American Economic Journal: Economic Policy 2021, 13(2): 207–240
https://doi.org/10.1257/pol.20180688

Racial Divisions and Criminal Justice: Evidence from Southern State Courts[†]

By Benjamin Feigenberg and Conrad Miller*

The US criminal justice system is exceptionally punitive. We test whether racial heterogeneity is one cause, exploiting cross-jurisdiction variation in punishment severity in four Southern states. We estimate the causal effect of jurisdiction on arrest outcomes using a fixed effects model that incorporates extensive charge and defendant controls. We validate our estimates using defendants charged in multiple jurisdictions. Consistent with a model of ingroup bias in electorate preferences, the relationship between local severity and Black population share follows an inverted U-shape. Within states, defendants are 27–54 percent more likely to be incarcerated in "peak" heterogeneous jurisdictions than in homogeneous jurisdictions. We estimate that confinement rates and race-based confinement rate gaps would fall by 15 percent if all jurisdictions adopted the severity of homogeneous jurisdictions within their state. (*JEL* H76, J15, K42)

The United States incarcerates residents at a higher rate than any other country in the world. While less than 5 percent of the world's population resides in the United States, nearly 25 percent of the world's prison population is held in US facilities (Walmsley 2016). Though differences in violent crime rates can in part explain this pattern, the United States is also exceptionally punitive (Pfaff 2014). Some observers have argued that race plays a key role in driving American criminal justice policy (Alexander 2010). There is prima facie evidence: Black Americans are incarcerated at six times the rate of whites and face longer sentences for similar crimes (Carson 2014, Rehavi and Starr 2014). Race may play a broader role, even influencing the incarceration rate for white Americans, which itself would rank near the top among developed nations (Gottschalk 2015). Just as racial heterogeneity predicts lower support for redistribution and public goods (Alesina, Baqir, and Easterly 1999), it may increase support for harsher punishment if, for example, voters prefer

* Feigenberg: Department of Economics, University of Illinois at Chicago, 601 South Morgan Street, Chicago, IL 60607 (email: bfeigenb@uic.edu); Miller: Haas School of Business, University of California, Berkeley, 2220 Piedmont Avenue, Berkeley, CA 94720, and NBER (email: ccmiller@berkeley.edu). Matthew Notowidigdo was coeditor for this article. Eric Hernandez and Colin Ash provided excellent research assistance. We thank Amanda Agan, Ernesto Dal Bo, Fred Finan, Anil Jain, Jonathan Leonard, Abhishek Nagaraj, Aurelie Ouss, Evan Rose, Yotam Shem-Tov, and seminar participants at Dartmouth College, Pomona College, UC Berkeley, UCLA, University of Illinois at Chicago, University of Chicago Crime Lab, Chicago Harris, Duke, and Brown for comments. We thank the Center for Science and Law for providing access to Alabama court data, and we thank Claire Lim, James Snyder, Jr., and David Strömberg for generously sharing voting data they collected on state ballot propositions.

[†] Go to https://doi.org/10.1257/pol.20180688 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

to punish outgroup members more severely. In this paper, we ask whether racial heterogeneity can in part explain US exceptionalism in criminal justice.

Empirical research on the role of race in criminal justice policy is complicated by the difficulty of separating the relative importance of policy versus underlying criminal conduct in generating cross-country variation in incarceration rates. Harmonized micro data covering the United States and a significant number of other countries do not exist, and differences in the definitions of crimes across countries would make harmonization difficult. Instead, we study the relationship between racial divisions and criminal justice policy by investigating cross-jurisdiction variation in punishment *within* US states. In doing so, we take advantage of harmonized data and fixed criminal codes within states and exploit the substantial within-state variation in how criminal law is enforced.

While much statutory criminal justice policy is driven by state-level legislation, localities have significant discretion in how they enforce those laws, and that discretion is tied to electorate preferences. Prosecutors and judges are often locally elected and influence outcomes at each stage of the criminal justice process: prosecutors decide what charges to file and negotiate plea bargains; judges make sentencing decisions after conviction. The electorate may affect adjudication outcomes by serving as jurors or influencing spending on indigent defense. A 2016 *New York Times* article illustrates the role of local politics in driving local punishment severity with a quote from the elected prosecutor in Dearborn County, Indiana:

> I am proud of the fact that we send more people to jail than other counties. … My constituents are the people who decide whether I keep doing my job. The governor can't make me. The legislature can't make me. (Keller and Pearce 2016)

In this paper, we evaluate the role that racial heterogeneity plays in determining criminal justice outcomes. We first estimate local *punishment severity*, the causal effect of jurisdiction on the outcome of a criminal arrest charge, using data from four Southern states. We then link variation in punishment severity to local racial heterogeneity in the population. Consistent with a simple model of ingroup bias in electorate preferences, we find that the relationship between local punishment severity and the Black population share follows an inverted U-shape: jurisdictions with the largest white and Black shares are relatively lenient while heterogeneous jurisdictions are more punitive.

To measure punishment severity, we use rich criminal justice administrative data that track criminal charges from arrest through sentencing. Our benchmark measures of punishment severity are the jurisdiction fixed effects we estimate in a regression of charge outcomes on an extensive set of covariates, including defendant demographics and criminal history, the specific arrest charge, and the year of the charge. Importantly, our data include arrest charges that are dropped by prosecutors and convictions that do not result in incarceration sentences. By contrast, many past studies of racial disparities in sentencing use data that only include

convictions that lead to incarceration sentences and so are subject to selection bias concerns.[1]

To the extent that the rich covariates included in benchmark models fully account for those determinants of charge outcomes that are correlated with jurisdiction, our estimates will provide unbiased causal measures of jurisdiction effects. To provide support for this causal interpretation, we develop and apply a quasi-experimental research design that exploits variation in outcomes for defendants arrested in multiple jurisdictions (Chetty, Friedman, and Rockoff 2014). We show that our benchmark punishment severity estimates accurately predict the *within-defendant* changes in charge outcomes coinciding with changes in jurisdiction. Throughout the analysis, our benchmark specifications focus on the share of charges that lead to incarceration sentences (the *confinement rate*) as the relevant measure of punitiveness, though we present supplementary analyses that confirm that our findings are similar if we use conviction or sentence length as the outcome or employ case-level rather than charge-level specifications.

The data cover charges from 2000 to 2014 in Alabama, North Carolina, Texas, and Virginia, with ranges varying by state (Alabama Administrative Office of Courts 2017, North Carolina Administrative Office of the Courts 2015, Texas Department of Public Safety 2015, Virginia Office of the Executive Secretary 2016). These states account for about 20 percent of all prisoners held under state jurisdiction in the United States. We focus on the South because there is substantial variation in racial composition across southern counties. In all four states, district attorneys are locally elected; in all but Virginia, judges are locally elected. The data reveal significant within-state variation in jail and prison admissions that is matched by substantial heterogeneity in punishment severity. A defendant charged in a jurisdiction in the top quartile by punishment severity is 1.8 to 3.6 times more likely to be incarcerated for a given charge than the same defendant charged in a jurisdiction in the bottom quartile. We find that 80–93 percent of the difference in confinement rates between top and bottom quartile jurisdictions is explained by the causal effect of jurisdiction. Interestingly, punishment severity estimates constructed separately by defendant race are highly correlated. Jurisdictions that are more punitive for Blacks are also more punitive for whites.

We next document the relationship between local punishment severity and racial heterogeneity. We motivate our analysis with a simple model of ingroup bias where voters prefer more severe punishment when offenders are more likely to belong to a different racial group. Prior work documents that common group membership is associated with declines in envy and punishment for misbehavior (Chen and Li 2009). This mechanism implies that the relationship between local punishment severity and the Black share of the population (or share of defendants) will follow an inverted U-shape. While white voters prefer more punitive policy as the Black share

[1] See, for instance, Miethe (1987). It is important to note that the extent of selection bias may be more limited in federal criminal cases than in state cases. Fischman and Schanzenbach (2012), for instance, conditions on conviction but argues that associated selection bias is limited because acquittals account for only 1 percent of the federal criminal cases that they analyze. A closely related paper to ours, Rehavi and Starr (2014), uses data tracking federal criminal cases from arrest through sentencing and finds that, conditional on an arrest charge, a prosecutor's initial court charge is an important driver of racial disparities in sentencing.

of defendants increases, for jurisdictions with sufficiently large Black populations, the pivotal voter is more likely to be Black and to support less harsh punishment.

Lacking a natural experiment that generates variation in racial composition across jurisdictions, we test for an inverted U-shaped pattern in the cross section. The predicted relationship is borne out in the data and the magnitude of the relationship is large. Our estimates imply that punishment severity peaks where the Black share of the population (defendants) is around 0.3 (0.4). At this peak, predicted confinement rates for a given offense are 24 (43) log points larger than in a jurisdiction with a Black share of the population (defendants) that is zero.[2] Notably, we do not find evidence of nonmonotonic relationships between punishment severity and other jurisdiction characteristics, and we estimate that selection on unobservables would have to be substantially larger in magnitude than selection on observables to explain the cross-sectional relationship between punishment severity and Black share.[3]

We conclude by simulating outcomes under a counterfactual in which more heterogeneous jurisdictions within a state adopt the punishment severity imposed by those at the tenth percentile of the predicted confinement rate distribution based on Black population share. Under this counterfactual, overall confinement rates and racial confinement rate gaps fall by approximately 15 percent, on average, once we account for both the static effect of lower punishment severity on confinement outcomes and the dynamic effect of lower punishment severity on defendants' criminal histories.

Our work contributes to a political economy literature that studies the association between local racial composition and policy preferences. Alesina, Baqir, and Easterly (1999) provide evidence that public goods spending is inversely related to ethnic fragmentation in US cities and argue that this finding is driven by cross-group policy preference heterogeneity. Luttmer (2001) shows that self-reported support for welfare spending is increasing in the share of local recipients from the respondent's own racial group and Dahlberg, Edmark, and Lundqvist (2012) find that plausibly exogenous increases in immigration to Swedish municipalities are associated with decreases in support for redistribution. We argue that the inverted U-shaped relationship between Black population share and severity of incarceration policy in our data can be explained by the same racial ingroup bias that drives the positive association between racial homogeneity and support for redistribution.

In emphasizing racial divisions as a key driver of electoral preferences and local punitiveness, we build on a large literature that highlights the racialized nature of crime policy in the United States (Muhammad 2010) and the role of "racial threat" in explaining policy and punishment preferences (Key 1949, Glaser 1994, Enos 2016, Unnever and Cullen 2007). The most recent and compelling evidence suggests that a larger minority population increases white voter turnout and support for conservative policies and candidates (Enos 2016). A related body of work finds that whites

[2] When we adjust for other jurisdiction characteristics, the difference between "peak" heterogeneous and homogeneous jurisdictions is reduced but remains substantial at 14–27 log points. By contrast, *within* jurisdictions, the "unexplained" Black-white gap in confinement rates varies from 11–19 log points across states. For comparison, Rehavi and Starr (2014) find a 10 percent unexplained gap in sentence length in federal courts.

[3] Our approach is similar in spirit to Finkelstein, Gentzkow, and Williams (2016) who decompose geographic variation in Medicare spending into location and patient effects by exploiting patient migration across markets and then correlate estimated location and patient effects with observable characteristics.

who express more racial resentment or are primed to consider the prison population as "more Black" are more likely to support harsh crime-control policies (Unnever and Cullen 2010, Hetey and Eberhardt 2014). While we cannot measure local preferences directly, we measure local policy in the form of punishment severity.[4]

Motivated by the racial threat hypothesis, several papers test for a relationship between state racial composition and imprisonment rates, with mixed results. Most relevant to our work, Keen and Jacobs (2009) finds an inverted U-shaped relationship between Black population share and racial *disparities* in state prison admissions per capita. In contrast with this past research, we focus on county-level criminal justice and use within-defendant variation in jurisdiction to credibly isolate the causal effect of charge location on sentencing. We find an inverted U-shaped relationship between county Black population share and punishment severity that applies to *all* defendants.

Our findings provide a potential explanation for a pattern that has been documented in several recent papers: courts and police officers appear to be more punitive in areas with larger nonwhite populations (Rehavi and Starr 2014; Raphael and Rozo 2018; Goncalves and Mello, forthcoming). Each of these papers is focused on measuring racial disparities in outcomes, and finds that observed gaps decrease with the inclusion of locality fixed effects. By contrast, we focus on estimating unbiased measures of locality punishment severity itself, and our research design is suited for this objective. Moreover, while the papers above find that local punitiveness is positively correlated with the Black share of the local population, a key prediction of our ingroup bias model is that the relationship is *nonmonotonic*, and we document this nonmonotonic relationship empirically.

