020. "Incarceration, Recidivism, and Employment." *Journal of Political Economy*, 128(4): 1269–1324.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky. 2026. "When is TSLS Actually LATE?" *The Review of Economic Studies*, rdag029.

Bonhomme, Stéphane, Koen Jochmans, and Jean-Marc Robin. 2016. "Estimating Multivariate Latent-Structure Models." *Annals of Statistics*, 44(2): 540–563.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa. 2022. "Discretizing Unobserved Heterogeneity." *Econometrica*, 90(2): 625–643.

Chan, David C., Matthew Gentzkow, and Chuan Yu. 2022. "Selection with Variation in Diagnostic Skill: Evidence from Radiologists." *Quarterly Journal of Economics*, 137(2): 729–783.

Chyn, Eric, Brigham Frandsen, and Emily Leslie. 2025. "Examiner and Judge Designs in Economics: A Practitioner's Guide." *Journal of Economic Literature*, 63(2): 401–439.

Collinson, Robert, John Eric Humphries, Nicholas Mader, Davin Reed, Daniel Tannenbaum, and Winnie van Dijk. 2024. "Eviction and Poverty in American Cities." *Quarterly Journal of Economics*, 139(1): 57–120.

Coulibaly, Mohamed, Yu-Chin Hsu, Ismael Mourifié, and Yuanyuan Wan. 2024. "A Sharp Test for the Judge Leniency Design." NBER Working Paper 32456.

Dahl, Gordon B., Andreas Ravndal Kostøl, and Magne Mogstad. 2014. "Family Welfare Cultures." *Quarterly Journal of Economics*, 129(4): 1711–1752.

Dobbie, Will, Jacob Goldin, and Crystal S. Yang. 2018. "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges." *American Economic Review*, 108(2): 201–240.

Farré-Mensa, Joan, Deepak Hegde, and Alexander Ljungqvist. 2020. "What Is a Patent Worth? Evidence from the U.S. Patent 'Lottery'." *Journal of Finance*, 75(2): 639–682.

Frandsen, Brigham, Lars Lefgren, and Emily Leslie. 2023. "Judging Judge Fixed Effects." *American Economic Review*, 113(1): 253–277.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár. 2026. "Leniency Designs: An Operator's Manual." *Journal of Economic Perspectives*, 40(3): 213–240.

Gross, Max, and E. Jason Baron. 2022. "Temporary Stays and Persistent Gains: The Causal Effects of Foster Care." *American Economic Journal: Applied Economics*, 14(2): 170–199.

Heckman, James J., and Edward J. Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica*, 73(3): 669–738.

Henry, Marc, Yuichi Kitamura, and Bernard Salanié. 2014. "Partial Identification of Finite Mixtures in Econometric Models." *Quantitative Economics*, 5(1): 123–144.

Hoshino, Tadao, and Takahide Yanagi. 2022. "Estimating Marginal Treatment Effects under Unobserved Group Heterogeneity." *Journal of Causal Inference*, 10(1): 197–216.

Hu, Yingyao. 2008. "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution." *Journal of Econometrics*, 144(1): 27–61.

Hu, Yingyao, and Susanne M. Schennach. 2008. "Instrumental Variable Treatment of Nonclassical Measurement Error Models." *Econometrica*, 76(1): 195–216.

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

Kamat, Vishal, Samuel Norris, and Matthew Pecenco. 2026. "Identification in Multiple Treatment Models under Discrete Variation." Revised January 2026.

Kasahara, Hiroyuki, and Katsumi Shimotsu. 2014. "Non-parametric Identification and Estimation of the Number of Components in Multivariate Mixtures." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1): 97–111.

Kruskal, Joseph B. 1977. "Three-Way Arrays: Rank and Uniqueness of Trilinear Decompositions, with Application to Arithmetic Complexity and Statistics." *Linear Algebra and its Applications*, 18(2): 95–138.

Lee, Sokbae, and Bernard Salanié. 2018. "Identifying Effects of Multivalued Treatments." *Econometrica*, 86(6): 1939–1963.

Maestas, Nicole, Kathleen J. Mullen, and Alexander Strand. 2013. "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt." *American Economic Review*, 103(5): 1797–1829.

Mao, Minghai, Pedro H. C. Sant'Anna, and Xiaojun Song. 2026. "Falsifying Marginal Treatment Effects." Working paper.

Mogstad, Magne, Alexander Torgovitsky, and Christopher R. Walters. 2021. "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables." *American Economic Review*, 111(11): 3663–3698.

Mogstad, Magne, and Alexander Torgovitsky. 2024. "Instrumental Variables with Unobserved Heterogeneity in Treatment Effects." In *Handbook of Labor Economics*, Vol. 5, 1–114. Elsevier.

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

Newey, Whitney K., and Daniel McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." In *Handbook of Econometrics*, ed. Robert F. Engle and Daniel L. McFadden, Vol. 4, 2111–2245. Elsevier.

Teicher, Henry. 1963. "Identifiability of Finite Mixtures." *Annals of Mathematical Statistics*, 34(4): 1265–1269.

van der Vaart, A. W. 1998. *Asymptotic Statistics.* Vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge: Cambridge University Press.

Vytlacil, Edward. 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica*, 70(1): 331–341.

Yakowitz, Sidney J., and John D. Spragins. 1968. "On the Identifiability of Finite Mixtures." *Annals of Mathematical Statistics*, 39(1): 209–214.

Appendix Contents

A Proofs 46

B Further Results on Identification from Outcome Mixtures 55

C Identification of Regimes with Gaussian Subcomponents 63

D Identification Using Pretreatment Measurements 65

E Further Results on Treatment Propensities 76

F Population Average Treatment Effects and IV Weights across Covariates 81

G Approximation and Sensitivity 83

H Parametric Estimation 89

I Inference for Fixed-Dimensional Mixtures 93

J Additional Simulation Details 103

K Additional Results for the Application to Pretrial Release 110

A Proofs

Proof of Lemma 2.1

Fix a covariate value x where the assumptions hold, a cell (d, z, k) with $\Pr(D = d, Z = z, S = k \mid X = x) > 0$, and y. By the consistency relation in (2.1), $Y = Y(d)$ on the event $\{D = d\}$, hence

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

Treatment consistency gives $D = D(z)$ on $\{Z = z\}$. By Assumption 2.3 and the selection equation,

$$Z \perp (Y(d), V) \mid (X, S), \qquad D(z) = \mathbf{1}\{g_S(X, z, V) \ge 0\} \quad \Longrightarrow \quad Z \perp (Y(d), D(z)) \mid (X, S).$$

Therefore,

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

where the second equality is Assumption 2.4. □

Proof of Lemma 2.2

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

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

By Lemma 2.1,

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

Bayes' rule and Assumption 2.2 give, whenever $\Pr(D = d \mid X = x, Z = z) > 0$,

$$\begin{aligned}\Pr(S = k \mid D = d, X = x, Z = z) &= \frac{\Pr(D = d \mid S = k, X = x, Z = z) w_k}{\Pr(D = d \mid X = x, Z = z)} \\ &= \begin{cases} w_k p_k(x, z)/p(x, z), & d = 1, \\ w_k\{1 - p_k(x, z)\}/\{1 - p(x, z)\}, & d = 0, \end{cases}\end{aligned}$$

as in (2.11). □

Proof of Proposition 3.1

Using Lemma 2.1, within regime k,

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

where $\mu_{0k}(x) := \mathbb{E}[Y(0) \mid X = x, S = k]$. Thus

$$\begin{aligned}m(x, z) &= \sum_{k=1}^{K_0} w_k(x) \mathbb{E}[Y \mid X = x, Z = z, S = k] \\ &= \sum_{k=1}^{K_0} w_k(x) \mu_{0k}(x) + \sum_{k=1}^{K_0} w_k(x) p_k(x, z) \tau_k(x),\end{aligned}$$

and

$$m(x, z) - m(x, z') = \sum_{k=1}^{K_0} \{\pi_k(x, z) - \pi_k(x, z')\} \tau_k(x).$$

Since $p(x, z) = \sum_k \pi_k(x, z)$,

$$p(x, z) - p(x, z') = \sum_{k=1}^{K_0} \{\pi_k(x, z) - \pi_k(x, z')\}.$$

For $p(x, z) \neq p(x, z')$,

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

$$\sum_{k=1}^{K_0} \omega_k^*(x; z, z') = \frac{\sum_k \{\pi_k(x, z) - \pi_k(x, z')\}}{p(x, z) - p(x, z')} = 1.$$

These are (3.2)–(3.3). □

Proof of Proposition 3.2

Taking covariance under $q(\cdot \mid x)$ in the conditional mean representation gives

$$\mathrm{Cov}_q\{m(x, Z), \zeta(Z)\} = \sum_{k=1}^{K_0} \mathrm{Cov}_q\{\pi_k(x, Z), \zeta(Z)\} \tau_k(x),$$

because $\sum_k w_k(x)\mu_{0k}(x)$ is constant in Z. Since $c(x) \neq 0$,

$$\beta^{q,\zeta}(x) = \sum_{k=1}^{K_0} \frac{\mathrm{Cov}_q\{\pi_k(x,Z), \zeta(Z)\}}{c(x)} \tau_k(x) = \sum_{k=1}^{K_0} \Omega_k^{q,\zeta}(x)\tau_k(x),$$
$$\sum_{k=1}^{K_0} \Omega_k^{q,\zeta}(x) = \frac{\mathrm{Cov}_q\{\sum_k \pi_k(x,Z), \zeta(Z)\}}{c(x)} = \frac{c(x)}{c(x)} = 1.$$

This proves (3.4). For any finite-support functions a and b, the covariance identity is

$$\frac{1}{2}\sum_{z,z'} q(z \mid x)q(z' \mid x)\{a(z) - a(z')\}\{b(z) - b(z')\} = \mathrm{Cov}_q\{a(Z), b(Z)\}.$$

Apply this identity to the reduced form and first stage. For each pair with $p(x,z) \neq p(x,z')$, use

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

Under the restriction on ζ,

$$p(x,z) = p(x,z') \quad \Longrightarrow \quad \zeta(z) - \zeta(z') = 0.$$

Thus, with sums over ordered pairs,

$$\beta^{q,\zeta}(x) = \sum_{\substack{z,z' \\ p(x,z)\neq p(x,z')}} \frac{q(z \mid x)q(z' \mid x)\{p(x,z) - p(x,z')\}\{\zeta(z) - \zeta(z')\}}{2c(x)} \psi(x;z,z'),$$
$$1 = \sum_{\substack{z,z' \\ p(x,z)\neq p(x,z')}} \frac{q(z \mid x)q(z' \mid x)\{p(x,z) - p(x,z')\}\{\zeta(z) - \zeta(z')\}}{2c(x)}.$$

For $\zeta(z) = p(x,z)$, each pair weight is

$$\frac{q(z \mid x)q(z' \mid x)\{p(x,z) - p(x,z')\}^2}{2\,\mathrm{Var}_q\{p(x,Z)\}} \geq 0.$$

□

Proof of Theorem 4.1

Fix a treatment arm d satisfying the assumptions.

Step (i): The observed outcome distributions span a space of dimension K_0. By (2.10), each g_{dz} lies in $V := \mathrm{span}\{f_{dk}\}_{k=1}^{K_0}$. Assumption 4.1 gives $\dim V = K_0$, so the observed mixtures span a space of dimension at most K_0. Conversely, choose $z_1, \ldots, z_L$ as in Assumption 4.2, where $L \geq K_0$ and $\Omega_d := [\omega_{dk}(z_\ell)]_{k,\ell}$ has rank K_0. Since $\mathrm{rank}(\Omega_d) = K_0$, there exist K_0

columns, 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 (2.10) implies

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

and Assumption 4.1 gives

$$\Omega_d^* c = 0 \quad \Longrightarrow \quad c = (\Omega_d^*)^{-1} 0 = 0.$$

Hence

$$K_0 = \dim \operatorname{span}\{g_{dz_{\ell_j}}\}_{j=1}^{K_0} \leq \dim \operatorname{span}\{g_{dz} : z\} \leq \dim V = K_0.$$

Step (ii): The component distributions remain unidentified. The observed mixtures can have another admissible representation in V. Write $f_d = (f_{d1}, \ldots, f_{dK_0})^\top$ and let $q = (q_1, \ldots, q_{K_0})^\top = A f_d$ for an invertible matrix A, and suppose the entries of q are probability distributions. If $\tilde{\omega}_d(z) = A^{-\top} \omega_d(z)$ is nonnegative and sums to one at every z, then

$$\tilde{\omega}_d(z)^\top q = \omega_d(z)^\top A^{-1} A f_d = \omega_d(z)^\top f_d = g_{dz},$$
$$\operatorname{span}\{q_k\}_{k=1}^{K_0} = V, \qquad \operatorname{rank}(A^{-\top} \Omega_d) = \operatorname{rank}(\Omega_d) = K_0.$$

The separation and rank assumptions therefore cannot distinguish these admissible representations.

For a finite set of moments or CDF thresholds, the uncentered matrix has rank K_0 if the evaluations preserve the linear independence of the component distributions and the mixing weight matrix has full row rank (Theorem 4.1). Differencing against a reference assignment z_0 gives

$$\mathbf{1}^\top \{\omega_d(z) - \omega_d(z_0)\} = 1 - 1 = 0, \qquad \operatorname{rank}[\omega_d(z) - \omega_d(z_0)]_z \leq K_0 - 1.$$

The matrix of evaluated differences therefore has rank at most $K_0 - 1$. □

Proof of Proposition 4.1

Fix x and suppress it from the notation. The potential outcome distribution within each regime is Gaussian, so identifying the Gaussian components identifies the regimes without the grouping condition in Theorem C.1. I first prove linear independence, which is also used in Theorem C.1.

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

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

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 $c_1, \dots, c_K$.

The Gaussian moment generating functions exist for every real t, so the assumed linear relation implies

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

Since the parameter pairs are distinct, there is a unique component m with maximal variance and, among ties, maximal mean. Thus

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

For $k \neq m$, as $t \to +\infty$,

$$(\mu_k - \mu_m)t + \tfrac{1}{2}(\sigma_k^2 - \sigma_m^2)t^2 \longrightarrow -\infty, \quad \begin{cases} \sigma_k^2 - \sigma_m^2 < 0, & \sigma_k^2 < \sigma_m^2, \\ \mu_k - \mu_m < 0, & \sigma_k^2 = \sigma_m^2. \end{cases}$$

Consequently,

$$c_m = -\lim_{t \to \infty} \sum_{k \neq m} c_k \exp\Big((\mu_k - \mu_m)t + \tfrac{1}{2}(\sigma_k^2 - \sigma_m^2)t^2\Big) = 0.$$

The family $\{\phi(y; \mu_k, \sigma_k^2)\}_{k=1}^K$ is therefore linearly independent: induction gives $c_k = 0$ for every k. This proof relies on the existence of moment generating functions at arbitrarily large positive t, so it cannot be extended directly to heavy-tailed families.

Step 2: Identification of the component parameters and weights. All components have strictly positive weight at $\bar{z}$. The finite Gaussian mixture

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

therefore identifies the mixing measure

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

up to permutation (Teicher, 1963; Yakowitz and Spragins, 1968). Fix any labeling of these components. Their means and variances are common across instrument values, so for any other z,

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

has unique mixture coefficients by Step 1:

$$\sum_{k=1}^K \{\omega_{dk}(z) - \tilde{\omega}_{dk}(z)\}\phi(y; \mu_{dk}, \sigma^2_{dk}) \equiv 0 \quad \Longrightarrow \quad \omega_{dk}(z) = \tilde{\omega}_{dk}(z) \quad \text{for every } k.$$

If the same conclusion holds for both arms and there is a unique permutation of the labels for the distributions under no treatment σ such that $p(z)\omega_{1k}(z) + \{1 - p(z)\}\omega_{0,\sigma(k)}(z)$ is constant in z for every k, then after applying this pairing,

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

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. □

Proof of Corollary 4.1

Fix x, a treatment status d, and an assignment value z with $\Pr(D = d \mid X = x, Z = z) > 0$. Let f and $\tilde{f}$ be two uncensored Gaussian mixture densities that give the same censored distribution. Then

$$\begin{aligned} f = \tilde{f} \text{ a.e. on } (\ell, u) &\implies f = \tilde{f} \text{ on } (\ell, u) && \text{by continuity,} \\ &\implies f = \tilde{f} \text{ on } \mathbb{R} && \text{by real analyticity.} \end{aligned}$$

The second implication is the identity theorem. Proposition 4.1 identifies the uncensored parameters and weights. The endpoint masses are

$$\Pr(Y = \ell \mid D = d, X = x, Z = z) = \int_{-\infty}^{\ell} f(y)\, dy,$$
$$\Pr(Y = u \mid D = d, X = x, Z = z) = \int_{u}^{\infty} f(y)\, dy.$$

When there is one component per regime, identify the components at $\bar{z}$ using Proposition 4.1, then use linear independence to recover their weights at the other assignments. For several subcomponents per regime, apply Theorem C.1 to the uncensored mixtures identified at each assignment: proportional subcomponent weight vectors identify which subcomponents belong to the same regime. If the regime distributions under treatment and no treatment have a unique pairing, the accounting identities then identify w_k and $p_k(z)$. The identified censored distributions give

$$\tau_k = \int y \, dF_{1k}(y) - \int y \, dF_{0k}(y).$$

□

Proof of Theorem 4.2

For a fixed treatment arm, omit the arm index. To prove linear independence, suppose distinct negative binomial parameter pairs $\{(r_j, q_j)\}_{j=1}^{J}$ satisfy

$$\sum_{j=1}^{J} c_j b(y; r_j, q_j) = 0 \qquad \text{for every } y \in \mathbb{N}_0.$$

The asymptotic expansion of the ratio of gamma functions gives

$$b(y; r, q) = \frac{(1-q)^r}{\Gamma(r)} y^{r-1} q^y \{1 + o(1)\} \qquad \text{as } y \to \infty.$$

If the relation has a nonzero coefficient, choose a component j among those with nonzero coefficients, taking the largest q_j and then the largest r_j over ties. For $k \neq j$ with $c_k \neq 0$,

$$\frac{b(y; r_k, q_k)}{q_j^y y^{r_j - 1}} = \frac{(1-q_k)^{r_k}}{\Gamma(r_k)} y^{r_k - r_j} \left(\frac{q_k}{q_j}\right)^y \{1 + o(1)\} \longrightarrow 0,$$

since either $q_k < q_j$ or $q_k = q_j$ and $r_k < r_j$. Hence

$$0 = \lim_{y \to \infty} \sum_{k=1}^{J} c_k \frac{b(y; r_k, q_k)}{q_j^y y^{r_j - 1}} = c_j \frac{(1-q_j)^{r_j}}{\Gamma(r_j)} \quad \Longrightarrow \quad c_j = 0,$$

a contradiction. Thus $c_1 = \cdots = c_J = 0$.

For two finite mixture representations at $\bar{z}$, let $\{(r_j, q_j)\}_{j=1}^{J}$ be the union of their distinct parameter pairs, with weight zero for any absent pair. Then

$$\sum_{j=1}^{J} \omega_j b(y; r_j, q_j) = \sum_{j=1}^{J} \tilde{\omega}_j b(y; r_j, q_j)$$

$$\implies \quad \sum_{j=1}^{J} (\omega_j - \tilde{\omega}_j) b(y; r_j, q_j) = 0 \quad \implies \quad \omega_j = \tilde{\omega}_j \text{ for all } j.$$

The positive weights identify the same components and parameters in both representations, up to permutation. The parameters are common across assignments, so linear independence also identifies the weights at every other assignment. If both treatment statuses satisfy the assumptions and their outcome mixtures have a unique pairing, the accounting identities in the proof of Theorem C.1 recover the regime shares and treatment propensities from those identified mixtures.

□

Proof of Corollary 5.1

For part (i), write the conditional reduced form as

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

Under Assumptions 2.2–2.4,

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

The potential outcome means enter only through $\tau_k(x)$, proving part (i).

For part (ii), hold the regime shares and treatment propensities fixed. The decomposition $\psi(x; z, z') = \sum_k \omega_k^*(x; z, z') \tau_k(x)$ of Proposition 3.1 then depends on potential outcome means only through the average treatment effects. If $\omega_k^* \geq 0$ for all k, then $\sum_k \omega_k^* = 1$ gives

$$\min_k \tau_k \leq \psi = \sum_k \omega_k^* \tau_k \leq \max_k \tau_k.$$

Conversely, if $\omega_{k^*}^* < 0$, take $M > 0$ and

$$Y(0) = 0, \qquad Y(1) = M \mathbf{1}\{S = k^*\}, \qquad \tau_k = M \mathbf{1}\{k = k^*\}.$$

These potential outcomes are constant within each regime and satisfy the maintained assumptions, but

$$Y(1) - Y(0) \geq 0 \quad \text{for every individual,} \qquad \psi = M\omega^*_{k^*} < 0.$$

Thus sign preservation in Definition 5.1 fails. □

Proof of Proposition 5.1

Statement (i) implies (ii) because centering removes the intercepts and leaves

$$P_x C_L = b(x)\{\lambda(z_1) - \bar{\lambda}, \ldots, \lambda(z_L) - \bar{\lambda}\},$$

and the nonconstant aggregate propensity gives

$$w' P_x C_L \neq 0 \quad \Longrightarrow \quad 1 \leq \operatorname{rank}(P_x C_L) \leq 1, \qquad w = (w_1(x), \ldots, w_{K_0}(x))'.$$

To show that (ii) implies (iii), write $P_x C_L = bc'$. Then

$$0 \neq w' P_x C_L = (w'b)c' \quad \Longrightarrow \quad w'b \neq 0,$$

so

$$P_x C_L = \frac{b}{w'b}\, w' P_x C_L.$$

Row by row, this equality gives

$$p_k(x,z) - \bar{p}_k(x) = \beta_k(x)\{p(x,z) - \bar{p}(x)\}, \qquad \beta_k(x) := \frac{b_k}{w'b},$$

and hence

$$p_k(x,z) = \alpha_k(x) + \beta_k(x) p(x,z), \qquad \alpha_k(x) := \bar{p}_k(x) - \beta_k(x)\bar{p}(x).$$

This is statement (iii).

Statement (iii) implies (i) by taking $\lambda(z) = p(x,z)$. It also implies (iv), since

$$\begin{aligned}
&p_k(x,z) - p_k(x,z') = \beta_k(x)\{p(x,z) - p(x,z')\},\\
&p(x,z) = p(x,z') \implies p_k(x,z) = p_k(x,z'),\\
&p(x,z) \neq p(x,z') \implies \omega^*_k(x;z,z') = \frac{w_k(x)\beta_k(x)\{p(x,z) - p(x,z')\}}{p(x,z) - p(x,z')} = w_k(x)\beta_k(x).
\end{aligned}$$

For the converse, fix z_0. Under (iv), define $\beta_k(x) = \bar{\omega}_k(x)/w_k(x)$. For every z with $p(x,z) \neq p(x,z_0)$, the common pairwise weights give

$$p_k(x,z) - p_k(x,z_0) = \beta_k(x)\{p(x,z) - p(x,z_0)\}.$$

For $p(x,z) = p(x,z_0)$, (iv) gives

$$p_k(x,z) - p_k(x,z_0) = 0 = \beta_k(x)\{p(x,z) - p(x,z_0)\}.$$

Thus, for every z,

$$p_k(x,z) = \underbrace{p_k(x,z_0) - \beta_k(x)p(x,z_0)}_{\alpha_k(x)} + \beta_k(x)p(x,z),$$

which proves (iii). By Proposition 3.1,

$$\psi(x;z,z') = \sum_{k=1}^{K_0} w_k(x)\beta_k(x)\tau_k(x), \qquad \beta_k(x) \geq 0 \implies w_k(x)\beta_k(x) \geq 0.$$

□

B Further Results on Identification from Outcome Mixtures

This appendix characterizes the regime shares when only the outcome distribution among treated cases identifies its mixture components and weights. It then examines matching components identified separately in both treatment groups and identifying regime-specific average treatment effects and the potential outcome distributions under no treatment from variation in the joint probabilities of treatment and regime membership.

