A Joint Bayesian Framework for Estimating Racial Disparity in Imprisonment When the Measures Disagree

Miura Meng

2026-07-11

Abstract

Every measure of racial disparity in imprisonment is a function of two quantities: the rate at which a state imprisons its Black residents and the rate at which it imprisons its white ones. Four summaries are in common use; because each collapses two dimensions into one, they can disagree on the same data about direction, not just magnitude. This paper makes the disagreement the object of study. It puts one question, whether more democratic U.S. states punish Black and white residents more equally, through the four measures and three further analytic choices: the comparison design, a clustering of states into types, and the reading of a voting-rights reform. The answer changes at nearly every fork, and the changes share one anatomy: white imprisonment falls with state democracy while Black imprisonment stays flat, so a one-number summary can deteriorate with nothing changing for Black residents. Rather than choose among summaries, I model the two rates jointly and read every measure off one posterior. Disagreement becomes an estimable quantity: the probability that stronger democracy goes with a higher ratio and lower white imprisonment is 0.97. Institutions that attach formula grants and sentencing law to these numbers are grading, in part, a denominator.

Introduction

Racial disparity in imprisonment is a property of two numbers, and it is almost always reported as one. The compression is convenient. It is also not innocent: the one-number summaries in common use, the ratio, the gap, the absolute rate, each discard a different part of the pair, and on the same data they can disagree about whether a place is becoming more equal or less, and sometimes about the direction itself. Criminology has long argued over which summary is the right one, with the answer known to hinge on the benchmark in the denominator, on the attribution of differential involvement, and on the data source itself (Beck & Blumstein, 2018; Cesario et al., 2019; Sabol et al., 2025). The argument does not end because it cannot: it is an argument about which dimension of a two-dimensional object to throw away. The arithmetic behind the disagreement is elementary, and this paper does not present it as news. What practice lacks is a way to tell when the arithmetic, rather than anything about Black residents, is moving the number; providing one is the contribution.

Nor does the choice stay inside scholarship. How equally states punish Black and white residents is among the most consequential questions in the study of race and punishment (Western, 2006), and public institutions answer it with these summaries every day. The federal juvenile justice system grades states with a minority-to-white rate ratio, with a fifth of each state's formula grant attached to the result; racial impact statements evaluate proposed sentencing law with the same family of measures; and the best-known national rankings of imprisonment disparity are orderings of the Black/White ratio. Under that ranking New Jersey runs the widest racial disparity in the country, while by its absolute Black imprisonment rate it sits below the national average (Nellis, 2021), and both readings are available to whoever prefers them.

The case in this paper makes the stakes concrete. Asked whether more democratic states punish Black and white residents more equally, the most common measure answers that inequality is worse where democracy is healthier. The underlying rates answer differently: white imprisonment collapses as state democracy rises while Black imprisonment barely moves, so the ratio climbs although nothing changes for Black residents, and the same reversal appears for poverty, urbanization, and the party of the governor. The disagreement is not noise to be cleaned up by choosing a better number. It is built into the fact that four summary statistics are being asked to stand in for two underlying quantities.

That observation points to a different approach. Rather than choose among the competing measures, I model the two quantities they all summarize: the rates at which a state imprisons its Black and its white residents. In a single Bayesian framework, every measure, the ratio, the gap, the absolute rates, emerges as a contrast of one posterior. The measures stop being rival summaries and become different views of the same estimated quantities. Their disagreement, which the conventional approach leaves the reader to reconcile, becomes something the model states directly, as a probability. This is the paper's central contribution: not a verdict on which measure of racial disparity is correct, but a way to avoid having to choose one.

The pieces of the diagnosis are individually known: Muller (2012) traced the North's wider ratios to white imprisonment long before the current policy era, and Sabol et al. (2025) document how measured disparity moves with data and definitions. What this paper adds is the constructive step, a single model that absorbs the choices and prices their disagreement as a probability, and a demonstration of how much that step matters on a live substantive question.

The model comes last because the demonstrations give it its content. The earlier sections carry the same question through four ordinary analytic choices and show how far each one moves the conclusion: the measure of disparity, the comparison between states and within them, a clustering of states into types, and the reading of a major voting-rights reform. Only with those choices on the table does the unifying model do its work. The substantive case, race, democracy, and punishment across the U.S. states, is worth studying for its own sake; it is also the clearest setting I know in which to show why measurement needs a model rather than a verdict.

Background

Measurement, framing, and what counts as a finding

A claim about inequality is never simply a reading of the evidence. It is also a product of the choices made in turning evidence into a number, and a number into a sentence. Scholars writing under banners such as data feminism and QuantCrit have pressed this point in recent years, arguing that data are not neutral and that statistics, far from speaking for themselves, carry the assumptions of whoever collected, classified, and summarized them (D'Ignazio & Klein, 2020; Gillborn et al., 2018). The complaint is not that quantification is worthless but that it hides its own premises. Measurement of inequality is a clear example: faced with the same data, an analyst can often choose the measure that best supports a preferred conclusion, and the decision between an absolute and a relative measure alone can send a trend in opposite directions (Kjellsson et al., 2015; Moonesinghe & Beckles, 2015). Framing does comparable work even when the numbers are held fixed. Describing a gap as one of "achievement" rather than "opportunity," for instance, shifts where readers locate its causes, away from structures and toward individuals, without altering a single statistic (Quinn, 2025). As a finding travels outward from the model that produced it, moreover, the qualifications that gave it meaning tend to drop away, so that a conditional finding begins to circulate as a settled fact (van der Bles et al., 2020; Merry, 2016). A more constructive response to the same problem is to stop hiding the choices and report a result across all of the defensible ones, the approach formalized as the garden of forking paths and the multiverse and specification-curve analyses it motivated (Gelman & Loken, 2014; Steegen et al., 2016; Simonsohn et al., 2020; Young & Holsteen, 2017). This paper works in that tradition.

What this tradition has lacked is a way past the choice itself. This paper offers one: a multiverse of measurement and design choices, usually presented as a row of separate results, can be reorganized as contrasts of a single Bayesian model that both explains the disagreements and quantifies how reliable they are. The components of that model are standard; what is new is the combination and the use it is put to. The second contribution is the demonstration that motivates the model: a single consequential question carried through the ordinary choices an analyst makes, with the answer shown to move at each step. The paper takes one case and walks it through four routine decisions, treating each decision as part of the analysis rather than as a preliminary to it. The insight that measurement and framing matter is old; the worked example shows how much they matter, and how readily they compound, once a real claim about racial inequality is at stake.

Measuring racial disparity in incarceration

The most common summary of racial disparity in imprisonment is a single ratio, the Black imprisonment rate divided by the white rate. Its appeal is obvious, but a ratio can move for more than one reason. A state can post a high ratio because it imprisons Black residents at an unusually high rate, or because it imprisons white residents at an unusually low one, and these are very different social facts that the ratio reports identically. Criminologists have long shown that conclusions about racial disproportionality hinge on such choices: on how much of a gap is attributed to differential involvement rather than to downstream processing (Beck & Blumstein, 2018), on which benchmark population is placed in the denominator (Cesario et al., 2019), and even on the data source and racial-classification scheme, which can widen or narrow the measured Black/White gap (Sabol et al., 2025). Rather than commit to one measure in advance, this paper carries three in parallel, the ratio, the absolute Black rate, and the gap between the Black and white rates, alongside the white rate that underlies them all, and treats their disagreement as part of the evidence.