Our paper also builds on a literature that documents the effects of electoral pressure on the composition and behavior of judges and, to a lesser extent, prosecutors (see, for instance, Huber and Gordon 2004; Berdejó and Yuchtman 2013; Lim 2013; Lim, Snyder, and Strömberg 2015a; Dyke 2007; and Nelson 2014). In our model, the predicted relationship between local punishment severity and racial composition that we document is mediated through electorate preferences. We provide support for this interpretation using data on local voting for statewide ballot measures aimed at increasing punishment harshness or limiting the rights of the accused (Lim, Snyder, and Strömberg 2015b).[5]

The remainder of the paper is structured as follows. Section I describes the data used for the analysis. Section II discusses our approach to characterizing cross-jurisdiction differences in punishment severity, including our validation strategy using defendants arrested in multiple jurisdictions, and provides estimates. Section III presents a model of racial ingroup bias to highlight the role that racial divisions may play in explaining this variation and empirically tests the predictions of the model. In Section IV, we summarize results from counterfactual confinement rate simulations and conclude.

[4] As discussed below, we also analyze data on local voting for criminal justice-related statewide ballot measures, a noisy proxy for local punishment preferences.

[5] Specifically, we find that electoral support for these measures predicts more severe punishment and also has an inverted U-shaped relationship with the local Black share.

I. Data

We use administrative criminal justice data from four states: Alabama, North Carolina, Texas, and Virginia.[6] The data source and years of data we analyze for each state are presented in Table 1. We summarize the content of the data here and discuss data construction and state-specific institutional context in greater detail in online Appendix A.

One key distinction across states is the data source. The data from Alabama, North Carolina, and Virginia are administrative court records, and include relatively detailed and complete information on criminal charges starting from the time they are filed in court. In principle, a limitation of these data is that they do not include information on criminal charges prior to court filing. Fortunately, in these states *all* arrests based on probable cause result in court charges, so we effectively have data on all valid arrests.[7] The Texas data are maintained by the Texas Department of Public Safety, and include data from arresting agencies (e.g., police departments), prosecutors, and courts. These data contain records for all qualifying arrests, including arrests that did not lead to a court charge. However, the data contain less detailed information on court processes and do not identify whether a charge was ever filed in court.

Though the data from each state differ in their exact content, they all track state felony and misdemeanor criminal charges from arrest through sentencing, and share important data elements. Critically, data for all states include arrest charges that are ultimately dropped. Data from all states include information on each criminal charge, including the original arrest charge, the date of arrest, the court where the charge is assigned, final court charge, charge disposition, and, if the charge results in conviction, the final sentence. The data allow us to group charges into cases. Defendant information includes date of birth (except Virginia, which does not include year of birth), gender, and race. Data from North Carolina and Texas also identify Hispanic defendants.

For all states, the data include property, violent, and drug offenses. We refer to offenses in these categories as "core" offenses. The data also include "crimes against society," including driving while intoxicated (DWI), writing bad checks, and trespassing. For all states, we drop non-DWI traffic offenses. We also exclude charges in which the final listed disposition is an intermediate outcome, such as a transfer between district and circuit courts or across jurisdictions. Lastly, we exclude technical probation and parole violations that do not result in new criminal charges. While

[6] We have also analyzed data from Arkansas and Maryland. However, we omit data from these states due to data quality issues. Including data from these states does not substantively affect any of the reported results.

[7] In these states, if an officer serves an arrest warrant or makes a warrantless arrest based on probable cause, the officer takes the arrested person before a magistrate. For a warrantless arrest, the magistrate determines whether there is probable cause for arrest. Once probable cause is determined, the magistrate sets conditions of release and issues the arrested person a court date for a first appearance before a court judge. At this stage, a court record is generated. Based on conversations with numerous court and law enforcement officials in each of these states, our understanding is that this process generally occurs without the involvement of prosecutors, except for some exceptionally high-profile cases. Note that, in some other states, after probable cause is determined prosecutors decide whether to file court charges.

Table 1—Data by State

State	Source	Year
Alabama	Alabama Administrative Office of the Courts	2000–2010
North Carolina	North Carolina Administrative Office of the Courts	2007–2014
Texas	Texas Department of Public Safety	2000–2010
Virginia	Virginia Office of the Executive Secretary	2006–2014

Note: Data sources are discussed in more detail in online Appendix A.

we include all remaining charges in the baseline analysis, we also explore limiting charges to core offenses as a robustness check.

We drop charges for defendants aged below 16, which are likely to be adjudicated within the juvenile justice system. We also exclude offenses with fewer than 100 occurrences in the data. These offenses are rare—this restriction removes many specific offense codes from the data, but only around 1 percent of charges. Lastly, we drop offenses that by statute cannot lead to an incarceration sentence and offenses with zero instances that result in confinement. This leaves us with about 400–600 unique offenses in each state.[8]

In Alabama and Virginia, we restrict to Black and white defendants. In Alabama, American Indian-, Asian-, and Hispanic-coded defendants account for less than 0.25 percent of charges. In Virginia, the same categories amount for about 2 percent of charges. In North Carolina and Texas, we restrict to Black, white, and Hispanic defendants. American Indian- and Asian-coded defendants account for less than 2 percent and 1 percent of charges in these states, respectively. In all states, we drop defendants with missing race codes. These account for about 1 percent or less of charges in all states. See online Appendix A for more details.

We use the county as our measure of jurisdiction for all states. In all but North Carolina, the most granular partition among prosecutor and judge electoral districts is the county. In any case, results are similar if we alternatively group counties into prosecutor or judge electoral districts.

A. *Confinement Sentence as a Benchmark Outcome Measure*

There are several potential outcomes to use for measuring punishment severity. A criminal charge can be pursued or dropped by the prosecution. Pursued charges can result in conviction, acquittal, deferred judgment, or some other outcome. Conviction can lead to probation or confinement sentences of varying lengths, or an alternative sentence.

For our measure of severity, we examine whether a given charge results in a jail or prison confinement sentence. This excludes alternative sentences, such as probation or suspended sentences, where the defendant may serve time in jail or prison if they violate the terms of their alternative sentence. We study confinement as our outcome given our particular interest in US exceptionalism in incarceration policy. We focus

[8] When we analyze court outcomes at the case level rather than the charge level as described below, we include excluded offenses when constructing controls if they are not the primary charge in the case.

on the extensive margin of confinement rather than sentence length in part because our data generally do not include information on the mapping between nominal sentence length and realized sentence, which may vary across jurisdictions in ways we cannot measure.[9]

As a robustness check, we also examine two alternative outcomes: conviction and (nominal) sentence length. As we show below, results are qualitatively similar for all three outcomes.

B. *Descriptive Statistics*

We tabulate descriptive statistics for charge data from each state in Table 2. We include information on defendant demographics, charge characteristics, and charge outcomes. The number of charges in our data ranges from 1.9 million in Alabama to 5.9 million in Texas. The number of charges per defendant ranges from 2.3 in Texas to 3.1 in North Carolina. Across states, 71.1 percent to 78.7 percent of charges are filed against male defendants. Defendants are disproportionately Black; while the Black share of the population ranges from 11.8 percent in Texas to 26.1 percent in Alabama, the Black share of defendants ranges from 24.4 percent in Texas to 43.1 percent in Virginia. In both Texas and North Carolina, the Hispanic share of defendants is *lower* than the Hispanic share of the population. However, there is evidence that law enforcement may underreport Hispanic status (Collister and Ellis 2015). Twenty-eight percent to 40 percent of charges are felonies. The distribution of offense types varies across states, though in each state a plurality of charges is for "Other" offenses.

Note that charge outcomes vary significantly across states. In Texas, 40.2 percent of charges result in confinement, and 28.1 percent result in a jail or prison sentence of at least 90 days; in North Carolina, those shares are 8.4 percent and 3.6 percent. This is due in part to variation in severity across states, but may also be due to differences in charging behavior across states. Across states, the same crime may result in a different set of arrests, which may in turn result in a different set of recorded charges.[10] Throughout the analysis, we focus on comparing jurisdictions within states.

We compare jail and prison admissions across jurisdictions within states in Table 3. We use three measures: jail and prison admissions per 100,000 residents (age 15 or above), jail and prison admissions per case, and the share of charges that lead to a jail or prison sentence. Throughout, we refer to the last measure as the *confinement rate*. While the first measure incorporates variation in number of cases and charges per capita across jurisdictions, the second and third measures come closer to capturing how a given case or charge is treated differently across jurisdictions.

There is substantial variation in all three measures. For admissions per 100,000 residents, the (unweighted) coefficient of variation varies from 38 percent in North Carolina to 72 percent in Texas. For admissions per case, the coefficient of variation

[9] The same nominal sentence in two counties may lead to different realized sentences, for example, due to parole board decisions. Notably, parole board members are not locally elected.

[10] Charging behavior may vary across jurisdictions within states, an issue we explore in more detail below.

Table 2—Charge-Level Descriptive Statistics

	Alabama	North Carolina	Texas	Virginia
Male	71.1	75.2	78.7	72.5
Black	37.2	42.3	24.4	43.1
Hispanic		4.2	30.5	
Age	32.9	31.6	31.0	
	(10.9)	(11.9)	(10.8)	
Felony	35.5	27.6	31.2	39.8
Property	17.0	31.0	22.1	33.6
Violent	10.0	13.6	12.2	11.1
Drug	17.3	19.5	21.8	15.0
Other	55.6	35.9	43.9	40.4
Dropped	40.4	61.1	22.3	43.4
Convicted	57.5	36.7	55.4	51.7
Probation	27.8	15.6	30.9	11.4
Confinement	21.2	8.4	40.2	18.9
Sentence $\geq$ 90 days	16.1	3.6	28.1	9.5
Number of defendants	727,419	1,840,251	2,588,641	1,108,911
Number of charges	1,854,208	5,742,283	5,876,448	2,613,297
Number of cases	1,221,317	3,984,894	4,931,314	1,777,549
Charges per defendant	2.5	3.1	2.3	2.4
	(4.3)	(5.4)	(2.4)	(4.1)
Cases per defendant	1.7	2.2	1.9	1.6
	(2.0)	(2.7)	(1.7)	(1.7)
Charges per case	1.5	1.4	1.2	1.5
	(1.8)	(1.6)	(0.6)	(2.0)

Notes: Missing values reflect characteristics that are unavailable for particular states. "Other" offenses include crimes against society and offenses we are unable to classify due to miscoding.

varies from 26 percent in Virginia to 52 percent in Alabama. For confinement sentence per charge, the coefficient of variation varies from 29 percent in Virginia to 59 percent in Alabama.[11]

II. Estimating Punishment Severity

We posit that jurisdictions vary in their *punishment severity*—they vary systematically in how they punish equivalent charges, so that there is a causal effect of jurisdiction on charge outcomes. A key objective of this paper is to measure and compare punishment severity across jurisdictions. Punishment severity reflects variation across jurisdictions in prosecutor and judge behavior, defense attorney quality, and jury preferences. The local electorate plays an important role by electing prosecutors and judges, serving as jurors, and by indirectly determining the level of funding for indigent defense.

To form our benchmark estimates of punishment severity, we estimate linear regression models where the dependent variable is the outcome of a charge and the explanatory variables are rich observable charge and defendant characteristics and

[11] Variation in admissions per case and confinement sentence per charge is not due to chance; if we randomly allocate cases to jurisdictions, maintaining the number of cases per jurisdiction, the coefficient of variation ranges from 2 percent to 5 percent.