B.1 Identification from Outcomes among Treated Cases

When outcomes are observed only for treated cases, as with pretrial misconduct upon release, defining $Y := DY(1)$ sets the outcome without treatment to zero. The distribution among untreated cases is then degenerate at zero and contains no information about its mixture weights. Identification of the potential outcome distributions under treatment and the regime proportions among treated cases still recovers the joint probabilities of treatment and regime membership $\pi_k(z) = w_k p_k(z)$. Each regime must contain at least as much population mass as is treated at any instrument value: $w_k \geq \max_z \pi_k(z)$. These bounds are sharp by Proposition B.1:

every vector of shares satisfying the bounds and summing to one can generate the observed distribution.

Proposition B.1 (Regime shares from joint probabilities of treatment and regime membership)**.** *Fix x and a finite instrument support $\mathcal{Z}$. Suppose the outcome mixture among treated cases identifies the joint probabilities of treatment and regime membership for K_0 regimes, each with positive probability of treatment at some instrument value, and define $\pi_k^{\max}(x) := \max_{z\in\mathcal{Z}} \pi_k(x,z) > 0$. Regime shares that are constant across assignments with $\pi_k(x,z) = w_k(x)p_k(x,z)$ and $p_k(x,z) \in [0,1]$ exist if and only if*

$$\sum_{k=1}^{K_0} \pi_k^{\max}(x) \leq 1. \tag{B.1}$$

The sharp identified set for the shares is

$$\mathcal{W}_x(\pi) = \{w \in \Delta_{K_0} : w_k \geq \pi_k^{\max}(x) \text{ for every } k\}, \qquad \Delta_{K_0} := \left\{ w \in [0,1]^{K_0} : \sum_k w_k = 1 \right\}. \tag{B.2}$$

For each $w \in \mathcal{W}_x(\pi)$, the regime-specific treatment propensity is $p_k(x,z) = \pi_k(x,z)/w_k$.

Proof. Fix x and suppress it in the notation. Necessity of (B.1) follows from

$$\pi_k(z) = w_k p_k(z) \leq w_k \quad \Longrightarrow \quad \sum_k \pi_k^{\max} \leq \sum_k w_k = 1.$$

If (B.1) holds, every vector in (B.2) has the form

$$w_k = \pi_k^{\max} + r_k, \qquad r_k \geq 0, \qquad \sum_k r_k = 1 - \sum_k \pi_k^{\max}.$$

For each such vector,

$$0 \leq p_k(z) := \frac{\pi_k(z)}{w_k} \leq 1, \qquad w_k p_k(z) = \pi_k(z).$$

Take $S \perp Z$ with $\Pr(S = k) = w_k$, set $Y(0) = 0$, and draw $Y(1)$ independently of Z conditional on S from its identified component distribution F_{1k}. Let

$$V \sim \text{Unif}[0,1], \qquad V \perp (S, Z, Y(1)), \qquad D(z) = \mathbf{1}\{V \leq p_S(z)\}.$$

For every measurable outcome set B,

$$\Pr(Y \in B, D = 1 \mid Z = z) = \sum_k w_k p_k(z) F_{1k}(B) = \sum_k \pi_k(z) F_{1k}(B),$$
$$\Pr(Y \in B, D = 0 \mid Z = z) = \left\{1 - \sum_k \pi_k(z)\right\} \mathbf{1}\{0 \in B\}.$$

Independence of V also gives $\Pr(Y(1) \le y \mid D(z) = 1, S = k) = F_{1k}(y)$ for every y whenever $p_k(z) > 0$. Thus every share vector in the proposed set attains the observed distribution and satisfies the model restrictions, establishing sharpness. □

With at least two regimes and lower bounds that sum to less than one, the remaining population mass can be allocated across regimes in several ways. The observed distribution is the same under these allocations, although the shares and treatment propensities differ. The share vector becomes unique when $\sum_k \pi_k^{\max}(x) = 1$, since each share then equals its lower bound, or when there is just one regime, whose share is one. Outside these cases, point identification requires information from identified component distributions and weights among untreated cases, auxiliary information about the shares, or another normalization.

Corollary B.1 (Wald weights and selection restrictions when observed outcomes without treatment are zero)**.** *Fix x and suppose the outcome mixture among treated cases identifies $\{\pi_k(x, z) : k \le K_0, z \in \mathcal{Z}\}$ up to a common permutation. For every pair (z, z') with $p(x, z) \ne p(x, z')$,*

$$\omega_k^*(x; z, z') = \frac{\pi_k(x, z) - \pi_k(x, z')}{p(x, z) - p(x, z')}, \qquad p(x, z) = \sum_{k=1}^{K_0} \pi_k(x, z).$$

Let $P_x = (p_k(x, z_\ell))_{k,\ell}$, $\Pi_x = (\pi_k(x, z_\ell))_{k,\ell}$, and $C_L = I_L - L^{-1}\mathbf{1}_L\mathbf{1}_L'$. Then

$$\Pi_x C_L = \operatorname{diag}\{w_1(x), \ldots, w_{K_0}(x)\} P_x C_L,$$

so $\operatorname{rank}(\Pi_x C_L) = \operatorname{rank}(P_x C_L)$ *and regime-specific increments in the two matrices have the same signs. The data therefore identify the sum of the absolute values of negative IV weights and whether common direction and Rank-1 hold. Separate identification of $w_k(x)$ and $p_k(x, z)$ is unnecessary for either conclusion.*

With $\pi_k(x, z)$ fixed, choosing another admissible value of $w_k(x)$ changes all treatment propensities in regime k by the same positive factor. Multiplying a row of P_x by a positive constant changes neither the signs of the propensity differences nor the rank of the centered matrix. Both restrictions can therefore be evaluated from Π_x, which is identified using the outcome mixture among treated cases and the overall treatment propensities.

Proof of Corollary B.1. When the outcome is defined to be zero without treatment ($Y(0) \equiv 0$), the outcome mixture among treated cases identifies $\pi_k(x,z)$ and $p(x,z) = \sum_k \pi_k(x,z)$. Proposition 3.1 gives

$$\omega_k^*(x;z,z') = \frac{w_k(x)\{p_k(x,z) - p_k(x,z')\}}{p(x,z) - p(x,z')} = \frac{\pi_k(x,z) - \pi_k(x,z')}{p(x,z) - p(x,z')}.$$

Since $w_k(x) > 0$ for every k,

$$\begin{aligned} \Pi_x C_L &= \operatorname{diag}\{w_1(x), \ldots, w_{K_0}(x)\} P_x C_L, \\ \operatorname{rank}(\Pi_x C_L) &= \operatorname{rank}(P_x C_L), \\ \operatorname{sgn}\{\pi_k(x,z) - \pi_k(x,z')\} &= \operatorname{sgn}\{p_k(x,z) - p_k(x,z')\}. \end{aligned}$$

Thus the identified $\pi_k(x,z)$ suffice to calculate the Wald weights and their negative mass and to check common direction and Rank-1. □

B.2 Matching Potential Outcome Distributions across Treatment Statuses

The matching condition depends on changes across instrument values in the joint probabilities of treatment status and component membership. Fix x and suppress it in the notation. For instrument values $z_1, \ldots, z_L$, let

$$t_k(z) = p(z)\omega_{1k}(z), \qquad u_\ell(z) = \{1 - p(z)\}\omega_{0\ell}(z)$$

be the joint probabilities of treatment status and component membership in the treated and untreated groups. These probabilities are zero whenever the treatment group has zero probability. Write $t_k = (t_k(z_1), \ldots, t_k(z_L))'$ and $u_\ell = (u_\ell(z_1), \ldots, u_\ell(z_L))'$, and center these vectors as

$$t_k^c = t_k - \bar{t}_k \mathbf{1}_L, \qquad u_\ell^c = u_\ell - \bar{u}_\ell \mathbf{1}_L,$$

where $\bar{t}_k = L^{-1}\sum_{j=1}^L t_k(z_j)$ and $\bar{u}_\ell = L^{-1}\sum_{j=1}^L u_\ell(z_j)$.

Proposition B.2 (Unique pairing of potential outcome distributions across treatment statuses)**.** *Suppose the outcome distributions for both treatment statuses identify K component distributions and their vectors of joint probabilities across instrument values, up to separate permutations for each treatment status, and each vector is nonzero at some instrument value. A permutation σ of the labels for the distributions under no treatment satisfies the accounting identities for shares invariant across instrument values if and only if*

$$t_k^c = -u_{\sigma(k)}^c \qquad \text{for every } k. \tag{B.3}$$

If at least one such permutation exists, it is unique if and only if the centered vectors of joint probabilities $t_1^c, \ldots, t_K^c$ are pairwise distinct.

When $t_i^c = t_j^c$ for distinct indices i and j, one can exchange the distributions and weights paired with them under no treatment without violating assignment invariance. This exchange can remain possible even when outcome distributions differ across regimes. Proposition B.2 concerns matching at a fixed covariate value using assignment invariance. Auxiliary information under the conditions in Section 4.2 can identify the matching directly.

Proof of Proposition B.2. For a permutation σ,

$$\begin{gathered} t_k^c = -u_{\sigma(k)}^c \quad \Longleftrightarrow \\ t_k(z) + u_{\sigma(k)}(z) = \bar{t}_k + \bar{u}_{\sigma(k)} =: w_k \quad \text{for every } z. \end{gathered}$$

Nonnegativity and the positive-probability condition give

$$\begin{gathered} w_k > 0, \qquad 0 \le t_k(z) \le w_k, \\ \sum_k w_k = \sum_k t_k(z) + \sum_k u_{\sigma(k)}(z) = p(z) + 1 - p(z) = 1. \end{gathered}$$

Thus the probabilities $p_k(z) := t_k(z)/w_k \in [0,1]$ satisfy

$$w_k p_k(z) = t_k(z), \qquad w_k\{1 - p_k(z)\} = w_k - t_k(z) = u_{\sigma(k)}(z).$$

Suppose σ and $\widetilde{\sigma}$ are valid. Then

$$t_k^c = -u_{\widetilde{\sigma}(k)}^c = t_{\sigma^{-1}(\widetilde{\sigma}(k))}^c.$$

If the vectors $t_1^c, \ldots, t_K^c$ are pairwise distinct, this equality implies

$$\sigma^{-1}(\widetilde{\sigma}(k)) = k \quad \Longrightarrow \quad \widetilde{\sigma}(k) = \sigma(k) \qquad \text{for every } k.$$

If instead $t_i^c = t_j^c$ for $i \neq j$, define

$$\widetilde{\sigma}(k) = \begin{cases} \sigma(j), & k = i, \\ \sigma(i), & k = j, \\ \sigma(k), & k \notin \{i, j\}. \end{cases}$$

Then $\widetilde{\sigma} \neq \sigma$ and $u_{\widetilde{\sigma}(k)}^c = -t_k^c$ for every k, so (B.3) admits a second valid permutation. □

B.3 Identification from Variation in Joint Probabilities of Treatment and Regime Membership

Sufficiently rich variation in the joint probabilities of treatment and regime membership identifies the regime-specific average treatment effects directly from changes in the observed outcome mean. This argument takes the joint probabilities $\pi_k(z)$ as identified up to a permutation common across instrument values, as they are when the outcome mixture among treated cases is identified. Let $\mu_{dk} := \mathbb{E}[Y(d) \mid X = x, S = k]$, $m(z) := \mathbb{E}[Y \mid X = x, Z = z]$, and define

$$\Pi := \big(\pi_k(z_\ell)\big)_{k,\ell}, \qquad C_L := I_L - L^{-1}\mathbf{1}_L\mathbf{1}_L', \qquad \mathbf{m} := \big(m(z_1), \ldots, m(z_L)\big)'.$$

Proposition B.3 (Average treatment effects within regimes from variation in joint probabilities of treatment and regime membership)**.** *Fix x and instrument values $z_1, \ldots, z_L$ in its conditional support. Suppose Assumptions 2.1–2.3 hold, $\mathbb{E}[|Y(d)| \mid X = x] < \infty$ for $d = 0, 1$, and*

$$\mathbb{E}[Y(d) \mid D(z) = d, X = x, S = k] = \mu_{dk} \tag{B.4}$$

whenever the selection event has positive conditional probability. If Π is identified up to one common permutation and $\operatorname{rank}(\Pi C_L) = K_0$, the vector of regime-specific average treatment effects $\tau = (\mu_{1k} - \mu_{0k})_{k=1}^{K_0}$ is identified with the same regime labeling as Π by

$$\tau = (\Pi C_L \Pi')^{-1} \Pi C_L \mathbf{m}. \tag{B.5}$$

If the mean potential outcomes under treatment μ_{1k} are also identified, then $\mu_{0k} = \mu_{1k} - \tau_k$ is identified for each regime.

The mean restriction gives

$$\begin{aligned} m(z) &= \sum_k w_k \mu_{0k} + \sum_k \pi_k(z)\tau_k, \\ m(z) - m(z_0) &= \sum_k \{\pi_k(z) - \pi_k(z_0)\}\tau_k. \end{aligned} \tag{B.6}$$

The rank condition requires at least $K_0 + 1$ instrument values and stronger variation than pairwise distinct centered vectors t_k^c sufficient for matching two separately identified outcome mixtures for the two treatment statuses. For example, two regimes with distinct, proportional vectors t_k^c can be matched even though their centered matrix of joint probabilities has rank one. The inversion in (B.5) therefore fails under exact Rank-1 proportionality when $K_0 > 1$, and cannot resolve a pairing ambiguity caused by identical vectors t_k^c.

Under the distributional equality in Assumption 2.4, the same rank condition also identifies every distribution F_{0k} from changes in the observed joint probabilities $\Pr(Y \le y, D = 0 \mid X =$

$x, Z = z$). Corollary B.2 recovers these distributions on the labels identified from the outcome distributions among treated cases without a Gaussian or negative binomial specification for F_{0k}.

Proof of Proposition B.3. Fix x and suppress it in the notation. Instrument exogeneity, consistency, and (B.4) give

$$\mathbb{E}[Y\mathbf{1}\{D = d\} \mid Z = z, S = k] = \Pr(D = d \mid Z = z, S = k)\mu_{dk}.$$

If an arm has zero probability in a regime, both sides are zero. Averaging over the instrument-invariant shares yields

$$\begin{aligned} m(z) &= \sum_k w_k\{p_k(z)\mu_{1k} + (1 - p_k(z))\mu_{0k}\} \\ &= b + \sum_k \pi_k(z)\tau_k, \qquad b := \sum_k w_k\mu_{0k}. \end{aligned}$$

Since $C_L\mathbf{1}_L = 0$ and $C_L = C_L' = C_L^2$,

$$\begin{aligned} \mathbf{m} &= b\mathbf{1}_L + \Pi'\tau, \\ C_L\mathbf{m} &= C_L\Pi'\tau, \\ \Pi C_L\mathbf{m} &= \Pi C_L\Pi'\tau, \qquad \Pi C_L\Pi' = (\Pi C_L)(\Pi C_L)'. \end{aligned}$$

The rank condition gives, for every nonzero $a \in \mathbb{R}^{K_0}$,

$$a'\Pi C_L\Pi' a = \|(\Pi C_L)'a\|^2 > 0, \qquad \tau = (\Pi C_L\Pi')^{-1}\Pi C_L\mathbf{m},$$

which proves (B.5). For a permutation matrix P,

$$\{(P\Pi)C_L(P\Pi)'\}^{-1}(P\Pi)C_L\mathbf{m} = P(\Pi C_L\Pi')^{-1}\Pi C_L\mathbf{m} = P\tau.$$

Thus the effects have the same labeling as Π. If μ_{1k} is identified, then $\mu_{0k} = \mu_{1k} - \tau_k$ is identified as well. □

For the distributional result, define the observed joint probabilities of an outcome below y and no treatment

$$H_{0z}(y) := \Pr(Y \le y, D = 0 \mid X = x, Z = z), \qquad \mathbf{H}_0(y) := \Big(H_{0z_1}(y), \ldots, H_{0z_L}(y)\Big)',$$

and let $\mathbf{F}_0(y) := (F_{01}(y \mid x), \ldots, F_{0K_0}(y \mid x))'$. These joint probabilities are defined even at instrument values where $\Pr(D = 0 \mid X = x, Z = z) = 0$.

Corollary B.2 (Potential outcome distributions under no treatment)**.** *Fix x and instrument values $z_1, \ldots, z_L$. Suppose Assumptions 2.1–2.4 hold, the matrix Π of joint probabilities of treatment and regime membership is identified up to a common permutation, and* $\operatorname{rank}(\Pi C_L) = K_0$*. Then the potential outcome distributions under no treatment are identified with the same regime labeling as Π, with*

$$\mathbf{F}_0(y) = -(\Pi C_L \Pi')^{-1} \Pi C_L \mathbf{H}_0(y) \qquad \text{for every } y. \tag{B.7}$$

If these distributions are linearly independent as finite measures, the regime shares w_k and propensities $p_k(z) = \pi_k(z)/w_k$ are also identified on this instrument support.

Proof. The mixture representation for joint probabilities is

$$\begin{aligned} H_{0z}(y) &= \sum_k \{w_k - \pi_k(z)\} F_{0k}(y \mid x), \\ \mathbf{H}_0(y) &= \mathbf{1}_L w' \mathbf{F}_0(y) - \Pi' \mathbf{F}_0(y). \end{aligned}$$

Since $C_L \mathbf{1}_L = 0$ and $\Pi C_L \Pi'$ is invertible,

$$\begin{aligned} \Pi C_L \mathbf{H}_0(y) &= -\Pi C_L \Pi' \mathbf{F}_0(y), \\ \mathbf{F}_0(y) &= -(\Pi C_L \Pi')^{-1} \Pi C_L \mathbf{H}_0(y), \end{aligned}$$

proving (B.7). For any fixed z_ℓ,

$$H_{0z_\ell}(y) + \sum_k \pi_k(z_\ell) F_{0k}(y \mid x) = \sum_k w_k F_{0k}(y \mid x)$$

is identified. If another share vector $\widetilde{w}$ satisfies this equation, then

$$\sum_k (w_k - \widetilde{w}_k) F_{0k}(y \mid x) = 0 \quad \text{for every } y \quad \Longrightarrow \quad w = \widetilde{w}$$

by linear independence of the component measures. Assumption 2.1 gives $w_k > 0$, so $p_k(z) = \pi_k(z)/w_k$ is also identified. □

Centering removes the regime shares from the equations for F_{0k}, so the rank condition identifies these distributions even if they coincide across regimes. Together with identified F_{1k} and finite potential outcome means, this suffices for each regime-specific average treatment effect. Recovering the shares from the equation in levels requires more: when all F_{0k} are equal, that equation depends on the shares only through $\sum_k w_k = 1$ and leaves the ambiguity in Proposition B.1. To obtain the population average treatment effect, one must therefore identify the shares or impose other restrictions that identify the weighted average of the regime-specific average treatment effects.

C Identification of Regimes with Gaussian Subcomponents

A regime's outcome distribution can contain several Gaussian subcomponents while retaining the same treatment propensity for all its subcomponents. Proposition 4.1 covers a Gaussian potential outcome distribution in each regime. The result below identifies a regime by grouping subcomponents whose weights vary proportionally across instrument values.

Suppose that each regime-specific potential outcome distribution is a finite Gaussian mixture. Standard mixture identification recovers the Gaussian subcomponents and their weights wherever they have positive probability (Teicher, 1963; Yakowitz and Spragins, 1968). Since subcomponents have fixed shares within a regime, their observed mixture weights vary proportionally across assignments. Grouping these proportional vectors identifies the regimes when the vectors for different regimes are nonproportional, and thus distinguishes additional outcome components from additional regimes. This pairwise nonproportionality condition is weaker, for $K_0 \geq 3$, than the full row rank required by the span argument.

Theorem C.1 (Identification with Gaussian mixtures within regimes)**.** *Fix a treatment status d and suppose each regime-specific potential outcome density is a finite Gaussian mixture*

$$f_{dk}(y) = \sum_{a=1}^{M_{dk}} \alpha_{dka}\, \phi(y; \theta_{dka}), \qquad \theta_{dka} = (\mu_{dka}, \sigma^2_{dka}),$$

where $\alpha_{dka} > 0$ and $\sum_a \alpha_{dka} = 1$, so that the observed outcome distribution conditional on $D = d$ can be written as $g_{dz}(y) = \sum_{k=1}^{K_0} \sum_{a=1}^{M_{dk}} \alpha_{dka}\, \omega_{dk}(z)\, \phi(y; \theta_{dka})$. Assume:

(i) *each Gaussian subcomponent belongs to one regime and has parameters θ_{dka} different from those of every other subcomponent;*

(ii) *within each regime, the subcomponent shares α_{dka} are the same at every instrument value;*

(iii) *no regime's vector of mixture weights $\omega_{dk}(\cdot)$ is proportional to that of another regime; and*

(iv) *each subcomponent appears with positive weight in the observed mixture for at least one instrument value.*

Then K_0, the regime-specific potential outcome densities $\{f_{dk}\}$, and the vectors of mixture weights $\{\omega_{dk}(\cdot)\}$ are identified from $\{g_{dz}\}_{z \in \mathcal{Z}}$ up to relabeling. If both treatment statuses satisfy these conditions, independence of regime membership and assignment (Assumption 2.2) *and a unique cross-arm pairing also identify w_k, $\pi_k(z) = p(z)\omega_{1k}(z)$, and $p_k(z)$.*

Proof of Theorem C.1. Fix the treatment status d and suppress its index. Index the individual Gaussian subcomponents by $r = (k, a)$ and write

$$\lambda_r(z) := \alpha_{ka}\omega_k(z), \qquad g_z(y) = \sum_r \lambda_r(z)\phi(y; \theta_r).$$

At each instrument value, finite Gaussian mixture identification recovers all parameters of subcomponents with positive weight and their weights up to permutation (Teicher, 1963; Yakowitz and Spragins, 1968). Every subcomponent has positive weight somewhere, so

$$\{\theta_r : r\} = \bigcup_{z \in \mathcal{Z}} \{\theta_r : \lambda_r(z) > 0\}$$

is identified. Distinct Gaussian densities are linearly independent, as shown in Step 1 of the proof of Proposition 4.1. Hence, at every z,

$$\begin{aligned} &\sum_r \{\lambda_r(z) - \widetilde{\lambda}_r(z)\}\phi(y; \theta_r) = 0 \quad \text{for every } y \\ &\qquad \Longrightarrow \quad \lambda_r(z) = \widetilde{\lambda}_r(z) \quad \text{for every } r, \end{aligned}$$

including subcomponents with zero weight.