The case: state democracy, disenfranchisement, and Florida

Criminology gives reasons to expect a link between state politics and race-specific punishment: racial threat theory predicts more punitive control where the Black population looms larger (Blalock, 1967), and state imprisonment rates respond to partisan control and conservative public opinion (Jacobs & Carmichael, 2001). State democracy itself can be measured and tracked. The State Democracy Index scores each state on features such as electoral participation, the fairness of district maps, and the accessibility of voting, and those scores move over time, including downward in a subset of states (Grumbach, 2022, 2023). Criminal disenfranchisement, meanwhile, is one of the oldest and most racially loaded boundaries on the vote. Many states wrote or expanded these laws in the decades after the Civil War, some in terms understood at the time as a way to limit Black political power, and the legacy persists: roughly four million Americans remain barred from voting by a felony record, with Black citizens barred at several times the rate of others (Behrens et al., 2003; Manza & Uggen, 2006; The Sentencing Project, 2024). Florida concentrated this history into a single recent episode. In 2018 its voters passed Amendment 4, restoring the vote to about 1.4 million people upon completion of their sentences. Within months, however, the legislature passed SB7066, which redefined “completion” to require paying all outstanding court fines and fees first, a condition critics likened to a poll tax and a federal appeals court allowed to stand in 2020.

Data and measures

The analysis combines three public datasets, each at the state-year level. The State Democracy Index 2.0 (Grumbach & Bitton, 2024) scores the fifty states from 2000 to 2023, with higher values meaning more democratic. The Vera Institute's Incarceration Trends (Vera Institute of Justice, 2024) provides race-specific imprisonment rates per 100,000, counting people held in state prisons rather than local jails, ages 15 to 64; the Black and white rates are used here from 2000 to 2022, the years the State Democracy Index covers. The UF Election Lab turnout series (McDonald, 2023) gives voting-age-population (VAP) turnout for each even-year election.

Because racial inequality has no single definition, the paper follows four measures side by side: the Black/White rate ratio, the gap between the Black and white rates, and the two absolute rates, Black and white, that underlie them. Over this period the national ratio sits near a median of 5.4. For the natural experiment the outcome is VAP turnout, whose denominator is the voting-age population and so does not move mechanically when people are re-enfranchised.

Demonstration 1: the measure decides the sign

Begin with the question in its most direct form. For each state I average the Black/White imprisonment ratio and the State Democracy Index over 2016 to 2020, and plot one against the other (Figure 1). The line slopes upward: the more democratic a state is, the wider its Black/White ratio tends to be. Read at face value, this is an uncomfortable result, because it seems to say that democratic health and racial fairness in punishment are in tension, and that the states doing democracy best are the ones treating Black and white residents most unequally behind bars.

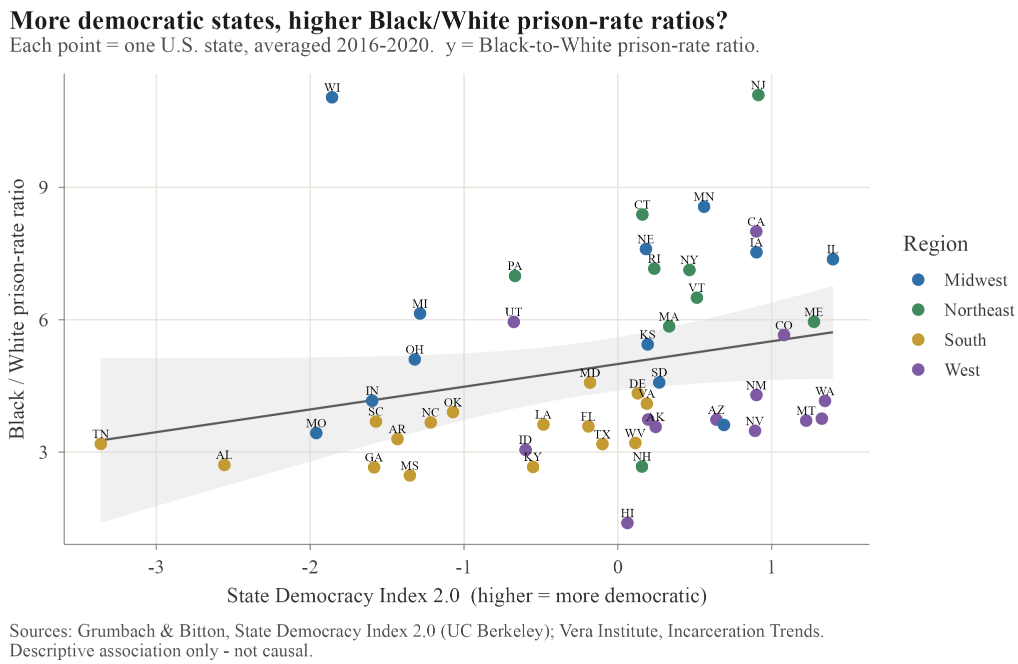

Figure 1: Across states, averaged over 2016 to 2020, the Black/White imprisonment ratio rises with the State Democracy Index.

Before accepting that, it is worth asking whether the result is a fact about the world or a fact about the ratio. I hold the horizontal axis fixed, so the measure of democracy never changes, and I substitute different measures of "racial inequality" on the vertical axis (Figure 2). If the relationship is real, it should survive the substitution. It does not. The Black/White ratio correlates positively with democracy, at about +0.26, which is the pattern we just saw. The absolute Black imprisonment rate, though, correlates slightly negatively, at about −0.14, so by that measure more democratic states imprison Black residents at marginally lower rates rather than higher ones. The gap between the Black and white rates is essentially flat, at about −0.04. The absolute white rate, for its part, correlates strongly negatively, at about −0.34. Four measures of the supposedly same inequality, and they do not merely differ in strength; they point in different directions.

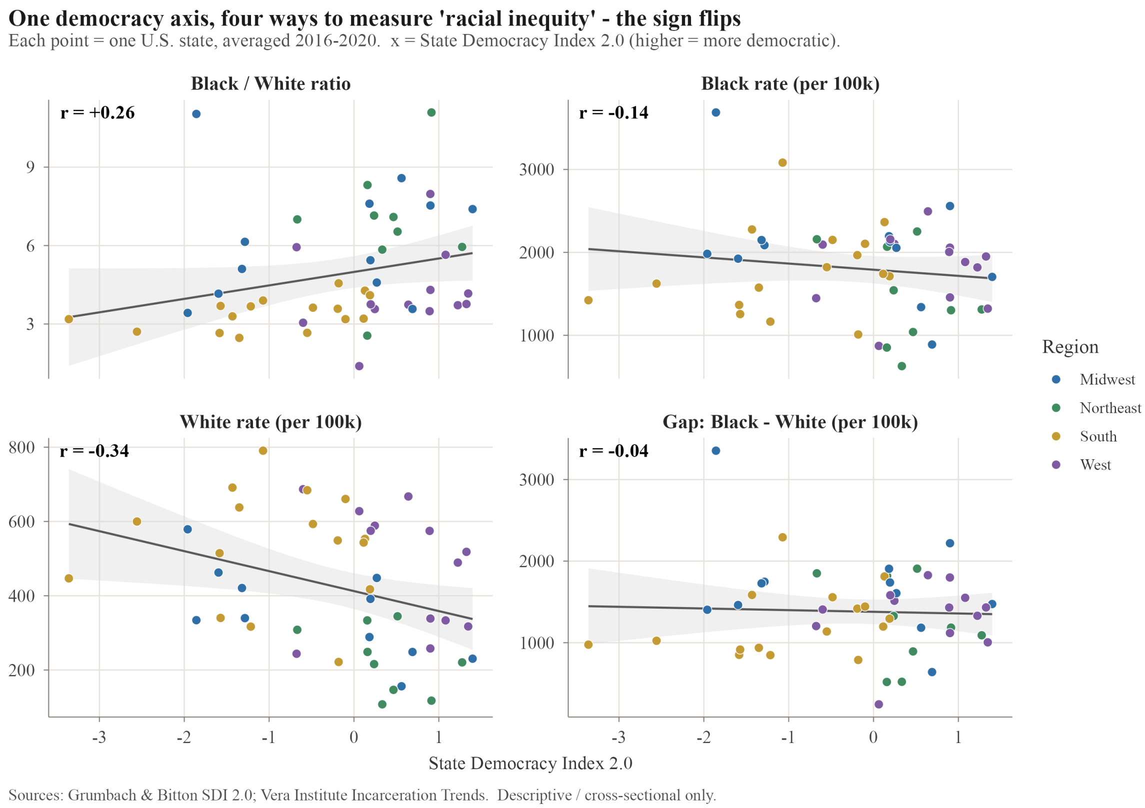

Figure 2: The same democracy axis against four measures of racial inequality. The sign of the relationship flips depending on the measure.

