Judging Disparities: Recidivism Risk, Image Motives, and In-Group Bias on Wisconsin Criminal Courts

Elliott Ash and Claudia Marangon*

July 1, 2024

Abstract

This paper studies racial in-group disparities in Wisconsin, which has one of the highest Black-to-White incarceration rates among all U.S. states. The analysis is motivated by a model in which a judge may want to incarcerate more due to three factors: (1) taste-based preferences about the defendant's group identity; (2) higher recidivism risk where the defendant is more likely to commit future crimes; and (3) image motives stemming from the defendant being in the same group as the judge. Further, a judge may have better information on recidivism risk due to two factors: (4) becoming more experienced, and (5) sharing the same group as the defendant. We take these ideas to new data on 1 million cases from Wisconsin criminal courts, 2005-2017. Looking at racial disparities between majority (White) and minority (Black) judges and defendants, we find no evidence for anti-out-group bias (1). Using a recidivism risk score that we construct using machine learning tools to predict reoffense, we find evidence that judges do tend to incarcerate defendants with a higher recidivism risk (2). Consistent with judge experience leading to better information on defendant recidivism risk (4), we find that more experienced judges are more responsive in jailing defendants with a high recidivism risk score. Consistent with image motives (3), we find that when the minority group is responsible for most crimes, minority-group judges are harsher on their in-group. Finally, consistent with judges having better information on recidivism risk for same-group defendants (5), we find that judges are more responsive to the recidivism risk score for defendants from the same group when that group makes up a relatively small share of defendants.

JEL codes: J15, J16, K4, C53

*Center for Law and Economics, ETH Zurich. Contact emails: ashe@ethz.ch; cmarangon@ethz.ch. We thank Clémentine Abed Meraim, Nianyun Li, and Peiyao Sun for helpful research assistance. We thank Magdalena Breyer, Talia Gillis, Avery Katz, Ranae Jabri, Bentley MacLeod, Suresh Naidu, and participants at ALEA, the Columbia Law School Law and Economics Workshop, ETH Center for Law and Economics Brownbag Seminar, Nottingham Interdisciplinary Centre for Economic and Political Research Conference, Workshop of the Economics of Crime for Junior Scholars, and Zurich Political Economy Seminar Series for helpful feedback. We acknowledge research funding from ETH Zurich and from ERC Starting Grant 101042554.

1 Introduction

An institutional ambition of liberal societies is that judges should decide based on the facts of the case, rather than the characteristics of the perpetrator. Yet today that ambition remains unfulfilled, in the judicial context as in others characterized by disparities by race and class (e.g. Fagan and Ash, 2017). There are many potential drivers of these disparities, from differences in criminality to statistical discrimination and stereotyping to taste-based discrimination and in-group preferences. Distinguishing between these mechanisms is crucial because they entail different policy responses. Nonetheless, evidence on these mechanisms remains scarce.

In the context of criminal courts in Wisconsin, this paper provides evidence of the interplay between in-group disparities, group-image concerns, and statistical discrimination. In Wisconsin, racial disparities in incarceration rates are among the highest in the United States, with 2,742 per 100,000 Black residents incarcerated, compared to 230 per 100,000 White residents.[1] To help understand the origins of these disparities, we have access to unusually detailed records from Wisconsin Circuit Court Access (WCCA), a government-run legal database. The records include information on the case (e.g. charges filed), the defendant (name, birth date, gender, race, address), the judge, and the associated decisions (jail sentence, etc.). The dataset is powerful not only in its detail but in its scale – 1 million criminal cases decided by 564 judges in 72 counties over 12 years (2005-2017).

To complement this case data, we link it to newly collected information on the characteristics of the judges – gender, race, and political ideology. Further, we estimate a proxy for defendants' recidivism risk through a machine learning model, which captures potential additional information that judges may use in sentencing decisions. This proxy allows us both to investigate disparities for defendants with similar levels of predicted risk and to explore judges' responses to this proxy for better information on the defendant.

Even with such rich data, there are several challenges to isolating in-group disparities in this context. First, it could be that judge characteristics are correlated with case characteristics. To account for this, we leverage judges quasi-random assignment within each court.