For $r = (k, a)$ and $s = (k, b)$ in the same regime,

$$\lambda_r(\cdot) = \frac{\alpha_{ka}}{\alpha_{kb}}\lambda_s(\cdot).$$

For $r = (k, a)$ and $s = (j, b)$ in different regimes, proportionality would give

$$\lambda_r(\cdot) = c\lambda_s(\cdot) \quad \Longrightarrow \quad \omega_k(\cdot) = \frac{c\alpha_{jb}}{\alpha_{ka}}\omega_j(\cdot),$$

contrary to assumption. Thus the equivalence classes $C_1, \ldots, C_{K_0}$ of proportional subcomponent weight vectors are the regimes. Within each class,

$$\begin{aligned} \sum_{r \in C_k} \lambda_r(z) = \omega_k(z) \sum_{r \in C_k} \alpha_r = \omega_k(z), \\ \alpha_r = \frac{\lambda_r(z)}{\omega_k(z)} \qquad \text{whenever } \omega_k(z) > 0, \\ f_k(y) = \sum_{r \in C_k} \alpha_r \phi(y; \theta_r). \end{aligned}$$

All terms on the right are identified.

If both treatment statuses satisfy the assumptions, apply the unique cross-arm pairing to their identified components. The accounting identities

$$w_k = p(z)\omega_{1k}(z) + \{1 - p(z)\}\omega_{0k}(z), \qquad \pi_k(z) = p(z)\omega_{1k}(z), \qquad p_k(z) = \frac{\pi_k(z)}{w_k}$$

then identify the common regime shares, joint probabilities of treatment and regime membership, and propensities. □

Remark C.1 (Estimation and scope)**.** With fixed numbers of regimes and subcomponents, the Gaussian subcomponent model is a special case of the joint likelihood in (H.1), so Theorem H.1 establishes consistency under Assumption H.1. The assumption imposes compactness, positive mixture weights, component separation and identification, including unique cross-arm pairing. The inference results in Appendix I also hold both numbers fixed.

The subcomponent shares must be fixed across assignments within each regime for this grouping argument to apply. A Gaussian subcomponent also cannot be shared by two regimes: its observed weight vector would add their contributions, which proportionality alone cannot separate. The grouping must be done within each treatment status, after which the regimes under treatment and no treatment still need to be paired. Using common labels in the likelihood imposes such a pairing, but its identification still requires the conditions in Section 4.1. Nor does good predictive fit ensure correct regime-specific potential outcome distributions under misspecification (Remark H.1).

D Identification Using Pretreatment Measurements

To prove Theorem 4.3 and Proposition 4.2, I first identify the measurement distributions and regime shares by simultaneous diagonalization, following Bonhomme, Jochmans and Robin (2016).

D.1 Setup and Assumptions

Throughout this appendix, I condition on $X = x$ for a fixed covariate value and suppress x except where needed. In addition to (Y, D, X, Z), the data contain three auxiliary pretreatment measurements

$$R = (R^{(1)}, R^{(2)}, R^{(3)}),$$

with all three observed for every case in the sample. They exist at assignment and are observed before the decision, and can be scalar or vector-valued, including administrative variables or case file text. For each r, the measurement space $(\mathcal{R}_r, \mathcal{B}_r)$ is standard Borel. The

identification argument uses no topology on $\mathcal{R}_r$. Define the regime-specific distribution of each measurement by

$$F_k^{(r)}(B) \;=\; \Pr\Big(R^{(r)} \in B \mid S = k, X = x\Big), \qquad B \in \mathcal{B}_r, \quad k = 1, \ldots, K_0, \quad r = 1, 2, 3.$$

These distributions carry no treatment status index because the measurements are observed before treatment. For the designated measurement in Assumption 4.5, write $Q := R^{(r^*)}$ and $F_k^Q := F_k^{(r^*)}$.

In addition to Assumptions 4.3, 4.4, and 4.5 of Section 4.2, I impose the following condition on the regime-specific measurement distributions.

Assumption D.1 (Linear independence and distinctness of measurement distributions)**.** *After reordering the measurements, the measures $F_1^{(r)}, \ldots, F_{K_0}^{(r)}$ are linearly independent as finite signed measures on $(\mathcal{R}_r, \mathcal{B}_r)$ for $r = 1, 2$, and the measures $F_1^{(3)}, \ldots, F_{K_0}^{(3)}$ are pairwise distinct.*

D.2 Mixture Representations

Under Assumption 4.3,

$$\Pr(R \in \cdot \mid X = x) = \sum_{k=1}^{K_0} w_k(x) \bigotimes_{r=1}^{3} F_k^{(r)}(\cdot \mid x).$$

For measurable sets $B_r \subseteq \mathcal{R}_r$, conditional independence within regimes gives

$$\Pr\left(\bigcap_{r=1}^{3} \{R^{(r)} \in B_r\} \,\middle|\, X = x\right) = \sum_k w_k(x) \prod_{r=1}^{3} F_k^{(r)}(B_r \mid x).$$

Write $w_k(x, z) := \Pr(S = k \mid Z = z, X = x)$. Under Assumption 4.4, for every z with positive conditional probability,

$$\Pr\left(\bigcap_{r=1}^{3} \{R^{(r)} \in B_r\} \,\middle|\, Z = z, X = x\right) = \sum_k w_k(x, z) \prod_{r=1}^{3} F_k^{(r)}(B_r \mid x).$$

These equalities on measurable rectangles establish the product-mixture representations before and after conditioning on assignment. If also $S \perp Z \mid X$, then

$$w_k(x, z) = \Pr(S = k \mid Z = z, X = x) = \Pr(S = k \mid X = x) = w_k(x),$$

so the second mixture is invariant in z.

Conditioning on treatment can change the measurement distributions within a regime. Assumption 4.5 rules out this change for the designated measurement Q; it imposes no product representation for all three measurements conditional on treatment and assignment.

D.3 Identification of Regime-Specific Measurement Distributions

The first lemma constructs finite vectors of test functions that preserve the linear independence of the first two measurements' component distributions. The second constructs one bounded function whose expectation distinguishes every pair of regimes in the third measurement.

Lemma D.1 (Nonsingular matrices of component moments)**.** *Let $\mu_1, \dots, \mu_K$ be linearly independent probability measures on a measurable space $(\mathcal{R}, \mathcal{B})$. Then there exist bounded measurable functions $\varphi_1, \dots, \varphi_K$, with $\varphi_1 \equiv 1$ and $\varphi_j = \mathbf{1}_{A_j}$ for $j \geq 2$, such that*

$$A := \left[\int \varphi_j \, d\mu_k\right]_{j,k=1}^{K}$$

is nonsingular.

Proof. For $B \in \mathcal{B}$, let $v(B) := (\mu_1(B), \dots, \mu_K(B))^\top$ and $\mathcal{V} := \operatorname{span}\{v(B) : B \in \mathcal{B}\} \subseteq \mathbb{R}^K$. If $\mathcal{V} \neq \mathbb{R}^K$, choose $a \in \mathcal{V}^\perp \setminus \{0\}$ and define $\nu := \sum_k a_k \mu_k$. Then

$$\nu(B) = \sum_k a_k \mu_k(B) = a^\top v(B) = 0 \quad \text{for every } B \in \mathcal{B},$$

so $\nu = 0$, contrary to linear independence. Thus $\mathcal{V} = \mathbb{R}^K$. Because $v(\mathcal{R}) = \mathbf{1}_K \neq 0$, choose $A_2, \dots, A_K$ so that $v(\mathcal{R}), v(A_2), \dots, v(A_K)$ form a basis. With $\varphi_1 \equiv 1$ and $\varphi_j = \mathbf{1}_{A_j}$,

$$A = \begin{pmatrix} v(\mathcal{R})^\top \\ v(A_2)^\top \\ \vdots \\ v(A_K)^\top \end{pmatrix}, \qquad \operatorname{rank}(A) = K.$$

□

Lemma D.2 (Distinct expectations from distinct measures)**.** *Let $\mu_1, \dots, \mu_K$ be pairwise distinct probability measures on a measurable space. There is a bounded measurable function h, expressible as a finite linear combination of indicator functions, such that $\int h \, d\mu_1, \dots, \int h \, d\mu_K$ are pairwise distinct.*

Proof. For $K = 1$, take $h \equiv 0$. For $K \geq 2$, choose a measurable set for each pair of measures on which their probabilities differ, and enumerate these sets as $E_1, \dots, E_J$. The vectors

$c_k := (\mu_k(E_j))_{j=1}^J$ are pairwise distinct. A finite union of proper hyperplanes cannot cover $\mathbb{R}^J$, so choose

$$\eta \notin \bigcup_{k<\ell} \{u \in \mathbb{R}^J : (c_k - c_\ell)^\top u = 0\}, \qquad h = \sum_{j=1}^J \eta_j \mathbf{1}_{E_j}.$$

Then h is bounded and measurable, and

$$\int h \, d\mu_k - \int h \, d\mu_\ell = \sum_{j=1}^J \eta_j \{\mu_k(E_j) - \mu_\ell(E_j)\} = (c_k - c_\ell)^\top \eta \neq 0 \qquad (k \neq \ell).$$

□

Proof of Theorem 4.3

Write $K = K_0$ and suppress x. Order the measurements as in Assumption D.1. By Lemma D.1, for $r = 1, 2$ there are K-vectors of bounded functions $\varphi^{(r)}$ with first coordinate equal to one such that

$$A = \left[\int \varphi_j^{(1)} dF_k^{(1)}\right]_{j,k}, \quad B = \left[\int \varphi_\ell^{(2)} dF_k^{(2)}\right]_{\ell,k}$$

are nonsingular. With $e_1 = (1, 0, \ldots, 0)^\top$, their normalizations are

$$e_1^\top A = e_1^\top B = \mathbf{1}_K^\top.$$

For a bounded measurable function g on the third measurement, define the observable $K \times K$ matrix

$$M_g := \mathbb{E}\left[\varphi^{(1)}(R^{(1)})\varphi^{(2)}(R^{(2)})^\top g(R^{(3)}) \mid X = x\right].$$

Conditional independence gives

$$M_g = A \operatorname{diag}\left(w_k \int g \, dF_k^{(3)}\right) B^\top. \tag{D.1}$$

Step 1: identification of the minimal number of components. Taking $g \equiv 1$ gives

$$M_0 := \mathbb{E}\left[\varphi^{(1)}(R^{(1)})\varphi^{(2)}(R^{(2)})^\top \mid X = x\right] = A \operatorname{diag}(w) B^\top.$$

Since $w_k > 0$ and A, B are nonsingular, $\operatorname{rank}(M_0) = K$. Any representation with K' product components gives vectors a_q, b_q and weights v_q satisfying

$$M_0 = \sum_{q=1}^{K'} v_q a_q b_q^\top,$$

$$K = \operatorname{rank}(M_0) \leq \sum_{q=1}^{K'} \operatorname{rank}(v_q a_q b_q^\top) \leq K'.$$

The maintained representation has K components and is therefore minimal. For any finite vectors of bounded functions $\psi^{(1)}$ and $\psi^{(2)}$,

$$\operatorname{rank}\Big(\mathbb{E}[\psi^{(1)}(R^{(1)})\psi^{(2)}(R^{(2)})^\top \mid X = x]\Big) \le K = \operatorname{rank}(M_0).$$

Thus the largest attainable rank identifies K.

Step 2: identification of regime-specific moments and population shares. Consider any other positive weight representation with exactly K product components,

$$\Pr(R \in \cdot \mid X = x) = \sum_{q=1}^{K} v_q G_q^{(1)} \otimes G_q^{(2)} \otimes G_q^{(3)}.$$

Applying the first two measurements' test functions gives matrices $\widetilde{A}, \widetilde{B}$, with

$$\begin{aligned} M_0 &= \widetilde{A}\,\operatorname{diag}(v)\widetilde{B}^\top, \\ K = \operatorname{rank}(M_0) &\le \min\{\operatorname{rank}(\widetilde{A}), \operatorname{rank}(\widetilde{B})\} \le K. \end{aligned}$$

Thus both matrices are nonsingular. For $K = 1$, $w_1 = v_1 = 1$ and $F_1^{(r)} = \Pr(R^{(r)} \in \cdot \mid X = x) = G_1^{(r)}$ for each r, so suppose $K \ge 2$.

Lemma D.2 gives a bounded function h with pairwise distinct $\rho_k := \int h\, dF_k^{(3)}$. By (D.1),

$$\begin{aligned} N := M_h M_0^{-1} &= A\,\operatorname{diag}(\rho_1, \ldots, \rho_K)A^{-1} \\ &= \widetilde{A}\,\operatorname{diag}(\widetilde{\rho}_1, \ldots, \widetilde{\rho}_K)\widetilde{A}^{-1}, \\ \widetilde{\rho}_q &:= \int h\, dG_q^{(3)}. \end{aligned}$$

The distinct eigenvalues give unique one-dimensional eigenspaces. After one common permutation of the candidate components, there are nonzero scalars c_k such that

$$\widetilde{A}_{\cdot k} = c_k A_{\cdot k}, \qquad 1 = e_1^\top \widetilde{A}_{\cdot k} = c_k e_1^\top A_{\cdot k} = c_k.$$

Thus $\widetilde{A} = A$. The unweighted matrix and the first-coordinate normalizations now give

$$\begin{aligned} A^{-1}M_0 &= \operatorname{diag}(w)B^\top = \operatorname{diag}(v)\widetilde{B}^\top, \\ A^{-1}M_0 e_1 &= \operatorname{diag}(w)\mathbf{1}_K = w = v, \\ \operatorname{diag}(w)^{-1}A^{-1}M_0 &= B^\top = \widetilde{B}^\top. \end{aligned}$$

The shares and both component moment matrices therefore agree under the same labeling.

Step 3: recovery of the full component measures. For any bounded measurable g on the third measurement,

$$A^{-1}M_g(B^\top)^{-1} = \operatorname{diag}\Big(w_k \int g\, dF_k^{(3)}\Big) = \operatorname{diag}\Big(w_k \int g\, dG_k^{(3)}\Big),$$
$$\int g\, dF_k^{(3)} = \frac{\{A^{-1}M_g(B^\top)^{-1}\}_{kk}}{w_k} = \int g\, dG_k^{(3)}.$$

For a bounded measurable function f on the first measurement, conditional independence gives

$$\begin{aligned}\mathbb{E}[\varphi^{(2)}(R^{(2)})f(R^{(1)}) \mid X = x] &= B\operatorname{diag}(w)\left(\int f\, dF_k^{(1)}\right)_{k=1}^K,\\ \left(\int f\, dF_k^{(1)}\right)_{k=1}^K &= \operatorname{diag}(w)^{-1}B^{-1}\mathbb{E}[\varphi^{(2)}(R^{(2)})f(R^{(1)}) \mid X = x]\\ &= \left(\int f\, dG_k^{(1)}\right)_{k=1}^K.\end{aligned}$$

For a bounded measurable function f on the second measurement,

$$\begin{aligned}\mathbb{E}[\varphi^{(1)}(R^{(1)})f(R^{(2)}) \mid X = x] &= A\operatorname{diag}(w)\left(\int f\, dF_k^{(2)}\right)_{k=1}^K,\\ \left(\int f\, dF_k^{(2)}\right)_{k=1}^K &= \operatorname{diag}(w)^{-1}A^{-1}\mathbb{E}[\varphi^{(1)}(R^{(1)})f(R^{(2)}) \mid X = x]\\ &= \left(\int f\, dG_k^{(2)}\right)_{k=1}^K.\end{aligned}$$

Taking indicators of measurable sets gives, under the same permutation,

$$F_k^{(r)}(B) = G_k^{(r)}(B) \qquad \text{for every } B \in \mathcal{B}_r, \quad k = 1, \dots, K, \quad r = 1, 2, 3. \qquad \square$$

D.4 Recovery of Treatment Propensities and Potential Outcome Distributions

A weaker restriction on Q. Retain the conditions of Proposition 4.2, but replace Assumption 4.5 by a restriction on conditional means. Fix bounded measurable functions $\varphi^* = (\varphi_1^*, \ldots, \varphi_{K_0}^*)^\top$ of Q, with $\varphi_1^* \equiv 1$, such that

$$A^* := \Big[\mathbb{E}[\varphi_j^*(Q) \mid S = k, X = x]\Big]_{j,k=1}^{K_0}$$

is nonsingular, and suppose that

$$\begin{gathered}\operatorname{Cov}(\varphi^*(Q), D \mid S = k, X = x, Z = z) = 0 \\ \text{for every } k \text{ and every } z \text{ with } \Pr(Z = z \mid X = x) > 0.\end{gathered}$$

Both parts of Proposition 4.2 continue to hold. Under Assumption 4.5, the covariances are zero for any bounded functions, and Lemma D.1 supplies functions that make A^* nonsingular, so the proof below also covers the original proposition.

Proof of Proposition 4.2

Part (i). Theorem 4.3 identifies $w(x)$ and the measurement distributions $F_1^Q, \ldots, F_{K_0}^Q$, and hence A^*, under one labeling of the regimes. Let $\pi(x, z) := (\pi_1(x, z), \ldots, \pi_{K_0}(x, z))^\top$, with $\pi_k(x, z) = \Pr(S = k, D = 1 \mid Z = z, X = x)$. Within regime k, Assumption 4.4 and the zero covariances give

$$\mathbb{E}[\varphi^*(Q)D \mid S = k, Z = z, X = x] = p_k(x, z)\,\mathbb{E}[\varphi^*(Q) \mid S = k, X = x].$$

Averaging over regimes,

$$\mathbb{E}[\varphi^*(Q)D \mid Z = z, X = x] = \sum_{k=1}^{K_0} \pi_k(x, z)\,\mathbb{E}[\varphi^*(Q) \mid S = k, X = x] = A^*\pi(x, z).$$

Since A^* is nonsingular, the joint probabilities are identified, and Assumption 2.2 gives the treatment propensities:

$$\pi(x, z) = (A^*)^{-1}\,\mathbb{E}[\varphi^*(Q)D \mid Z = z, X = x], \qquad p_k(x, z) = \frac{\pi_k(x, z)}{w_k(x)}.$$

Bayes' rule then gives the mixture weights in each treatment arm with positive conditional probability:

$$\omega_{1k}(x,z) = \frac{\pi_k(x,z)}{p(x,z)} \qquad \text{if } p(x,z) > 0,$$
$$\omega_{0k}(x,z) = \frac{w_k(x) - \pi_k(x,z)}{1 - p(x,z)} \quad \text{if } p(x,z) < 1.$$

The labels are common across all combinations of treatment status and assignment because the same identified A^* applies to every cell.

Part (ii). Fix $d \in \{0,1\}$ and instrument values $z_1, \ldots, z_L$ with positive probabilities $\Pr(D = d, Z = z_\ell \mid X = x)$. Let

$$\Omega_d(x) := [\omega_{dk}(x, z_\ell)]_{k=1,\ldots,K_0;\, \ell=1,\ldots,L},$$
$$F^Y_{dk}(\cdot \mid x) := \Pr(Y(d) \in \cdot \mid S = k, X = x),$$
$$F^Y_{dk}((-\infty, y] \mid x) = F_{dk}(y \mid x).$$

Under consistency and Assumptions 2.3 and 2.4, Lemma 2.1 gives, for each ℓ,

$$\Pr(Y \in \cdot \mid D = d, Z = z_\ell, X = x) = \sum_{k=1}^{K_0} \omega_{dk}(x, z_\ell) F^Y_{dk}(\cdot \mid x).$$

The weights are identified in part (i). Let G_d be the L-vector of observed outcome distributions conditional on treatment status d and each assignment, and let $F_d = (F^Y_{d1}, \ldots, F^Y_{dK_0})^\top$. Full row rank gives

$$G_d = \Omega_d(x)^\top F_d,$$
$$L_d := \{\Omega_d(x)\Omega_d(x)^\top\}^{-1}\Omega_d(x),$$
$$L_d\Omega_d(x)^\top = I_{K_0}, \qquad F_d = L_d G_d.$$

For any other vector $\widetilde{F}_d$ of finite signed measures satisfying the mixture equations,

$$\Omega_d(x)^\top(\widetilde{F}_d - F_d) = 0 \quad \Longrightarrow \quad \widetilde{F}_d - F_d = L_d\Omega_d(x)^\top(\widetilde{F}_d - F_d) = 0.$$

Equivalently, for every bounded measurable h,

$$\left(\int h\, dF^Y_{dk}\right)_{k=1}^{K_0} = L_d \left(\mathbb{E}[h(Y) \mid D = d, Z = z_\ell, X = x]\right)_{\ell=1}^{L}.$$

In particular,

$$\left(F^Y_{dk}((-\infty, y] \mid x)\right)_{k=1}^{K_0} = L_d\Big(\Pr(Y \le y \mid D = d, Z = z_\ell, X = x)\Big)_{\ell=1}^{L}.$$

The true potential outcome distributions are probability measures satisfying the mixture equations, so the unique solution among signed measures identifies them. The labels are fixed by Theorem 4.3 and common across treatment arms. When both arms satisfy the rank condition and the means are finite,

$$\tau_k(x) = \int y \, dF^Y_{1k}(y \mid x) - \int y \, dF^Y_{0k}(y \mid x).$$

□

Identification requires no separation or parametric restrictions on the outcome distributions and no conditional independence between Y and R beyond the maintained assumptions.

The identified joint probabilities $\pi_k(x, z)$ give the Wald weights, NegMass, and the rank of the centered propensity matrix. Conditional on (S, X, Z), treatment may depend on the other two measurements. It may also depend on Q, provided that, within each regime, treated and untreated cases have the same means of the test functions.

Remark D.1 (Assignment-specific regime shares)**.** Retain the conditions of part (i) of Proposition 4.2, except Assumption 2.2. For every z with $\Pr(Z = z \mid X = x) > 0$, let

$$w_k(x, z) := \Pr(S = k \mid Z = z, X = x), \qquad m^*_{z,j}(x) := \mathbb{E}[\varphi^*_j(Q) \mid Z = z, X = x].$$

Assumption 4.4 gives

$$\Pr(Q \in \cdot \mid Z = z, X = x) = \sum_{k=1}^{K_0} w_k(x, z) F^Q_k.$$

Hence

$$m^*_z = A^* w(x, z), \qquad w(x, z) = (A^*)^{-1} m^*_z.$$

For $w_k(x, z) > 0$, Bayes' rule gives

$$\pi_k(x, z) = w_k(x, z) p_k(x, z), \qquad p_k(x, z) = \frac{\pi_k(x, z)}{w_k(x, z)}.$$

Under consistency and Assumptions 2.3 and 2.4, part (ii) of Proposition 4.2 applies whenever $\operatorname{rank}\{\Omega_d(x)\} = K_0$. Constancy of the recovered $w_k(x, z)$ is therefore an overidentifying implication of independence of regime membership and assignment conditional on covariates within the maintained model for the auxiliary measurements.

With three regimes, a binary Q supplies only one probability and the adding-up restriction for three unknown mixture weights. Pairwise distinct distributions therefore do not suffice for recovering these weights. The distributions of Q must be linearly independent, even when it is the third measurement in Assumption D.1.