Table 3—Jail and Prison Admissions across Jurisdictions

	Alabama	North Carolina	Texas	Virginia
Admissions per 100,000:				
Mean (weighted)	605	600	787	729
Mean	627	569	509	770
Standard deviation	(420)	(219)	(366)	(485)
Admissions per case:				
Mean (weighted)	0.251	0.116	0.406	0.236
Mean	0.223	0.107	0.231	0.234
Standard deviation	(0.117)	(0.034)	(0.120)	(0.062)
Confinement sentence per charge:				
Mean (weighted)	0.213	0.085	0.408	0.193
Mean	0.189	0.077	0.235	0.191
Standard deviation	(0.111)	(0.024)	(0.121)	(0.056)
Number of jurisdictions	67	100	253	118

Notes: "Admissions per 100,000" is the total number of cases resulting in a jail or prison sentence in a county and year divided by county population that is age 15 or above in that year, averaged across years, and multiplied by 100,000. "Admissions per case" is the rate that *cases* result in a jail or prison sentence. "Confinement sentence per charge" is the rate that *charges* result in a jail or prison sentence. Weighted means are weighted by jurisdiction population in 2000.

jurisdiction fixed effects. Specifically, we estimate models of the following form, *separately by state*:

$$y_{ict} = \tau_{cth(i,t)} + x_i\gamma^x + z_{it}\gamma^z + \theta_{j(i,c,t)} + \epsilon_{ict}, \tag{1}$$

where i indexes individuals, c indexes initial charge, t indexes year, $h(i,t)$ is the criminal history for individual i at time t, and $j(i,c,t)$ is the court jurisdiction. y_{ict} is an indicator for any confinement sentence, our primary charge outcome of interest; $\tau_{cth(i,t)}$ are specific arrest offense code by defendant criminal history by year fixed effects; x_i is a vector of time invariant individual controls (defendant race and gender); z_{it} is a vector of time-varying individual controls (age). Finally, $\theta_{j(i,c,t)}$ is a jurisdiction fixed effect, which we use to construct our punishment severity measure.

Our objective is to measure the causal effect of each jurisdiction on charge outcomes. Equation (1) includes rich controls; there are 400–600 unique arrest offense codes per state and several criminal history categories, which we describe below. For equation (1) to recover the causal effects of interest, it must satisfy a *selection on observables* assumption: conditional on $\tau_{cth(i,t)}$, x_i, and z_{it}, unobserved determinants of charge outcomes must be uncorrelated with jurisdiction. It is plausible that this assumption is satisfied given the extensive set of included covariates. Nonetheless, there may remain unobserved determinants of charge outcomes that we cannot measure, e.g., the quality of the evidence possessed by the prosecutor. Further note that we model punishment severity as additively separable from other charge characteristics. That is, we assume that jurisdictions that are punitive for one type of charge (e.g., a violent crime) are also punitive for other types of charges (e.g., a property crime).

We assess these assumptions below. In Section IIB, we use the subset of defendants that are arrested in multiple jurisdictions to validate our baseline punishment severity estimates and address concerns that defendants sort across jurisdictions on (time-invariant) unobservables. We also present evidence that defendants do not sort on time-varying unobservables by examining pre-trends in punishment prior to their changes in arrest jurisdiction. In Section IIA, we show that punishment severity estimates derived from various subsets of charges are highly correlated, supporting the assumption that punishment severity can be modelled additively.

Returning to our benchmark model, to construct criminal history $h(i,t)$ we rely on state-specific sentencing legislation that defines mandatory or suggested sentencing enhancements based on the severity of current charges in combination with the number and severity of prior convictions. The number of resultant categories ranges from 2 (for misdemeanor defendants in Texas) to 20 (for defendants charged with larceny in Virginia). A more detailed description of state-specific criminal history classification is provided in online Appendix B. Though the criminal history classifications are in some instances quite coarse (particularly for misdemeanor defendants in Texas and North Carolina), we have verified that results are robust to defining criminal history based on federal statute. This alternative approach, detailed in online Appendix B, generates a more continuous measure that incorporates number of past convictions, number of past incarceration sentences, and length of past incarceration sentences and allows for a consistent criminal history classification across states.

To construct punishment severity from the θ_j estimates, we add a state-specific constant so that the result is the predicted confinement rate for each jurisdiction using the overall composition of charges in that state.[12] This procedure ensures that log transformations of punishment severity are well-defined, which we use when making cross-state comparisons of punishment severity in Section III.

The coefficient estimates for equation (1) are presented in panel A of Table 4. The pattern of coefficients is consistent with past research (for example, Rehavi and Starr 2014). Conditional on offense charge, criminal history, year, and jurisdiction, Black and male defendants are more likely to receive confinement sentences. Where the data are available, Hispanic defendants are also more likely to receive confinement sentences. The relationship between punishment and defendant age is nonmonotonic, increasing in age at younger ages and decreasing at older ages.

Punishment severity estimates are summarized in panel B of Table 4 and displayed on state maps in Figure 1. Notably, controlling for observable offense and defendant characteristics does not substantially mute cross-jurisdiction variation in confinement rates.[13] Panel B of Table 4 also includes the average punishment severity for jurisdictions in the top and bottom quartiles of jurisdictions, ranked

[12] That is, we average predicted values for each charge in that state derived from equation (1) but omitting the jurisdiction effect *corresponding to the location of the charge*, and then add the estimated jurisdiction effect, θ_j, to construct the punishment severity for jurisdiction j.

[13] Variation in estimated punishment severity is not due to chance; if we randomly allocate cases to jurisdictions, maintaining the number of cases per jurisdiction, the standard deviation of pseudo punishment severity ranges from 0.2 percent in North Carolina to 1.7 percent in Texas. Outside of Texas, there are more than 600 charges in each jurisdiction. In Texas, there are 25 counties with fewer than 600 charges, and 10 with fewer than 100. Excluding these counties from the analysis has no meaningful effect on any of the results presented in this paper.

Table 4—Punishment Severity Model Estimates

	Alabama	North Carolina	Texas	Virginia
Panel A. Coefficient estimates from punishment severity models				
Outcome: Confinement				
Black	0.033	0.020	0.072	0.027
	(0.001)	(0.000)	(0.000)	(0.001)
Hispanic		0.033	0.056	
		(0.001)	(0.000)	
Male	0.045	0.027	0.100	0.040
	(0.001)	(0.000)	(0.000)	(0.001)
Age	0.001	0.005	0.013	
	(0.000)	(0.000)	(0.000)	
$\text{Age}^2 \times 100$	−0.002	−0.005	−0.015	
	(0.000)	(0.000)	(0.000)	
Criminal history × charge × year fixed effects	✓	✓	✓	✓
Jurisdiction fixed effects	✓	✓	✓	✓
Number of charges	1,854,208	5,742,283	5,876,448	2,613,297
Adjusted R^2	0.153	0.093	0.187	0.163
Mean confinement	0.212	0.084	0.402	0.189
Panel B. Summary of baseline punishment severity estimates				
Average confinement rate (percent)	18.9	7.7	23.6	19.1
Standard deviation of punishment severity	10.7	2.0	11.2	4.8
Number of jurisdictions	67	100	253	118
Adjusted Q1 rate	10.3	6.1	13.2	13.3
Adjusted Q4 rate	37.0	10.8	41.3	25.6

Notes: Panel A presents coefficients from state-specific estimates of equation (1). The outcome is an indicator for any confinement sentence. Missing values reflect characteristics that are unavailable for particular states. Standard errors are clustered by defendant in parentheses. Panel B punishment severity estimates are derived by estimating equation (1) separately by state and then adding a state-specific constant as described in Section II. As above, the outcome is an indicator for any confinement sentence. Further details on the estimation of punishment severity are discussed in Section II.

by punishment severity. The differences in punishment severity between quartiles is substantial. Across states, defendants are 1.8 to 3.6 times more likely to face a confinement sentence in fourth-quartile jurisdictions than in first-quartile jurisdictions.

A. *Robustness Checks and Extensions*

In this section, we assess the robustness of our benchmark punishment severity estimates in several ways. First, we scope the potential for "match effects"—interactions between charge characteristics and punishment severity. Second, we assess whether variation in the mapping of crimes to arrests can account for the variation in punishment severity we observe. Third, we analyze arrests at the case level rather than the charge level. Fourth, we estimate punishment severity using alternative charge outcomes: conviction and sentence length. In Section IIB, we use the subset of defendants that are arrested in multiple jurisdictions to further probe the robustness of our estimates.

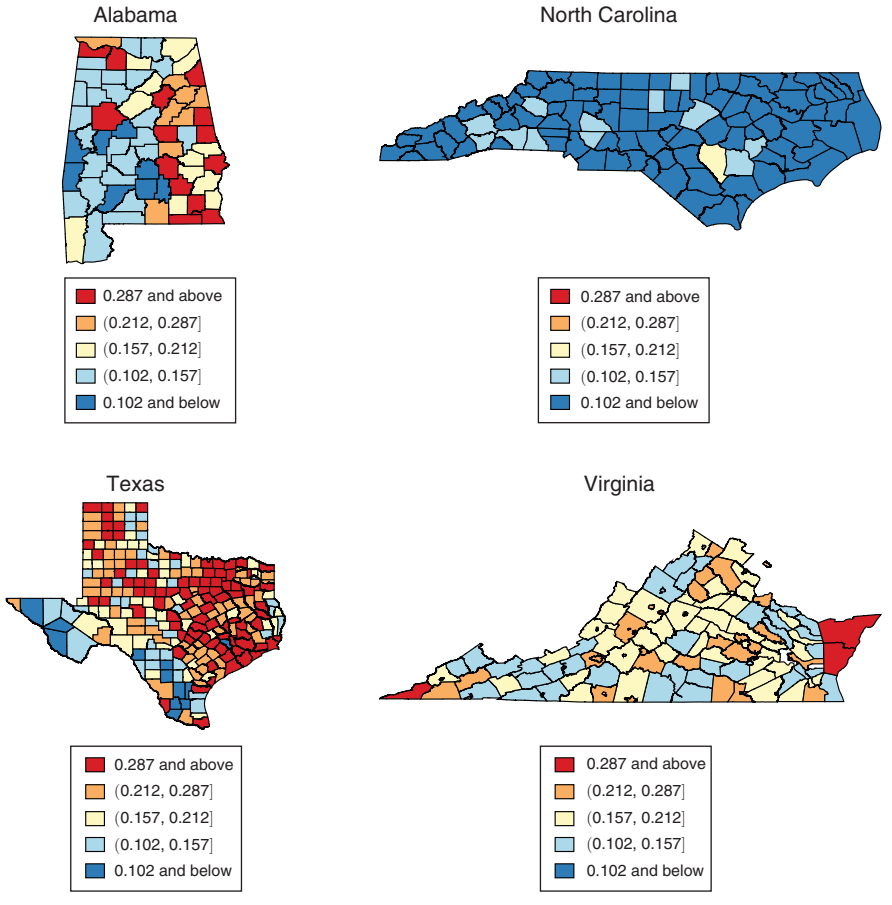

FIGURE 1. MAPS OF PUNISHMENT SEVERITY

Notes: These maps depict punishment severity estimates by county. Punishment severity estimates are derived by estimating equation (1) separately by state and then adding a state-specific constant as described in Section II. The outcome is an indicator for any confinement sentence. Further details on the estimation of punishment severity are discussed in Section II.

Match Effects.—Our estimating equation (1) models punishment severity as separable from other charge characteristics such as crime type or defendant race. This may obscure heterogeneity in punishment severity across types of charges or defendants. For example, a jurisdiction that we characterize as moderately punitive may be lenient with property crimes but harsh with violent crimes. In online Appendix C (panel A of Table C1), we gauge whether such match effects are empirically important. We reestimate punishment severity separately for different types of charges: by defendant race (Black versus white), by criminal history (first-time versus repeat offenders), and by crime category. We estimate punishment severity separately for property, violent, and drug charges, and for those three core categories pooled together. In summary, we find that jurisdictions that are punitive for one type of defendant or charge are also punitive for other types. Moreover, we will show

that the patterns in punishment severity that we document below are quantitatively similar for each subcategory of charges.

Selection into Arrest and Arrest Charge.—Another measurement concern that could bias cross-jurisdiction comparisons is that the threshold that determines (i) whether an arrest is made and (ii) which specific charge is filed may vary across jurisdictions. For example, some police departments may be more lenient than others in deciding whether to arrest a suspect. In that case, jurisdictions with fewer marginal arrests may appear more severe in part because the composition of offenses that actually lead to an arrest may be (unobservably) more serious. Among arrests, some police departments may pursue more severe charges, conditional on the underlying criminal conduct. Because we control flexibly for the initial court charge as our measure of underlying conduct, jurisdictions with more (unobserved) charge upgrading by police officers may consequently appear less punitive in part because the composition of offenses that actually lead to a given initial charge may be (unobservably) less serious. In online Appendix C, we address both selection into arrest and selection into specific arrest charge.

To evaluate selection into arrest, we investigate how a proxy for selection into the court data correlates with estimated punishment severity. To proxy for selection, we calculate the ratio of charges in the court data for a given county and year to crimes reported in the FBI Uniform Crime Reports (UCR) for the same county and year, and then average that ratio across years by county.[14] Jurisdictions that we measure as more punitive have somewhat fewer recorded charges relative to the number of reported crimes. However, conditional on the characteristics we consider in Section IIIB—population density, in particular—we find no relationship between punishment severity and the charge to crime ratio.