The reason comes into focus once we look at what the ratio is made of. It is the Black rate divided by the white rate, so its value depends as much on the denominator as on the numerator. The point is an identity rather than an observation. On the log scale the ratio is a difference of the two log rates,

$$\log R = \log B - \log W, \tag{1}$$

so a regression of the log ratio on any state characteristic splits exactly into two slopes,

$$\beta_R = \beta_B - \beta_W, \tag{2}$$

and a characteristic correlated with white imprisonment and unrelated to Black imprisonment will manufacture a disparity finding by arithmetic alone. (The identity is exact for log-scale regression coefficients; raw-scale correlations, like those reported in Figure 2 and Appendix A, obey no such additivity.) Democracy turns out to be exactly such a characteristic. More democratic states imprison white residents at strikingly low rates, and that small denominator does the arithmetic work, inflating the ratio even in a state where Black imprisonment is no higher than average. The South shows the reverse. Southern states post lower ratios, but not because they treat Black and white residents more evenly; they imprison white residents heavily too, and the large white denominator pulls the ratio down. The regional averages put numbers on this. White imprisonment runs about 230 per 100,000 in the more democratic Northeast and about 535 in the South, more than twice as high, while the Black rate stays within a much narrower band, roughly 1,460 to 2,060 across the four regions. The ratio therefore runs highest on average across the low-white-rate Northeast and lowest across the high-white-rate South, even though the South is the less democratic region. The regional pattern itself is old: northern states have posted wider Black/White imprisonment disparities than southern ones since long before the current policy era (Muller, 2012). The measure reported most often, then, is also the one most exposed to movement that has nothing to do with how Black people are treated, since a state can climb the ratio simply by incarcerating fewer white people. Judged by the absolute Black rate, more democratic states look slightly better; judged by the gap, they look no different; judged by the ratio, they look worse. Nothing about the data has changed between those three sentences. Only the measure has.

This is not a quirk of incarceration data. The divergence between absolute and relative measures of disparity is a known formal result, demonstrated in health research where the two can move in opposite directions over the same period (Moonesinghe & Beckles, 2015), and the dependence of measured racial disparity on such choices has been shown directly for imprisonment (Sabol et al., 2025). What the figure adds is the size of the turn: one decision, made before any modeling begins, has already settled whether democracy looks like a friend or an enemy of racial equality. Nor is the flip a special property of the democracy index. Poverty, urbanization, and the governor's party, common structural determinants of incarceration and political economy, reproduce the same pattern across the four measures. By the ratio, states with Republican governors rank among the most racially equal, because the governor's party correlates with white imprisonment and is essentially unrelated to Black imprisonment; Appendix A reports the full grid.

The same arithmetic runs through the field's most-quoted trend. Nationally, the Black/White prison-rate ratio narrowed from about 7.4 in 2000 to 4.2 in 2019, a development widely and rightly read as progress. Decomposed on the log scale, though, only 71 percent of that narrowing came from Black imprisonment falling; the other 29 percent came from white imprisonment rising, by about 17 percent over those years, the years of the opioid crisis. More than a quarter of the improvement in this celebrated trend is the denominator moving in a direction no one celebrates. Extending the window through 2022 shrinks the white-rate share to about 5 percent, because pandemic-era declines pulled the white rate back down; Appendix B gives the decomposition.

Demonstration 2: comparing states, or comparing a state with itself

Comparing different states with one another folds democracy together with everything else that distinguishes them, including region, demographics, history, and the local economy, and the first demonstration has already shown how badly that folding can distort an answer. A cleaner question is longitudinal. When a single state's democracy rises or falls, does its incarceration move with it? Posed this way, each state serves as its own control, and the fixed differences between states drop out of the comparison.

To carry this out I split the democracy score into two parts. The first is each state's long-run average, which holds the between-state variation that the earlier figures relied on; the second is each year's deviation from that average, which holds the within-state variation that tracks change over time. The split follows Mundlak (1978). I then fit a linear mixed model with a random intercept for each state and a common year trend, using the lme4 package (Bates et al., 2015), so that the two kinds of variation are estimated side by side rather than confused with each other (Figure 3). In equation form,

$$y_{st} = \alpha_s + \beta_{\mathrm{b}}\,\bar{D}_s + \beta_{\mathrm{w}}\,(D_{st} - \bar{D}_s) + \gamma\,t + \varepsilon_{st}, \qquad \alpha_s \sim N(\alpha_0, \sigma_\alpha^2),$$

where y_{st} is one log outcome (the Black rate, the white rate, or their ratio) in state s and year t, $\bar{D}_s$ is the state's long-run mean democracy score, and β_{b} and β_{w} are the between- and within-state coefficients plotted in Figure 3.

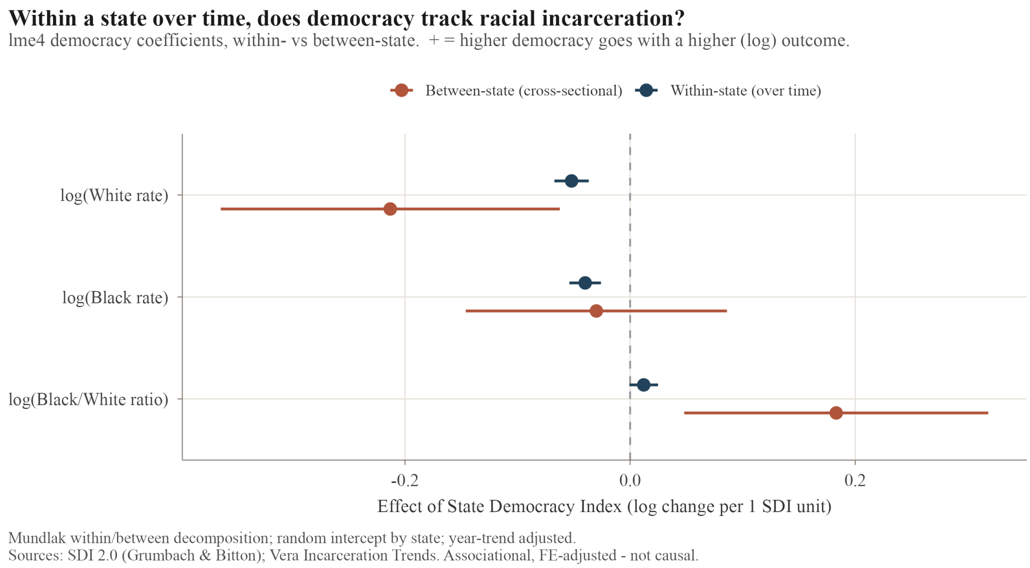

Figure 3: Within-state versus between-state democracy coefficients. The cross-sectional ratio effect (orange) disappears within states (blue).

The alarming cross-sectional result for the ratio turns out to be entirely between-state, at about +0.18, and to vanish within states, at about +0.01. In other words, it reflects which states are which, not what happens when a given state's democracy changes. Within states, stronger democracy goes with lower imprisonment of both groups, by roughly −0.04 for the Black rate and −0.05 for the white rate in log points per index point, and because the two rates fall together the ratio barely moves, exactly as the first demonstration would predict. The largest term in every model, however, is neither within nor between democracy but the year trend: the ratio fell about 2.6 percent a year and the Black rate about 2.7 percent, a nationwide decline that has little to do with state democracy. An analysis that left time out would mistake that tide for its own discovery. So the within-state link points consistently downward, but it is modest, and, as the later model shows, how real it looks depends on whether the strong year-to-year persistence is modeled. One choice of comparison, cross-section against within-state, has already turned a striking effect into a small one.