[1] As the report Nellis (2021) states, this is particularly remarkable data given that the Black population in Wisconsin amounts to only 6% of the total population.

Therefore, using court-time fixed effects, we are able to exploit quasi-exogenous judge assignment to cases to rule out that these differences are due to judges with specific characteristics selecting into cases based on defendant type. Moreover, using charge fixed effects, we can rule out that the disparities are due to differences in the severity of crimes committed by different groups, the severity of charging behavior by prosecutors, or case assignment to judges based on charges. Further, we add flexible controls for recidivism risk in the regressions investigating in-group disparities to compare defendants with a similar level of predicted risk.

We start out by assessing in-group disparities in sentencing decisions – that is, are White judges favoring White defendants, or Black judges favoring Black defendants? In line with some of the literature, like Ash et al. (2022), we find mixed evidence of in-group bias on average. If anything, we see that Black judges are harsher on Black defendants, rather than more lenient. That result is related to previous evidence, for example by Depew et al. (2017), showing anti-in-group bias in some contexts.

To explore more deeply the connections among preferences, information, and image concerns in judicial disparities, we propose a model of a judge's decision on whether to incarcerate or release a criminal defendant. In the model, the judge observes a signal about the defendant's true recidivism risk. Judges decide on whether to jail or release defendants based on a threshold rule where the threshold value includes a bias component due to taste-based preferences (Phelps 1972; Becker 1957). Further, the precision of the signal increases with judges' experience and when the judge and defendant belong to the same race group. Finally, judges have image concerns that might cause them to punish in-group members more harshly when they are committing many crimes (Marques and Yzerbyt, 1988).

Using these insights, we explore additional dimensions of the data. Consistent with a role for information, judges are more likely to give jail time to defendants who are more likely to recidivate. Further, more experienced judges have a steeper risk-jail gradient than less experienced judges.

Next, we show that there is a robust, significant anti-in-group bias among Black judges, but the effect is concentrated among judges who rule on cases with mostly Black defendants. One potential mechanism for this result is image concerns among judges, where they want

to punish members of the same group if they are responsible for most crimes. That would be consistent with group-image theory from social psychology, which highlights how misbehaving in-group members are judged more harshly than members of the out-group when this behavior becomes salient (Marques and Yzerbyt, 1988).

Finally, we assess in-group heterogeneity in response to recidivism risk. We find that when judges have seen relatively few defendants from the minority group, they are more responsive to recidivism risk among defendants from the same group – that is, judges are relatively lenient for same-race defendants that have low recidivism risk, but relatively harsh on same-race defendants with high recidivism risk. We interpret this evidence as in line with the prediction of the model, assuming that judges get a higher quality signal for defendants with similar identity characteristics.

These results are relevant to a large social-science literature and policy apparatus on disparities in criminal justice (Fagan and Ash, 2017). Previous work has shown large disparities towards racial minorities (Arnold et al., 2018), even when controlling for recidivism risk (Arnold et al., 2022; Jung et al., 2024). There is more mixed evidence on in-group bias, with most papers showing positive in-group bias (e.g. Shayo and Zussman, 2011), some showing a null effect (Ash et al., 2022; Lim et al., 2016), and others showing *anti-in-group* bias (Depew et al., 2017). Finally, there is some work on image motives driving judge decision-making (Guo et al., 2023). This paper combines these three elements of judge decision-making, to show that there are differential responses to recidivism risk when judge and defendant share a race identity, which can be further moderated by the population share of the minority group. A differential in-group response to riskiness could lead to positive, null, or negative in-group disparities on average, even in the absence of racial animus. Previous work in this area should be reviewed in light of these insights, especially since many datasets lack information on judge characteristics, defendant recidivism risk, or both.

On the policy side, this evidence has a number of implications. First, if disparities are not due to racial animus at the sentencing level, then policies designed to adjust implicit racial attitudes or prejudices of judges are unlikely to reduce observed disparities. However, if these disparities are due to incorrect racial stereotypes, then providing additional information and statistics on recidivism risk across racial groups could be of help. Further, the use of

individualized risk scoring, if done responsibly (Chouldechova, 2017), could help reduce racial gaps due to stereotypes and low information.