To evaluate selection into specific arrest charge, we replace the granular arrest charges used to control for underlying conduct in our baseline regression models with a *coarse* measure of initial court charges. The motivation for using a coarse charge type is that, conditional on underlying criminal conduct that leads a charge to be filed, police and prosecutors have little discretion over whether the charges filed are categorized as violent, property, drug, or other. While we have over 400 types of court charges across our states, for our coarsened measure, we group offenses into those four categories. The correlation between the baseline punishment severity estimates and punishment severity estimates derived using these coarsened arrest charges ranges from 0.98 to 0.99. Thus, while the mapping of underlying conduct to specific arrest charge may vary across jurisdictions, this distinction is unlikely to bias our punishment severity estimates.

Charges versus Cases.—Although we conduct our baseline analysis at the charge level rather than the case level for simplicity, this may introduce bias if co-charges contribute to charge outcomes and charge composition within cases varies by jurisdictions.

[14] Summary statistics for the charge-to-crime ratio are reported in online Appendix Table A1.

In case-level specifications, we redefine y_{ict} as an indicator for whether a case results in any confinement sentence. Rather than control for arrest charge interacted with criminal history and arrest year $(\tau_{cth(i,t)})$, we control for both the most severe arrest charge in the case and the number of additional misdemeanor and felony charges in the case, interacted with criminal history and arrest year. We also look at cases that consist of only a *single charge*, where there is no distinction between charge and case.

The coefficient estimates for case-level and single charge versions of equation (1) are presented in online Appendix Table A2 and online Appendix Table A3. We also correlate our baseline punishment severity estimates with case-level and single charge analogs in online Appendix Table A4. Estimates are very similar across approaches, with correlations ranging from 0.89 to 0.99.

Using Alternative Charge Outcomes.—We next examine alternative measures of punishment severity based on two different charge outcomes: whether the charge results in a conviction, and the sentence length associated with the charge. We again estimate equation (1) separately by state, but replace the outcome variable. Given the skewed distribution of sentence length and the frequency of zero values, we use two transformations of sentence length as outcomes: an indicator for a sentence of at least 90 days, and an inverse hyperbolic sine (asinh) transformation of sentence length.[15] We measure sentence length in days. For charges that do not result in a jail or prison sentence, we record the sentence length as zero.

A key advantage of our data is that they include charges that are dropped or result in no incarceration sentence. By comparison, many studies use data that only include convictions or charges that lead to incarceration sentences. These more limited data can lead to misleading conclusions about the relative punishment severity of jurisdictions if the conviction or incarceration margin is an important source of variation across jurisdictions. For the sake of comparison, we also include a punishment severity measure derived using the inverse hyperbolic sine of sentence length, but limited to charges that result in *any incarceration sentence*.

Correlations between severity measures are presented separately by state in panel B of online Appendix Table C1. The correlation between confinement- and conviction-based severity measures ranges from 0.51 to 0.64 across states. Outside of Texas, the confinement-based severity measure and measures derived from sentence length are highly correlated: for the measure based on sentences that are at least 90 days, correlations range from 0.72 to 0.83; for the measure based on transformed sentence length, they range from 0.95 to 0.97. In Texas, the correlations are more modest: 0.38 for the measure based on sentences that are at least 90 days, and 0.60 for the measure based on transformed sentence length.

In general, there is a strong correlation between these measures, where the correlation is stronger between the baseline confinement-based measure and sentence length-based measures. Moreover, when we examine jurisdiction characteristics

[15] The asinh function closely parallels the natural logarithm function, but is well defined at zero (Card and DellaVigna 2020).

that correlate with severity in Section III, the patterns we identify are qualitatively similar across severity measures.

By contrast, our conditional sentence length measure is weakly and *negatively* correlated with the baseline confinement-based measure. Without data on charges that do not lead to an incarceration sentence, we would substantively mischaracterize punishment severity by jurisdiction. This illustrates the importance of using data that includes charges that are dropped or result in no incarceration sentence.

B. *Validating Estimates Using Multijurisdiction Defendants*

In the analysis above we control for rich offense and charge characteristics that should account for a substantial portion of factors other than jurisdiction-specific punishment severity that determine charge outcomes. However, it is possible that there are critical unobservable determinants that vary across jurisdictions. For example, we do not have direct measures of defendant socioeconomic status, which may affect outcomes directly or through defense attorney quality. We may also miss unobservable severity of the offense or other characteristics of the defendant (e.g., perceived crime risk) that may have important implications for charge outcomes. If these unobservables vary across jurisdictions, they will bias our estimates of punishment severity.

We test for whether unobservables bias our punishment severity estimates by exploiting the fact that many defendants are arrested multiple times and in *multiple jurisdictions.* We use the change in the jurisdiction in which a defendant is arrested as a quasi-experiment for validating our benchmark punishment severity estimates. Within-defendant comparisons net out time invariant defendant characteristics that contribute to charge outcomes, and we can assess the importance of time-varying unobservable factors by exploiting the timing of the defendant's "move" from one jurisdiction to another.[16] If our benchmark punishment severity estimates are unbiased measures of the causal effect of jurisdiction, then those estimates should provide unbiased forecasts for *changes* in confinement rates for a given defendant that is arrested in multiple jurisdictions. Our approach is inspired by methods developed in the teacher value-added (Chetty, Friedman, and Rockoff 2014), worker-firm wage decomposition (Abowd, Kramarz, and Margolis 1999; Card, Heining, and Kline 2013; Card, Cardoso, and Kline 2016), and health care spending (Finkelstein, Gentzkow, and Williams 2016) literatures. Our approach is most similar to Chetty, Friedman, and Rockoff (2014), who validate benchmark measures of teacher value-added using teachers moving from one school to another as quasi-experiments.

In online Appendix Table A5, we compare charge and individual characteristics for multijurisdiction (MJ) defendants, those who have been arrested in multiple jurisdictions, and single-jurisdiction (SJ) defendants, those who have only been arrested in one jurisdiction. Among SJ defendants, we also look separately at defendants who have faced multiple cases. Twenty-six percent to 40 percent of defendants

[16] When we refer to defendants "moving" from one jurisdiction to another, we are referring to changes in the jurisdiction where they are *arrested*, not necessarily changes in *residence*.

have multiple cases in our data, accounting for 58 percent to 75 percent of charges. Among defendants with multiple cases, 26 percent to 41 percent are arrested in multiple jurisdictions, accounting for 19 percent to 33 percent of all charges. MJ defendants are more likely to face confinement sentences than all SJ defendants, and more likely to face confinement sentences than SJ defendants with multiple cases in all states but Texas. They are less likely to be Black than all SJ defendants and SJ defendants with multiple cases.

For MJ defendants and SJ defendants with multiple cases, we also compare pre- and post-move case pairs for MJ defendants and sequential pairs of cases for SJ defendants in online Appendix Table A6, focusing on the main charge. For 37.5 percent to 50.4 percent of MJ defendant case pairs, the main charge is of the same crime type in each case. This range is 40.9 percent to 69.3 percent for SJ defendant pairs. For MJ defendants, 53.5 percent to 68.7 percent of post-move cases are in counties adjacent to the pre-move case.

To implement our test, we use a split-sample procedure. We first randomly partition defendants in each state into 10 equal-sized subsets. For each subset, we estimate equation (1) using the other 9 subsets. To avoid overfitting, we use these (subset-specific) estimates to forecast confinement outcomes for MJ defendants in the selected subset. For these MJ defendants we compare the actual change in the confinement rate before and after the change in jurisdiction of arrest to the forecasted change, adjusting for offense and criminal history. That is, for a defendant who faces one charge in county A and one charge in county B, we compare the forecasted difference in outcomes between the two charges to the actual difference in outcomes. For a regression of the actual difference in outcomes on the predicted difference, a slope coefficient of one would indicate that the punishment severity estimates are unbiased. For more details on estimation and testing, see Appendix D.

In panel A of Figure 2, we plot these actual changes against forecasted changes separately by state, pooling by origin and destination punishment severity quartile.[17] The data points fall roughly on the 45° line. We estimate a slope of 1.00 and intercept of 0.00. We cannot formally reject the null hypothesis that punishment severity estimates provide unbiased forecasts. We also cannot reject *symmetry* for moves to more punitive and less punitive jurisdictions.[18] However, the data points deviate sufficiently from the 45° line that we reject the null hypothesis that our punishment severity estimates have the same predictive validity for every group of moves (Angrist et al. 2017). Yet the deviations are small and, reassuringly, the results we present below are unchanged if we use punishment severity estimates derived from a variant of equation (1) that includes defendant fixed effects.[19]

These findings have two important implications. First, we can forecast within-defendant changes in confinement remarkably well using data on all defendants. This indicates that punishment severity estimates for all defendants are similar to punishment severity estimates for MJ defendants. Second, these forecasts

[17] This follows an analogous specification check developed in Card, Cardoso, and Kline (2016).

[18] In particular, if we fit a two-piece linear spline with the knot set at zero, we cannot reject that the two slopes are equal.

[19] Estimation of this variant is discussed in online Appendix D. Robustness of results on the relationship between punishment severity and racial heterogeneity is discussed in Section IIIB.

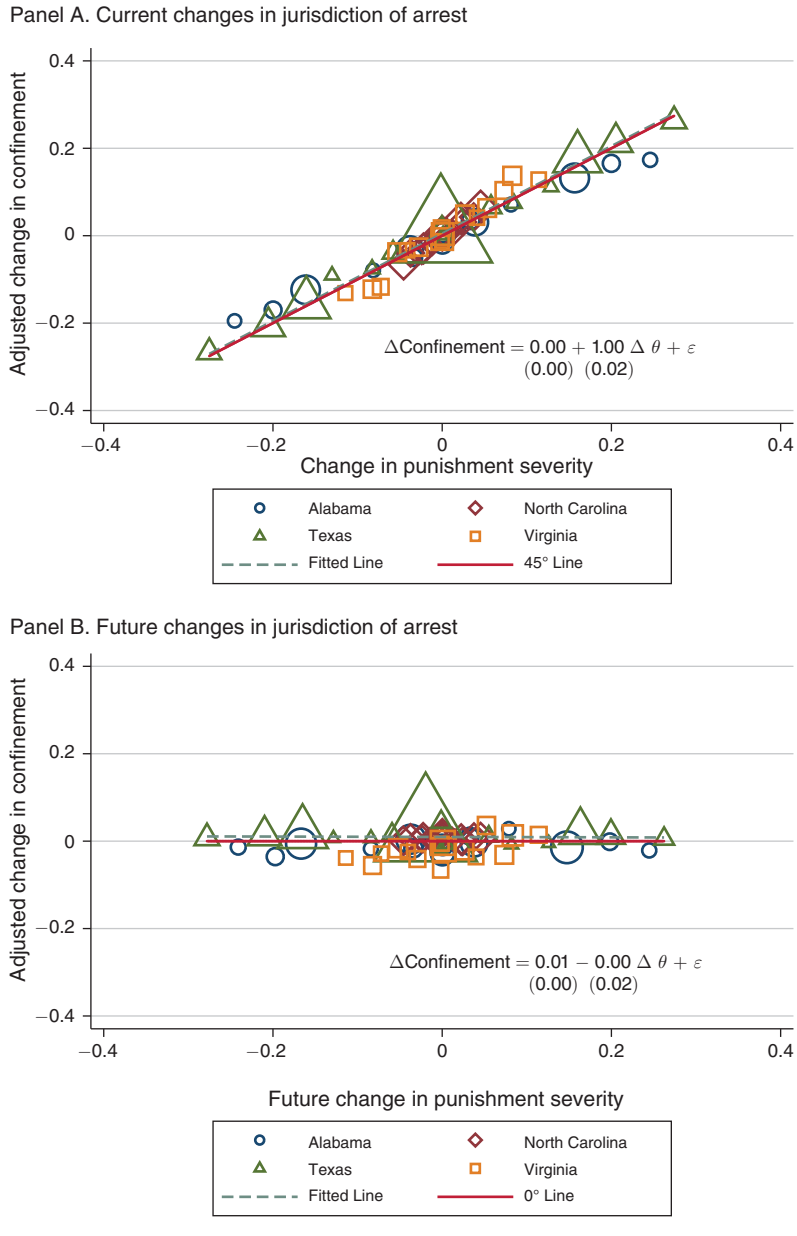

FIGURE 2. FORECASTED VERSUS ACTUAL CHANGE IN OUTCOMES FOR MULTIJURISDICTION DEFENDANTS

Notes: In panel A, we plot adjusted realized changes in confinement rates before and after the change in jurisdiction against forecasted changes by state, adjusting for offense and criminal history, and pooling by origin and destination punishment severity quartile. Marker size is proportional to the number of charges represented in the origin quartile by destination quartile by state cell. The solid line is the 45-degree line, while the dashed line is the linear best fit, weighted by cell size. In panel B, we plot within-jurisdiction changes in confinement rates against predicted changes based on future changes in jurisdiction. The solid line is a horizontal line overlapping with the horizontal axis, while the dashed line is the linear best fit, weighted by cell size.

are reasonably accurate for a variety of defendants as defined by their origin and destination jurisdictions.