Demonstration 3: a method that reads in more than the data hold

An average describes the typical state but conceals the differences between states, and those differences are part of the story. So the next question is whether states fall into recognizable types according to how democracy and incarceration moved together. Group-based trajectory modeling, developed in criminology by Nagin and Land (1993) to classify offending over the life course, is built for exactly this. Using the gbmt package (Magrini, 2022), it sorts the fifty states by their joint path from 2000 to 2022 on two series at once, democracy and the logged Black imprisonment rate. A BIC search over two to five classes is lowest at five (4186, 3924, 3869, 3815), the boundary of the searched range, with an average assignment probability near 0.99 (Figure 4).

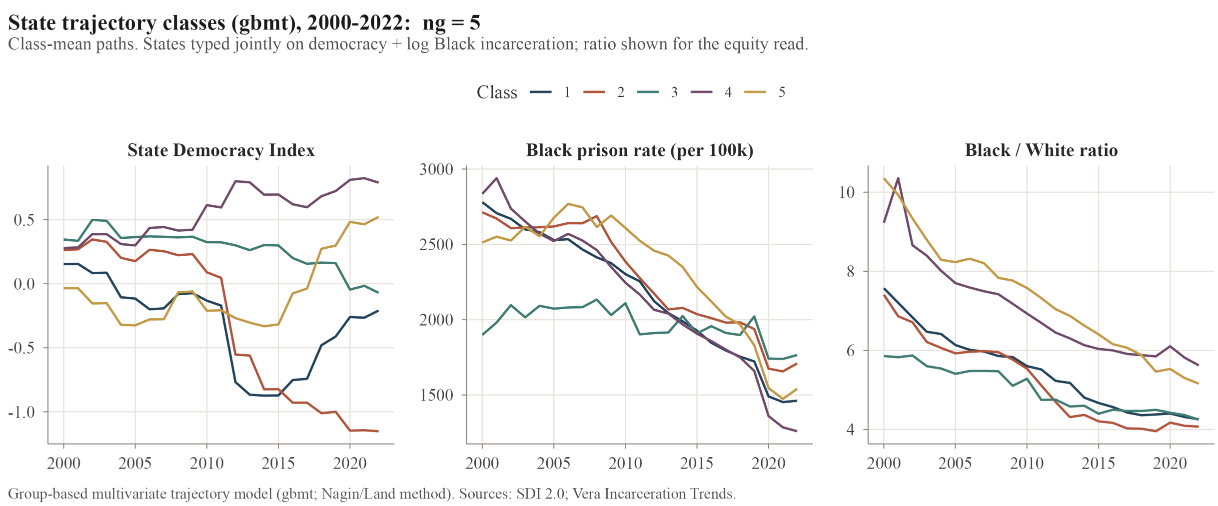

Figure 4: Five state trajectory classes from the gbmt model. Each line is a class-mean path for democracy, Black imprisonment, and the Black/White ratio, on raw scales; the clustering itself uses two inputs, democracy and the logged Black rate, so the ratio panel is descriptive rather than an input.

At the extremes the classes echo the earlier pattern. The class that ended most democratic cut Black imprisonment the most, from about 2,835 to 1,260 per 100,000, and its white rate fell the most as well; the class that backslid the furthest, the group Grumbach (2023) calls the laboratories of backsliding, cut Black imprisonment by about a third and saw its white prison rate rise, from about 387 to 466 (the white series is not shown in the figure). Between the extremes the ordering is loose: the flattest Black-rate path belongs to a class in the middle of the democracy range, not to the deepest backslider. The method invites overreading, though, and the details argue for restraint. Every class decarcerated, because the national decline runs through all of them, so the groups differ in degree rather than in direction. The ratio converged toward four or five across the board. The BIC, moreover, keeps improving as classes are added, so the figure of five is not firmly pinned down; highly autocorrelated series like these can produce classes that are not really there. The classes are useful summaries, not natural kinds, and they remain entangled with region. A method can hand back a clean typology that looks more solid than the data warrant, and choosing to read it as description rather than discovery is itself part of the analysis.

Demonstration 4: a reform by its text, or by its effect

The state-level patterns are associations, not causes, and to get closer to cause I turn to a natural experiment. In 2018 Florida passed Amendment 4, restoring the vote to about 1.4 million people with felony convictions, the largest such restoration in modern U.S. history. With a single treated state I use synthetic control (Abadie et al., 2010), implemented with the tidysynth package (Dunford, 2023). The method assembles a synthetic Florida from a weighted combination of states that did not change their felon-voting laws, chosen so that the combination reproduces Florida's turnout history before the reform, and then reads the reform's effect as the gap that opens afterward. The outcome is VAP turnout, whose denominator does not move when people are re-enfranchised, and a placebo test provides inference by asking how unusual Florida's gap is against the same gaps computed for every donor state.

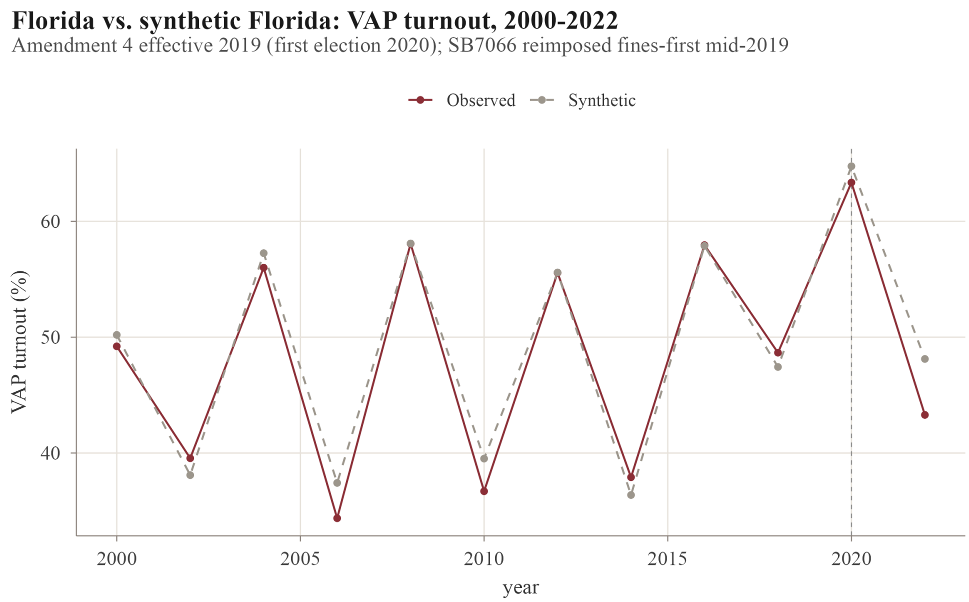

Figure 5: Florida compared with synthetic Florida on VAP turnout. The observed series (red) follows the synthetic series (grey) through 2018, then falls below it after the reform, most sharply in 2022.