The rest of this paper is organized as follows. Section 2 provides details on the institutional setting. Section 3 describes the data sources and construction of the risk score. Section 4 outlines the empirical strategy and provides the baseline results on in-group bias. Section 5 introduces a simple model of judges' decisions with different signal precisions that helps rationalize potential mechanisms. Section 6 provides additional empirical analysis motivated by the model, while Section 7 provides additional robustness checks. Section 8 concludes.

2 Institutional Setting

We analyze sentencing decisions from criminal cases in the Circuit Courts of Wisconsin. There are 69 of these courts, most serving a single county and three serving two counties. They handle most felony, misdemeanor, and criminal traffic crimes, with a few exceptions like cases involving drugs, drunk driving, or veterans, for which counties sometimes have specialized courts. Each circuit court is divided into branches, to each of which a judge is assigned.[2] For large counties with many branches, judges rotate across divisions (e.g. criminal, civil) every 2 to 4 years, while for small counties, judges handle all types of cases.

Most criminal cases are initiated by a prosecutor filing a complaint. At this point, cases are assigned to a branch according to rules that change across courts. In most courts, cases are randomly assigned to a branch within the criminal division by a computer, either unconditionally or conditional on the caseload of each branch.

In the initial appearance, the defendant appears in front of the court and is informed about the charges filed against him[3] and about his rights to an attorney. If the defendant is in custody, the court determines whether the defendant can be released on bail, and, in case he is released, the bail conditions.[4] Afterward, an arraignment is held where the complaint's

[2]The only exceptions are the courts of Buffalo and Pepin, Florence and Forest, and Shawano and Menominee, which are paired off and share judges.

[3]For clarity, we will use he/him to refer to the defendant and she/her to refer to the judge.

[4]In case of a felony, the defendant also has the right to a preliminary examination of the case, during which the prosecutor has to provide evidence that the defendant has committed a felony. If the prosecutor fails to do so, then the court dismisses the complaint, and the defendant is released.

information is read aloud, and the defendant makes a plea. With pleas of "guilty" or "no contest", the judge can immediately pronounce a sentence, including probation. With a plea of "not guilty", the case is scheduled for trial.

The Wisconsin Supreme Court has issued detailed guidelines for judges in deciding about criminal sentences. As outlined in *McCleary v. State*, 49 Wis. 2d 263 (1971), judges should take into account multiple factors during sentencing and should report how such considerations affect their decision. The key precedent listing the factors that judges should use in sentencing is *State v. Gallion*, 270 Wis. 2d 535 (2004). *Gallion* outlines the main objectives when deciding on sentences, including the protection of the community from potential further crimes, punishment of the defendant, rehabilitation of the defendant, and deterrence. *Gallion* goes through the factors that the courts may take into account to pursue these objectives, mentioning, among others, past criminal records, the nature of the crime, and defendants' characteristics such as education and employment records.

3 Data

This section provides background on the data. Section 3.1 discusses the main data source for cases. Section 3.2 describes how we build a recidivism risk score using machine learning methods.

3.1 Wisconsin Circuit Courts Access

The case data come from the API service of Wisconsin Circuit Courts Access (WCCA, available at wcca.wicourts.gov), which includes detailed records on trial court cases, with data starting from 1970 (Li et al. 2022; Ash et al. 2023).[5] We use data on criminal cases filed from 2005 to 2017 in 72 county courts in Wisconsin, resulting in more than 1 million cases[6]. In particular, we conduct our empirical analysis on a restricted sample that includes only Black and White defendants – around 900,000 cases.

[5]This same data source is used by Berdejó (2018, 2019) to examine racial and gender differences in plea bargaining.

[6]It is worth pointing out that the cases we have access to are subject to some limitations outlined in Appendix Section A

The case records include information about the charges, the type of offense (felony, misdemeanor, and criminal traffic), defendants' demographic information (e.g., gender, race, and date of birth), and names of judges, attorneys, and prosecutors. Additionally, the records contain information about bail hearings, bail amounts, sentencing (including probation), records about incarceration and parole, and many other events. The records contain information about judges' decisions on the case (whether the defendant is incarcerated or not) which is the main outcome of our analysis, and the number of days of jail assigned, which we will use as a secondary outcome.