The mixture weights remain identified when the component distributions are nearly identical and linearly independent, but small errors in the conditional moments can lead to large errors in the recovered weights. For a fixed normalized collection of test functions, $\|(A^*)^{-1}\|_\infty$ bounds this amplification, though its value depends on the functions chosen.

In estimation, the constant first test function imposes the adding-up normalization, and estimated proportions can be restricted to the simplex using constrained least squares or projection. Proposition 4.2 identifies regime shares and treatment propensities from the auxiliary measurements even when $Y(0) = 0$. For treatment and assignment cells with positive probability, it also identifies $\Omega_d(x)$, so its full row rank can be checked before recovering the potential outcome distributions, including distributions that coincide across regimes. Exact Rank-1 proportionality with regime shares constant across assignments limits $\text{rank}\{\Omega_d(x)\}$ to at most two, ruling out full row rank when $K_0 > 2$. Identification at each x is up to a common permutation of regime labels. Quantities invariant to these permutations can be averaged directly over X, whereas following a particular regime across X requires a labeling restriction that holds throughout the support, such as a common ordering by a statistic of the measurement distributions. Continuity and distinctness alone permit different permutations on disconnected parts of the support.

D.5 Auxiliary Pretreatment Information in Leniency Designs

This section provides a small, stylized example based on a real dataset to illustrate how to implement the nonparametric route proposed in Section 4.2 in practice. The first requirement is that the measurements in Section 4.2 must be determined before decision makers are assigned and mutually independent conditional on the regime and observed covariates. One of the measurements, designated as Q, is required to satisfy an additional restriction: among cases from the same regime and with the same observed characteristics, the cases a decision maker treats must have the same distribution of Q as the cases not treated by the decision maker. A measurement may be a collection of related variables instead of one variable, analogous to the definition of "collection" in Allman, Matias and Rhodes (2009).

An example: Norwegian disability appeals. This illustration is a proposed construction based on the disability insurance appeals studied by Dahl, Kostøl and Mogstad (2014). In that setting, a parent whose initial claim for disability insurance is denied can appeal, and a judge decides whether or not to grant the appeal based solely on written evidence. Within each department of the appeals court, cases are assigned to judges by rotation, without regard to their content. The treatment D is whether the judge grants the parent's appeal and awards disability benefits; Z is the assigned judge; and X contains interacted year and department indicators, within which assignment is random, together with the ages of the parent and child

and the parent's educational attainment. The outcome Y is whether the appellant's adult child receives disability insurance five years after the parent's appeal decision.

In this setting, we can think of each regime as a latent family type. One family type would have more steady work histories, musculoskeletal claims, and children who completed high school, while another would have more intermittent work histories, mental disorders, and higher school dropout rates. The framework recovers these types, the probability that each judge grants the appeal for each type, and each type's average effect of a granted appeal on the child's later benefit receipt, along with the weights IV places on these effects.

To keep things simple, construct a sample by pairing each appellant with one adult child, say the oldest.[8] Suppose these pairs have all the relevant information below or otherwise restrict the sample accordingly. The following are some proposed candidate measurements from the disability insurance appeals data:

- $R^{(1)}$: the parent's attachment to work: the number of years, from zero to five, in which the parent had any earnings during the sixth to tenth years before applying;
- $R^{(2)}$: type of diagnosis on the parent's claim: mental disorders, musculoskeletal disorders or other conditions;[9]
- $R^{(3)} = Q$: the child's educational attainment in the year before the parent's application: less than high school, high school or any college education.

Theorem 4.3 requires the distributions of parent's attachment to work and type of diagnosis to be linearly independent across regimes, and the third measurement to be only pairwise distinct. However, in this example, the third measurement is also designated as Q, so Proposition 4.2 requires its regime-specific distributions to be linearly independent as well. With only three categories for the type of diagnosis and the child's educational attainment, the linear independence restriction holds for at most three regimes. Measurements with richer support can support more latent family types. Since identification holds for a fixed covariate cell, it may fail in some cells even if it holds in others. In cells with young children, for example, few may have college education before the application, which limits the number of regimes Q can pick out in those cells.

Within a regime and covariate cell, the parent's earlier work history, the parent's diagnosis, and the child's education must be mutually independent (Assumption 4.3). Clearly, the parent's work history, the diagnosis and the child's educational attainment are related. A long illness shortens the parent's work history and shows up in the diagnosis. A parent who worked steadily while the child was in school is also more likely to have a child who stayed longer in school. What Assumption 4.3 requires is that this relationship is mediated only

[8] Any other pairing that does not depend on the judge or the appeal outcome would also work.

[9] Dahl, Kostøl and Mogstad (2014) measure appellants' diagnoses before the appeal and include them in their test of random assignment, so the diagnoses do not depend on the judge assignment.

through the covariates and the family's latent type such that controlling for the latent type and the covariates makes the measurements mutually independent.

Assumption 4.5, on the other hand, requires that within a latent family type and covariate cell, the "special" measurement, Q, which is the child's educational attainment in this example, should have the same distribution under every judge for parents that are granted and denied appeal. This effectively requires that the judge should not use the child's educational attainment in the decision within any latent family type. Educational attainment could correlate with other factors the judge uses in the decision, such as the parent's educational attainment, but if those factors are included in X, their threat to the validity of this assumption is neutralized. Notice that the child's education is allowed to predict whether the child later receives disability benefits. The restriction is only about how education relates to the judge's decision on the parent's appeal.

Under Proposition 4.2, the distribution of the child's education among appellants whose appeals are granted and those whose appeals are denied identifies the mixture weights in each treatment arm, and hence the treatment propensities and IV weights. Under Assumption 2.4, observed child benefit rates are weighted averages of regime-specific benefit probabilities, and sufficient variation in the identified mixture weights across judges in each treatment arm identifies those probabilities in both treatment arms. Their difference is the regime-specific average treatment effect.

Other settings. Datasets used in other studies with leniency designs suggest that similar administrative data can be put together in a variety of settings. In US disability insurance, Maestas, Mullen and Strand (2013) observe applicants' annual earnings histories and diagnoses. In pretrial bail, Dobbie, Goldin and Yang (2018) link defendants' prior offenses to tax records on earnings, household income and benefit receipt. In Norway, Bhuller et al. (2020) observe defendants' criminal charges, prison spells, employment, earnings and family circumstances. In Harris County in the US, Mueller-Smith (2015) observes defendants' criminal histories, quarterly earnings and receipt of food stamps and cash assistance. In Michigan child protection, Gross and Baron (2022) observe prior maltreatment investigations together with school attendance and test scores. In New York City eviction cases, Collinson et al. (2024) observe earnings, shelter stays, and hospital visits. These studies document potential sources of the measurements required in Section 4.2. Applying this framework to such datasets requires carefully arguing for the validity of the candidate measurements and the plausibility of the other conditions provided in this paper. This task is left to empirical researchers interested in using the method in such settings.

E Further Results on Treatment Propensities

Throughout this appendix, I hold x fixed.

E.1 Observable Restrictions on Joint Treatment and Outcome Distributions

Define the subdistribution functions, which give the joint probability of an outcome below y and treatment status d:

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

I also write H_{dz} for the associated finite measure on outcome sets. Under Assumptions 2.2–2.4,

$$H_{1z}(y) = \sum_k w_k p_k(z) F_{1k}(y), \qquad H_{0z}(y) = \sum_k w_k \{1 - p_k(z)\} F_{0k}(y).$$

Proposition E.1 (Observable implications of proportionality and common direction)**.** *Fix x and suppose Assumptions 2.2–2.4 hold.*

(i) If (5.2) holds and $\sum_k w_k b_k \neq 0$, then for each arm d there are functions A_d and B_d such that

$$H_{dz}(y) = A_d(y) + B_d(y) p(z)$$

for every y and z. Subtracting the subdistribution under any common reference assignment from those under the other assignments gives a matrix of changes, with instrument values in rows and outcome thresholds in columns. This matrix has rank at most one.

(ii) If the potential outcome distributions under treatment $F_{11}, \ldots, F_{1K_0}$ are linearly independent, then for any reference value z_0,

$$\dim \operatorname{span}\{H_{1z} - H_{1z_0} : z \in \mathcal{Z}\} = \operatorname{rank}(\Pi_x C_L). \tag{E.1}$$

If the aggregate propensity is nonconstant, Rank-1 holds if and only if these subdistribution changes span one dimension.

(iii) Let $z_1, \ldots, z_L$ be ordered so that $p(z_{\ell+1}) \geq p(z_\ell)$. Suppose $p_k(z_{\ell+1}) \geq p_k(z_\ell)$ for every k and ℓ. Then $H_{1,z_{\ell+1}} - H_{1,z_\ell}$ is a nonnegative measure and $H_{0,z_{\ell+1}} - H_{0,z_\ell}$ is a nonpositive measure.

A finite grid recovers the rank in (E.1) only when the chosen thresholds distinguish the linear combinations of the subdistribution changes. Two such combinations may agree on the grid and differ elsewhere, in which case evaluating them only on the grid gives a smaller rank. Knowing the correct rank does not suffice to recover the coefficients of the regime-specific treatment propensities or to calculate the distance statistic used in the subsequent test.

For the treated arm, a nonnegative subdistribution difference at every y does not by itself establish part (iii). If that difference falls between thresholds a and b, the probability difference on $(a, b]$ is negative. Part (iii) therefore has to be checked using interval increments

or a sufficiently rich class of nonnegative test functions, since it applies to every measurable outcome set.

Failure of either observable restriction can also reflect a failure of the distributional equality in Assumption 2.4, which both arguments require. In Section 6, I develop estimation and inference for the identified treatment propensities.

Proof of Proposition E.1. Under the mixture representation,

$$H_{1z}(y) = \sum_k w_k p_k(z) F_{1k}(y), \qquad H_{0z}(y) = \sum_k w_k \{1 - p_k(z)\} F_{0k}(y).$$

For part (i), $p_k(z) = a_k + b_k \lambda(z)$ gives

$$\bar{a} := \sum_k w_k a_k, \qquad \bar{b} := \sum_k w_k b_k \neq 0, \qquad p(z) = \bar{a} + \bar{b}\lambda(z), \qquad \lambda(z) = \frac{p(z) - \bar{a}}{\bar{b}}.$$

Hence $H_{dz}(y) = A_d(y) + B_d(y)p(z)$, where

$$\begin{aligned} A_1(y) &= \sum_k w_k \left(a_k - \frac{b_k \bar{a}}{\bar{b}} \right) F_{1k}(y), & B_1(y) &= \sum_k \frac{w_k b_k}{\bar{b}} F_{1k}(y), \\ A_0(y) &= \sum_k w_k \left(1 - a_k + \frac{b_k \bar{a}}{\bar{b}} \right) F_{0k}(y), & B_0(y) &= -\sum_k \frac{w_k b_k}{\bar{b}} F_{0k}(y). \end{aligned}$$

For any thresholds $y_1, \ldots, y_m$ and reference assignment z_0,

$$\left[H_{dz_\ell}(y_j) - H_{dz_0}(y_j) \right]_{\ell,j} = \begin{pmatrix} p(z_1) - p(z_0) \\ \vdots \\ p(z_L) - p(z_0) \end{pmatrix} \begin{pmatrix} B_d(y_1) & \cdots & B_d(y_m) \end{pmatrix},$$

so this matrix has rank at most one.

For part (ii), linear independence of the component measures gives

$$T : a \mapsto \sum_k a_k F_{1k}, \qquad \ker T = \{0\}, \qquad T\{\pi(z) - \pi(z_0)\} = H_{1z} - H_{1z_0}.$$

Therefore,

$$\begin{aligned} \dim \operatorname{span}\{H_{1z} - H_{1z_0} : z \in \mathcal{Z}\} &= \dim \operatorname{span}\{\pi(z) - \pi(z_0) : z \in \mathcal{Z}\} \\ &= \operatorname{rank}(\Pi_x C_L). \end{aligned}$$

If the aggregate propensity is nonconstant, some z satisfies

$$(H_{1z} - H_{1z_0})(\mathbb{R}) = p(z) - p(z_0) \neq 0, \qquad \operatorname{rank}(\Pi_x C_L) \geq 1.$$

Since $\Pi_x = \mathrm{diag}(w)P_x$ and the regime shares are positive,

$$\mathrm{rank}(\Pi_x C_L) = \mathrm{rank}(P_x C_L),$$

and Proposition 5.1 gives the equivalence with Rank-1.

For part (iii), let $\Delta_{\ell k} = p_k(z_{\ell+1}) - p_k(z_\ell) \geq 0$. For every measurable outcome set B,

$$(H_{1,z_{\ell+1}} - H_{1,z_\ell})(B) = \sum_k w_k \Delta_{\ell k} F_{1k}(B) \geq 0,$$
$$(H_{0,z_{\ell+1}} - H_{0,z_\ell})(B) = -\sum_k w_k \Delta_{\ell k} F_{0k}(B) \leq 0.$$

□

E.2 Decision Makers and Leniency Groups

The instrument can distinguish individual decision makers or group them by a leniency measure. Grouping replaces the treatment propensities in each regime by averages within groups. Opposing or nonproportional changes can cancel under this averaging, so restrictions that fail across individual decision makers may hold across the groups. Let J denote decision maker identity and define the leniency measure and grouped instrument by

$$L_x(j) := \Pr(D = 1 \mid X = x, J = j), \qquad Z = B_x\{L_x(J)\},$$

where B_x partitions the leniency measure into ordered bins. In applications, $L_x(j)$ is often estimated by a leave-one-out treatment rate. Write

$$p_k^J(x, j) := \Pr(D = 1 \mid X = x, S = k, J = j)$$

for the treatment propensity in regime k under assignment to decision maker j.

Proposition E.2 (Preservation of selection restrictions under grouping by leniency)**.** *Suppose $J \perp S \mid X$ and $Z = B_x\{L_x(J)\}$. Then*

$$p_k(x, z) = \mathbb{E}\{p_k^J(x, J) \mid X = x, Z = z\}.$$

If there is a scalar function λ_x of decision maker identity, common to all regimes, such that $p_k^J(x, j) = a_k(x) + b_k(x)\lambda_x(j)$ for every j in the conditional support of J given $X = x$, then

$$p_k(x, z) = a_k(x) + b_k(x)\mathbb{E}\{\lambda_x(J) \mid X = x, Z = z\},$$

so the centered matrix of these treatment propensities has rank at most one, and has rank one whenever the aggregate treatment propensity varies across groups. Further, whenever each $p_k^J(x, j)$ is weakly increasing in $L_x(j)$ and B_x partitions the leniency measure into ordered, nonoverlapping intervals, each $p_k(x, z)$ is weakly increasing across the ordered bins.

The weights used to average decision makers within a group are common across regimes. Indeed, $J \perp S \mid (X, Z)$ follows from $J \perp S \mid X$ because Z is a function of (X, J). Define these within-group assignment weights by

$$W_x(j, z) := \Pr(J = j \mid X = x, Z = z).$$

Let Π_x^J have regimes as rows, decision makers as columns, and joint probabilities of treatment and regime membership $\pi_k^J(x, j)$ as entries. Averaging the columns of Π_x^J with the common assignment weights gives

$$\Pi_x^{\text{group}} = \Pi_x^J W_x, \qquad \operatorname{rank}(\Pi_x^{\text{group}}) \leq \operatorname{rank}(\Pi_x^J).$$

A rejection using leniency groups rules out the relevant restriction on the original decision makers' regime-specific treatment propensities under the conditions of Proposition E.2. Under those conditions, averaging preserves both common direction and the affine representation. Rank-1 is also preserved if the aggregate treatment propensity is nonconstant across groups. A test based on the groups can fail to reject because averaging has concealed violations of common direction or proportionality. Averaging can also reduce the dimension of the span of observed outcome distributions, limiting the number of regimes that this span can identify.

A mixture weight matrix for L instrument groups has L columns and hence rank at most L. The span argument in Theorem 4.1 can therefore identify no more than L regimes. Centering the matrix of joint probabilities of treatment and regime membership lowers the maximum rank to $L - 1$. With ten groups, the first argument allows identification of at most ten regimes, and the centered matrix of regime-specific treatment propensities has no more than nine linearly independent directions of variation.

Proof of Proposition E.2. Because Z is a function of J and $J \perp S \mid X$,

$$\Pr(J = j \mid X = x, S = k, Z = z) = \Pr(J = j \mid X = x, Z = z) = W_x(j, z).$$

Iterated expectations therefore give

$$p_k(x, z) = \sum_j p_k^J(x, j) W_x(j, z).$$

Under Rank-1 proportionality across judges, define $\bar{\lambda}_x(z) := \mathbb{E}\{\lambda_x(J) \mid X = x, Z = z\}$. Then

$$\begin{aligned} p_k(x, z) &= \sum_j \{a_k(x) + b_k(x)\lambda_x(j)\} W_x(j, z) \\ &= a_k(x) + b_k(x)\bar{\lambda}_x(z), \\ P_x C_L &= b(x)\Big(\bar{\lambda}_x(z_1), \ldots, \bar{\lambda}_x(z_L)\Big) C_L, \end{aligned}$$

so $\operatorname{rank}(P_x C_L) \leq 1$. If the aggregate treatment propensity varies across groups,

$$w(x)' P_x C_L \neq 0 \quad \Longrightarrow \quad \operatorname{rank}(P_x C_L) = 1.$$

For common direction, ordered nonoverlapping bins and monotonicity in $L_x(j)$ imply

$$p_k^J(x, j) - p_k^J(x, j') \geq 0 \quad \text{whenever } W_x(j, z_{\ell+1}) W_x(j', z_\ell) > 0.$$

Consequently,

$$\begin{aligned} p_k(x, z_{\ell+1}) - p_k(x, z_\ell) &= \sum_{j,j'} W_x(j, z_{\ell+1}) W_x(j', z_\ell) \{p_k^J(x, j) - p_k^J(x, j')\} \\ &\geq 0. \end{aligned}$$

□

F Population Average Treatment Effects and IV Weights across Covariates

This appendix extends the comparison between IV and the population average treatment effect to averages over covariate values. At a fixed covariate value x, let $a = (z, z')$ be a comparison with $\Delta p(x; a) := p(x, z) - p(x, z') \neq 0$, and define

$$r_k(x; a) := \frac{p_k(x, z) - p_k(x, z')}{\Delta p(x; a)}, \qquad A(x) := \sum_k w_k(x) \tau_k(x).$$

The regime IV weight is $w_k(x) r_k(x; a)$, and $\sum_k w_k(x) r_k(x; a) = 1$.

Under the conditions of Proposition 3.1, the Wald decomposition in (3.2) gives

$$\psi(x; a) = A(x) + \Gamma(x; a), \qquad \Gamma(x; a) := \operatorname{Cov}_{w(x)}(r_k(x; a), \tau_k(x)). \tag{F.1}$$

Both terms are invariant to a permutation of the regime labels.

The identity allows signed IV weights. Identifying its two terms requires regime shares and average treatment effects within regimes as well as joint probabilities of treatment and regime membership. Identifying the outcome mixture among treated cases can leave the regime shares set identified, as in (B.2).

For the IV estimand using all instrument values in Proposition 3.2, with scalar instrument $\zeta(Z)$ and assignment probabilities $q(z \mid x)$, write

$$\kappa_k(x) := \operatorname{Cov}_{q(\cdot|x)}\{\pi_k(x, Z), \zeta(Z)\}, \qquad c(x) := \sum_k \kappa_k(x).$$

If $c(x) \neq 0$, replace $r_k(x;a)$ by $r_k^{q,\zeta}(x) := \kappa_k(x)/\{w_k(x)c(x)\}$ to obtain

$$\beta^{q,\zeta}(x) = A(x) + \Gamma^{q,\zeta}(x), \qquad \Gamma^{q,\zeta}(x) := \mathrm{Cov}_{w(x)}\Big(r_k^{q,\zeta}(x), \tau_k(x)\Big). \tag{F.2}$$

The regime-specific IV weights are

$$\Omega_k^{q,\zeta}(x) = \frac{\kappa_k(x)}{c(x)} = w_k(x) r_k^{q,\zeta}(x). \tag{F.3}$$

Aggregation over covariates. Let ν be the target distribution of X and suppose $C := \mathbb{E}_\nu[c(X)] \neq 0$. The scalar IV estimand after conditioning on X is

$$\beta := \frac{\mathbb{E}_\nu[\mathrm{Cov}_{q(\cdot|X)}\{m(X,Z), \zeta(Z)\}]}{C}.$$

Its covariate weights are $\lambda(x) := c(x)/C$, with $\mathbb{E}_\nu[\lambda(X)] = 1$; these weights may be signed.

Proposition F.1 (Population average treatment effects and weighting across covariates)**.** *Under Assumptions 2.1–2.4, suppose $C \neq 0$ and all displayed moments are finite. Then*

$$\begin{aligned}\beta = & \underbrace{\mathbb{E}_\nu[A(X)]}_{\text{population average treatment effect}} + \underbrace{\mathrm{Cov}_\nu\{\lambda(X), A(X)\}}_{\text{covariate weighting}} \\ & + \underbrace{\frac{1}{C}\mathbb{E}_\nu\left[\sum_k \{\kappa_k(X) - w_k(X)c(X)\}\tau_k(X)\right]}_{\text{IV weights relative to population shares}}.\end{aligned} \tag{F.4}$$

Each term is invariant to regime labels at every covariate value.

Proof. The conditional mean representation gives

$$\begin{aligned} m(x,z) &= \sum_k w_k(x)\mu_{0k}(x) + \sum_k \pi_k(x,z)\tau_k(x), \\ \mathrm{Cov}_q\{m(x,Z), \zeta(Z)\} &= \sum_k \mathrm{Cov}_q\{\pi_k(x,Z), \zeta(Z)\}\tau_k(x) = \sum_k \kappa_k(x)\tau_k(x). \end{aligned}$$

Since $A(x) = \sum_k w_k(x)\tau_k(x)$ and $\mathbb{E}_\nu[\lambda(X)] = 1$,

$$\begin{aligned}\beta &= \frac{1}{C}\mathbb{E}_\nu\left[\sum_k \kappa_k(X)\tau_k(X)\right] \\ &= \mathbb{E}_\nu[\lambda(X)A(X)] + \frac{1}{C}\mathbb{E}_\nu\left[\sum_k \{\kappa_k(X) - w_k(X)c(X)\}\tau_k(X)\right] \\ &= \mathbb{E}_\nu[A(X)] + \mathrm{Cov}_\nu\{\lambda(X), A(X)\} \\ &\quad + \frac{1}{C}\mathbb{E}_\nu\left[\sum_k \{\kappa_k(X) - w_k(X)c(X)\}\tau_k(X)\right].\end{aligned}$$

□

If $c(X) \neq 0$ almost surely, the last term equals $\mathbb{E}_\nu[\lambda(X)\Gamma^{q,\zeta}(X)]$. The covariance expression in (F.4) also covers covariate values with $c(x) = 0$, where regime contributions to the first stage can offset while contributing to the reduced form.