The identifying assumption that underpins this validation strategy is that MJ defendants do not sort across jurisdictions in a manner that relates to: (i) time-varying unobservable defendant-level or jurisdiction-level determinants of charge outcomes; or (ii) match effects—interactions between punishment severity and defendant characteristics.

For example, if defendants that move to a particular jurisdiction are also committing increasingly (and unobservably) more severe crimes, then we would mistakenly identify the jurisdiction as punitive. If a jurisdiction is particularly lenient for drug cases but not other cases, and defendants are more likely to commit drug crimes in that jurisdiction, then we would mistakenly identify this jurisdiction as lenient, when in fact it is only lenient for a particular type of case.[20] We assess these two assumptions in the next section.

Do Defendants Sort on Time-Varying Unobservables or Match Effects?—First, we test whether defendants sort on time-varying unobservables using a placebo test adapted from Card, Heining, and Kline (2013). In particular, we test for pre-trends in MJ defendant confinement rates prior to the defendant's change in jurisdiction. To do this, we focus on confinement rates for defendants that are charged in multiple cases in one jurisdiction and subsequently in at least one case in a different jurisdiction.[21] As an illustrative example, consider a defendant that faces criminal cases 1 and 2 in county A, and criminal case 3 in county B. If defendants are sorting on time-varying unobservables, we may see pre-trends in punishment *prior* to the defendant's change in jurisdiction, conditional on observable case and defendant characteristics. To test for such pre-trends, we can thus check whether the identity of county B predicts the difference in outcomes between cases 1 and 2.[22] If sorting on time-varying unobservables is not a factor, then future changes in jurisdiction should not predict changes in outcomes between cases 1 and 2.

In panel B of Figure 2, we plot within-jurisdiction changes in confinement rates against forecasted changes based on future changes in jurisdiction of arrest. The points roughly fall on the horizontal line at zero, and we cannot formally reject the null hypothesis that the slope is zero. This indicates that future changes in jurisdiction of arrest do not predict earlier changes in confinement rates. Note that this test is not definitive, however; it is possible that defendants sort on time-varying unobservables in a way that coincides precisely with changes in the jurisdiction in which they are charged.

Second, in online Appendix D we test whether MJ defendants sort on match effects across jurisdictions. We find that they do not, at least on the basis of jurisdiction by crime type or jurisdiction by criminal history match effects. We also documented in Section IIA that the scope for match effects appears to be limited.

[20] There may also be match effects that are specific to MJ defendants. For example, some jurisdictions may be more punitive with "out of town" defendants than long-term residents. However, if punishment severity estimates forecast MJ defendant confinement rates well, this would imply this type of match effect is not important empirically.

[21] In online Appendix Figure A1, we replicate panel A of Figure 2 for this sample of defendants.

[22] Sorting across jurisdictions based on time-varying unobservables would introduce bias, for example, if defendants that committed increasingly (unobservably) serious crimes were also more likely to relocate to less punitive locations.

Decomposing Punishment Severity.—In this section, we quantify the role of punishment severity in explaining cross-jurisdiction variation in confinement rates by decomposing that variation into several components. Our approach follows Finkelstein, Gentzkow, and Williams (2016), and we provide a detailed description of our methodology in online Appendix D. In sum, we first estimate a variant of equation (1) that includes defendant fixed effects. We summarize punishment severity estimates derived from this approach in panel A of Table 5. Consistent with Figure 2, panel A, these estimates are very similar to the benchmark estimates. The correlation between estimates within states ranges from 0.82 in Alabama to 0.96 in Texas.[23]

In panel B of Table 5 we present an additive decomposition of the difference between the top-quartile and bottom-quartile jurisdictions by confinement rate, separately by state. We find that jurisdiction effects, rather than differences in defendant or charge effects, explain the bulk of the difference, ranging from 80.2 percent in Alabama to 93.1 percent in North Carolina. In online Appendix D, we show that if jurisdiction effects were equalized across jurisdictions, cross-jurisdiction variation in confinement rates would be reduced by 64–93 percent.

III. Racial Divisions and Punishment Severity

We have provided evidence in support of a causal interpretation of our punishment severity estimates and established the robustness of our severity measure across alternative outcomes and approaches. We next identify jurisdiction-level characteristics that predict punishment severity. To guide this analysis, we sketch a simple model of preferences for punishment based on racial ingroup bias to derive a predicted relationship between punishment severity and local racial heterogeneity.

A. *A Simple Model*

For the purposes of our model, we assume that local residents have to choose an optimal level of punishment, but are constrained to choose an overall punishment severity rather than separate punishment severities by race.[24] Given this restriction, we model the utility of individual i as follows:

$$u_i\big(s;p(r_i)\big) \;=\; s \times \Big[\alpha\big(1-p(r_i)\big)+\beta p(r_i)\Big] - c(s),$$

where r_i is the racial group of individual i, $p(r_i)$ is the probability that an offender arrested in individual i's home jurisdiction is a member of individual i's racial group, and $c(s)$ is a strictly increasing and convex function (with $c(0) = 0$)

[23] The variation is slightly larger for estimates using defendant fixed effects, due at least in part to added measurement error.

[24] This assumption is justified empirically by the findings that (i) incarceration policy severity in a given jurisdiction is highly correlated across racial groups and (ii) there is no consistent relationship in our sample between those jurisdiction characteristics that predict overall jurisdiction-level severity and the gap between within-jurisdiction Black and white defendant-specific severity parameters. The latter finding is discussed in more depth below.

TABLE 5—SUMMARY OF PUNISHMENT SEVERITY ESTIMATES: OVERALL VERSUS WITHIN-DEFENDANT

	Alabama	North Carolina	Texas	Virginia
Avgerage confinement rate (percent)	18.9	7.7	23.6	19.1
σ (Overall)	10.7	2.0	11.2	4.8
σ (Defendant fixed effects)	11.1	2.5	11.8	5.9
Correlation	0.82	0.89	0.96	0.91
Decomposition: Q1 versus Q4 by punishment severity				
Difference in confinement rate				
Overall	28.3	5.8	30.3	14.2
Jurisdiction	22.7	5.4	27.9	12.4
Defendants	5.3	0.7	2.8	1.2
Charges	0.3	−0.3	−0.5	0.6
Share (percent) of difference due to				
Jurisdiction	80.2	93.1	92.1	87.3
Defendants	18.7	12.1	9.2	8.5
Charges	1.1	−5.2	−1.7	4.2
Number of jurisdictions	67	100	253	118

Notes: The top panel compares punishment severity estimates derived with and without defendant fixed effects. "Overall" punishment severity estimates are derived by estimating equation (1) separately by state and then adding a state-specific constant as described in Section II. "Defendant fixed effects" punishment severity estimates are derived by estimating a variant of equation (1) that includes defendant fixed effects separately by state and then adding a state-specific constant. This specification is described in more detail in online Appendix D (see equation (D.5)). The bottom panel decomposes differences in confinement rates between the top and bottom quartile jurisdictions by punishment severity ($Q1$ and $Q4$), separately by state. The first row reports the difference in average confinement rates between the two sets of jurisdictions ($\hat{Y}_{Q1} - \hat{Y}_{Q4}$); the second row reports the difference due to jurisdiction ($\hat{\theta}_{Q1} - \hat{\theta}_{Q4}$); the third row reports the difference due to defendants ($\hat{\gamma}_{Q1} - \hat{\gamma}_{Q4}$); the fourth row reports the difference due to charge and defendant criminal history ($\hat{\tau}_{Q1} - \hat{\tau}_{Q4}$). The next three rows report the share of the difference in confinement rates due to jurisdiction, defendants, and charge and criminal history. See online Appendix D for additional details regarding variable definitions and the decomposition methodology.

characterizing the fiscal and nonpecuniary costs associated with higher severity s.[25] In the expression for individual utility, α and β reflect the relative utility gains associated with punishing outgroup members versus punishing ingroup members (i.e., a negative-valued β implies disutility associated with punishing ingroup members). Based on the existing literature related to racial group ingroup bias, we make the assumptions that $\alpha > 0$ and $\alpha > \beta$.[26]

To characterize how predicted punishment preferences vary as a function of local racial composition, first consider a jurisdiction in which a substantial majority of offenders are white (i.e., $p_w \gg 1/2$). In this case, the punishment

[25] For example, increased punishment s may impose an additional nonpecuniary cost to the extent that an increase in the likelihood of type II errors, whereby innocent individuals are incorrectly punished, decreases utility (due either to fairness concerns or an individual's self-interested concern that he/she may be erroneously convicted of a crime).

[26] Luttmer (2001) and Chen and Li (2009) provide observational and experimental support for these assumptions. Anwar, Bayer, and Hjalmarsson (2012) finds that all-white jury pools convict Black defendants significantly more often than white defendants, and this gap in conviction rates is eliminated when the jury pool includes at least one Black member. These findings are consistent with jurors preferring to punish outgroup defendants over ingroup defendants.

severity preferred by white residents, $c'^{-1}\big(\alpha(1-p_w)+\beta p_w\big)$, will be lower than $c'^{-1}\big(\alpha(p_w)+\beta(1-p_w)\big)$, the punishment severity preferred by Black residents. Now, suppose that there is a pivotal (median) voter whose preferences determine the jurisdiction-specific punishment severity. Since racial population shares are highly correlated with the share of defendants of each race, the likelihood that the pivotal voter is white is increasing in the share of defendants that is white, and so white punishment preferences will determine local severity. Next, note that as the Black share of offenders $(1-p_w)$ increases, the punishment severity preferred by white residents will also increase given that $\alpha > \beta$ and that $c'^{-1}(\cdot)$ is a strictly increasing function by construction. Hence, the punishment severity chosen by the median voter is increasing in Black offender share until the median voter switches from a white to Black resident. By the symmetry of the model, the punishment severity preferred by Black residents is falling as the Black share of offenders continues to rise. Consequently, the model predicts that local punishment severity as a function of the Black share of offenders will follow an inverted U-shape.

B. *Testing the Model*

Our model predicts a particular nonmonotonic causal relationship between local racial composition and punishment severity. To test the model, we would ideally identify a source of exogenous variation in racial composition across jurisdictions, and use that variation to test whether the causal relationship between racial composition and punishment severity exhibits the inverse U-shaped pattern the model predicts. Unfortunately, we are unaware of any natural experiment that would provide suitable variation. Instead, we test for an inverted U-shaped pattern in the cross section and adjust for other covariates. An important concern with this approach is omitted variable bias—unobserved differences across jurisdictions may drive any observed relationship between racial composition and punishment severity. Despite this, we believe our "selection on observables" test is compelling, particularly due to the specific inverse U-shaped pattern we are testing for. As we will argue, it is not clear what alternative explanation would be consistent with this pattern.

As an initial test of the prediction derived from the model, panels A and B of Figure 3 plot transformed punishment severity for each county as a function of its racial composition. To measure racial composition, we use both the Black share of the population in 2000 (panel A) and the Black share of defendants in that county (panel B).[27] To make our punishment severity measure comparable across states in this analysis, we transform the measure and express it in terms relative to each jurisdiction's state average. We begin with punishment severity estimates derived from equation (1) using the full data. We then divide this predicted confinement rate by the same severity measure averaged across jurisdictions within the state and take the log of this ratio.[28] The transformed measure is approximately the proportional

[27] In the model, individual preferences depend on the racial composition of offenders, but the identity of the pivotal voter depends on the composition of the electorate. In practice, the Black share of defendants and the Black share of the population are highly correlated.

[28] Since regression models include state fixed effects, this normalization does not alter regression results but facilitates data visualization by demeaning logged values separately by state.