We construct a number of additional variables from the case records. We use a combination of first name, last name, and date of birth as a unique identifier for a defendant. This identifier allows us to conduct a search in the database of case records to match the defendant across multiple cases and construct the additional variables. We obtain the defendant's prior count of each of the three crime types – felony, misdemeanor, and criminal traffic – for each case. We infer age at judgment from the defendant's date of birth and judgment disposition date of the case. Age at first offense is the the age when the defendant committed the first crime.

Table 1 presents summary statistics of the main dataset. There are five ethnic groups, with around 66% Caucasian and 21% African American. The proportion of male criminals (80%) is significantly higher than females, and most cases are committed at a younger age (below 30). The recidivism rate is the highest (55%) among Native Americans. Misdemeanors are the most frequent crime type except for Hispanics, with criminal traffic (46%) being the most common crime type. Finally, the average incarceration rate is around 40% and the average sentence length is 407 days; both are highest for African Americans.

Additional summary statistics are reported in the appendix. First, more detailed case tabulations by crime type and charge severity are reported in Appendix Tables B.3, B.4, B.5, B.6, and B.7. Appendix Figure B.1 shows variation in judge and defendant characteristics by charge severity. Appendix Figure B.2 shows a time series of the main and secondary outcomes over time (jail rate and log length of jail sentence).

Next, we collected data on 564 judges in these courts. We code judges' race and gender by consulting the judge websites and Ballotpedia. We compute judge experience/tenure by

Table 1: Summary of Wisconsin Circuit Courts Dataset

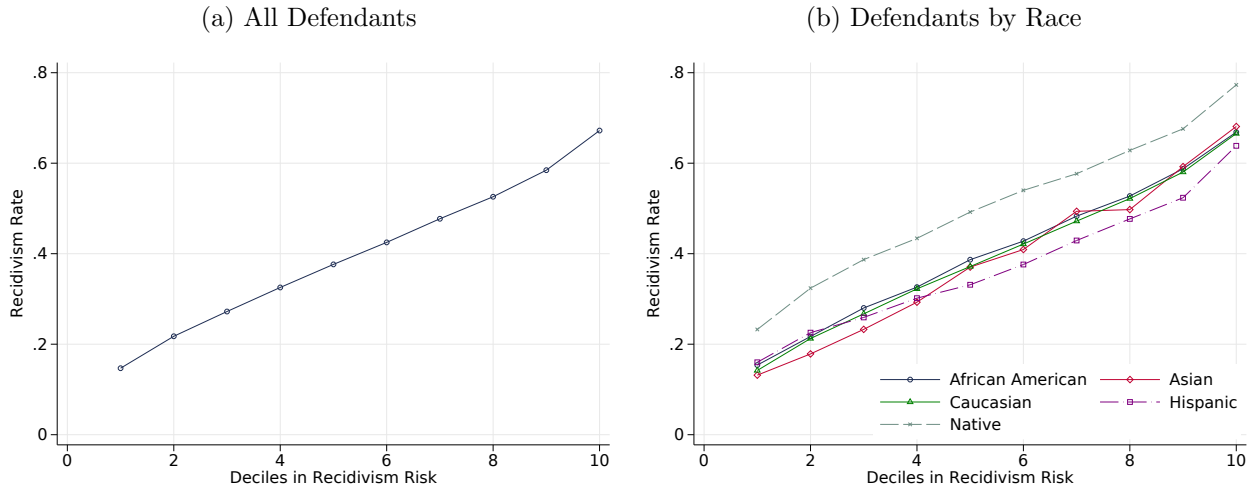
	Full sample	Caucasian	African American	Hispanic	Native American	Asian
Sample size	1,029,314	674,474	219,837	77,893	47,079	10,031
Sample share		0.66	0.21	0.08	0.05	0.0097
Incarcerated	0.39	0.37	0.48	0.36	0.38	0.36
Recidivism Rate	0.40	0.39	0.43	0.36	0.55	0.36
Sentence Length	299	253	407	265	217	373
Sex						
Female	0.20	0.22	0.16	0.12	0.32	0.13
Age						
Below 30	0.53	0.51	0.56	0.57	0.56	0.68
30 to 60	0.47	0.49	0.43	0.43	0.44	0.32
Case type						
Felony	0.33	0.32	0.43	0.21	0.31	0.37
Misdemeanor	0.43	0.44	0.44	0.33	0.48	0.40
Criminal Traffic	0.23	0.24	0.23	0.46	0.21	0.23