With regime labels aligned across covariate values, define the population share and average treatment effect for regime k, integrating over covariates by $W_k := \mathbb{E}_\nu[w_k(X)]$ and $T_k := \mathbb{E}_\nu[w_k(X)\tau_k(X)]/W_k$. Their product contributes to the population average treatment effect, $\sum_k W_k T_k = \mathbb{E}_\nu[A(X)]$. The aggregate IV weights and contributions are instead

$$\Omega_k := \frac{\mathbb{E}_\nu[\kappa_k(X)]}{C}, \qquad B_k := \frac{\mathbb{E}_\nu[\kappa_k(X)\tau_k(X)]}{C}, \qquad \sum_k \Omega_k = 1, \quad \beta = \sum_k B_k.$$

In general, $B_k \neq \Omega_k T_k$ because T_k and B_k use different weights across covariate values. Comparing Ω_k with $\mathbb{E}_\nu[\lambda(X)w_k(X)]$ uses the same covariate weights for both quantities and isolates the difference due to weighting regimes differently from their population shares.

G Approximation and Sensitivity

Assumption 2.4 equates the potential outcome distribution in each treatment group with that of its regime population. Even with exogenous assignment, decision makers can select different values of V within a regime, and those values may predict potential outcomes. The components of the observed outcome mixture can then vary across assignments. Equality of the mean of $Y(d)$ among cases assigned status d and in the full regime population suffices for the Wald decomposition, whereas the distributional equality justifies identifying its components as regime-specific potential outcome distributions.

Finite regimes can also approximate continuous latent heterogeneity. A related use of discrete groups to approximate continuous heterogeneity appears in the panel models of Bonhomme, Lamadon and Manresa (2022). If potential outcome distributions vary little among cases grouped in the same regime, changing their composition through treatment selection changes the outcome distribution among cases assigned a given treatment status by little. The following argument formalizes this interpretation. It does not establish that additional components chosen to improve likelihood fit form such a refinement.

G.1 Finite Regimes as an Approximation

Fix x. Let A be a pretreatment case type in a compact metric space $(\mathcal{A}, \rho)$, and let S_K index a finite partition $\{C_{1K}, \ldots, C_{KK}\}$ whose largest cell diameter is

$$h_K := \max_k \sup_{a,a' \in C_{kK}} \rho(a, a').$$

Maintain consistency, $D(z) = d(z, X, A, V)$, and

$$Z \perp (A, V, Y(1), Y(0)) \mid X, \qquad (Y(1), Y(0)) \perp V \mid (X, A).$$

Every partition index is a function of A and is therefore independent of assignment conditional on X. Although selection can change the distribution of A within a cell, it leaves the potential outcome distribution at each A unchanged.

Denote the potential outcome distribution at $(X, A) = (x, a)$ by $F_d(\cdot \mid x, a)$. I compare distributions using the integral probability metric $d_{\mathcal{F}}(P, Q) := \sup_{\varphi \in \mathcal{F}} |\int \varphi \, dP - \int \varphi \, dQ|$, assuming that the integrals are finite. For cells and selection events of positive conditional probability, define

$$F_{dk,K}(y \mid x) := \Pr(Y(d) \leq y \mid X = x, S_K = k),$$
$$F^{\text{sel}}_{dkz,K}(y \mid x) := \Pr(Y(d) \leq y \mid D(z) = d, X = x, S_K = k).$$

Let $\mathcal{I}_K(x)$ index the triples (d, k, z) for which both conditioning events above have positive probability.

Assumption G.1 (Approximate equality of potential outcome distributions)**.** *For every K, the following bound holds for some finite $\delta_K(x) \geq 0$:*

$$\sup_{(d,k,z) \in \mathcal{I}_K(x)} d_{\mathcal{F}}\Big(F^{\text{sel}}_{dkz,K}(\cdot \mid x), F_{dk,K}(\cdot \mid x)\Big) \leq \delta_K(x).$$

Assumption 2.4 gives zero error. Conversely, zero error implies equality of the distributions only when $\mathcal{F}$ distinguishes probability distributions, as does the class of half-line indicators. A smaller class may impose equality only of the moments indexed by $\mathcal{F}$.

Lemma G.1 (Continuity of potential outcome distributions)**.** *Suppose $h_K \to 0$ and, for $d = 0, 1$,*

$$d_{\mathcal{F}}\Big(F_d(\cdot \mid x, a), F_d(\cdot \mid x, a')\Big) \leq \varpi_d\Big(\rho(a, a')\Big),$$

where ϖ_d is a modulus of continuity with $\varpi_d(t) \to 0$ as $t \downarrow 0$. Under the assignment and conditional independence restrictions above, Assumption G.1 holds with $\delta_K(x) \leq \max_d \varpi_d(h_K)$.

Proof. Fix $(d, k, z) \in \mathcal{I}_K(x)$ and $\varphi \in \mathcal{F}$, and write $m_{\varphi d}(a) := \mathbb{E}[\varphi(Y(d)) \mid X = x, A = a]$. The conditional independence restrictions imply

$$\int \varphi \, dF^{\text{sel}}_{dkz,K} = \mathbb{E}[m_{\varphi d}(A) \mid D(z) = d, X = x, S_K = k],$$
$$\int \varphi \, dF_{dk,K} = \mathbb{E}[m_{\varphi d}(A) \mid X = x, S_K = k].$$

Both distributions of A are supported on C_{kK}, so

$$\begin{aligned}\left|\int \varphi \, dF^{\text{sel}}_{dkz,K} - \int \varphi \, dF_{dk,K}\right| &\leq \sup_{a,a' \in C_{kK}} |m_{\varphi d}(a) - m_{\varphi d}(a')| \\ &\leq \sup_{a,a' \in C_{kK}} d_{\mathcal{F}}\{F_d(\cdot \mid x, a), F_d(\cdot \mid x, a')\} \\ &\leq \varpi_d(h_K).\end{aligned}$$

Therefore,

$$\sup_{(d,k,z) \in \mathcal{I}_K(x)} d_{\mathcal{F}}(F^{\text{sel}}_{dkz,K}, F_{dk,K}) \leq \max_d \varpi_d(h_K).$$

□

Proposition G.1 (Approximation by mixtures of regime-specific potential outcome distributions)**.** *Let* $\omega_{dk,K}(x,z) := \Pr(S_K = k \mid D = d, X = x, Z = z)$*. Under Assumption G.1, for every* (d,z) *with* $\Pr(D = d, Z = z \mid X = x) > 0$*,*

$$d_{\mathcal{F}}\left(G_{dz}(\cdot \mid x), \sum_k \omega_{dk,K}(x,z) F_{dk,K}(\cdot \mid x)\right) \leq \delta_K(x),$$

where the sum includes only regimes with $\omega_{dk,K}(x,z) > 0$*.*

Proof. Consistency and assignment independence give

$$G_{dz} = \sum_k \omega_{dk,K} F^{\text{sel}}_{dkz,K}.$$

For the regimes with positive weights, Assumption G.1 implies

$$\begin{aligned}d_{\mathcal{F}}\left(G_{dz}, \sum_k \omega_{dk,K} F_{dk,K}\right) &= \sup_{\varphi \in \mathcal{F}} \left|\sum_k \omega_{dk,K}\left(\int \varphi \, dF^{\text{sel}}_{dkz,K} - \int \varphi \, dF_{dk,K}\right)\right| \\ &\leq \sum_k \omega_{dk,K} d_{\mathcal{F}}(F^{\text{sel}}_{dkz,K}, F_{dk,K}) \\ &\leq \delta_K(x) \sum_k \omega_{dk,K} = \delta_K(x).\end{aligned}$$

□

The bound $\delta_K(x)$ also applies to selection bias in each regime's mean potential outcome whenever $d_{\mathcal{F}}$ controls mean differences: the mean of $Y(d)$ among cases with $D(z) = d$ differs from the mean in the full regime by at most $\delta_K(x)$. Including the identity function in $\mathcal{F}$, with a suitable integrable envelope, is sufficient for this implication. The next section derives bounds on errors in the IV decomposition from these bounds on potential outcome means.

G.2 Sensitivity to Differences in Potential Outcome Means within Regimes

Hold the regime partition, shares, and treatment propensities fixed, and maintain Assumptions 2.1–2.3 and finite potential outcome means. Suppress x where convenient. Write

$$\mu_{dk} := \mathbb{E}[Y(d) \mid X = x, S = k], \qquad \mu_{dk}^{\text{sel}}(z) := \mathbb{E}[Y(d) \mid D(z) = d, X = x, S = k]$$

whenever the selection event has positive probability. Selection bias is the difference between the mean potential outcome among cases assigned status d and its mean in the whole regime. Comparing the observed means across treatment statuses therefore combines $\tau_k = \mu_{1k} - \mu_{0k}$ with selection bias in both groups.

Assumption G.2 (Bounded range of conditional potential outcome means)**.** *For each treatment status d and regime k, the conditional mean $\mathbb{E}[Y(d) \mid X = x, S = k, V = v]$ varies by at most $\delta_{dk}(x) \geq 0$ over the conditional support of V.*

Lemma G.2 (Bounds on differences in potential outcome means)**.** *Under Assumption G.2,*

$$|\mu_{dk}^{\text{sel}}(z) - \mu_{dk}| \leq \delta_{dk}(x)$$

for every selection event of positive probability.

Proof. Write $m_{dk}(v) := \mathbb{E}[Y(d) \mid X = x, S = k, V = v]$ and let $\mathcal{V}_{xk}$ be the conditional support of V given $(X, S) = (x, k)$. Since $D(z)$ is a function of (x, k, V),

$$\mu_{dk}^{\text{sel}}(z) = \mathbb{E}[m_{dk}(V) \mid D(z) = d, X = x, S = k],$$
$$\mu_{dk} = \mathbb{E}[m_{dk}(V) \mid X = x, S = k].$$

Assumption G.2 then gives

$$|\mu_{dk}^{\text{sel}}(z) - \mu_{dk}| \leq \sup_{v, v' \in \mathcal{V}_{xk}} |m_{dk}(v) - m_{dk}(v')| \leq \delta_{dk}(x).$$

□

The case $\delta_{dk} = 0$ imposes mean independence from V, which is sufficient for $\mu_{dk}^{\text{sel}}(z) = \mu_{dk}$. That equality also holds if the conditional mean varies with V but every selection event leaves its average unchanged. The bounds below use only the inequality in Lemma G.2, so they also apply to the partition approximation when its metric controls means.

Sensitivity with IV Weights Held Fixed

For each treatment status and regime, let $a_{dkz} \geq 0$ sum to one over instrument values with positive selection probability, and suppose these averages exist. Define

$$\bar{\mu}_{dk}^{\text{sel}} := \sum_z a_{dkz} \mu_{dk}^{\text{sel}}(z),$$
$$\bar{\tau}_k^{\text{sel}} := \bar{\mu}_{1k}^{\text{sel}} - \bar{\mu}_{0k}^{\text{sel}}, \qquad \delta_k^\tau := \delta_{1k} + \delta_{0k}.$$

Then $|\bar{\tau}_k^{\text{sel}} - \tau_k| \leq \delta_k^\tau$.

Proposition G.2 (Sensitivity with IV weights held fixed)**.** *Fix a comparison (z, z') with $p(x, z) \neq p(x, z')$, and let ω_k^* be its IV weights. Set $\psi^0 := \sum_k \omega_k^* \tau_k$, the Wald ratio under Assumption 2.4 with the same population means and treatment propensities. Replacing the regime-specific average treatment effects by the contrasts $\bar{\tau}_k^{\text{sel}}$ gives $\psi^{\text{sel}} := \sum_k \omega_k^* \bar{\tau}_k^{\text{sel}}$, with*

$$|\psi^{\text{sel}} - \psi^0| \leq \sum_k |\omega_k^*| \delta_k^\tau.$$

Proof. The bound on each contrast gives

$$\begin{aligned} |\psi^{\text{sel}} - \psi^0| &= \left| \sum_k \omega_k^* (\bar{\tau}_k^{\text{sel}} - \tau_k) \right| \\ &\leq \sum_k |\omega_k^*| \, |\bar{\tau}_k^{\text{sel}} - \tau_k| \\ &\leq \sum_k |\omega_k^*| \delta_k^\tau. \end{aligned}$$

□

Corollary G.1 (Approximation error with fixed IV weights)**.** *Under Assumption G.1, if $d_{\mathcal{F}}$ controls means and the averages $\bar{\mu}_{dk}^{\text{sel}}$ exist, the error with IV weights held fixed for partition S_K is bounded by*

$$2\delta_K(x) \sum_k |\omega_{k,K}^*(x; z, z')|.$$

If $Y(0) \equiv 0$, the factor $2\delta_K(x)$ can be replaced by $\delta_K(x)$.

Proof. Since $d_{\mathcal{F}}$ controls means,

$$\begin{aligned} |\bar{\mu}_{dk}^{\text{sel}} - \mu_{dk}| &\leq \sum_z a_{dkz} |\mu_{dk}^{\text{sel}}(z) - \mu_{dk}| \\ &\leq \delta_K(x) \sum_z a_{dkz} = \delta_K(x), \\ |\bar{\tau}_k^{\text{sel}} - \tau_k| &\leq |\bar{\mu}_{1k}^{\text{sel}} - \mu_{1k}| + |\bar{\mu}_{0k}^{\text{sel}} - \mu_{0k}| \leq 2\delta_K(x). \end{aligned}$$

If $Y(0) \equiv 0$, then

$$\bar{\mu}_{0k}^{\text{sel}} = \mu_{0k} = 0, \qquad |\bar{\tau}_k^{\text{sel}} - \tau_k| = |\bar{\mu}_{1k}^{\text{sel}} - \mu_{1k}| \leq \delta_K(x).$$

Proposition G.2 therefore gives

$$|\psi^{\text{sel}} - \psi^0| \leq \begin{cases} 2\delta_K(x) \sum_k |\omega_{k,K}^*(x; z, z')|, & \text{in general,} \\ \delta_K(x) \sum_k |\omega_{k,K}^*(x; z, z')|, & \text{if } Y(0) \equiv 0. \end{cases}$$

□

When $\delta_k^\tau = \delta$ for all regimes, nonnegative weights give the bound δ, while negative weights amplify it by $\sum_k |\omega_k^*| = 1 + 2\text{NegMass}$. This bound cannot be applied directly to a refitted mixture when Assumption 2.4 fails, because refitting may change the IV weights and give means other than $\bar{\mu}_{dk}^{\text{sel}}$. After aligning the regimes, the difference between the IV coefficient implied by the fitted model and the benchmark Wald estimand decomposes as

$$\widetilde{\psi} - \psi^0 = \sum_k \omega_k^*(\widetilde{\tau}_k - \tau_k) + \sum_k (\widetilde{\omega}_k^* - \omega_k^*)\widetilde{\tau}_k.$$

Proposition G.2 neither bounds the second term nor establishes that the errors in the estimated average treatment effects within regimes in the first are bounded by δ_k^τ.

The Wald Estimand under Within-Regime Selection Bias

The Wald estimand responds to assignment-specific selection bias in both arms. Define $b_{dkz} := \mu_{dk}^{\text{sel}}(z) - \mu_{dk}$ on positive-probability selection events, and set $b_{dkz} = 0$ otherwise, since its coefficient in the observed mean is zero.

Proposition G.3 (Sensitivity of the Wald estimand to within-regime selection bias)**.** *Suppose* $|b_{dkz}| \leq \delta_{dk}$ *and* $\Delta p := p(x, z) - p(x, z') \neq 0$. *With*

$$e(z) := \sum_k w_k \{p_k(z) b_{1kz} + [1 - p_k(z)] b_{0kz}\},$$

the Wald estimand satisfies

$$\psi^{\text{obs,sel}} - \psi^0 = \frac{e(z) - e(z')}{\Delta p},$$

and

$$|\psi^{\text{obs,sel}} - \psi^0| \leq \frac{1}{|\Delta p|} \sum_k w_k \Big[\{p_k(z) + p_k(z')\}\delta_{1k} + \{2 - p_k(z) - p_k(z')\}\delta_{0k} \Big].$$

Proof. The observed conditional mean satisfies

$$\begin{aligned} m(z) &= \sum_k w_k \{p_k(z)[\mu_{1k} + b_{1kz}] + [1 - p_k(z)][\mu_{0k} + b_{0kz}]\} \\ &= \sum_k w_k \mu_{0k} + \sum_k w_k p_k(z) \tau_k + e(z). \end{aligned}$$

Hence

$$\begin{aligned} \psi^{\text{obs,sel}} &= \frac{m(z) - m(z')}{\Delta p} \\ &= \sum_k \frac{w_k \{p_k(z) - p_k(z')\}}{\Delta p} \tau_k + \frac{e(z) - e(z')}{\Delta p} \\ &= \psi^0 + \frac{e(z) - e(z')}{\Delta p}. \end{aligned}$$

Using $|b_{dkz}| \leq \delta_{dk}$ gives

$$\begin{aligned} |\psi^{\text{obs,sel}} - \psi^0| &= \frac{|e(z) - e(z')|}{|\Delta p|} \\ &\leq \frac{1}{|\Delta p|} \sum_k w_k \Big[p_k(z)|b_{1kz}| + p_k(z')|b_{1kz'}| \\ &\qquad + [1 - p_k(z)]|b_{0kz}| + [1 - p_k(z')]|b_{0kz'}| \Big] \\ &\leq \frac{1}{|\Delta p|} \sum_k w_k \Big[\{p_k(z) + p_k(z')\}\delta_{1k} + \{2 - p_k(z) - p_k(z')\}\delta_{0k} \Big]. \end{aligned}$$

□

The bound for the observed Wald estimand allows selection bias to vary across assignments, whereas Proposition G.2 holds the IV weights fixed and bounds the error from replacing the regime-specific average treatment effects with $\bar{\tau}_k^{\text{sel}}$. In either calculation, a small aggregate first stage can make the departure from ψ^0 large.

Evaluating the bounds over a prespecified grid of the generally unidentified parameters δ_{dk} gives a sensitivity analysis for Assumption 2.4, which is imposed in estimation. Improved predictive fit alone does not establish that fitted components are the partition cells in Lemma G.1.

H Parametric Estimation

I use the joint likelihood of treatment and observed outcomes to estimate regime shares, treatment propensities and potential outcome distributions. The number of regimes and the dimension of the outcome models are held fixed. For the consistency result, each regime may have a negative binomial outcome distribution or a Gaussian or censored Gaussian mixture

with the number of components fixed. I use a common dominating measure for the densities f_{dk}, with point masses at the censoring limits included in that measure.

H.1 Joint Likelihood for Both Treatment Statuses

For $d \in \{0, 1\}$, define

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

Let $\theta = (\alpha, \beta, \eta)$, where the three parameter vectors govern regime shares, regime-specific treatment propensities and potential outcome distributions under both treatment statuses, respectively. When the instrument is discrete, treatment propensities can be estimated separately for each regime and instrument value, with a finite-dimensional specification for how regime shares, propensities and outcome parameters depend on covariates.

Assumption H.1 (Parametric estimation)**.** (E1) *The observations are drawn independently from a common distribution under which (Y, D), conditional on (X, Z), follows* (H.1) *at a parameter θ_0.*

(E2) *Restrict Θ to a compact set on which regime shares, within-regime component weights, Gaussian variances and negative binomial dispersions are uniformly bounded below by positive constants, and treatment propensities are uniformly bounded away from zero and one. The Gaussian means and variances and the negative binomial means and dispersions range over compact sets.*

(E3) *For the chosen specification, the candidate component distributions satisfy the conditions of Proposition 4.1, Corollary 4.1, Theorem 4.2, or Theorem C.1. With regime and subcomponent labels fixed, the conditional functions identify the regression coefficients. Distinct outcome components remain at least a fixed positive distance apart, and the pairing across treatment statuses is unique. Any two parameter values that give the same observed distribution differ only by permutations of regime labels or of subcomponent labels within regimes. The same permutations must apply for P_X-almost every x and respect the restrictions across treatment statuses.*

(E4) *Almost surely, $\theta \mapsto \log h(y, d \mid x, z; \theta)$ is continuous, with an integrable envelope that is uniform over Θ.*

Theorem H.1 (Consistency of the joint likelihood estimator)**.** *Under Assumption H.1, let*

$$\widehat{\theta}_n \in \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log h(Y_i, D_i \mid X_i, Z_i; \theta).$$

The relabelings of θ_0 allowed by (E3) *form a finite set, and the distance of $\widehat{\theta}_n$ from that set converges to zero in probability. For P_{XZ}-almost every (x, z), the fitted conditional distribution converges in probability to the true conditional distribution in total variation. Plug-in estimators are consistent for every continuous label-invariant functional of θ_0, including the Wald ratio and* NegMass *for pairs with a nonzero aggregate first stage and the normalized rank-one distance when its normalization is nonzero.*

Assumption H.1 bounds component weights away from zero and keeps distinct component distributions separated. In applications, I examine fitted weights, distances between component parameters, expected cell counts and stability across starting values.

Because identification at one value of x does not fix the labels at another, a common labeling across x must be imposed in the specification. One possible restriction orders regimes strictly by their mean potential outcome under one treatment status throughout the covariate support.

Proof. Let $Q_n(\theta) = n^{-1}\sum_i \log h(Y_i, D_i \mid X_i, Z_i; \theta)$, $Q(\theta) = \mathbb{E}[Q_n(\theta)]$, and let $[\theta_0]$ be the finite set of relabelings permitted by (E3). Uniform convergence follows from Newey and McFadden (1994, Lemma 2.4); correct specification and (E3) identify the population maximizers:

$$\sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0, \qquad \arg\max_{\theta\in\Theta} Q(\theta) = [\theta_0].$$

The compactness argument in Newey and McFadden (1994, Theorem 2.1) applies with distance to this finite maximizer set, giving

$$\operatorname{dist}(\widehat{\theta}_n, [\theta_0]) \xrightarrow{p} 0.$$

For P_{XZ}-almost every (x, z), pointwise convergence of the fitted densities and Scheffé's lemma yield

$$\left\| h(\cdot \mid x, z; \widehat{\theta}_n) - h(\cdot \mid x, z; \theta_0) \right\|_{L^1} \xrightarrow{p} 0,$$

and hence convergence in total variation. For every continuous label-invariant functional Ψ, the continuous mapping theorem gives

$$\Psi(\widehat{\theta}_n) \xrightarrow{p} \Psi(\theta_0).$$

□

H.2 Special Case: Outcomes Equal to Zero without Treatment

When $Y = DY(1)$ and $Y(0) = 0$, parameterize the identified joint probabilities of treatment and regime membership directly. The likelihood is

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

where $\theta = (\gamma, \eta)$ has fixed dimension. In addition to probability restrictions at each value of the instrument and controls, the candidate joint probabilities must satisfy $\sum_k \max_z \pi_k(x, z) \leq 1$ for regime shares that are constant across assignments to exist (Proposition B.1).

Corollary H.1 (Consistency when observed outcomes without treatment equal zero)**.** *Suppose the observations are i.i.d. and (H.2) is correctly specified at θ_0 in a compact parameter space satisfying the feasibility restriction above. Assume the log likelihood is almost surely continuous and has an integrable envelope uniformly over that space. Suppose the component class and regression specification identify θ_0 up to a single common regime permutation across covariate values. Then a global maximum likelihood estimator converges in probability to that equivalence class, and continuous label-invariant functionals are consistently estimated.*

Proof. Use the uniform-convergence and argmax argument in the proof of Theorem H.1, with likelihood (H.2) and this corollary's identification condition. The population maximizers are the common regime permutations of θ_0; continuous mapping gives consistency of the label-invariant functionals. □

A finite-sample estimate can define a valid observed mixture while failing the feasibility restriction for regime shares that are constant across assignments. If that restriction is uniformly slack at the truth and the estimated joint probabilities converge uniformly, the unconstrained global maximizer is feasible with probability approaching one. Imposing the restriction then leaves the estimate unchanged with probability approaching one, provided the labels have a common normalization and the maximizer is locally unique.