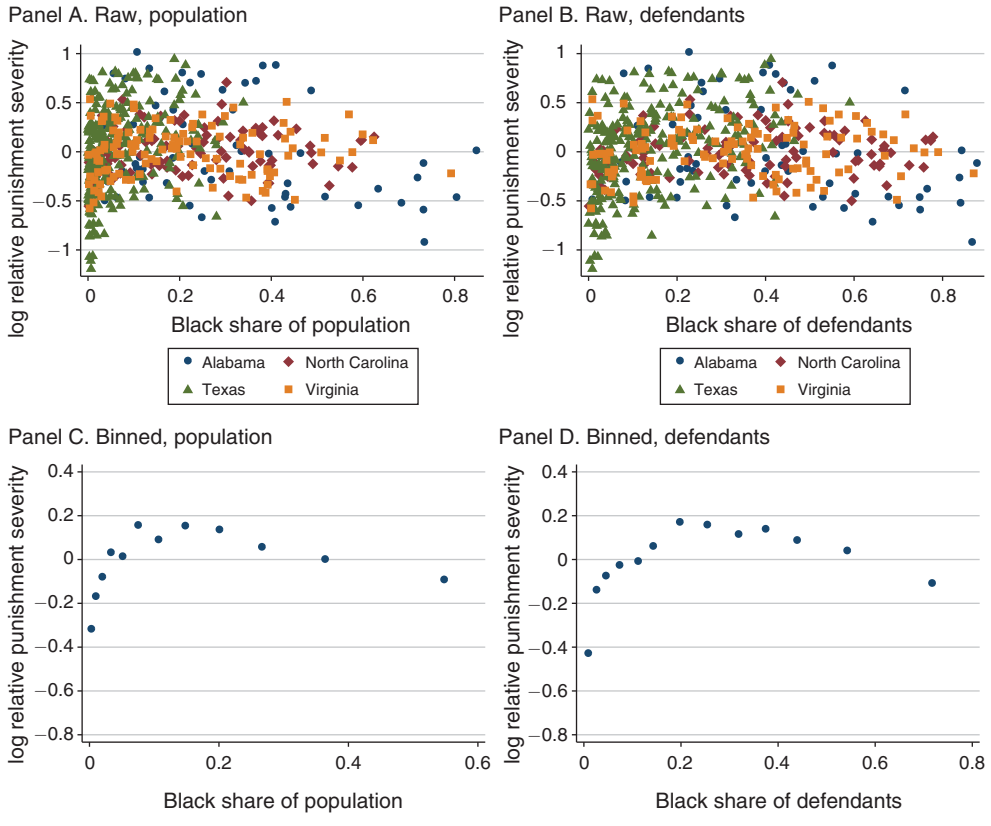

FIGURE 3. PUNISHMENT SEVERITY AND RACIAL HETEROGENEITY

Notes: In this figure, log relative punishment severity is constructed by dividing the predicted confinement rate for each jurisdiction based on the overall composition of charges within the state by the overall state confinement rate and then taking the log of this ratio. In panels C and D, we bin jurisdictions using the data-driven approach developed in Cattaneo et al. (2019).

difference in confinement rates between a jurisdiction and the average jurisdiction in a state, holding other charge characteristics fixed. Given cross-state differences in average predicted confinement rates, we study proportional differences to facilitate cross-state comparisons. Below we denote this transformed punishment severity by $\log \theta_j'$ and refer to this measure as log *relative punishment severity*.

The plot reveals that the inverted U-shaped relationship predicted by our model is indeed borne out in the data. For an initial range of values for the Black share of the population or defendants, punishment severity is increasing in the Black share. After this range, the sign of the relationship flips.[29]

[29] Since punishment severity is estimated with controls for defendant demographics, including race, comparisons across jurisdictions reflect a weighted average of differences in the severity of treatment of Black and white offenders (with weights determined by jurisdiction-specific offender shares). This approach eliminates the mechanical relationship between local severity and local Black defendant share that would otherwise bias cross-jurisdictional comparisons.

To clarify this relationship, we pool jurisdictions into bins using the data-driven approach developed in Cattaneo et al. (2019).[30] For each bin we then plot the average log relative punishment severity measure, $\log \theta_j'$. The results are presented in panels C and D of Figure 3. Note that the span of the vertical axes is substantially narrower in these panels. There is a clear nonmonotonic relationship between the Black share of the population or defendants and punishment severity, where punishment severity is initially increasing in Black share and then the sign of the relationship flips. We use regression models below to demonstrate that this nonmonotonic relationship is robust to the inclusion of additional jurisdiction-level covariates and to measure the implied "peak" value for the Black share. In online Appendix Figure A2 we show that the inverted U-shaped relationship remains visible after conditioning on these covariates.

Note that if population and defendant shares are equal, voting rates are uniform, voters have unidimensional preferences that are homogeneous by race, and all voters are either white or Black, then the model predicts a peak where the Black share of the population is equal to one half. In practice, it is not surprising that we find a peak where the Black share of the population is below 0.5. Existing research documents less punitive preferences among Blacks than whites (Bobo and Johnson 2004). Then, to the extent that there is preference heterogeneity such that some white residents have less punitive preferences and do not exhibit ingroup bias, we should anticipate a peak below 0.5.[31]

We next move to a more thorough analysis of the relationship between local punishment severity and racial composition. Absent any source of plausibly exogenous cross-sectional variation in racial composition, we introduce a series of additional jurisdiction-level covariates into a regression of log adjusted punishment severity on a quadratic in the Black share of the population (or defendants) to assess the extent to which alternative mechanisms may drive the observed relationship. Specifically, we estimate models of the following form:

$$\log \theta_j' = x_j \beta + \tau_s + \epsilon_j, \tag{2}$$

where $\log \theta_j'$ is the log adjusted punishment severity described above, x_j is a vector of jurisdiction characteristics, and τ_s is a set of state fixed effects.

Researchers studying US trends in crime and punishment have highlighted the important role that historical violent crime rates played in driving the increased severity of punishment over recent decades and in generating cross-state variation in punishment severity (see, for instance, Western 2006). To test whether local variation in past crime rates is associated with differences in punishment severity *within* states, we control for measures of growth in violent crime rates between 1970 and

[30] Cattaneo et al. (2019) reframe binscatter as a nonparametric estimator for the conditional expectation function and select the number of bins that minimizes integrated mean square error.

[31] Differences in voting rates by race, in the share of the population categorized as "Other race," and the multidimensionality of policy preferences would also generate uncertainty in the precise level of the Black population share at which punishment severity peaks. Although Republican Party support is undoubtedly an imperfect proxy for punishment preferences and does not capture preference intensity, a back-of-the-envelope calculation relying on race-specific party affiliation reveals that the Black population share at which we would expect to observe the median voter change from Republican to Democrat ranges from 0.23 in North Carolina to 0.40 in Alabama.

1990 and the 2000 violent crime rate, both measured at the jurisdiction level.[32] Each measure is standardized to have a mean of zero and a standard deviation of one. The crime measures are derived from FBI UCR data. In addition to a quadratic in the Black share of the population or defendants, we include log average household income, the Gini index of income inequality, the fraction of prime-aged males in the population, and log population density, all measured in 2000. Descriptive statistics for these county characteristics are reported in online Appendix Table A7. There is one observation per jurisdiction. As noted in online Appendix Table A7, we are missing data on crime and the Gini index for some counties. In the regression models, we set missing values to zero and include indicators for missing data for each of these covariates as additional controls.

Regression estimates are presented in Table 6. In columns 1–5 we use the Black share of the population and its square to measure a jurisdiction's racial composition. In columns 6–8, we use the Black share of defendants and its square. The results are similar for both measures. We discuss the results using the Black share of the population first and then discuss the differences in results between the two measures.

Column 1 presents the regression equivalent of Figure 3, with no controls other than the Black share of the population, its square, and state fixed effects. Point estimates are consistent with an inverted U-shaped relationship between local severity and Black share of the population and imply that punishment severity is highest in jurisdictions with a Black share of the population equal to 0.3. At this maximum, the predicted value of θ is 24 log points larger than the predicted value where Black share is set to zero. This implies that predicted punishment severity is 27 percent higher in jurisdictions with this level of heterogeneity relative to all-white jurisdictions.[33]

Columns 2 and 3 add our set of jurisdiction-level controls: log population density, log average household income, the Gini index of income inequality, and the fraction of prime-aged males in the population. The only difference between the two models is the measure of local crime that we include as a control. Column 2 uses the growth in the violent crime rate from 1970 to 1990, and column 3 uses the violent crime rate in 2000.[34] In both specifications, the inverted U-shaped relationship between punishment severity and Black population share remains highly significant, though is somewhat muted in magnitude. The peak value for the Black share of the population moves up to 0.33 in column 2 and to 0.37 in column 3. At these peak values for columns 2 and 3, the predicted value of θ is 14 and 17 log points larger than the predicted value where Black share is set to zero, respectively. The coefficient on the

[32] We calculate the growth in violent crime as

$$r_{growth} = \frac{r_{1990} - r_{1970}}{0.5\, r_{1990} + 0.5\, r_{1970}},$$

where r_{1990} and r_{1970} are the local violent crime rates in 1990 and 1970.

[33] We employ two alternative approaches to testing for an inverted U-shaped relationship between Black population share and punishment severity. First, we estimate two piece linear splines and test for a positive initial slope and negative final slope. If we set the knot point to 0.3, we estimate an initial slope of 0.928 (standard error 0.184) and final slope of -1.151 (0.224). Second, we test directly for an inverse U-shape using the approach outlined in Lind and Mehlum (2010). We reject the null hypothesis of a monotone or U-shaped relationship against an inverse U-shaped alternative (the p-value on this test is 1.0×10^{-7}).

[34] Results are similar if we use total Part I crime rates rather than restricting to violent crime.

Table 6—Punishment Severity and Racial Heterogeneity

Outcome:	log relative punishment severity							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black share of population	1.644	0.853	0.946	0.859	0.895			
	(0.297)	(0.257)	(0.257)	(0.257)	(0.279)			
Black share of population, squared	−2.764	−1.283	−1.279	−1.014	−1.495			
	(0.502)	(0.405)	(0.396)	(0.383)	(0.451)			
Black share of defendants						2.148	1.177	1.425
						(0.270)	(0.265)	(0.280)
Black share of defendants, squared						−2.707	−1.376	−1.907
						(0.353)	(0.348)	(0.389)
log population density		0.085	0.099	†	x		†	x
		(0.020)	(0.021)					
log average household income		0.395	0.362	0.345	x		0.256	x
		(0.163)	(0.164)	(0.171)			(0.169)	
Gini coefficient		−0.194	−0.070	−0.011	x		0.076	x
		(0.220)	(0.226)	(0.228)			(0.226)	
Fraction males aged 15–29		0.077	0.054	0.233	x		0.135	x
		(0.745)	(0.733)	(0.742)			(0.735)	
Violent crime rate growth, 1970–1990		−0.017		−0.027			−0.024	
		(0.016)		(0.016)			(0.015)	
Violent crime rate, 2000			−0.044	−0.042	x		−0.038	x
			(0.018)	(0.017)			(0.017)	
Black share at "peak" severity	0.30	0.33	0.37	0.42	0.30	0.40	0.43	0.37
	(0.019)	(0.048)	(0.056)	(0.078)	(0.044)	(0.015)	(0.040)	(0.027)
State fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Adjusted R^2	0.049	0.202	0.206	0.222	0.238	0.111	0.239	0.266
Observations	538	538	538	538	538	538	538	538

Notes: In this table, log relative punishment severity is constructed by dividing the predicted confinement rate for each jurisdiction based on the overall composition of charges within the state by the overall state confinement rate and then taking the log of this ratio. For covariates that are missing for some jurisdictions (crime rates and Gini index), we set missing values to zero and include indicators for missing data for each of these covariates as additional controls. In each column, Black share at "peak" severity is estimated from the corresponding quadratic term coefficients on Black share of population/defendants. Corresponding standard errors are constructed using the delta method. "†" denotes inclusion of a five-piece linear spline in log population density. "x" denotes inclusion of the covariate interacted with state fixed effects. Robust standard errors are in parentheses.

growth in violent crime in column 2 is close to zero and statistically insignificant, while the coefficient on violent crime in 2000 in column 3 is *negative* and small in magnitude, though statistically significant. Results from these specifications lend little support to the hypothesis that within-state variation in present-day severity is explained by historical crime waves or current crime patterns. Turning to the remaining covariates, population density also consistently predicts higher confinement rates. A jurisdiction with 10 percent higher population density is predicted to be about 1 percent more punitive.

Given that population density is a strong predictor of severity, one concern is that the relationship we identify between racial composition and punishment severity is driven in part by a nonlinear relationship between population density and severity. Column 4 repeats the specification in column 3 but adds a five-piece linear spline in log population density as controls. Controlling for population density in this more

flexible manner has little effect on the coefficient estimates for the Black share of the population and its square.