Notes: Summary statistics for case and defendant characteristics, in aggregate and by race of the defendant. The unit of observation is defendant-case. *Incarcerated* represents the share of judge's decision to sentence the defendant to jail in that case, while *Recidivism Rate* is the share of observed episodes of recidivism. *Sentence Length* represents the harshness of the sentence and is measured in number of days.

the number of years since they show up in the dataset. Finally, we infer judges' partisanship and ideology by matching judges to personal campaign donations to candidates from the DIME Dataset (Bonica, 2016). Summary statistics on these characteristics are reported in Appendix Tables B.1 and B.2. In particular, in Appendix Table B.1 we show judges' characteristics separately for White and Black judges. Judges from different races are overall similar with respect to sentencing harshness, tenure, and political preferences as measured by contributions.

Finally, we merge local demographic variables from the 2010 census data. That includes population density, the share of people with a college education, the share of people eligible for food stamps, African American population share, Hispanic population share, male population share, population share who live in rural and urban areas, and median household income. We merge in these variables for each case twice – once at the county level (for the court) and once at the zip code level (for the defendant's address).

Figure 1: Recidivism Rate by Risk Decile in Held-Out Test Set



Notes: Average share of defendants who re-offend within a two-year period, plotted by recidivism risk deciles. Recidivism risk deciles are centered by county-year and charge severity. Panel (a) plots all defendants together, while panel (b) plots defendants separately by race.

3.2 Recidivism Risk Score

Following Li et al. (2022), we construct a measure for recidivism risk. We train a machine learning model to predict whether the defendant commits another crime within two years from when he is released (see Appendix Figure C.3)[7] on the whole sample – i.e, including defendants of all ethnicities.

The target variable of interest is whether a defendant recidivates or not[8]. We use gender, type of offense, prior criminal count (separately by type), and age at judgment and at first offense as features to predict the target variable. We use a gradient-boosted classifier from the Python package XGBoost implementation (Chen and Guestrin, 2016). We train the model, including both L1 and L2 regularization, and tune the hyperparameters via grid search and 5-fold cross-validation.[9]

The quality of the recidivism risk score is evaluated in held-out test data that the model did not observe during training. Appendix Table C.8 reports a number of standard classifier

[7]We define this two years threshold following Larson et al. (2016), who defined recidivism as a new offense within a two year period, mainly because Northpointe, the company that designed the COMPAS tool, indicates that its recidivism score is based on that timeline. Further, a study (Hunt and Dumville, 2016) by the U.S. Sentencing Commission showed that most recidivists re-offend within two years after release (if they re-offend at all).

[8]See Appendix C for further details on the construction of this variable.

[9]See Appendix Section C.2 for a detailed explanation of XGBoost.

metrics: accuracy, area under the ROC curve (AUC), false positive rate (FPR), and false negative rate (FNR). At about 67% accuracy and AUC=0.7, the classifier achieves a similar performance to other work predicting recidivism in other jurisdictions (e.g. Lakkaraju et al., 2017; Chouldechova, 2017; Kleinberg et al., 2018). Further, the model ranks defendants well by recidivism risk, as shown in Figure 1. For example, in the top decile of the algorithm's score, around 70% of defendants re-offend. In the bottom decile, under 20% re-offend. Finally, the algorithm has similar error rates across race of the defendant (Appendix Table C.8).