H.3 Likelihood Regularity and Misspecification

For mixtures of Gaussian components, the log likelihood has an integrable quadratic envelope when means and variances lie in compact sets, variances have a positive lower bound, mixture probabilities are bounded away from zero, and $\mathbb{E}[Y^2] < \infty$.

Remark H.1 (Misspecification and the interpretation of regimes)**.** With a misspecified mixture, a good fit to the observed data can still give incorrect estimates of regime-specific

average treatment effects and IV weights, because the fitted distributions can differ from the regime-specific potential outcome distributions. With compactness, continuity and the envelope condition, the global likelihood maximizer still converges to the set of parameters maximizing expected log likelihood, and those parameters define the Kullback–Leibler projection onto the fitted class if they are unique up to relabeling.

I hold the regime partition and treatment propensities fixed when deriving the approximation and sensitivity bounds in Appendix G, so these bounds do not establish recovery of the partition or control changes in fitted parameters under misspecification.

H.4 Computation and Specification Choice

The EM algorithm alternates between calculating posterior probabilities of regimes and outcome subcomponents and maximizing the expected complete-data log likelihood over the selection and outcome parameters. I use numerical maximization for count likelihoods and specifications with covariates, and bound Gaussian variances below by positive constants to prevent arbitrarily concentrated components. In the simulations with outcomes under both treatment statuses, I follow EM with constrained maximization of the observed-data likelihood.

I initialize the algorithm from several starting values and choose the converged fit with the highest likelihood, or the highest penalized likelihood when a penalty is specified.

I compare specifications with different numbers of regimes or within-regime outcome components using estimated regime-specific average treatment effects, IV weights and predictive likelihood. Both numbers may vary across specifications; inference applies to a correctly specified model with both held fixed.

I Inference for Fixed-Dimensional Mixtures

With the numbers of regimes and outcome components fixed, I base inference for smooth functionals and the two specification tests on a local asymptotic expansion of the estimator and a bootstrap that consistently estimates its sampling distribution. I assume that the mixture is correctly specified and identified after normalizing its labels, with positive regime shares and component weights and no duplicate components.

The number of regimes K may be specified in advance or chosen using an independent training sample. When the instrument is constructed, I condition on the training data and the grouping rule.

Let $O_i = (Y_i, D_i, X_i, Z_i, H_i)$, where H_i gives the design stratum used in resampling. After label normalization, the finite-dimensional parameter is $\theta \in \Theta \subset \mathbb{R}^p$. In the special case where $Y(0) \equiv 0$ and the mixture of outcomes among treated cases is identified, the joint probabilities of treatment and regime membership can be used as parameters in these results.

Assumption I.1 (Regularity of the fixed-K estimator)**.** *Fix the working number of regimes K. For a constructed instrument, impose the following conditions conditional on the training data and grouping rule:*

(i) (Interior regularity.) *θ_0 is an interior, regular point of the label-normalized parameter space. The joint probabilities of treatment and regime membership and outcome parameters are bounded away from their boundaries. In a parameterization using outcome distributions for both treatment statuses, the same condition applies to regime shares and treatment propensities. For a parameterization using only the outcome distribution among treated cases, the inequality ensuring that these joint probabilities admit regime shares constant across instrument values is uniformly slack:*

$$\inf_{x\in\mathcal{X}}\left\{1-\sum_{k=1}^{K}\max_{z\in\mathcal{Z}}\pi_k^0(x,z)\right\}>0.$$

Labels are aligned across covariate values within the maintained parameterization.

(ii) (Local label normalization.) *All statements below use the unique representative selected by the normalization in a neighborhood of the permutation equivalence class $[\theta_0]$. For permutation-invariant functionals, working on the quotient space gives equivalent statements.*

(iii) (Asymptotic linearity.) *The global maximizer $\widehat{\theta}$ of the fixed-K likelihood satisfies $\sqrt{n}(\widehat{\theta}-\theta_0)=n^{-1/2}\sum_{i=1}^{n}\varphi_\theta(O_i)+o_p(1)$ with mean-zero summands and Gaussian limit $N(0,V_\theta)$. For observations dependent within assignment clusters, V_θ is the cluster covariance limit.*

(iv) (Resampling regularity.) *For the within-stratum case bootstrap or the nonnegative cluster multiplier bootstrap used by the assignment design, the conditional law of $\sqrt{n}(\widehat{\theta}^*-\widehat{\theta})$ given the data consistently estimates the law of $\sqrt{n}(\widehat{\theta}-\theta_0)$.*

(v) (Regularity for parametric resampling under the null.) *For any sequence $\theta_n\to\theta_0$ in either restricted model, a sample generated from P_{θ_n} and re-estimated without the restriction has the same asymptotic linear expansion, with covariance converging to V_θ.*

Assumption I.1(iii)–(v) specifies the asymptotic expansion and bootstrap conditions for the likelihood and sampling design. The results apply to numerical estimates with the same local label normalization and optimization error negligible in that expansion.

I.1 Generated Instruments

In the application in Appendix K, the training sample T provides the leniency measures and group boundaries. Leniency measures for cases in the training sample use leave-defendant-out

calculations, while cases in the inference sample use the leniency measure estimated on the training sample. Given T, the leniency measures and group boundaries are fixed, and the instrument $Z = g_T(X, J)$ is a function of observed controls and judge identity. Assumption I.1 must hold conditional on T, allowing for dependence among defendants in the same assignment cluster even when they fall in different samples.

Write $g(x, j) := B\{\ell(x, j)\}$ for the grouping formed from the population leniency measure $\ell(x, j)$ and population grouping rule B. To use the population grouping in place of the estimated grouping in the asymptotic argument, I require

$$\Pr\{g_T(x, j) = g(x, j) \text{ for every supported judge–control pair } (x, j)\} \to 1.$$

Estimation error in the leniency measures and cutoffs can change the group of a case near a boundary. Trimming pairs with small first-stage differences leaves these group assignments unchanged.

For a target that averages over training samples, re-estimating the leniency measures and group boundaries in each bootstrap draw accounts for their uncertainty, but the resampling must also allow for dependence within and between the training and inference samples.

I.2 Bootstrap Approximation for Smooth Functionals

With $\mathcal{A}$ denoting the finite set of candidate instrument pairs, define $\Delta_\theta(a) := p_\theta(z) - p_\theta(z')$ for $a = (z, z')$. Trimming at $\delta > 0$ leaves the set $\mathcal{P}_\delta(\theta) := \{a \in \mathcal{A} : |\Delta_\theta(a)| \geq \delta\}$. The bootstrap for functionals evaluated on $\mathcal{P}_\delta(\widehat{\theta})$ must account for sampling variation from estimating the functionals and from choosing pairs using the estimated first stages.

Assumption I.2 (Stability of the set of instrument pairs after trimming)**.** *There exists $\eta > 0$ with $\min_{a\in\mathcal{A}} \big| |\Delta_{\theta_0}(a)| - \delta \big| \geq \eta$, that is, no candidate pair sits on the trimming boundary.*

Replacing the estimated pair set by $\mathcal{P}_\delta(\theta_0)$ leaves the functionals' asymptotic behavior unchanged because Assumption I.2 gives $\Pr\{\mathcal{P}_\delta(\widehat{\theta}) = \mathcal{P}_\delta(\theta_0)\} \to 1$. For pairs chosen in advance and held fixed throughout the bootstrap, Assumption I.2 is unnecessary provided their aggregate first stages are bounded away from zero.

Proposition I.1 (Bootstrap validity for smooth plug-in functionals)**.** *Suppose Assumptions I.1 and I.2 hold, and let $\Psi : \Theta \to \mathbb{R}^q$ be continuously differentiable at θ_0 with Jacobian $\nabla\Psi(\theta_0)$. Then*

$$\sqrt{n}\{\Psi(\widehat{\theta}) - \Psi(\theta_0)\} \Rightarrow \nabla\Psi(\theta_0)' Z_\theta, \qquad Z_\theta \sim N(0, V_\theta),$$

and the conditional law of $\sqrt{n}\{\Psi(\widehat{\theta}^*) - \Psi(\widehat{\theta})\}$ given the data consistently estimates the same limit law. For a scalar contrast with positive asymptotic variance, percentile bootstrap intervals are valid to first order pointwise at θ_0, as are studentized and Wald intervals when the variance is consistently estimated.

Proof. Assumption I.2 gives $\mathcal{P}_\delta(\widehat{\theta}) = \mathcal{P}_\delta(\theta_0)$ with probability approaching one. With this pair set fixed, the sample and bootstrap delta methods give (van der Vaart, 1998, Theorems 3.1 and 23.9)

$$\sqrt{n}\{\Psi(\widehat{\theta}) - \Psi(\theta_0)\} = \nabla\Psi(\theta_0)'\sqrt{n}(\widehat{\theta} - \theta_0) + o_p(1),$$
$$\sqrt{n}\{\Psi(\widehat{\theta}^*) - \Psi(\widehat{\theta})\} = \nabla\Psi(\theta_0)'\sqrt{n}(\widehat{\theta}^* - \widehat{\theta}) + o_{p^*}(1).$$

Here $o_{p^*}(1)$ converges to zero conditionally in probability. Assumption I.1(iii)–(iv) gives the two limit laws in the proposition. □

Proposition I.2 (Smoothness of the reported plug-in functionals)**.** *Under Assumption I.1 and for pairs $a \in \mathcal{P}_\delta(\theta_0)$, the following are continuously differentiable in a neighborhood of θ_0:* (i) *the first stage $p_\theta(z)$, reduced form $m_\theta(z)$, and joint probabilities of treatment and regime membership $\pi_{k,\theta}(z)$;* (ii) *the pairwise Wald ratio $\psi_\theta(a)$;* (iii) *the IV weights $\omega^*_{k,\theta}(a)$;* (iv) *the regime-specific average treatment effects $\tau_{k,\theta}$; and* (v) *when the maintained parameterization separately identifies regime shares and average treatment effects within regimes, the population average treatment effect $A(x)$ and the difference between IV and population weighting, $\Gamma(x;a)$ in* (F.1)*. The scalar IV estimand and the versions averaged over the target covariate distribution in Appendix F are continuously differentiable if their first-stage covariance denominators are bounded away from zero.*

Proof. Suppress x. The component means, mixture probabilities and joint probabilities $\pi_{k,\theta}(z)$ are smooth in the interior. The identities

$$\pi_{k,\theta}(z) = w_{k,\theta}p_{k,\theta}(z), \qquad p_\theta(z) = \sum_k \pi_{k,\theta}(z),$$
$$\tau_{k,\theta} = \mu_{1k,\theta} - \mu_{0k,\theta}, \quad m_\theta(z) = \sum_k \{w_{k,\theta}\mu_{0k,\theta} + \pi_{k,\theta}(z)\tau_{k,\theta}\}$$

establish (i) and (iv) by finite sums and products. Setting $\mu_{0k,\theta} = 0$ gives the case $Y(0) \equiv 0$, in which the first stage and reduced form use only the joint probabilities and treated-outcome means.

For pairs in $\mathcal{P}_\delta(\theta_0)$ and the scalar IV estimand with a nonzero first-stage covariance, continuity gives a neighborhood U of θ_0 on which

$$\inf_{\theta\in U} |\Delta_\theta(a)| \geq \delta/2 > 0, \qquad \inf_{\theta\in U} |\mathrm{Cov}_q\{p_\theta(Z), \zeta(Z)\}| > 0.$$

The Wald ratio, IV weights and scalar IV estimand are therefore smooth ratios. When the shares are separately identified as required in (v), $A_\theta = \sum_k w_{k,\theta}\tau_{k,\theta}$ and $\Gamma_\theta(a) = \sum_k\{\omega^*_{k,\theta}(a) - w_{k,\theta}\}\tau_{k,\theta}$ are also smooth finite sums and products. □

Remark I.1 (Quantities unchanged by relabeling the regimes)**.** For a permutation π of regime labels, $p_{\pi\theta}(z) = p_\theta(z)$ and $m_{\pi\theta}(z) = m_\theta(z)$, so the Wald ratio is label invariant. The IV weight vector is permuted, so NegMass $= \sum_k [-\omega^*_k]_+$ is invariant. A regime permutation

reorders the rows of the matrix of joint probabilities of treatment and regime membership and therefore leaves the Rank-1 distance unchanged. The terms $A(x)$ and $\Gamma(x; a)$ are also invariant because a permutation only reorders their summands. These quantities are unaffected by label switching within an identified collection of component distributions. For the estimator using outcome mixtures for both treatment statuses considered here, average treatment effects within individual regimes require identified component distributions and a unique cross-arm pairing. Proposition B.3 and the auxiliary measurement results give alternative identification conditions, but the inference results in this section retain the assumptions on the estimator in use.

I.3 Tests of Common Direction and Rank-1

The identified joint probabilities $\pi_k(x, z) = w_k(x)p_k(x, z)$ suffice to evaluate both restrictions, including when the regime-specific outcome distributions and mixture weights are identified only among treated cases. Write

$$\Pi_x := \big(\pi_k(x, z_\ell)\big)_{k,\ell}, \qquad C_L := I_L - L^{-1}\mathbf{1}_L\mathbf{1}_L'.$$

Both tests use a bootstrap that approximates the distribution under the relevant null. For either bootstrap, I start from a feasible estimate satisfying the null and follow the assignment design when simulating new samples or recentering resampling perturbations. In the parametric implementation, I draw samples from this estimate and refit the unrestricted mixture. Alternatively, I resample cases or apply cluster multiplier weights according to the assignment design and refit the unrestricted mixture. For Rank-1, this second implementation adds the estimated sampling perturbation to the restricted estimate. For common direction, both implementations center the perturbations and impose equality on the moments in an estimated contact set, defined below.

Resampling and any reconstruction of a generated leniency measure must respect the assignment design. The test of common direction additionally requires an instrument order such that $p(x, z_{\ell+1}) \geq p(x, z_\ell)$. I hold a prespecified ordering fixed throughout the bootstrap. If the ordering is estimated, I impose the following condition.

Assumption I.3 (Stable instrument order)**.** *The ordered instrument support is finite. If its order is estimated from aggregate propensities, distinct ordered values satisfy*

$$\min_{\ell \neq m:\, p(x,z_\ell) \neq p(x,z_m)} |p(x, z_\ell) - p(x, z_m)| > \eta$$

for some $\eta > 0$, and equal propensity values are pooled before the test based on differences between adjacent instrument values is computed.

When aggregate propensities are tied, the test of common direction uses the grouped instrument. On this support, nonnegative adjacent changes are equivalent to common direction for all pairs of instrument values with a nonzero aggregate first stage. Grouping can, however, conceal differences in regime-specific treatment propensities among decision makers with the same aggregate propensity. There is no such loss when aggregate propensities are distinct or when all regime-specific treatment propensities agree within each tied group.

Test A. Common-direction test. The null is $H_0^N : \omega_k^*(x; z, z') \geq 0$ for all k and pairs of instrument values (z, z') with a nonzero aggregate first stage. For interpretation, I report the estimated sum of absolute values of negative IV weights for each pair

$$\widehat{\mathrm{NegMass}}(x; z, z') = \sum_k [-\widehat{\omega}_k^*(x; z, z')]_+,$$

and summarize it by the maximum over trimmed pairs of instrument values. Define the changes in joint probabilities $\pi_k(x, z)$ between adjacent instrument values

$$g_{k\ell}(x) := \pi_k(x, z_{\ell+1}) - \pi_k(x, z_\ell).$$

For consistent asymptotic standard deviations $\widehat{\sigma}_{k\ell}$, the formal statistic is

$$T_n^{\max} = \max_{k,\ell} \left[\frac{-\sqrt{n}\, \widehat{g}_{k\ell}(x)}{\widehat{\sigma}_{k\ell}} \right]_+ .$$

The test rejects for a value of $T_n^{\max}$ strictly above its bootstrap critical value. Adjacent changes suffice because a change between any ordered pair is their sum. For a parametric bootstrap, draw from a restricted estimate satisfying $\widetilde{g}_{k\ell} \geq 0$ and center the unrestricted refits to obtain $\sqrt{n}\{\widehat{g}_{k\ell}^* - \widetilde{g}_{k\ell}\}$. Under resampling, use $\sqrt{n}\{\widehat{g}_{k\ell}^* - \widehat{g}_{k\ell}\}$ instead. Hold the standard deviation estimates $\widehat{\sigma}_{k\ell}$ fixed across bootstrap draws. Substitute these perturbations for $\sqrt{n}\, \widehat{g}_{k\ell}$ in the studentized maximum and compute that maximum over

$$\widehat{\mathcal{C}}_n = \left\{ (k, \ell) : \frac{\sqrt{n}\, \widehat{g}_{k\ell}}{\widehat{\sigma}_{k\ell}} \leq \kappa_n \right\}, \qquad \kappa_n \to \infty, \quad \kappa_n / \sqrt{n} \to 0.$$

The bootstrap maximum is zero when $\widehat{\mathcal{C}}_n$ is empty.

Test B. Rank-1 proportionality test. The null is $H_0^R : \mathrm{rank}(\Pi_x C_L) = 1$. Let $\widehat{M}_x = \widehat{\Pi}_x C_L$. The statistic is the normalized distance from $\widehat{M}_x$ to its best rank-one approximation:

$$D_n^{\mathrm{rank}} = \left(\frac{\sum_{j \geq 2} \sigma_j^2(\widehat{M}_x)}{\sum_{j \geq 1} \sigma_j^2(\widehat{M}_x)} \right)^{1/2}, \tag{I.1}$$

where $\sigma_j(\widehat{M}_x)$ are singular values. A null estimate can be obtained from the leading singular component of $\widehat{M}_x$ after restoring the row means, provided the probability constraints hold, or by restricted likelihood. The parametric bootstrap draws from this estimate and refits the unrestricted model. With resampling, add $\widehat{\Pi}_x^* - \widehat{\Pi}_x$ to the null estimate before evaluating the statistic. Permuting the instrument values leaves the centered statistic unchanged.

Restricted estimates used to generate bootstrap samples must also satisfy the condition for the existence of regime shares that are constant across assignments in Proposition B.1. The theorem below additionally requires these estimates to be root-n consistent under the null. In the finite-support parameterization of Appendix J.3, the full restricted estimate combines feasible projected joint probabilities with the unrestricted outcome parameters. The singular-value projection is locally smooth at a nonzero rank-one matrix, and isotonic projection is Lipschitz, so both preserve the root-n rate under the null when the probability constraints and numerical margins are strictly slack at the truth. A parameterization that links probabilities across covariate values instead requires a restricted estimate within that parameter space.

Theorem I.1 (Validity of the recentered bootstrap tests)**.** *Suppose Assumptions I.1 and I.3 hold for a fixed K. Assume the joint probabilities of treatment and regime membership are smooth in the locally normalized parameter and their estimation errors, scaled by $\sqrt{n}$, have a Gaussian limit that is nondegenerate on the tested coordinates. For the test of common direction, assume $\widehat{\sigma}_{k\ell} \xrightarrow{p} \sigma_{k\ell} > 0$. For the resampling implementation, suppose Assumption I.1*(iv) *holds. For the parametric implementation, suppose Assumption I.1*(v) *holds uniformly over $n^{-1/2}$ neighborhoods in the restricted models. For $H \in \{N, R\}$, let Θ_H be the parameter set satisfying H_0^H. Whenever an implementation uses a restricted estimate $\widetilde{\theta}_H$, assume that, under a fixed $\theta_0 \in \Theta_H$,*

$$\widetilde{\theta}_H \in \Theta_H, \qquad \sqrt{n}\,(\widetilde{\theta}_H - \theta_0) = O_p(1).$$

(i) *Under H_0^R, suppose* $\operatorname{rank}(\Pi_x C_L) = 1$*, its leading singular value is positive, and the feasible rank-one projection is locally unique and interior. If the limiting distribution of $\sqrt{n}\, D_n^{\text{rank}}$ is continuous at its $(1-\alpha)$ quantile, the recentered Rank-1 test has asymptotic rejection probability α.*

(ii) *Fix a regular distribution in H_0^N and let $\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$. When at least one inequality binds, suppose the limiting distribution is continuous at its $(1-\alpha)$ quantile and the bootstrap critical value converges in probability to that quantile. The contact-set bootstrap test based on $T_n^{\max}$ rejects with limiting probability α when at least one*

inequality binds. If all inequalities are strictly slack, its rejection probability converges to zero. Thus, in both cases,

$$\limsup_{n\to\infty} \Pr(\text{reject } H_0^N) \le \alpha.$$

The theorem holds pointwise for a correctly specified mixture with a fixed number of components, local regularity after label normalization and an instrument order that is pre-specified or consistently estimated (Appendix I.3). It does not establish uniform validity along sequences in which component distributions merge, regime shares vanish or the instrument order becomes unstable. Approximating the conditional bootstrap quantiles with B draws also requires $B \to \infty$.

Joint rejection rule. Under $H_0^N \cup H_0^R$, at least one component null is true. I reject this union only when both tests reject, using $p_{\text{joint}} = \max\{p_N, p_R\}$. A joint rejection therefore rejects a true component null, which gives pointwise asymptotic rejection probability at most α by the intersection–union argument (Berger, 1982). The proportional common-direction benchmark requires $H_0^N \cap H_0^R$, which is not established by failure to reject the union.