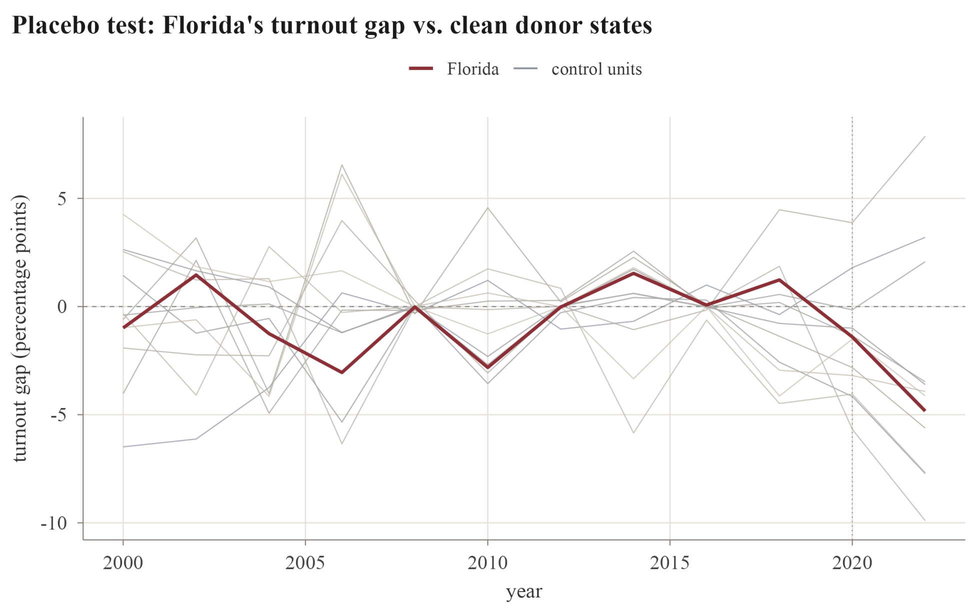

Figure 6: Placebo test. Florida's post-treatment gap (red) sits inside the cloud of donor-state placebos.

The pre-reform fit is reasonable rather than sharp: the pre-period root mean squared prediction error is 1.6 percentage points, with gaps of +1.5 in 2014, +0.1 in 2016, and +1.2 in 2018 (Figure 5). After the reform, the gaps are −1.4 points in 2020 and −4.8 in 2022, both negative rather than positive. The placebo test returns a permutation p-value of about 0.14: Florida ranks third of the twenty-one units on the post-to-pre prediction error ratio, high in the placebo distribution but short of conventional significance (Figure 6; Appendix C reports the donor pool, weights, and fit diagnostics). Amendment 4 produced no measurable increase in turnout.

Two further checks indicate the null is not an artifact of a single specification. An in-time placebo that moves the treatment back to 2016, before the reform, yields only small gaps in 2016 and 2018, the years before the real reform, about +1.0 points and +2.1; these are modest and positive, the opposite sign of the post-reform gaps, so the design does not manufacture the result, though the 2018 gap signals a mild pre-existing drift, plausibly including mobilization by the Amendment 4 campaign itself, which was on the November 2018 ballot. A leave-one-out analysis that re-fits the model dropping each weighted donor in turn, with synthetic Florida drawing mainly on Arizona and New Hampshire, leaves the 2020 gap between −1.6 and +1.1 points and the 2022 gap between −5.4 and +1.3, straddling zero in every case, so no single donor drives the estimate. The negative 2022 gap is therefore best read as a fluctuation around zero rather than as turnout suppression: it is not statistically distinguishable from zero and reverses when individual donors are removed.

A null result is only worth reporting if its source is visible, and here it is: as Morse (2021) documents, the reform barely took effect. SB7066 required people to pay all outstanding court fines, fees, and restitution before regaining the vote, and unpaid court debt kept most of the people the amendment covered from regaining it. The requirement works as a wealth test on the franchise and falls hardest on poor and disproportionately Black citizens (Meredith & Morse, 2017). Its effect is visible in the counts: in 2024, six years after the amendment passed, more than 700,000 Floridians who had completed their sentences remained disenfranchised for unpaid legal financial obligations, more than in any other state (The Sentencing Project, 2024). On paper the reform enfranchised more than a million people. In practice it reached few, and whether you call it sweeping or empty depends on which number you read.

From multiverse to model

The four demonstrations are a small multiverse: one question carried through choices that should not change the answer, so that the disagreements among the answers can show. This section turns the core of that multiverse into a single Bayesian model, in which the measure and comparison choices reappear as contrasts of one posterior. The other two forks, the clustering and the reading of the reform, are different designs rather than different summaries, and stay outside the model.

Before collapsing the choices into a model, it helps to see them all at once. Figure 7 is the full specification curve: the democracy-inequality association estimated under every crossing of the four inequality measures, the two comparison designs, and the year trend in or out, sixteen commonly run specifications in all, with each outcome standardized so the effects are comparable. The estimates run from clearly negative to clearly positive. The most negative are the white imprisonment rate compared across states; the most positive are the Black/White ratio compared across states; the within-state specifications sit near zero. The sign of the headline is chosen at the same moment as the specification. This is the multiverse the rest of the section sets out to discipline.

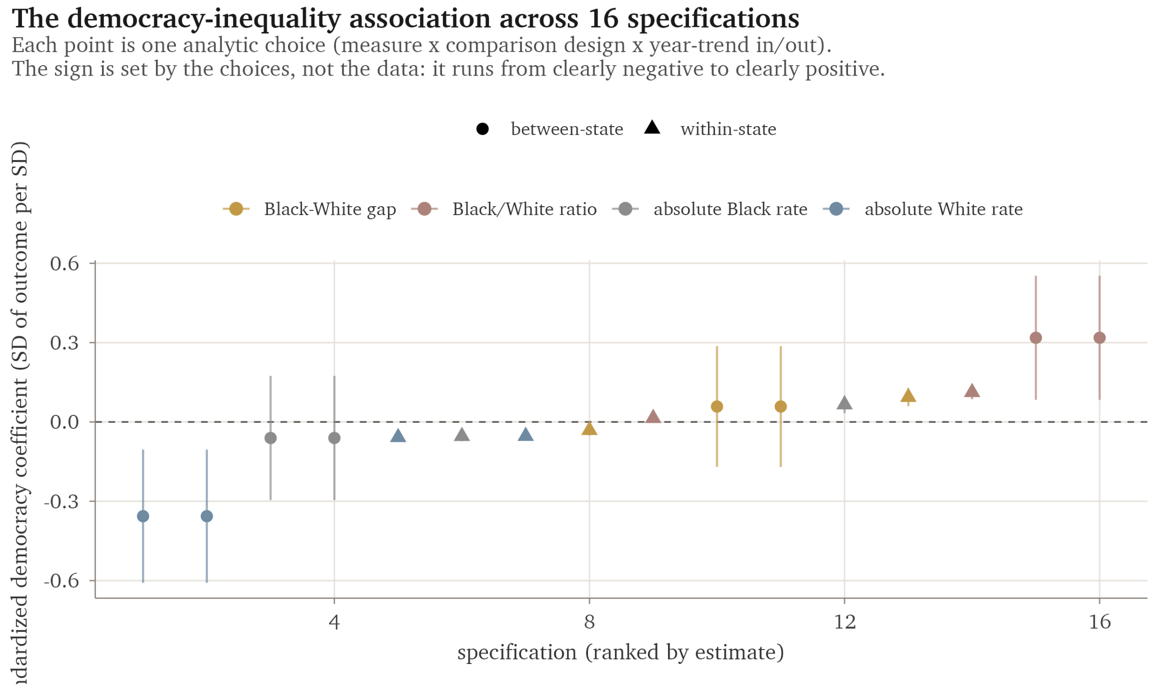

Figure 7: The democracy-inequality association across sixteen specifications (four inequality measures, two comparison designs, year trend in or out), each outcome standardized. The sign runs from clearly negative (the white rate compared across states) to clearly positive (the Black/White ratio compared across states); within-state specifications sit near zero.

The first two demonstrations treated the choice of measure and the choice of comparison as separate forks in the analysis. They are not separate. The four measures and the within and between contrasts are all summaries of the same two underlying quantities, how heavily a state imprisons its Black residents and how heavily it imprisons its white residents, and how each of those moves with democracy. Rather than walk the forks one at a time, then, we can fit a single model of those two quantities and read every fork off the one set of estimates it implies.