Using this trained model, we compute out-of-sample predictions of the probability of recidivism (recidivism risk score) for each defendant. Appendix Figure C.6 shows that there are different distributions of risk by race and gender (see also Appendix Figure C.7). The distribution for Black defendants is shifted right, indicating a higher predicted recidivism risk on average (reflecting higher arrest rates in the dataset). The gender difference is even starker, with male defendants having a much higher average recidivism risk than female defendants.

Judge decisions are correlated with recidivism risk. Appendix Figure C.4 shows that as recidivism risk increases, the jail rate increases. This relationship holds when conditioning on court-year fixed effects and charge fixed effects (indicators for the severity of the most severe charge).[10]

4 Judicial In-Group Bias

This section presents the main analysis of in-group bias in criminal justice outcomes in the Wisconsin courts. Section 4.1 describes our empirical strategy, detailing how we leverage random assignment. Section 4.2 outlines our regression model and how we deal with further caveats to identification. Section 4.3 reports the main results on in-group bias.

[10]Appendix Figure C.5 shows the jail rate by recidivism risk, separately by judge race and defendant race. As will be explored further below, Black judges are harsher on Black defendants, but treat all White defendants the same.

4.1 Empirical Strategy

Our empirical strategy is a regression analysis that estimates the effect of defendant and judge characteristics on judge release decisions. A major concern in this setting is that judges may select into cases based on their preferences, expertise, and defendant characteristics. If this is the case, then the disparities we observe across judge/defendant groups could be due to case and/or judges' characteristics, like case severity or judge leniency, rather than actual in-group preferences. For instance, suppose that Black judges are strongly against thefts of automobiles; in turn, they would try to pick cases with such offenses and rule harshly on these. Suppose that, meanwhile, accused automobile thieves are mainly African Americans. Then, we would observe greater harshness by Black judges toward Black defendants, but that would be due to the selection of the harsh judge into auto theft cases rather than due to in-group bias.

We can address this concern in the setting of Wisconsin Courts, thanks to procedural rules that prevent trial judges within the same court from selecting into different types of cases (conditional on observables). In most courts, cases are randomly assigned to judges by a computer program. In most others, we were able to verify that case assignment is not intentionally random but "quasi-random" in the sense of being arbitrary – that is, judges cannot intentionally select defendants, and defendants cannot intentionally select judges. These are official rules, so it is possible that they are not being followed with perfect compliance. Further, there are some exceptions where, for example, judges could recuse themselves from a case, or defendants could formally request a new judge. Finally, in some courts, we could not verify that assignment was (quasi-)random. Our results are robust to dropping those courts from the dataset.

To check that judge assignment is orthogonal to case characteristics, we run a set of balance checks using our data. Specifically, we run the following regression to check if judges' race is correlated with defendants' and cases' characteristics:

$$x^k_{ijct} = \beta_0 BlackJudge_j + \alpha_{ct} + \alpha^s_i + \varepsilon_{ijct} \tag{1}$$

where x^k_{ijct} is a characteristic k of case/defendant i (e.g., charge severity, ethnicity, gender)

assigned to judge j in court/county c at year/month t. On the right-hand side, $BlackJudge_j$ is an indicator for the judge being Black, while α_{ct} and α_i^s include county-year and charge severity fixed effects to allow for block randomization of judges to cases. As judges are (quasi-)randomly assigned to cases within a court and particular time period, court fixed effects are interacted with time fixed effects (court-year or court-month). Further, there can be selection of judges into cases based on the severity of the crime (see Appendix Figure B.1), which can happen, for example, when new judges do not sit on capital cases. Hence α_{ijct} also includes charge severity fixed effects.

Summary statistics on balance and associated regression checks are reported in Table 2. First, Columns 1 and 2 report the unadjusted means of each defendant's characteristics separately for White and Black judges. Column 3, reporting the unadjusted differences in means, shows that without the FE Black and White judges handle cases with different types of defendants. For instance, Black judges handle cases with more severe charges, more Black defendants, and defendants from poorer places. Finally, Column 4 reports the regression estimates for the black-white difference after adjusting for the FE. Most of these differences become small and non-significant when we include court-year and charge severity fixed effects.