Proof of Theorem I.1

Write

$$h_R(\theta) = \text{vec}\{\Pi_x(\theta) C_L\}, \qquad \mathcal{M}_R = \{\text{vec}(M) : M\mathbf{1}_L = 0,\ \text{rank}(M) \le 1\}.$$

The best rank-one approximation to $\widehat{M}_x$ is centered, so

$$\text{dist}\{h_R(\widehat{\theta}), \mathcal{M}_R\}^2 = \sum_{j\ge 2} \sigma_j^2(\widehat{M}_x),$$

$$D_n^{\text{rank}} = \frac{\text{dist}\{h_R(\widehat{\theta}), \mathcal{M}_R\}}{\|h_R(\widehat{\theta})\|}.$$

No probability constraints are imposed on the approximating matrix. At a matrix of rank exactly one with a positive leading singular value, the rank-one set is a smooth manifold within the space of centered matrices. Under H_0^R, the restricted estimate satisfies

$$\widetilde{\theta}_R \in \Theta_R, \qquad \sqrt{n}(\widetilde{\theta}_R - \theta_0) = O_p(1).$$

A local expansion of distance to a smooth manifold gives

$$\sqrt{n}\ \text{dist}\{h_R(\widehat{\theta}), \mathcal{M}_R\} = \left\| P_N D h_R(\theta_0) \sqrt{n}(\widehat{\theta} - \theta_0) \right\| + o_p(1),$$

where P_N projects onto the normal space within the space of centered matrices at $h_R(\theta_0)$. Since

$$\|h_R(\widehat{\theta})\| \xrightarrow{p} \|h_R(\theta_0)\| > 0,$$

the sample limit is

$$\sqrt{n}\, D_n^{\text{rank}} \Rightarrow \frac{\|P_N Dh_R(\theta_0) Z_\theta\|}{\|h_R(\theta_0)\|}.$$

For the bootstrap, let

$$h_n^* = \begin{cases} h_R(\widehat{\theta}^*), & \text{under parametric sampling,} \\ h_R(\widetilde{\theta}_R) + h_R(\widehat{\theta}^*) - h_R(\widehat{\theta}), & \text{under resampling.} \end{cases}$$

Assumption I.1(v), applied at $\widetilde{\theta}_R$, or Assumption I.1(iv), respectively, gives

$$\sqrt{n}\{h_n^* - h_R(\widetilde{\theta}_R)\} \Rightarrow_p^* Dh_R(\theta_0) Z_\theta.$$

Because $h_R(\widetilde{\theta}_R) \in \mathcal{M}_R$, expansion at this restricted estimate yields

$$\sqrt{n}\, D_n^{\text{rank},*} = \frac{\left\| \widetilde{P}_N \sqrt{n}\{h_n^* - h_R(\widetilde{\theta}_R)\} \right\|}{\|h_R(\widetilde{\theta}_R)\|} + o_{p^*}(1),$$

where $\widetilde{P}_N$ is the normal projection at $h_R(\widetilde{\theta}_R)$ and the remainder converges to zero conditionally in probability. Thus

$$\widetilde{P}_N \xrightarrow{p} P_N, \qquad \|h_R(\widetilde{\theta}_R)\| \xrightarrow{p} \|h_R(\theta_0)\|,$$
$$\sqrt{n}\, D_n^{\text{rank},*} \Rightarrow_p^* \frac{\|P_N Dh_R(\theta_0) Z_\theta\|}{\|h_R(\theta_0)\|}.$$

Conditional quantile consistency and continuity of this limit law at its $(1-\alpha)$ quantile give

$$\Pr(\text{reject } H_0^R) \to \alpha.$$

For the common direction test, collect the changes in joint probabilities of treatment and regime membership between adjacent instrument values in the finite vector $g(\theta)$ and let

$$\mathcal{C}_0 = \{j : g_j(\theta_0) = 0\}.$$

For $j \in \mathcal{C}_0$, root-n consistency and $\kappa_n \to \infty$ give

$$\frac{\sqrt{n}\, \widehat{g}_j}{\widehat{\sigma}_j} = O_p(1),$$
$$\Pr(j \notin \widehat{\mathcal{C}}_n) = \Pr\left(\frac{\sqrt{n}\, \widehat{g}_j}{\widehat{\sigma}_j} > \kappa_n \right) \to 0.$$

For $g_j(\theta_0) > 0$,

$$\frac{\widehat{g}_j}{\widehat{\sigma}_j} \xrightarrow{p} \frac{g_j(\theta_0)}{\sigma_j} > 0, \qquad \frac{\kappa_n}{\sqrt{n}} \to 0,$$

and hence

$$\Pr(j \in \widehat{\mathcal{C}}_n) = \Pr\left(\frac{\widehat{g}_j}{\widehat{\sigma}_j} \le \frac{\kappa_n}{\sqrt{n}}\right) \to 0.$$

There are finitely many tested coordinates, so

$$\Pr(\widehat{\mathcal{C}}_n \neq \mathcal{C}_0) \le \sum_{j \in \mathcal{C}_0} \Pr(j \notin \widehat{\mathcal{C}}_n) + \sum_{j \notin \mathcal{C}_0} \Pr(j \in \widehat{\mathcal{C}}_n) \to 0.$$

For the parametric implementation, feasibility and root-n consistency of $\widetilde{\theta}_N$ permit applying Assumption I.1(v) at that estimate. Together with the joint central limit theorem and Assumption I.1(iv) for resampling, this gives

$$\begin{aligned}
\sqrt{n}\{g(\widehat{\theta}) - g(\theta_0)\} &\Rightarrow Z_g, \\
\sqrt{n}\{g(\widehat{\theta}^*) - g(\widetilde{\theta}_N)\} &\Rightarrow^*_p Z_g \quad \text{under parametric sampling}, \\
\sqrt{n}\{g(\widehat{\theta}^*) - g(\widehat{\theta})\} &\Rightarrow^*_p Z_g \quad \text{under resampling}.
\end{aligned}$$

Here Z_g is the Gaussian limit at the fixed null parameter. Let $T_n^{\max,*}$ denote the bootstrap maximum over $\widehat{\mathcal{C}}_n$. If $\mathcal{C}_0 \neq \varnothing$,

$$\begin{aligned}
T_n^{\max} &\Rightarrow \max_{j \in \mathcal{C}_0} [-Z_{g,j}/\sigma_j]_+, \\
T_n^{\max,*} &\Rightarrow^*_p \max_{j \in \mathcal{C}_0} [-Z_{g,j}/\sigma_j]_+.
\end{aligned}$$

Continuity is needed at the $(1-\alpha)$ quantile, so an atom at zero in the distribution of the positive-part maximum is allowed. This continuity condition and consistency of the bootstrap critical value imply

$$\Pr(\text{reject } H_0^N) \to \alpha \qquad (\mathcal{C}_0 \neq \varnothing).$$

If all inequalities are strictly slack,

$$\begin{gathered}
\Pr(T_n^{\max} = 0) \to 1, \qquad \Pr(\widehat{\mathcal{C}}_n = \varnothing) \to 1, \\
T_n^{\max,*} = 0 \text{ on } \{\widehat{\mathcal{C}}_n = \varnothing\}.
\end{gathered}$$

The strict rejection rule then gives

$$\Pr(\text{reject } H_0^N) \le \Pr(T_n^{\max} > 0) \to 0.$$

See Andrews and Soares (2010) for moment selection methods. □

J Additional Simulation Details

This appendix specifies the experiments recovering potential outcome distributions under both treatment statuses in Section 7 and reports additional estimation results. The final subsection describes the experiments for the common direction and Rank-1 tests.

J.1 Data-Generating Process and Estimation

Judge assignment is uniform on $z_1, \ldots, z_{10}$ and independent of the regime, with $(w_1, w_2) = (0.6, 0.4)$. Let $\lambda(z_j) = (j-1)/9$. Treatment is Bernoulli with probability $p_k(z_j)$ conditional on $(S = k, Z = z_j)$. The three patterns of treatment propensities are

Treatment propensities across judges	$p_1(z)$	$p_2(z)$
Increase proportionally in both regimes	$0.15 + 0.65\lambda(z)$	$0.25 + 0.25\lambda(z)$
Increase in both regimes, but not proportionally	$0.15 + 0.65\lambda(z)$	$0.20 + 0.35\lambda(z) + 0.15\lambda(z)^2$
Increase in one regime and decrease in the other	$0.15 + 0.65\lambda(z)$	$0.75 - 0.50\lambda(z) - 0.10\lambda(z)^2$.

Conditional on regime membership, $Y(d) = \mu_{dk} + \sigma_d \varepsilon_d$, where the disturbances are independent standard normals and independent of treatment and assignment. The means are $(\mu_{01}, \mu_{02}) = (0.8, 0.2)$ and $(\mu_{11}, \mu_{12}) = (1.0, 1.6)$. The three outcome specifications have $(\sigma_0, \sigma_1) = (0.2, 0.2)$, $(0.4, 0.4)$, or $(0.6, 0.2)$. In estimation, component variances may differ across regimes as well as treatment statuses. Each simulated sample consists of $Y = DY(1) + (1-D)Y(0)$, D, and Z at fixed covariates.

I fit the joint Gaussian likelihood from twelve starting values, estimating regime shares, treatment propensities at each judge, and potential outcome means and variances. I vary the initial component labels across treatment arms over all relative permutations and, from each start, run EM followed by likelihood optimization subject to $w_k \geq 0.01$, $p_k(z) \in [0.005, 0.995]$, $\mu_{dk} \in [-10, 10]$, and $\sigma_{dk} \in [0.03, 5]$. I choose the fit with the largest likelihood, without using the true regime labels or parameter values at any stage of estimation. I order the regimes by their fitted mean potential outcomes under treatment and use the same regime labels under no treatment.

The IV calculations use $q(z_j) = 1/10$ and the scalar instrument $\zeta(z) = p(z) = \sum_k w_k p_k(z)$, fixed at its population value. Equation (3.4) then gives

$$\Omega_k = \frac{w_k \operatorname{Cov}_q\{\zeta(Z), p_k(Z)\}}{\operatorname{Cov}_q\{\zeta(Z), p(Z)\}}, \qquad \beta_{\mathrm{IV}} = \sum_k \Omega_k(\mu_{1k} - \mu_{0k}).$$

I substitute fitted shares, treatment propensities, and means into these expressions, keeping ζ and q fixed. This scalar instrument enters the decomposition but not likelihood estimation. The population values $(\Omega_2, \beta_{\text{IV}})$ are $(0.204, 0.445)$ when treatment propensities increase proportionally in both regimes, $(0.340, 0.607)$ when they increase in both regimes but not proportionally, and $(-1.585, -1.703)$ when they increase in one regime and decrease in the other.

I run 500 independent replications with $K = 2$ fixed for every combination of treatment propensities, outcome specification and $n \in \{2500, 10{,}000\}$. The six experiments selecting K use the same $n = 10{,}000$ samples as the fixed-K experiments with separated component distributions or overlapping component distributions under both treatment statuses.

J.2 Additional Estimation Results

Tables 5 and 6 distinguish estimation of the two potential outcome means from estimation of their difference. With overlap confined to the potential outcome distributions under no treatment, the mean potential outcomes under treatment remain precisely estimated, while estimation errors for the mean potential outcomes under no treatment account for most of the treatment-effect estimation error. All error calculations with $K = 2$ include every replication. Correct ordering in Table 2 means that, after ordering labels by the estimated mean potential outcomes under treatment, $\widehat{\mu}_{01} > \widehat{\mu}_{02}$, as in the data-generating process. This measure does not distinguish incorrect pairing from estimation error in the means.

Table 7 evaluates the predictive rule for selecting K. Each candidate is fitted on 80 percent of the sample, compared by log likelihood on the remaining 20 percent, and the winner is refitted on the full sample. The candidates have $K \in \{1, 2, 3\}$. The correct number is selected in 87.0–89.2 percent of samples with separated distributions. With overlapping component distributions under both treatment statuses, the frequencies are 79.6 percent when treatment propensities increase proportionally in both regimes, 63.6 percent when they increase in both regimes but not proportionally, and 88.6 percent when they increase in one regime and decrease in the other. I include all selected models when calculating estimation errors for the population ATE and IV estimand. The other tables hold $K = 2$ fixed when reporting regime-specific errors, so those errors are not conditional on selecting the correct number of regimes.

All 9,000 fixed-K final fits satisfy the numerical first-order conditions. Five of the 3,000 final fits after selecting K fail the stationarity check, all with $K = 3$; these fits are included in the tables. Excluding these five fits changes the aggregate RMSEs by less than 0.2 percent in each affected design. Of all 9,000 candidate fits used to select K, 41 fail the stationarity check, including five selected candidates. Treatment propensity bounds are reached in 1,375 final fits with K fixed and 393 after selecting K; the component standard deviation bounds

Table 5: Recovery of potential outcome means, regime shares, and treatment propensities

Treatment propensities across judges	Overlap between component distributions	RMSE					
		$\widehat{\mu}_{01}$	$\widehat{\mu}_{02}$	$\widehat{\mu}_{11}$	$\widehat{\mu}_{12}$	$\widehat{w}_1$	$\widehat{p}_k(z)$
$n = 2{,}500$							
Increase proportionally in both regimes	Separated	0.013	0.016	0.014	0.024	0.018	0.049
	Both statuses	0.316	0.434	0.140	0.229	0.182	0.192
	No treatment	0.205	0.541	0.014	0.025	0.120	0.163
Increase in both regimes, but not proportionally	Separated	0.064	0.070	0.014	0.022	0.022	0.047
	Both statuses	0.406	0.484	0.147	0.234	0.256	0.151
	No treatment	0.296	0.543	0.015	0.022	0.084	0.108
Increase in one regime and decrease in the other	Separated	0.011	0.015	0.011	0.017	0.016	0.046
	Both statuses	0.049	0.075	0.044	0.070	0.063	0.094
	No treatment	0.086	0.124	0.012	0.016	0.079	0.104
$n = 10{,}000$							
Increase proportionally in both regimes	Separated	0.007	0.008	0.007	0.011	0.009	0.024
	Both statuses	0.226	0.300	0.080	0.144	0.113	0.118
	No treatment	0.121	0.353	0.007	0.012	0.085	0.108
Increase in both regimes, but not proportionally	Separated	0.007	0.009	0.007	0.010	0.010	0.023
	Both statuses	0.360	0.436	0.103	0.161	0.196	0.086
	No treatment	0.260	0.491	0.007	0.011	0.070	0.085
Increase in one regime and decrease in the other	Separated	0.006	0.008	0.005	0.008	0.008	0.023
	Both statuses	0.025	0.037	0.023	0.033	0.030	0.046
	No treatment	0.053	0.079	0.005	0.008	0.054	0.065

Notes: Each row includes all 500 replications with $K = 2$ fixed. The population means are $(\mu_{01}, \mu_{02}) = (0.8, 0.2)$ and $(\mu_{11}, \mu_{12}) = (1.0, 1.6)$, and $w_1 = 0.6$. The last column is the square root of the mean squared error averaged over replications, both regimes, and all ten judges. Regime labels are ordered once by the estimated mean potential outcomes under treatment and applied to both arms. Outcome specifications are defined in Table 2.

Table 6: Bias in estimated treatment effects and recovery of the population average treatment effect

		Bias (Monte Carlo SE)		Population ATE	
Treatment propensities across judges	Overlap between component distributions	$\widehat{\tau}_1$	$\widehat{\tau}_2$	Bias	RMSE
$n = 2{,}500$					
Increase proportionally in both regimes	Separated	-0.0010 (0.0009)	-0.0019 (0.0013)	0.0003	0.014
	Both statuses	0.1452 (0.0098)	-0.3136 (0.0234)	-0.0585	0.118
	No treatment	0.0975 (0.0081)	-0.2665 (0.0212)	-0.0333	0.088
Increase in both regimes, but not proportionally	Separated	0.0069 (0.0028)	-0.0079 (0.0035)	0.0014	0.015
	Both statuses	0.2273 (0.0142)	-0.3388 (0.0226)	-0.0228	0.060
	No treatment	0.1987 (0.0099)	-0.3643 (0.0181)	-0.0205	0.059
Increase in one regime and decrease in the other	Separated	-0.0002 (0.0007)	-0.0009 (0.0010)	0.0007	0.014
	Both statuses	-0.0043 (0.0029)	-0.0054 (0.0046)	0.0025	0.022
	No treatment	-0.0121 (0.0038)	-0.0129 (0.0055)	0.0047	0.023
$n = 10{,}000$					
Increase proportionally in both regimes	Separated	0.0002 (0.0004)	0.0003 (0.0006)	-0.0001	0.007
	Both statuses	0.0764 (0.0069)	-0.1585 (0.0178)	-0.0334	0.082
	No treatment	0.0443 (0.0050)	-0.1354 (0.0147)	-0.0115	0.062
Increase in both regimes, but not proportionally	Separated	-0.0007 (0.0005)	-0.0008 (0.0007)	-0.0003	0.007
	Both statuses	0.1847 (0.0114)	-0.3164 (0.0193)	-0.0277	0.047
	No treatment	0.1684 (0.0089)	-0.3159 (0.0168)	-0.0246	0.051
Increase in one regime and decrease in the other	Separated	0.0005 (0.0003)	0.0004 (0.0005)	0.0002	0.007
	Both statuses	-0.0014 (0.0015)	-0.0011 (0.0023)	0.0008	0.010
	No treatment	-0.0026 (0.0024)	-0.0044 (0.0035)	0.0012	0.012

Notes: Each row includes all 500 replications with $K = 2$ fixed. The true average treatment effects are $(0.2, 1.4)$ and the population average treatment effect is 0.68. Parentheses contain Monte Carlo standard errors of the reported biases. Outcome specifications and regime labels are as in Table 2.

Table 7: Selection of the number of regimes and estimation of the population ATE and IV estimand

Treatment propensities across judges	Overlap between component distributions	Selected number (%)			RMSE	
		$K = 1$	$K = 2$	$K = 3$	ATE	IV
Increase proportionally in both regimes	Separated	0.0	87.8	12.2	0.030	0.022
	Both statuses	4.8	79.6	15.6	0.088	0.060
Increase in both regimes, but not proportionally	Separated	0.0	87.0	13.0	0.015	0.017
	Both statuses	21.4	63.6	15.0	0.043	0.040
Increase in one regime and decrease in the other	Separated	0.0	89.2	10.8	0.009	0.257
	Both statuses	0.0	88.6	11.4	0.018	0.296

Notes: Each row includes all 500 replications with $n = 10{,}000$ and generating $K_0 = 2$. Models with $K = 1, 2, 3$ are fitted on 80 percent of each sample, compared by log likelihood on the remaining 20 percent, and the selected model is refitted on the full sample. ATE and IV RMSEs include every selected model, including $K \neq 2$. The population ATE is 0.68. Monte Carlo standard errors for the correct-selection percentages range from 1.39 to 2.15 percentage points.

are reached in one and ten fits, respectively. All estimated regime shares exceed their lower bound. I include the estimates at these parameter bounds in the reported results.

J.3 Size and Power of the Tests

The test experiments use a Gaussian outcome mixture among treated cases, with $Y(0) = 0$, $(w_1, w_2) = (0.5, 0.5)$, $(\tau_1, \tau_2) = (0.15, 0.35)$, and component standard deviations of 0.05. Their treatment propensities are

Treatment propensities across judges	$p_1(z)$	$p_2(z)$
Increase proportionally in both regimes	$0.10 + 0.60\lambda(z)$	$0.30 + 0.40\lambda(z)$
Increase in both regimes, but not proportionally	$0.10 + 0.70\sqrt{\lambda(z)}$	$0.05 + 0.90\lambda(z)^3$
Increase in one regime and decrease in the other for some judge pairs	$0.20 + 0.60\lambda(z)$	$0.30 + 0.50\sin\{\pi\lambda(z)\}$
Constant propensity in regime 2	$0.10 + 0.60\lambda(z)$	0.30.

Each design has $J = 10$ instrument values, with λ equally spaced from zero to one. Judges are assigned uniformly and independently of regimes, and ordered by the population aggregate treatment propensity. I hold $K = 2$ fixed and use 399 replications per design with $n = 5000$ and seed 20260829. Table 8 includes all 399 samples from each design.

Table 8: Rejection frequencies of the tests of common direction and Rank-1 at five percent

Treatment propensities across judges	Test A: common direction	Test B: Rank-1	Both reject
Increase proportionally in both regimes	0.00 (0)	4.01 (16)	0.00 (0)
	[0.00, 0.95]	[2.48, 6.41]	[0.00, 0.95]
Increase in both regimes, but not proportionally	0.00 (0)	100.00 (399)	0.00 (0)
	[0.00, 0.95]	[99.05, 100.00]	[0.00, 0.95]
Increase in one regime and decrease in the other for some judge pairs	100.00 (399)	100.00 (399)	100.00 (399)
	[99.05, 100.00]	[99.05, 100.00]	[99.05, 100.00]
Constant treatment propensity in one regime	1.25 (5)	3.26 (13)	0.50 (2)
	[0.54, 2.90]	[1.91, 5.49]	[0.14, 1.81]

Notes: The table reports rejection frequencies from 399 completed replications for each design, using $n = 5000$, $J = 10$ and known $K = 2$, with 1999 bootstrap draws per test. For each rejection frequency (in percent), the table gives the number of rejections in parentheses and a pointwise 95 percent Wilson interval for the rejection probability in brackets. For Test A, I take the maximum studentized negative increment between adjacent assignments in the joint probabilities of treatment and regime membership. Test B uses the normalized distance from rank one of the centered matrix of these joint probabilities. The "Both reject" column reports rejection of the union null, which requires both tests to reject and has $p_{\text{joint}} = \max\{p_N, p_R\}$. Both restrictions hold when treatment propensities increase proportionally in both regimes and when one regime's treatment propensity is constant. Common direction holds but Rank-1 fails when treatment propensities increase in both regimes but not proportionally; both restrictions fail when treatment propensities increase in one regime and decrease in the other for some judge pairs.

I evaluate power at fixed alternatives in which treatment propensities increase in both regimes but not proportionally, or increase in one regime and decrease in the other for some judge pairs. The fourth design instead makes the treatment propensity in the second regime constant across instrument values, so all nine inequalities for adjacent joint probabilities of treatment and regime membership in that regime bind.

I calculate rejection frequencies using an unrestricted EM estimator of the joint probabilities $\pi_k(z)$, with eight initializations and at most 250 iterations per start. The fitted matrix has positive entries and column sums below one, but constant regime shares across instrument values also require $\sum_k \max_z \pi_k(z) \leq 1$. I do not impose this restriction in optimization. Imposing this constraint would leave the first-order limit unchanged, because the restriction

holds strictly in every simulation design and a consistent unrestricted estimator satisfies it with probability approaching one.

For each null, I construct a restricted estimate by changing the fitted joint probabilities while holding the fitted outcome means and variances fixed. I impose common direction by applying isotonic regression to each row in the fixed instrument order. For Rank-1, I center the matrix and replace it by its leading singular component, then add back the row means. If an entry of either projection is below 0.02 or a column sum exceeds 0.98, I scale the projection toward its row means until both bounds are satisfied.

All 1,596 observed fits and the two null projections of each fit satisfy $\sum_k \max_z \pi_k(z) < 1$. From each null projection I generate 1999 bootstrap samples and re-estimate the unrestricted model on each sample.

The test of common direction takes the maximum studentized negative change in a joint probability between adjacent instrument values. For studentization, I estimate each $\sigma_{k\ell}$ by the bootstrap standard deviation of $\sqrt{n}\,\widehat{g}^*_{k\ell}$ and hold that estimate fixed in the observed statistic and every bootstrap statistic. I center the bootstrap changes at the restricted estimate and take their maximum over the estimated contact set, using $\kappa_n = \sqrt{2\log\log n}$. For both tests, the bootstrap p-value is $(1 + \#\{T^*_b \geq T\})/2000$, and I reject at $p \leq 0.05$.

Even when the population increment is zero, its estimate can remain positive after isotonic projection. To remove this estimation error from the bootstrap null distribution, I center the bootstrap increments before taking the maximum over the estimated contact set, following generalized moment selection (Andrews and Soares, 2010). If all population increments are strictly positive, consistency instead implies that the observed negative-part statistic is zero with probability approaching one.

When treatment propensities increase in both regimes but not proportionally, common direction still holds, so the union null is true. The Rank-1 test rejects in every replication and the test of common direction never rejects. Consequently, the joint rule never rejects (Table 8). When some judges are more lenient toward one regime and stricter toward the other, both restrictions fail, and both tests reject in every replication.

Both Monte Carlo intervals in the two designs satisfying Rank-1 contain five percent, with rejection frequencies of 4.01 and 3.26 percent. For common direction, rejection is conservative in the design with binding inequalities: the frequency is 1.25 percent and its interval is 0.54–2.90 percent.

I estimate the contact set using $\kappa_n = \sqrt{2\log\log n}$, which is approximately 2.07 at $n = 5000$. Only nine inequalities bind in the population, whereas the estimated contact set contains 12–18 per sample, averaging 15.52. The bootstrap therefore treats strictly slack inequalities as binding in every sample, which can raise the critical value and contribute to conservative rejection.