To do this I fit one joint multilevel model with two outcomes at once, the log Black imprisonment rate and the log white rate, each allowed to depend on democracy. The estimation sample is 1,149 state-year observations across the fifty states, 2000 to 2022. Democracy enters in the Mundlak form of the second demonstration, split into a between-state part, a state's long-run average, and a within-state part, its year-to-year deviation; each state keeps its own intercept, and the two outcomes' intercepts and residuals are free to correlate; within each state the residuals follow a first-order autoregressive process, so the strong year-to-year persistence the second demonstration warned about is modeled rather than assumed away. Written out,

$$\log B_{st} = \alpha_s^B + \beta_{\mathrm{b}}^B\,\bar{D}_s + \beta_{\mathrm{w}}^B\,\tilde{D}_{st} + \gamma^B\,t + \varepsilon_{st}^B, \tag{3}$$

$$\log W_{st} = \alpha_s^W + \beta_{\mathrm{b}}^W\,\bar{D}_s + \beta_{\mathrm{w}}^W\,\tilde{D}_{st} + \gamma^W\,t + \varepsilon_{st}^W, \tag{4}$$

with $\bar{D}_s$ and $\tilde{D}_{st}$ the standardized between- and within-state democracy terms, t a scaled year trend, state intercepts (α_s^B, α_s^W) drawn from a bivariate normal with a freely estimated correlation, first-order autoregressive residuals $\varepsilon_{st}^k = \rho^k\,\varepsilon_{s,t-1}^k + u_{st}^k$ within each state, for each outcome $k \in \{B, W\}$, and the contemporaneous shocks u_{st}^B and u_{st}^W also free to correlate. Regression slopes carry Normal(0, 1) priors; the remaining priors are the brms defaults: Student-t(3, 0, 2.5) on the random-effect and residual standard deviations, LKJ(1) on both correlation matrices, Student-t intercepts centered on the sample medians, and a flat prior on the autoregressive coefficients. The between-state ratio contrast in Figure 8 is the posterior difference $\beta_{\mathrm{b}}^B - \beta_{\mathrm{b}}^W$. The model is fit with the brms package (Bürkner, 2017) and converged cleanly, with every R-hat below 1.01 and no divergent transitions; the autoregressive parameters are estimated at $\rho^B = 0.95$ (0.92 to 0.98) and $\rho^W = 0.98$ (0.96 to 1.00), which is just how serially dependent these series are. The coefficients below are scaled per one standard deviation of democracy rather than per index point, so their sizes are not directly comparable to those in the second demonstration. The signs match; the within-state contrasts, as reported below, no longer clear zero once the serial dependence is priced.

Once the model is fit, three of the four measures are no longer separate analyses but contrasts of one posterior, and the fourth, the raw Black-minus-white gap, a simple function of it (Figure 8). Between states, a one standard deviation increase in democracy goes with white imprisonment lower by about 0.13 log points (95 percent credible interval −0.25 to −0.01) and Black imprisonment that is essentially flat (−0.01, −0.11 to +0.09). The ratio contrast is just the difference between these two, and it comes out positive, about +0.12 (+0.01 to +0.23), for exactly the reason the first demonstration gave: the ratio rises not because Black imprisonment rises but because the white denominator falls. The cross-sectional reading that suggests worsening disparity and the one that suggests none are the same posterior, read two ways. Within states, once that serial dependence is modeled, the year-to-year effects shrink toward zero and their intervals straddle it; the modest within-state link of the second demonstration is itself sensitive to whether that dependence is acknowledged.

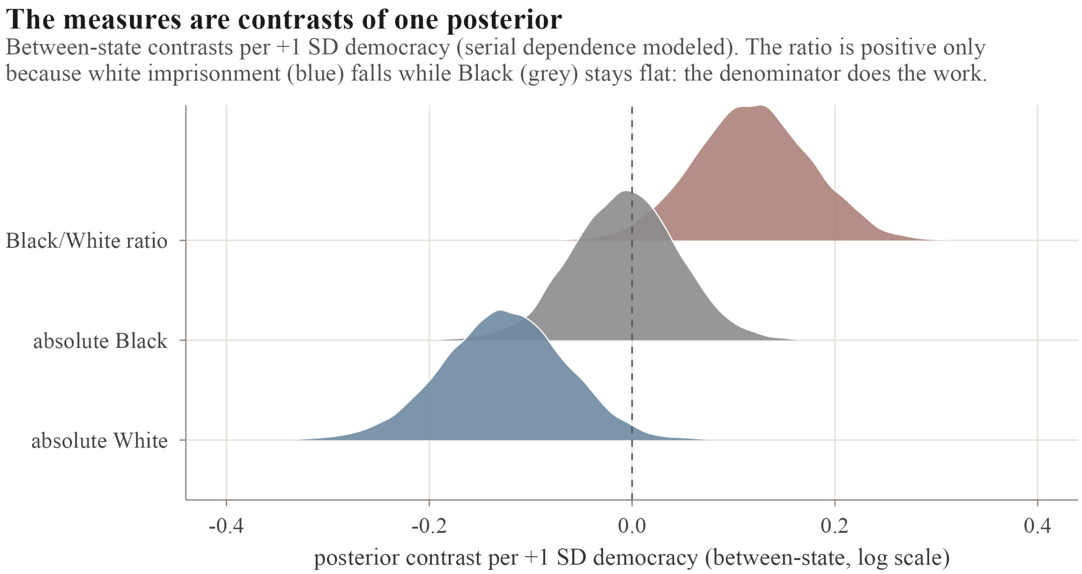

Figure 8: The competing measures of racial inequality as contrasts of one posterior. Each density is the between-state contrast associated with a one standard deviation difference in democracy; the raw Black-minus-white gap (not shown) is a simple function of the two rates. The Black/White ratio (red) sits above zero only because white imprisonment (blue) falls while the Black rate (grey) stays essentially flat; a single posterior produces a rising ratio and a falling white rate at the same time.

Holding the measures inside one model buys something the side-by-side comparison cannot. Because every contrast is computed from the same draws, we can ask not just whether the measures disagree but how reliably they do. The posterior probability that stronger democracy goes with a higher ratio and lower white imprisonment, the exact shape of the first demonstration's reversal, is about 0.97 once the serial dependence is modeled. The same quantity is 0.99 when that dependence is ignored, 0.98 when census region enters the simpler model as a fixed effect, and 0.86 on the 38-state subsample of the recording screen (Appendix D); the modeled figure is the honest headline. The weaker version, that the ratio rises while the absolute Black rate falls, is no better than a coin toss, about 0.54, because the Black rate is too close to flat to sign. So the dependable claim is the specific one: more democratic states post a higher ratio while imprisoning white residents at lower rates. That reversal is not a fragile accident of one arbitrary comparison. It is a robust feature of how the two rates move, reported as a probability that has already paid the price of the autocorrelation.

The reversal also survives the other obvious checks, run as variants of the model without the autoregressive term. A negative-binomial count model on the raw counts gives a between-state ratio contrast of +0.15 (95 percent credible interval +0.04 to +0.26) with a white-rate contrast of −0.17 (−0.28 to −0.05); census region as a fixed effect gives +0.11 (+0.00 to +0.22) with −0.14 (−0.26 to −0.03). Each is a positive ratio contrast carried by the falling white rate.

None of this makes the relationship causal. The model rearranges the same observational evidence more economically; it does not identify an effect. Its value is methodological. The earlier demonstrations show, in the accessible form of a multiverse, that the answer moves with the analyst's choices; this model shows why, by placing those choices as contrasts of a single set of estimates, and it puts a number on how much each choice matters. A multiverse is the better way to reveal the problem, and a model that absorbs the sources of variation is the more disciplined way to study it (Gelman et al., 2012). The two are complementary, and the counsel to carry more than one measure only gains force once the measures can be shown to be one model seen from different sides.

Discussion

The four demonstrations make a single argument. In each, a routine decision settles the result: a measure of inequality reverses the sign, a comparison design shrinks a striking effect to a small one, a clustering method offers more structure than the data hold, and a reform reads as sweeping or negligible depending on whether one consults its text or its effect. That such choices can decide the answer is not, by itself, a new point. The contribution is what the unifying model does with it. Rather than argue over which measure of racial disparity is correct, it fits one Bayesian model of the two underlying rates, the rates at which a state imprisons its Black and its white residents, and reads each competing measure off that model as a contrast of its posterior. The disagreement among the measures, left elsewhere for the reader to settle, becomes a quantity the model reports, as a probability. The move, throughout, is to stop choosing among measures and model what they summarize.