In column 5 we allow each of the nonrace covariates to vary by state, interacting each with state indicator variables. The inverted U-shaped relationship between punishment severity and Black population share remains highly significant and unchanged in this specification that controls more flexibly for the full set of nonrace covariates.

The pattern of coefficients is similar in columns 6–8, which are analogous to columns 1, 4, and 5 except that we replace the Black share of the population with the Black share of defendants. There are two noticeable differences. First, the implied peak moves to about 0.4. Second, the difference between predicted θ at "peak" heterogeneous jurisdictions and all-white jurisdictions increases to 43 log points without additional controls and to 24 log points with controls. Both findings are consistent with what we see graphically in Figure 3.

Robustness Checks and Extensions.—In this section we explore the robustness of our results along a number of dimensions.

First, we examine whether the relationship between punishment severity and racial composition that we identify is present for our alternative measures of severity based on conviction rates and sentence length. We estimate equation (2) but replace the outcome used to measure punishment severity. The results are presented in Table 7. In columns 1 and 2, the outcome is conviction. In columns 3 and 4, the outcome is a sentence above 90 days. In columns 5 and 6, the outcome is inverse hyperbolic sine-transformed sentence length. In odd columns, we use the Black share of the population and its square to measure a jurisdiction's racial composition, while in even columns we use the Black share of defendants and its square. Across outcomes, we see a similar inverted U-shaped relationship between punishment severity and Black share that peaks for Black share values in the 0.27 to 0.39 range.

In columns 7 and 8, the outcome is inverse hyperbolic sine-transformed sentence length, but limited to charges with any incarceration sentence. We include this measure to see what we would have concluded if our data excluded dropped charges and those not leading to an incarceration sentence. Strikingly, we see little to no relationship between this measure and the Black share of the population or defendants.

Second, we address the concern raised in Section IIA that the type of offenses that lead to charges may vary across counties. For example, jurisdictions with fewer marginal charges may appear more severe in part because the composition of offenses that actually lead to a charge may be (unobservably) more serious. We estimate versions of equation (2) that include a jurisdiction's charge to crime ratio as an additional control. To match the coverage of the UCR crime data, we also replace the baseline punishment severity measure with a measure derived from only violent and property crimes in some specifications. The results are presented in online Appendix Table A8. We find that, conditional on the jurisdiction covariates we include, the charge to crime ratio is uncorrelated with punishment severity and its inclusion has no effect on the coefficients for Black share.

Third, we address the concern that the inverted U-shaped relationship identified in Table 6 can be explained by endogenous migratory responses to local punishment

Table 7—Punishment Severity and Racial Heterogeneity, Alternative Outcomes

Outcome:	Convictions		Sentence $\geq$ 90 days		asinh(sentence length)		asinh(cond. sentence length)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black share of population	0.247		0.683		0.973		−0.049	
	(0.131)		(0.298)		(0.260)		(0.065)	
Black share of population, squared	−0.455		−1.131		−1.414		0.029	
	(0.183)		(0.440)		(0.382)		(0.096)	
Black share of defendants		0.375		0.832		1.131		−0.133
		(0.124)		(0.296)		(0.249)		(0.064)
Black share of defendants, squared		−0.553		−1.151		−1.436		0.134
		(0.164)		(0.381)		(0.327)		(0.079)
Black share at "peak" severity	0.27	0.34	0.30	0.36	0.34	0.39		
	(0.065)	(0.041)	(0.058)	(0.044)	(0.048)	(0.035)		
State fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Adjusted R^2	0.038	0.048	0.120	0.127	0.143	0.155	0.528	0.531
Observations	537	537	536	536	538	538	538	538

Notes: Each specification includes the following covariates: log population density, log average household income, Gini coefficient, fraction males aged 15–29, and violent crime rate in 2000. For covariates that are missing for some jurisdictions (crime rate and Gini coefficient), we set missing values to zero and include indicators for missing data for each of these covariates as additional controls. In each column, "Black share at 'peak' severity" is estimated from the corresponding quadratic term coefficients on "Black share of population/defendants." Corresponding standard errors are constructed using the delta method. Columns 1 and 2 exclude one jurisdiction (Austin County, Texas) with corresponding punishment severity estimate (in this case, the predicted conviction rate) below zero. Columns 3 and 4 exclude two jurisdictions (Brooks County, Texas, and Duval County, Texas) with corresponding punishment severity estimates (predicted rate of sentences $\geq$ 90 days) below zero. Robust standard errors are in parentheses.

severity or to other correlated community characteristics. In online Appendix Table A9, we replace the Black share of the population measure with the 1860 county-level share of the population that was enslaved. Despite the fact that historical data is only available for two-thirds of the jurisdictions in the sample, we identify a similarly robust inverted U-shaped relationship between 1860 slave share and contemporaneous punishment severity.

Fourth, to assess the validity of the assumption that jurisdiction residents' preferences determine average local punishment severity rather than *race-specific* punishment severity, we reestimate the specifications included in Table 6 in online Appendix Table A10 but use the Black-white difference in log adjusted local severity as our outcome measure. While the coefficients on Black population share and its square are statistically significant in more sparse specifications, these coefficients are no longer statistically significant when we allow for state-specific slopes for nonrace jurisdiction characteristics. When we measure Black share using the composition of defendants, the coefficients on Black share and its square are small in magnitude, statistically insignificant, and of inconsistent sign across specifications. Overall, we do not find robust evidence that race-based gaps follow the same inverted U-shaped pattern as overall punishment severity. Moreover, as described below, when we construct separate punishment severity measures for Black and white defendants, we find an inverted U-shaped pattern for both measures.

Fifth, we examine whether the nonmonotonic relationship we identify between punishment severity and Black share is present for other jurisdiction characteristics. In online Appendix Figure A3, we plot the relationship between each covariate included in Table 6 and punishment severity. To the extent that racial divisions indeed explain the inverted U-shaped relationship between punishment severity and Black population or defendant share, we should not expect to see a similar nonmonotonic relationship between local severity and any of the other included covariates. Reassuringly, there is indeed no evidence of a nonmonotonic relationship between any of the other included covariates and punishment severity.

Sixth, we test whether our results are robust to using punishment severity derived from the variant of equation (1) that includes defendant fixed effects or subgroup-based estimates explored in Section IIA. In online Appendix Table A11 we show the same inverted U-shaped relationship is present for each alternative measure of punishment severity.[35]

Seventh, to assess the degree of potential bias due to unobservables, we use the approach outlined in Oster (2019). We show in online Appendix Table A13 that selection on unobservables would need to be over two times as large as selection on observables to explain the measured relationship between punishment severity and racial composition. These estimates are notably above the upper bound of one suggested in Oster (2019) for calculating bias-adjusted treatment effects. Moreover, if we include population density in our baseline model, the implied degree of selection on unobservables that would be required to explain our estimates increases to between 3.8 and 12.4 times as large as selection on observables.

An alternative explanation for the relationship we identify between racial heterogeneity and punishment severity is that (i) a higher share of defendants in racially heterogeneous communities are paired with judges or prosecutors of another race and (ii) judges or prosecutors treat outgroup members more severely than ingroup members. Given the paucity of Black prosecutors, ingroup bias seems unlikely to explain the pattern we observe. In 2014, only 6.6 percent of chief prosecutors are Black in our sample states, and that drops to 2.5 percent if we exclude Virginia (Reflective Democracy Campaign 2018). While Shayo and Zussman (2011) documents robust evidence of judicial ingroup bias in Israel, findings from the United States are mixed and suggest that ingroup bias among judges may be limited. Cohen and Yang (2019) find that among Republican-appointed federal judges, white judges differentially punish Black defendants more severely. However, the authors do not find differential gaps in punishment among Democratic-appointed judges and note that the vast majority of Black federal judges are Democratic-appointed. Schanzenbach (2015) finds that federal judges do not exhibit ingroup bias, and Arnold, Dobbie, and Yang (2018) find no evidence that racial bias varies with judge race among bail judges in Philadelphia and Miami-Dade counties. While Abrams, Bertrand, and Mullainathan (2012) find that Black judges impose relatively short sentences on Black defendants, they find that the judges are not less likely to impose confinement sentences on Black defendants. Our own finding that the Black-white gap in punishment severity does

[35] We also show in online Appendix Table A12 that the same inverted U-shaped relationship is present when observations are weighted by jurisdiction population.

not vary in a consistent manner with local racial composition also suggests that judicial ingroup bias is unlikely to explain the relationship between racial heterogeneity and overall punishment severity that we identify. If, for instance, white-majority jurisdictions elected white judges who punished Black defendants more severely, we should identify a negative relationship between the Black share of the population and the Black-white gap in local punishment severity.

To provide support for the hypothesis that local racial composition affects punishment severity through the preferences of the local electorate, online Appendix Table A14 employs jurisdiction-level data on support for statewide ballot measures related to the punishment of criminals and the rights of the accused. We find that increased local support for harsher punishment is strongly associated with higher punishment severity and has the same inverse U-shaped relationship with the Black share of the population and with the Black share of defendants (though the quadratic term is imprecise when controls are included).[36]

A natural question is whether the relationship that we identify between local racial composition and punishment severity generalizes outside of our sample states. Given that the estimation of jurisdiction-specific punishment severity requires rich defendant- and charge-level data that are not widely-available outside of our sample, it is not feasible to answer this question conclusively. However, we can utilize comparable data from the State Court Processing Statistics Data series, which includes a sample of cases from the nation's largest counties, to make progress in assessing generalizability (US Department of Justice 1990–2009). Most included states have coverage for two or fewer counties, so we focus on cross-state (as opposed to within-state) analyses. Our findings, presented in Appendix Table A15, reveal an inverse U-shaped relationship between the *state-level* Black population share and punishment severity. This pattern is also shown graphically in online Appendix Figure A4. Point estimates associated with the county-level Black share of the population are comparable to the estimates from Table 6 for the South-only sample, but are imprecise in both the South-only and nationwide samples.[37] The relative magnitudes of these estimates indicate that state-level racial composition may play a stronger role in explaining cross-state variation in punishment severity than local racial composition plays in explaining within-state variation.[38] We speculate that the central role of state-level racial composition may reflect the influence of racial dynamics on state laws, and we hope these suggestive findings motivate future research aimed at better understanding this relationship.

[36] Consistent with Cohen and Yang (2019), we also show in online Appendix Table A14 that Republican Party support is a strong predictor of punishment severity.

[37] In the full-sample specification, we identify a similarly imprecise inverse U-shaped relationship when the explanatory variables characterizing the state-level Black share of the population are excluded.

[38] Estimates imply that punishment severity is highest in states with a Black share of the population equal to 0.17. At this maximum, predicted severity is 82 percent higher in jurisdictions with this level of heterogeneity relative to all-white jurisdictions. For reference, the measured difference in punishment severity between the most lenient and harshest states in our sample is approximately 250 percent. In the South-only sample, the peak occurs where the Black share of the population is 0.18, though the implied difference in punishment severity between jurisdictions with this level of heterogeneity relative to all-white jurisdictions is much larger (685 percent). We note, however, that these South-only estimates are based on only nine data points and confidence intervals are wide. Moreover, the most homogeneous southern state included in the analysis, Kentucky, has a Black population share over 7 percent and so this calculation is particularly reliant on out-of-sample extrapolation.

IV. Discussion

We study the role that racial divisions play in explaining the punitiveness of US criminal justice policy by collecting and analyzing administrative criminal justice data from four Southern states. We identify substantial variation in punishment severity across jurisdictions within a given state and show that this variation persists even when we include a rich set of charge- and defendant-level covariates or compare arrest outcomes for defendants arrested in multiple jurisdictions. We proceed to write down a simple model of racial ingroup bias that predicts an inverse U-shaped relationship between local Black share of the population and punishment severity. This prediction is borne out in the data.

We assess the quantitative importance of our findings by simulating the share of charges leading to an incarceration sentence and the race-based gap in this share under a counterfactual in which more punitive jurisdictions adopt the punishment severity imposed by the jurisdiction in their state that, based on Black population share, would have a predicted punishment severity at the tenth percentile of the state's distribution. Specifically, we take jurisdictions with actual punishment severity above this predicted level and reassign their punishment severity to this level. Table 8 presents a comparison of actual confinement outcomes to the simulated confinement outcomes for whites versus Blacks in the four states in our sample.[39] In the simulation, we account for the fact that reduced punishment severity interacts dynamically with our criminal history measures, which are a function of past charge dispositions. In order to do so, we adjust confinement probability to account for the fact that simulated criminal histories will be made shorter than actual criminal histories by the reduction in conviction rates (and confinement rates, in Virginia) imposed.