K Additional Results for the Application to Pretrial Release

K.1 Outcome, Leniency Groups and Model Specification

The outcome for each case is the longest sentence recorded for any of the included charges, capped at 3,650 days. To form this maximum, I leave sentences recorded in days unchanged, convert the rest using 24 hours per day, 30 days per month and 365 days per year, and give life sentences values above the cap. It measures recorded sentence length, which may differ from time served.

I split the data at the defendant level, keeping each defendant's cases in the same sample. Requiring at least 20 cases per judge in the training sample removes 118 of the initial 28,152 cases in that sample and 294 of the initial 66,058 cases in the inference sample. This leaves 28,034 cases from 19,795 defendants in the training sample and 65,764 cases from 46,215 defendants in the inference sample.

The regression for the leniency measure and the mixture likelihood use the same additive controls: indicators for combinations of court, year and day of week, and indicators for combinations of court, month and day of week. Centering the 84 indicators and removing linear dependencies leaves 79 columns. For IV, I instead demean the scalar instrument within the 432 observed covariate cells, each defined by the joint values of all controls, whereas the fitted models do not have an intercept for each cell.

The leniency measure is the judge contribution from a regression of treatment on judge indicators and controls in the training sample, with the judge indicators residualized on the controls. For fixed training data T, cases in the inference sample with the same judge and controls have the same leniency measure and belong to the same group $Z = g_T(X, J)$. A change in controls can change a judge's group: this occurs for 129 of the 164 judges. The number of groups observed within a covariate cell ranges from two to nine, with a median of four, out of ten groups overall.

Within each regime, a group's treatment propensity averages the propensities of its decision makers. The affine Rank-1 representation is preserved when the conditional assignment weights in this average are common across regimes. Preserving common direction also requires the ordering condition in Proposition E.2. Because estimated residualized leniency measures can misorder the true conditional treatment propensities, grouping cases by those leniency measures does not ensure the proposition's ordering condition.

I estimate the leniency measure for each defendant in the training sample after excluding that defendant from the regression. The coefficients can therefore differ between regressions, giving different leniency measures to cases in the training sample with the same judge and controls before assignment to the common groups. Dependence among defendants assigned in the same shift can extend across the training and inference samples even though all of each

defendant's cases stay together; inference with estimated leniency measures and boundaries must allow for that dependence.

Write $h_{da}(y \mid x)$ for the two-limit censored-normal distribution obtained by censoring a normal variable with mean $\nu_{da} + x'\xi_d$ and standard deviation σ_{da} at zero and $\log(3651)$. Its likelihood contribution uses probabilities at the two endpoints and the normal density between them. All fitted standard deviations are constrained to $[0.02, 20]$. Regime-specific treatment propensities have the form

$$p_k(x, z) = \Lambda(c_{kz} + x'\gamma),$$

where Λ is the logistic distribution function, each regime has an intercept for every group, and the covariate slopes are common. No monotonicity restriction is imposed across groups.

For the two-regime model with three outcome components, let $u_a(x), u_b(x), u_c(x)$ be multinomial-logit component shares. The two regime shares and outcome distributions are

$$\begin{aligned} w_1(x) &= u_a(x) + u_b(x), & w_2(x) &= u_c(x), \\ f_{d1}(y \mid x) &= \alpha_d(x) h_{da}(y \mid x) + \{1 - \alpha_d(x)\} h_{db}(y \mid x), & f_{d2}(y \mid x) &= h_{dc}(y \mid x). \end{aligned}$$

This specification combines two outcome components into one regime by requiring them to have the same treatment propensity at every value of the instrument and controls. For components a and b, setting $\alpha_d(x) = u_a(x)/\{u_a(x) + u_b(x)\}$ in both arms also gives equal within-regime component weights across treatment statuses. In the estimated model with this restriction, components a and b are the small short-sentence component and the component concentrated near zero. The component with the larger reduction in mean sentence length forms the other regime and has its own treatment propensities. The population share of this first regime is the sum of two multinomial-logit shares, allowing it to vary with covariates more flexibly than the binary logit in the two-regime model with one outcome component per regime.

I also estimate a version in which the within-regime component weights can differ under release and detention, with

$$\operatorname{logit} \alpha_1(x) - \operatorname{logit} \alpha_0(x) = \delta_0 + x'\delta,$$

while the regime shares and treatment propensities are common to both potential outcome distributions. Within each regime, I still require the outcome distribution for a given treatment status to be the same across assignments conditional on X. This specifies a marginal mixture for each potential outcome, but does not identify the joint distribution of their component membership within the regime.

I estimate each two-regime model with three outcome components from several initial component pairings, keeping the specification fixed, and report the fit with the highest

likelihood on the estimation sample. The projected gradient of the mean negative log likelihood is below 10^{-6} for all four reported fits.

K.2 Model Comparison

Table 9 reports likelihoods and BIC for the two estimation samples. The joint likelihood includes treatment and the observed outcome conditional on treatment. BIC uses the estimation sample size and the number of free parameters. The two- and three-regime models with one outcome component per regime have 345 and 439 parameters, respectively. The two-regime model with three outcome components has 429 parameters when within-regime component weights are equal across treatment statuses and 509 when these weights can differ.

Table 9: Likelihood and BIC for the conditional models of recorded sentence length

Specification	Estimation log likelihood	BIC	Predictive log likelihood
Panel A: Fitted on the inference sample			
Two regimes	−101,955.93	207,739.24	–
Three regimes	−97,269.70	199,409.60	–
Two regimes, three components: within-regime component weights equal across treatment statuses	−97,275.62	199,310.48	–
Two regimes, three components: within-regime component weights allowed to differ across treatment statuses	−97,185.01	200,016.78	–
Panel B: Fitted on the training sample			
Two regimes	−43,261.34	90,055.89	−102,564.35
Three regimes	−41,127.00	86,749.88	−98,854.24
Two regimes, three components: within-regime component weights equal across treatment statuses	−41,132.83	86,659.13	−98,838.66
Two regimes, three components: within-regime component weights allowed to differ across treatment statuses	−41,070.68	87,354.12	−99,391.83

Notes: Estimation samples contain 65,764 cases in the inference sample or 28,034 cases in the training sample. Predictive log likelihoods are calculated on cases in the inference sample using parameters estimated on the training sample. BIC is $-2\ell + r \log n$, with number of parameters r and estimation sample size n; it is not calculated for prediction on cases in the inference sample. All rows use the same observed-data likelihood and controls. The three-component specifications each have two regimes and three outcome components under each treatment status. The two specifications differ in whether the component weights within a regime must be equal under both treatment statuses; the row labels identify this restriction.

The model with equal within-regime component weights across treatment statuses fits almost as well as the three-regime model while using two regimes. Relative to the two-regime model with one outcome component per regime, it captures 99.87 percent of the three-regime

model's likelihood gain on the inference sample. When estimated on the training sample, it has a higher log likelihood on cases in the inference sample than the three-regime model, by 15.58.

I use equal within-regime component weights across treatment statuses because allowing them to differ raises BIC despite improving the estimation likelihood. Fitting the model with different component weights on the training sample also gives a lower log likelihood on cases in the inference sample: 553.17 below the model with equal component weights and 537.59 below the three-regime model.

With one outcome component per regime, the three-regime model has a lower BIC than the two-regime model even when I raise the minimum component standard deviation to 0.05 or 0.10, or group positive sentences into broad ranges. The third normal component can form its own regime or share treatment propensities with another component. I therefore compare the three-regime model with a two-regime model containing the same number of outcome components. Each likelihood comparison uses the same observation scheme for both models: either both use the recorded sentence lengths or both use the coarsened outcomes.

K.3 Average Treatment Effects and IV Weights

Table 10 compares treatment effects and IV weights across specifications.

I report treatment effects in days of recorded sentence length, evaluating every specification on the $n = 65{,}764$ cases in the inference sample. Let L_i be capped recorded sentence length and $\mu_{dk}(x)$ its mean under treatment status d in regime k. Because the model is fitted to $Y = \log(1 + L)$, I estimate this mean by integrating $\exp(y) - 1$ over the fitted censored outcome distribution, including both endpoint masses. The difference $\tau_k(x) = \mu_{1k}(x) - \mu_{0k}(x)$ is then a difference in mean days; exponentiating a difference in mean log outcomes would give a different quantity. Averaging over covariates gives estimates of the population share and average treatment effect within regime k:

$$\widehat{W}_k = \frac{1}{n}\sum_i \widehat{w}_k(X_i), \qquad \widehat{T}_k = \frac{\sum_i \widehat{w}_k(X_i)\widehat{\tau}_k(X_i)}{\sum_i \widehat{w}_k(X_i)}.$$

The fitted population average treatment effect is $\sum_k \widehat{W}_k \widehat{T}_k$.

Let $\zeta(Z_i)$ be the mean training-sample leniency measure in case i's group and let s_i be that leniency measure minus its mean among cases in the inference sample with the same covariate cell. Define

$$\widehat{C} = \sum_i s_i \sum_k \widehat{w}_k(X_i)\widehat{p}_k(X_i, Z_i).$$

For nonzero $\widehat{C}$, the estimated aggregate IV weight and contribution for regime k are

$$\widehat{\Omega}_k = \frac{\sum_i s_i \widehat{w}_k(X_i)\widehat{p}_k(X_i, Z_i)}{\widehat{C}}, \qquad \widehat{B}_k = \frac{\sum_i s_i \widehat{w}_k(X_i)\widehat{p}_k(X_i, Z_i)\widehat{\tau}_k(X_i)}{\widehat{C}}.$$

Table 10: Effects of release on recorded sentence length and IV weights across specifications

	Regime with the largest reduction				
Specification	Effect	Share (%)	IV weight (%)	ATE	Model-implied IV
Panel A: Fitted on the inference sample					
Two regimes	−204.47	26.44	13.20	−54.05	−27.08
Three regimes	−193.02	19.12	8.52	−37.94	−17.11
Two regimes, three outcome components	−192.48	19.11	8.46	−37.82	−17.15
Panel B: Fitted on the training sample					
Two regimes	−218.22	26.11	−2.77	−56.98	10.30
Three regimes	−215.75	18.77	−8.01	−41.27	16.17
Two regimes, three outcome components	−215.79	18.77	−7.89	−41.28	16.42
Observed IV: +169.10 days (defendant-cluster SE: 207.80)					

Notes: Effects, ATEs and model-implied IV coefficients are in days of maximum recorded sentence length across included charges, capped at 3,650. All economic quantities are evaluated on the same 65,764 cases in the inference sample, including those from models estimated on the 28,034 cases in the training sample. Fitted mean sentences and probabilities of a zero sentence within regime k are averages over the inference sample with weights $\widehat{w}_k(X_i)/\sum_j \widehat{w}_k(X_j)$, where $\widehat{w}_k(X_i)$ is the fitted probability of regime membership conditional on the controls. The first two rows of each panel use one outcome component per regime in each arm; the model with two regimes and three components imposes equal within-regime component weights across treatment statuses. For each fit, the table shows the regime with the largest estimated average sentence reduction in days. Those regimes can describe different latent populations across specifications. The scalar instrument is each group's mean training-sample leniency measure, demeaned within covariate cells in the inference sample. To calculate a regime's contribution to IV, I average the product of its conditional treatment effect and IV weight over covariates. The product of the displayed average treatment effect and IV weight can differ from this contribution. I compute the observed IV with the same cases, leniency measure and centering within covariate cells. Its standard error conditions on the instrument estimated from the training sample and treats defendant clusters as independent. Table 13 reports confidence intervals for the two-regime model with three outcome components fitted on the inference sample.

The model-implied IV estimate is $\sum_k \widehat{B}_k$. Under assignment invariance, conditional centering removes the term involving mean potential outcomes under no treatment. In general, $\widehat{B}_k \neq \widehat{\Omega}_k \widehat{T}_k$, because $\widehat{T}_k$ and $\widehat{B}_k$ use different weights across covariate values. The instrument in this calculation is the group mean leniency measure from the training sample, rather than the fitted conditional treatment propensity or all judge indicators.

For a covariate pattern x, define its estimated IV weight by

$$\widehat{A}_x = \frac{\sum_{i:X_i=x} s_i \widehat{p}(X_i, Z_i)}{\widehat{C}}, \qquad \widehat{p}(x,z) = \sum_k \widehat{w}_k(x)\widehat{p}_k(x,z).$$

Using $\sum_x \widehat{A}_x \widehat{w}_k(x)$ as the share benchmark removes the difference due to covariate weighting. For the regime with the larger estimated sentence reduction in the model with equal within-regime component weights across treatment statuses, this benchmark is 17.54 percent in the model estimated on the inference sample and 14.25 percent in the model estimated on the training sample, compared with IV weights of 8.46 and −7.89 percent. Both fits allowing component weights to differ under release and detention also have IV weights below this adjusted share. The direction of underweighting therefore holds under either benchmark in these point estimates. The weights $\widehat{A}_x$ and $\widehat{\Omega}_k$ can be signed. Table 13 reports confidence intervals for the gaps in the model estimated on the inference sample with equal within-regime component weights across treatment statuses.

In the specification with different component weights under release and detention, the model estimated on the inference sample assigns 24.64 percent of the population and 14.26 percent of the IV weight to the regime with the largest estimated sentence reduction. Its estimated average treatment effect is −212.02 days. Across the population, the estimated average treatment effect is −52.25 days and the model-implied IV estimate is −29.92 days.

In the model estimated on the training sample, the regime with the largest sentence reduction has an estimated average treatment effect of −239.47 days and a population share of 24.54 percent, but its IV weight is −0.66 percent. The estimated population average treatment effect is −58.76 days, whereas the model-implied IV estimate is +2.42 days. With three outcome components and two regimes, the two specifications estimate the largest reductions for different latent populations. I therefore use each specification's own estimated regime in the comparison.

Table 11 reports the first-stage and reduced-form covariances behind the observed and model-implied IV coefficients. All three first-stage covariances are positive; the signs of the IV coefficients follow the signs of their reduced-form covariances.

K.4 Sensitivity to Leniency Groups and the Instrument

The 20- and 40-group specifications refine the ten groups using quantiles of the same leniency measures estimated from the training sample. They hold the samples, controls, outcome

Table 11: Observed and model-implied first-stage and reduced-form covariances

	First-stage covariance	Reduced-form covariance	IV (days)
Observed	0.0003843	0.064986	169.10
Model fitted on the inference sample	0.0004510	−0.007734	−17.15
Model fitted on the training sample	0.0002282	0.003747	16.42

Notes: All calculations use the same 65,764 cases in the inference sample and the same leniency measure, demeaned within covariate cells. The observed covariances are $n^{-1}\sum_i s_i D_i$ and $n^{-1}\sum_i s_i L_i$, where L_i is capped recorded sentence length in days. Model-implied covariances replace treatment and sentence length by their fitted conditional expectations in the two-regime model with three outcome components. Within-regime component weights are equal across treatment statuses. The IV coefficient is the ratio of the reduced-form covariance to the first-stage covariance.

specification and number of regimes fixed, while increasing the number of parameters from 429 to 449 and 489. I calculate the IV weights using three scalar instruments: the ten groups' mean leniency measures from the training sample, each refined grouping's own mean leniency measures from the training sample, and the continuous leniency measure. Each is demeaned within covariate cells. The first and third instruments remain fixed across fitted specifications; using each grouping's own means changes the scalar IV estimand. The underweighting reported in the main text holds for all three instruments and also when the regime share uses the same covariate weights as IV.

Table 12 reports the estimates and likelihood comparisons.

K.5 Confidence Intervals

In both two-regime specifications with three outcome components fitted on the training sample, one normal component has a standard deviation at its lower bound, so I report those fits as point estimates. The confidence intervals below are for the specification with equal within-regime component weights across treatment statuses, fitted on the inference sample.

The delta-method standard errors are based on the full sandwich covariance matrix of the observed-data likelihood estimator, with scores summed within each of the 46,215 defendants and a CR1 correction. The standard errors allow for estimation of the regime shares, treatment propensities, outcome parameters and covariate-adjusted share benchmark through this joint covariance matrix. I invert the full likelihood Hessian without ridge regularization or truncation of small eigenvalues. In the normal approximation for the conventional 95 percent intervals, I treat the component standard deviation estimated at 0.02042 on the $\log(1 + L)$ scale as interior, without adjustment for its lower bound of 0.02000.

I condition these intervals on the specification, leniency measures estimated from the training sample, group boundaries, and covariates and assignments in the inference sample and assume independence across defendants, so they exclude uncertainty from choosing the specification.

Table 12: Sensitivity to the number of leniency groups and the scalar instrument

	Fitted on the inference sample			Fitted on the training sample		
Number of leniency groups	10	20	40	10	20	40
Panel A: Regime with the largest estimated sentence reduction						
Average treatment effect (days)	−192.48	−192.45	−192.47	−215.79	−215.86	−215.66
Population share (%)	19.11	19.11	19.11	18.77	18.77	18.77
Panel B: IV weight (%)						
Mean leniency measure in the original ten groups	8.46	7.97	6.20	−7.89	−8.57	−7.06
Mean leniency measure in each group	8.46	9.03	6.46	−7.89	−7.08	−12.36
Continuous leniency measure	8.20	8.82	6.48	−7.14	−7.03	−12.19
Panel C: IV weight minus covariate-adjusted population share (percentage points)						
Mean leniency measure in the original ten groups	−9.08	−9.39	−10.17	−22.14	−22.35	−22.12
Mean leniency measure in each group	−9.08	−8.43	−9.60	−22.14	−21.01	−25.84
Continuous leniency measure	−8.95	−8.41	−9.63	−21.13	−20.73	−25.68
Panel D: Model fit relative to ten groups						
Estimation log likelihood difference	0.00	18.57	56.48	0.00	41.97	107.10
BIC difference	0.00	184.73	552.68	0.00	120.88	400.27
Predictive log likelihood difference	–	–	–	0.00	−82.27	−197.22

Notes: All models have two regimes and three censored-normal outcome components for $\log(1+L)$, where L is recorded sentence length capped at 3,650 days. Within-regime component weights are equal under release and detention. Estimation samples contain 65,764 cases in the inference sample and 28,034 in the training sample. Panels A–C refer to the regime with the largest estimated sentence reduction in each model and are evaluated on the same 65,764 cases in the inference sample. Group means use training-sample leniency measures; all instruments are demeaned within covariate cells. The adjusted population share uses the same covariate weights as each IV estimand. Panel D subtracts the ten-group value within each estimation sample. Predictive log likelihoods evaluate parameters estimated on the training sample using cases in the inference sample.

The conventional and Bonferroni-adjusted intervals appear in Table 13. The candidate family for this adjustment consists of six specifications fitted to the inference sample: the two- and three-regime models with one outcome component per regime, the two specifications with three outcome components and two regimes with ten leniency groups, and the model with equal within-regime component weights across treatment statuses using 20 or 40 leniency groups. The comparison in Figure 2 uses three specifications fitted on two samples. For each type of regime-specific quantity, there are 13 candidate estimates, including the choice of regime, compared with six for aggregate quantities. The critical values are $\Phi^{-1}(1-0.05/26)=2.891$ and $\Phi^{-1}(1-0.05/12)=2.638$.

The adjustment accounts for selecting a target from the fixed candidate family: under the regularity conditions for the estimators, the adjusted intervals cover all targets of a given type with joint asymptotic probability at least 95 percent at each fixed data-generating

distribution. This probability is not conditional on the selected model, and the guarantee applies only to searches within the family.

Table 13: Confidence intervals for the two-regime model with three outcome components

Quantity	Estimate	Conventional 95%	Bonferroni 95%
Regime with the larger reduction			
Average treatment effect (days)	−192.48	[−233.70, −151.26]	[−253.27, −131.69]
Population share (%)	19.11	[18.78, 19.43]	[18.63, 19.58]
IV weight (%)	8.46	[−3.92, 20.84]	[−9.79, 26.71]
IV weight minus population share (pp)	−10.64	[−23.03, 1.74]	[−28.91, 7.62]
IV weight minus adjusted share (pp)	−9.08	[−19.29, 1.13]	[−24.14, 5.98]
Regime with the smaller reduction			
Average treatment effect (days)	−1.29	[−1.34, −1.25]	[−1.36, −1.22]
Population comparisons			
Population average treatment effect (days)	−37.82	[−45.78, −29.85]	[−48.54, −27.10]
Model-implied IV (days)	−17.15	[−40.60, 6.30]	[−48.72, 14.42]
Model-implied IV minus ATE (days)	20.67	[−3.21, 44.55]	[−11.48, 52.82]

Notes: I fit the two-regime model with three censored-normal outcome components to 65,764 cases in the inference sample, imposing equal within-regime component weights under release and detention. I calculate both sets of intervals from the same defendant-clustered sandwich standard errors, treating covariates, assignments, leniency measures estimated from the training sample and group boundaries as fixed. For the conventional intervals, I also hold the specification fixed, and for the Bonferroni intervals in the last column, I use the critical values given above. For the adjusted share, I use the same covariate weights as IV. Shares and IV weights are reported as percentages, and their differences are in percentage points (pp). The estimated model first-stage covariance is 0.000451 (clustered standard error 0.0000644).

Sensitivity to the Minimum Component Standard Deviation. I first recalculate the covariance matrix holding fixed the estimated standard deviation of the low-variance component under no treatment. Every reported standard error changes by less than 0.23 percent. I then raise the minimum component standard deviation to 0.05 and 0.10 and refit the specification used for these intervals from three starting values for each minimum. In both cases, the fit with the highest likelihood has a projected gradient below 10^{-6}, and only the low-variance component's standard deviation reaches its bound. I hold that standard deviation fixed in the sandwich covariance calculation, so the additional confidence intervals are conditional on its value.

Table 14: Sensitivity to the minimum component standard deviation

	Minimum component standard deviation		
Quantity	0.02	0.05	0.10
Average treatment effect (days)	−192.48 (21.03)	−205.59 (21.36)	−212.99 (21.54)
Population share (%)	19.11 (0.16)	18.83 (0.16)	18.70 (0.17)
IV weight (%)	8.46 (6.31)	6.74 (6.34)	6.50 (6.36)
IV weight minus adjusted population share (percentage points)	−9.08 (5.21)	−10.32 (5.26)	−10.44 (5.27)

Notes: All columns use the two-regime model with three outcome components fitted to log one plus days, with equal within-regime component weights under release and detention, and refer to the regime with the largest estimated reduction. Parentheses report defendant-clustered sandwich standard errors. All 429 parameters are estimated when the minimum standard deviation is 0.02. At minimum standard deviations of 0.05 and 0.10, the bound binds for the low-variance component under no treatment. Its standard deviation is held fixed in the covariance calculation, leaving 428 estimated parameters. These standard errors therefore condition on that value. The adjusted population share uses the same covariate weights as IV. Samples, controls, assignment groups and the scalar instrument are unchanged.

The larger minimum component standard deviations increase the estimated reduction from 192.48 days to 205.59 and 212.99 days, with standard errors still close to 21 days (Table 14). The IV weight remains below the population share, while the conventional 95 percent interval for its difference from the adjusted population share changes from $[-19.29, 1.13]$ percentage points with the minimum standard deviation set to 0.02 to $[-20.62, -0.02]$ and $[-20.78, -0.10]$ in the two conditional calculations.