What is easy to lose, stated only as a principle, is how concretely and how quickly the choices bite, and how they accumulate. Here they are shown one after another, with numbers, on a case where the stakes are real: the most common measure of racial disparity in incarceration points the wrong way for a mechanical reason, and a celebrated voting-rights reform reads as a triumph or a non-event depending on the frame. The practical counsel is simple to state: report more than one measure, and say so when they disagree; keep within-state change separate from between-state differences; treat the groups a clustering routine hands back as convenient summaries rather than discoveries; and judge a policy by its effect rather than its text.

The substantive estimates are deliberately unspectacular, and their size is part of the argument: across these analyses, the movement produced by the analyst's choices exceeds the movement produced by any predictor in the data. State democracy and racial inequality in punishment are linked weakly at best; any within-state link is small beside a national decline in incarceration and is sensitive to how serial dependence is handled; and re-enfranchisement on paper can be undone by a fee. One result travels beyond this case: the Black/White imprisonment ratio worsens with democracy only because the white rate falls, not because Black imprisonment moves. A widely reported equity metric can move for reasons that have nothing to do with how Black people are treated, and the institutions of the introduction put their weight on exactly this arithmetic: the Relative Rate

Index behind the federal juvenile justice system's disparity requirement (Office of Juvenile Justice and Delinquency Prevention [OJJDP], n.d.), the racial impact statements that evaluate proposed sentencing law (Mauer, 2007), and the rankings on which New Jersey, Wisconsin, and Minnesota, the low-white-denominator states of Figure 1, sit at the top (Nellis, 2021). Each grades states with a measure that can deteriorate when white imprisonment falls. Taken together, the demonstrations show how easily the convenient reading fails.

Limitations

Several limits bound the claims. Aggregate turnout is a blunt instrument: even a fully effective Amendment 4, involving about 8 percent of the voting-age population, most of whom never register or vote (Uggen & Manza, 2002), might move statewide turnout by less than a point, below what this design can pick up. The null means the effect was too small for this test to see, not that re-enfranchisement does nothing, and a sharper test would follow the formerly incarcerated themselves in voter files. The first three demonstrations are observational. They help frame the question; they do not answer it causally. State democracy and incarceration are themselves estimated quantities. The synthetic control rests on a single treated state, an imperfect midterm pre-fit, a small pool of clean donor states, and a 2020 treatment year disrupted by the pandemic; the 2022 gap also coincides with an unusually uncompetitive Florida cycle while the heavily weighted donors, Arizona and New Hampshire, held close races, so part of that gap is plausibly electoral competition rather than the reform.

The analysis is also narrower than the question. The racial contrast is binary: Black and white imprisonment rates are the series Vera reports most completely, so Latino, Native American, and other groups, and the states' uneven recording of Hispanic ethnicity, fall outside the analysis.

The recording problem is the sharper worry: states that fold Hispanic prisoners into the white count inflate their white rates, and those states skew southern, so the practice could mimic the denominator finding. Appendix D screens them out and refits the model; the pattern survives in sign and attenuates in strength. The state, finally, is the unit throughout. State rates average over counties that differ sharply in policing and prosecution, so nothing here speaks to disparity within a state.

The "democracy" coefficient itself needs careful reading. The State Democracy Index moves with party control of state government, the very lever the punishment literature emphasizes (Jacobs & Carmichael, 2001), with region, and with a state's broader political culture, and the analysis does not try to pull these apart. The robustness check that adds census region as a fixed effect preserves the reversal, but a "democracy" coefficient here is best read as tracking a bundle of liberal state governance rather than democratic institutions in isolation. Distinguishing the two would take a design built for that question.

Conclusion

This paper has used one case, race and democracy and punishment in the U.S. states, to make a methodological point: the answer depends on how we measure and frame the question, often more than on the data. Change the measure and a disparity reverses; change the comparison and an association shrinks; change the frame and a reform that restored the vote, on paper, to 1.4 million people did almost nothing. The data are not the weak link; the choices made around them deserve the same scrutiny as the findings themselves. When the measures disagree, the better response is to model what they share rather than to crown one.

Data and code availability

All data are public and documented in the repository. Code and data to rebuild the merged dataset, every model, and every figure are available at https://github.com/haomeng797-ship-it/dem-crime-equity.

Appendix A: the sign flip generalizes beyond democracy

Demonstration 1 varies the measure while holding the democracy axis fixed. This appendix varies the axis. Three common structural determinants of incarceration and political economy take the place of the State Democracy Index: the poverty rate (Census SAIPE, averaged over 2016 to 2020), the percent of the population living in urban areas (2020 Census), and the governor's party (January 2019, the midpoint of the window). The four disparity measures are state means over 2016 to 2020, as in Figure 2. The correlations across the fifty states:

State characteristic	Ratio	Black rate	White rate	Gap
State Democracy Index	+.26	−.14	−.34	−.04
Poverty rate	−.40	+.18	+.52	+.02
Percent urban	+.27	−.12	−.27	−.04
Republican governor	−.42	+.02	+.38	−.10

Every axis reproduces the flip. Poorer states look more equal by the ratio only because they imprison white residents at much higher rates; their Black rates are slightly higher, not lower. Urbanization mirrors the democracy index. The governor's party is the cleanest case: it correlates with white imprisonment and is essentially unrelated to Black imprisonment, so by the ratio, states with Republican governors rank among the most racially equal. These are descriptive snapshots;

the governor's party stands in for a bundle of policy and regional differences, and no partisan effect is claimed. Recoding the party variable as the share of the window's years with a Republican governor gives the same pattern (ratio −.27, white rate +.27, Black rate +.08). The covariate file and the script that produces the table are in the repository.

Appendix B: decomposing the national trend

National Black and white prison rates are aggregated from the Vera counts and populations (prison, ages 15 to 64), holding the panel to states with data in both endpoint years (all fifty qualify) so that changes in which states report cannot masquerade as trend. From 2000 to 2019 the Black rate fell from about 2,516 to 1,697 per 100,000, a 33 percent decline, while the white rate rose from about 342 to 400, a 17 percent increase. On the log scale the ratio's decline of 0.55 splits into 0.39 from the Black rate and 0.16 from the white rate, a 71/29 split. Extending the window from 2000 to 2022 gives a 95/5 split, because pandemic-era declines left the white rate only 3 percent above its 2000 level. The script that produces these figures is in the repository.

Appendix C: synthetic control details

Twenty-nine states have a complete VAP-turnout series across the twelve even-year elections from 2000 to 2022: Florida, the treated unit; eight states that also changed their felon-voting rules between 2016 and 2022 and are therefore dropped (California, Louisiana, Nevada, New Jersey, New Mexico, New York, Virginia, and Washington; four further reformers, Colorado, Connecticut, Iowa, and Kentucky, already fail the completeness screen); and twenty clean donors: Alaska, Arizona, Hawaii, Idaho, Illinois, Indiana, Kansas, Maryland, Massachusetts, Michigan, Montana, New Hampshire, North Dakota, Ohio, Oregon, Rhode Island, South Dakota, Utah, Vermont, and Wyoming.

Synthetic Florida is built from mean pre-period turnout over 2000 to 2018 and turnout in the presidential years 2008, 2012, and 2016. The optimizer places almost all weight on four donors: Arizona (0.49), New Hampshire (0.36), Utah (0.07), and Illinois (0.06); every other donor receives less than 0.01. The pre-period fit has a root mean squared prediction error of 1.59 percentage points, and the 2016 gap is 0.07 points.