Across all four states in the sample, the magnitude of the race-based confinement gap declines in level terms when we simulate outcomes. Importantly, this is not a mechanical consequence of the adjusted jurisdiction-specific punishment severity. Instead, this finding reflects the fact that Black residents of these states disproportionately reside in high-severity jurisdictions. Across states, the Black-specific measure of confinement sentences per capita declines by 15–20 percent, with an average decline of 16 percent, and the white-specific measure of confinement sentences per capita declines by 17–27 percent, with an average decline of 19 percent. Declines in punishment severity correspondingly reduce the magnitude of the gap in confinement sentences per capita by 12–16 percent, with an average decline of 14 percent. There are two caveats related to this simulation exercise that are worth highlighting. First, we abstract away from any endogenous changes in the degree (or location) of criminal behavior in response to adjustments in local punishment severity, including more mechanical incapacitation-driven responses. Second, we ignore any general equilibrium state-level statutory responses to changes in sentencing behavior. Nonetheless, our estimates provide insight into the significant role

[39] Simulation-based confinement sentences per capita measures do not line up precisely with population-based measures given the additive relationship between defendant covariates and charge dispositions that is assumed when constructing simulated outcomes under alternative counterfactual scenarios.

Table 8—Simulation Results

	Alabama	North Carolina	Texas	Virginia
Confinement sentences per 100,000				
White (actual)	592	481	1,673	644
Black (actual)	1,595	1,658	2,555	2,065
White (simulation)	429	391	1,396	516
Black (simulation)	1,280	1,405	2,171	1,714
Number of jurisdictions	67	100	253	118

Notes: Simulated confinement sentences per 100,000 age 15 or above are derived as described in Section IV. Statistics weighted by jurisdiction population.

that local discretion plays in explaining aggregate confinement rates and race-based confinement rate gaps.

While a large literature has documented the connection between racial stratification and support for public goods and redistribution, this research offers novel evidence that racial heterogeneity can be similarly linked to preferences for a "public bad": more punitive criminal justice policy. In the states in our sample, Blacks are more likely to reside in racially heterogeneous communities. As our simulation results demonstrate, this finding has important implications for the severity of criminal justice policy faced by the average white versus Black resident of these states. Moreover, our findings suggest that large race-based gaps in criminal justice outcomes may persist even in the absence of discriminatory treatment within any given jurisdiction.

REFERENCES

Abowd, John, Francis Kramarz, and David N. Margolis. 1999. "High Wage Workers and High Wage Firms." *Econometrica* 67 (2): 251–333.

Abrams, David S., Marianne Bertrand, and Sendhil Mullainathan. 2012. "Do Judges Vary in Their Treatment of Race?" *Journal of Legal Studies* 41 (2): 347–83.

Alabama Administrative Office of Courts. 2017. "Criminal Court Records, 2000–2010." (accessed November 3, 2016 via mail).

Alesina, Albert, Reza Baqir, and William Easterly. 1999. "Public Goods and Ethnic Divisions." *Quarterly Journal of Economics* 114 (4): 1243–84.

Alexander, Michelle. 2010. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness.* New York: New Press.

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters. 2017. "Leveraging Lotteries for School Value-Added: Testing and Estimation." *Quarterly Journal of Economics* 132 (2): 871–919.

Anwar, Shamena, Patrick Bayer, and Randi Hjalmarsson. 2012. "The Impact of Jury Race in Criminal Trials." *Quarterly Journal of Economics* 127 (2): 1017–55.

Arnold, David, Will Dobbie, and Crystal S. Yang. 2018. "Racial Bias in Bail Decisions." *Quarterly Journal of Economics* 133 (4): 1885–1932.

Berdejó, Carlos, and Noam Yuchtman. 2013. "Crime, Punishment, and Politics: An Analysis of Political Cycles in Criminal Sentencing." *Review of Economics and Statistics* 95 (3): 741–56.

Bobo, Lawrence D., and Devon Johnson. 2004. "A Taste for Punishment: Black and White Americans' Views on the Death Penalty and the War on Drugs." *Du Bois Review* 1 (1): 151–80.

Card, David, Ana Rute Cardoso, and Patrick Kline. 2016. "Bargaining, Sorting, and the Gender Wage Gap: Quantifying the Impact of Firms on the Relative Pay of Women." *Quarterly Journal of Economics* 131 (2): 633–86.

Card, David, and Stefano DellaVigna. 2020. "What Do Editors Maximize? Evidence from Four Leading Economics Journals." *Review of Economics and Statistics* 102 (1): 195–217.

Card, David, Jörg Heining, and Patrick Kline. 2013. "Workplace Heterogeneity and the Rise of West German Wage Inequality." *Quarterly Journal of Economics* 128 (3): 967–1015.

Carson, E. Ann. 2014. *Prisoners in 2013*. Washington, DC: Office of Justice Programs, US Department of Justice.

Cattaneo, Matias, Richard Crump, Max Farrell, and Yingjie Feng. 2019. "On Binscatter." https://cattaneo.princeton.edu/papers/Cattaneo-Crump-Farrell-Feng_2019_Binscatter.pdf.

Chen, Yan, and Sherry Xin Li. 2009. "Group Identity and Social Preferences." *American Economic Review* 99 (1): 431–57.

Chetty, Raj, John Friedman, and Jonah Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104 (9): 2593–2632.

Cohen, Alma, and Crystal S. Yang. 2019. "Judicial Politics and Sentencing Decisions." *American Economic Journal: Economic Policy* 11 (1): 160–91.

Collister, Brian, and Joe Ellis. 2015. "Texas Troopers Ticketing Hispanic Drivers as White." *American Renaissance*, November 9. https://www.amren.com/news/2015/11/texas-troopers-ticketing-hispanic-drivers-as-white/.

Dahlberg, Matz, Karin Edmark, and Helene Lundqvist. 2012. "Ethnic Diversity and Preferences for Redistribution." *Journal of Political Economy* 120 (1): 41–76.

Dyke, Andrew. 2007. "Electoral Cycles in the Administration of Criminal Justice." *Public Choice* 133 (3–4): 417–37.

Enos, Ryan D. 2016. "What the Demolition of Public Housing Teaches Us about the Impact of Racial Threat on Political Behavior." *American Journal of Political Science* 60 (1): 123–42.

Feigenberg, Benjamin, and Conrad Miller. 2021. "Replication data for: Racial Divisions and Criminal Justice: Evidence from Southern State Courts." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E119001V1.

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams. 2016. "Sources of Geographic Variation in Health Care: Evidence from Patient Migration." *Quarterly Journal of Economics* 131 (4): 1681–1726.

Fischman, Joshua B., and Max M. Schanzenbach. 2012. "Racial Disparities under the Federal Sentencing Guidelines: The Role of Judicial Discretion and Mandatory Minimums." *Journal of Empirical Legal Studies* 9 (4): 729–64.

Glaser, James. 1994. "Back to the Black Belt: Racial Environment and White Racial Attitudes in the South." *Journal of Politics* 56 (1): 21–41.

Goncalves, Felipe, and Steve Mello. Forthcoming. "A Few Bad Apples? Racial Bias in Policing." *American Economic Review*.

Gottschalk, Marie. 2015. *Caught: The Prison State and the Lockdown of American Politics.* Princeton, NJ: Princeton University Press.

Hetey, Rebecca C., and Jennifer L. Eberhardt. 2014. "Racial Disparities in Incarceration Increase Acceptance of Punitive Policies." *Psychological Science* 25 (10): 1949–54.

Huber, Gregory A., and Sanford C. Gordon. 2004. "Accountability and Coercion: Is Justice Blind When It Runs for Office?" *American Journal of Political Science* 48 (2): 247–63.

Keen, Bradley, and David Jacobs. 2009. "Racial Threat, Partisan Politics, and Racial Disparities in Prison Admissions: A Panel Analysis." *Criminology* 47 (1): 209–38.

Keller, Josh, and Adam Pearce. 2016. "This Small Indiana County Sends More People to Prison than San Francisco and Durham, N.C., Combined. Why?" *New York Times*, September 2. https://www.nytimes.com/2016/09/02/upshot/new-geography-of-prisons.html.

Key, Valdimer Orlando, Jr. 1949. *Southern Politics in State and Nation.* New York: Alfred A. Knopf.

Lim, Claire S.H. 2013. "Preferences and Incentives of Appointed and Elected Public Officials: Evidence from State Trial Court Judges." *American Economic Review* 103 (4): 1360–97.

Lim, Claire S.H., James M. Snyder, Jr., and David Strömberg. 2015a. "The Judge, the Politician, and the Press: Newspaper Coverage and Criminal Sentencing across Electoral Systems." *American Economic Journal: Applied Economics* 7 (4): 103–35.

Lim, Claire S.H., James M. Snyder, Jr., and David Strömberg. 2015b. "The Judge, the Politician, and the Press: Newspaper Coverage and Criminal Sentencing across Electoral Systems." OpenICPSR. https://www.openicpsr.org/openicpsr/project/113609/version/V1/view (accessed July 30, 2019).

Lind, Jo Thori, and Halvor Mehlum. 2010. "With or Without U? The Appropriate Test for a U-Shaped Relationship." *Oxford Bulletin of Economics and Statistics* 72 (1): 109–18.

Luttmer, Erzo F.P. 2001. "Group Loyalty and the Taste for Redistribution." *Journal of Political Economy* 109 (3): 500–28.

Miethe, Terance D. 1987. "Charging and Plea Bargaining Practices under Determinate Sentencing: An Investigation of the Hydraulic Displacement of Discretion." *Journal of Criminal Law and Criminology* 78 (1): 155–76.

Muhammad, Khalil Gibran. 2010. *The Condemnation of Blackness: Race, Crime, and the Making of Modern Urban America.* Cambridge, MA: Harvard University Press.

Nelson, Michael J. 2014. "Responsive Justice? Retention Elections, Prosecutors, and Public Opinion." *Journal of Law and Courts* 2 (1): 117–52.

North Carolina Administrative Office of the Courts. 2015. "Criminal Court Records, 2007–2014." (accessed August 6, 2019 via mail).

Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business and Economic Statistics* 37 (2): 187–204.

Pfaff, John F. 2014. "Escaping from the Standard Story: Why the Conventional Wisdom on Prison Growth Is Wrong, and Where We Can Go from Here." *Federal Sentencing Reporter* 26 (4): 265–70.

Raphael, Steven, and Sandra V. Rozo. 2018. "Racial Disparities in the Acquisition of Juvenile Arrest Records." https://www.sandravrozo.com/uploads/2/9/3/0/29306259/raphael_and_rozo_accepted_draft.pdf.

Reflective Democracy Campaign. 2018. "Justice for All? A Project by the Reflective Democracy Campaign on Who Prosecutes in America." Washington, DC: Reflective Democracy Campaign.

Rehavi, M. Marit, and Sonja B. Starr. 2014. "Racial Disparity in Federal Criminal Sentences." *Journal of Political Economy* 122 (6): 1320–54.

Schanzenbach, Max M. 2015. "Racial Disparities, Judge Characteristics, and Standards of Review in Sentencing." *Journal of Institutional and Theoretical Economics* 171 (1): 27–47.

Shayo, Moses, and Asaf Zussman. 2011. "Judicial Ingroup Bias in the Shadow of Terrorism." *Quarterly Journal of Economics* 126 (3): 1447–84.

Texas Department of Public Safety. 2015. "Texas Computerized Criminal History System." (accessed July 27, 2015 via secure electronic transfer).

US Department of Justice, Office of Justice Programs. 1990–2009. "State Court Processing Statistics, 1990–2009: Felony Defendants in Large Urban Counties." OpenICPSR. https://doi.org/10.3886/ICPSR02038.v5 (accessed July 30, 2019).

Unnever, James, and Francis Cullen. 2007. "Reassessing the Racial Divide in Support for Capital Punishment: The Continuing Significance of Race." *Journal of Research in Crime and Delinquency* 44 (1): 124–58.

Unnever, James D., and Francis T. Cullen. 2010. "The Social Sources of Americans' Punitiveness: A Test of Competing Models." *Criminology* 48 (1): 99–129.

Virginia Office of the Executive Secretary. 2016. "Criminal Court Records, 2006–2014." (accessed November 13, 2016 via secure electronic transfer).

Walmsley, Roy. 2016. *World Prison Population List: Eleventh Edition*. London: International Centre for Prison Studies.

Western, Bruce. 2006. *Punishment and Inequality in America.* New York: Russell Sage Foundation.