Year-by-year gaps between Florida and synthetic Florida are +1.5 points in 2014, +0.1 in 2016, +1.2 in 2018, then −1.4 in 2020 and −4.8 in 2022. The post-to-pre mean squared prediction error ratio is 9.2; the routine dates the intervention at the 2020 election and scores only 2022 as post, with 2020 itself in neither window (post MSPE 23.3 against pre MSPE 2.5 over 2000 to 2018), and the same convention applies to every placebo. The ratio places Florida third among the twenty-one units; permutation inference on that rank gives p = 3/21, about 0.14. The in-time placebo that moves treatment back to 2016 fits the genuinely pre-treatment years with an RMSPE of 1.64 (gaps of +0.6 points in 2012 and +1.1 in 2014) and, in the placebo's own post-period, produces only the small positive gaps reported in the text, +1.0 in 2016 and +2.1 in 2018, the opposite sign of the true post-reform gaps. Dropping each weighted donor in turn leaves the 2020 gap between −1.6 and +1.1 points and the 2022 gap between −5.4 and +1.3.

Appendix D: the reversal without the misclassifying states

States differ in whether Hispanic prisoners are recorded as Hispanic or folded into the white count, and the folding states skew southern and less democratic, so the practice could mimic the finding that the white denominator drives the ratio. A transparent screen flags states where Latinos are at least 4 percent of the working-age population yet the recorded Latino prison rate is missing or below 60 percent of the white rate, a value far below the national pattern and a strong sign that

Hispanic prisoners are being counted elsewhere. Twelve states fail the screen: Alabama, Arkansas, Florida, Georgia, Hawaii, Kentucky, Louisiana, Michigan, Missouri, Montana, Tennessee, and Virginia.

Dropping them leaves the sign flip intact in the cross-section: across the remaining thirty-eight states the ratio's correlation with democracy is +.13 while both absolute rates stay negative (Black −.25, white −.27). Refitting the AR(1) joint model on the subsample gives a between-state ratio contrast of +0.08 (95 percent credible interval −0.05 to +0.21) and a white-rate contrast of −0.12 (−0.28 to +0.03); the posterior probability that democracy goes with a higher ratio and lower white imprisonment is 0.86, against 0.97 in the full sample. The pattern survives in direction and attenuates in strength. Part of the attenuation is mechanical, since the screen removes most of the South and with it much of the between-state variation in democracy; part may be real, and is worth knowing. Either way the reversal is not a bare artifact of where Hispanic prisoners are recorded, and the sensitivity of its sharpness to a recording screen is itself one more instance of the paper's point.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association, 105*(490), 493–505. https://doi.org/10.1198/jasa.2009.ap08746

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beck, A. J., & Blumstein, A. (2018). Racial disproportionality in U.S. state prisons: Accounting for the effects of racial and ethnic differences in criminal involvement, arrests, sentencing, and time served. *Journal of Quantitative Criminology, 34*(3), 853–883. https://doi.org/10.1007/s10940-017-9357-6

Behrens, A., Uggen, C., & Manza, J. (2003). Ballot manipulation and the "menace of Negro domination": Racial threat and felon disenfranchisement in the United States, 1850–2002. *American Journal of Sociology, 109*(3), 559–605. https://doi.org/10.1086/378647

Blalock, H. M. (1967). *Toward a theory of minority-group relations*. Wiley.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Cesario, J., Johnson, D. J., & Terrill, W. (2019). Is there evidence of racial disparity in police use of deadly force? Analyses of officer-involved fatal shootings in 2015–2016. *Social Psychological and Personality Science, 10*(5), 586–595. https://doi.org/10.1177/1948550618775108

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press.

Dunford, E. (2023). *tidysynth: A tidy implementation of the synthetic control method* [R package].

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460–465. https://doi.org/10.1511/2014.111.460

Gillborn, D., Warmington, P., & Demack, S. (2018). QuantCrit: Education, policy, "Big Data" and principles for a critical race theory of statistics. *Race Ethnicity and Education, 21*(2), 158–179. https://doi.org/10.1080/13613324.2017.1377417

Grumbach, J. M. (2022). *Laboratories against democracy: How national parties transformed state* *politics*. Princeton University Press.

Grumbach, J. M. (2023). Laboratories of democratic backsliding. *American Political Science Review, 117*(3), 967–984. https://doi.org/10.1017/S0003055422000934

Grumbach, J. M., & Bitton, F. (2024). *State Democracy Index 2.0* [Data set]. Democracy Policy Lab, University of California, Berkeley.

Jacobs, D., & Carmichael, J. T. (2001). The politics of punishment across time and space: A pooled time-series analysis of imprisonment rates. *Social Forces, 80*(1), 61–89. https://doi.org/10.1353/sof.2001.0070

Kjellsson, G., Gerdtham, U.-G., & Petrie, D. (2015). Lies, damned lies, and health inequality measurements: Understanding the value judgments. *Epidemiology, 26*(5), 673–680. https://doi.org/10.1097/EDE.0000000000000319

Magrini, A. (2022). *gbmt: Group-based multivariate trajectory modeling* [R package].

Manza, J., & Uggen, C. (2006). *Locked out: Felon disenfranchisement and American democracy*. Oxford University Press.

Mauer, M. (2007). Racial impact statements as a means of reducing unwarranted sentencing disparities. *Ohio State Journal of Criminal Law, 5*(1), 19–46.

McDonald, M. P. (2023). *United States Elections Project: Voter turnout data, 1980–2022* [Data set]. UF Election Lab, University of Florida.

Meredith, M., & Morse, M. (2017). Discretionary disenfranchisement: The case of legal financial obligations. *The Journal of Legal Studies, 46*(2), 309–338. https://doi.org/10.1086/694323

Merry, S. E. (2016). *The seductions of quantification: Measuring human rights, gender violence, and sex trafficking*. University of Chicago Press.

Moonesinghe, R., & Beckles, G. L. A. (2015). Measuring health disparities: A comparison of absolute and relative disparities. *PeerJ, 3*, e1438. https://doi.org/10.7717/peerj.1438

Morse, M. (2021). The future of felon disenfranchisement reform: Evidence from the campaign to restore voting rights in Florida. *California Law Review, 109*(3), 1143.

Muller, C. (2012). Northward migration and the rise of racial disparity in American incarceration, 1880–1950. *American Journal of Sociology, 118*(2), 281–326. https://doi.org/10.1086/666384

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica, 46*(1), 69–85. https://doi.org/10.2307/1913646

Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model. *Criminology, 31*(3), 327–362. https://doi.org/10.1111/j.1745-9125.1993.tb01133.x

Nellis, A. (2021). *The color of justice: Racial and ethnic disparity in state prisons*. The Sentencing Project.

Office of Juvenile Justice and Delinquency Prevention. (n.d.). *National racial and ethnic disparities (R/ED) databook: Measuring disparity*. https://ojjdp.ojp.gov/statistical-briefing-book/data-analysis-tools/r-ed-databook/measuring-disparity

Quinn, D. M. (2025). Experimental effects of "opportunity gap" and "achievement gap" frames. *Sociology of Education*. https://doi.org/10.1177/00380407251321372

Sabol, W. J., Johnson, T. L., & Lynch, J. P. (2025). Discrepancies in measures of racial and ethnic disparities: Implications for research, policy, and practice. *American Journal of Criminal Justice*. https://doi.org/10.1007/s12103-025-09810-1

The Sentencing Project. (2024). *Locked out 2024: Four million denied voting rights due to a felony conviction*.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature*

Human Behaviour, 4(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Uggen, C., & Manza, J. (2002). Democratic contraction? Political consequences of felon disenfranchisement in the United States. *American Sociological Review, 67*(6), 777–803. https://doi.org/10.2307/3088970

van der Bles, A. M., van der Linden, S., Freeman, A. L. J., & Spiegelhalter, D. J. (2020). The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences, 117*(14), 7672–7683. https://doi.org/10.1073/pnas.1913678117

Vera Institute of Justice. (2024). *Incarceration trends* [Data set].

Western, B. (2006). *Punishment and inequality in America*. Russell Sage Foundation.

Young, C., & Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research, 46*(1), 3–40. https://doi.org/10.1177/0049124115610347