4.2 Regression Specification

To analyze racial in-group bias, we estimate the following regression equation:

$$y_{ijct} = \mathbf{x}_i'\delta_x + \mathbf{x}_i'\mathbf{w}_j'\delta_{xw} + \beta BlackJudge_j \times BlackDefendant_i + \alpha_{ct} + \alpha_i^s + \alpha_i^v + \alpha_j + \varepsilon_{ijct} \quad (2)$$

where i denotes a defendant-case, j the judge assigned to that case, c is the court/county, and t is year. The main outcome y_{ijct} is an indicator variable taking value 0 if the defendant in i is released and 1 if he is incarcerated. We also look at the effect on days spent in jail as a proxy for the harshness of the sentence. To leverage quasi-random judge assignment, we include court-time fixed effects (α_{ct}) and thus address issues of harsher judges selecting into/out of cases with same-race defendants. Similarly, we include charge severity fixed effects (α_i^s) to

Table 2: Randomization Checks for Judge Assignment

	Mean		Difference in Means	
	White Judges	Black Judges	Without FE	With FE
	(1)	(2)	(3)	(4)
Charge Severity	9.894	10.431	-0.537**	-0.217
	(2.705)	(3.025)	(0.024)	(0.189)
Recid. Risk	0.427	0.406	0.020**	0.00523
	(0.172)	(0.162)	(0.001)	(0.00352)
Black Defendant	0.241	0.579	-0.339**	0.0149
	(0.428)	(0.494)	(0.004)	(0.0107)
Female Defendant	0.208	0.170	0.038**	0.00109
	(0.406)	(0.375)	(0.004)	(0.00776)
Defendant Age	31.616	30.669	0.948**	0.277
	(11.242)	(10.810)	(0.098)	(0.336)
Prior Offense	0.772	0.732	0.040**	0.0160+
	(0.420)	(0.443)	(0.004)	(0.00861)
Misdemeanor	0.437	0.384	0.053**	-0.0172
	(0.496)	(0.486)	(0.004)	(0.0147)
Felony	0.347	0.454	-0.107**	-0.0319+
	(0.476)	(0.498)	(0.004)	(0.0192)
Criminal Traffic	0.216	0.162	0.054**	0.0490**
	(0.411)	(0.369)	(0.004)	(0.0108)
Zip Shr. Black	0.104	0.317	-0.213**	0.00905+
	(0.208)	(0.303)	(0.002)	(0.00513)
Zip Shr. Male	0.499	0.487	0.012**	0.000301
	(0.035)	(0.030)	(0.000)	(0.000552)
Zip Shr. Urban	0.578	0.912	-0.334**	0.00242
	(0.458)	(0.243)	(0.004)	(0.00396)
Zip Shr. College	0.230	0.261	-0.031**	0.00247
	(0.109)	(0.149)	(0.001)	(0.00187)
Zip Shr. Food Stamps	0.122	0.185	-0.063**	0.00131
	(0.078)	(0.109)	(0.001)	(0.00174)
Zip Median Income (Log)	10.766	10.626	0.140**	-0.016*
	(0.270)	(0.358)	(0.002)	(0.007)

Notes: Balance test of defendants' characteristics in the sample with only White and Black defendants (N=894,311). Recid. Risk is the predicted probability of recidivism. Columns 1 and 2 report the mean and standard deviation (in parentheses) of defendants' characteristics separately for White and Black judges. Column 3 reports the unadjusted difference in means between White and Black judges, with standard errors of the mean in parentheses. Column 4 reports the difference in means after taking out county-year and charge severity fixed effects, with standard errors (in parentheses) clustered at the county-year level): + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

address the issue that judges handle same-race defendants who commit more severe crimes. We also include risk ventile fixed effects (α_i^v) that allow us to compare defendants with similar misconduct potential. These ventile fixed effects are calculated starting from the residualized recidivism risk score, after taking out county-year-judge fixed effects, and are designed to address the issue that recidivism risk is correlated with race (Appendix Figure C.4). In particular, Black defendants have much higher recidivism rates and predicted recidivism risk than other defendants. Hence, if judges from different ethnicities respond differently to recidivism, variation in same-race judge decisions across races could be a response to correlated recidivism risk. When including the risk ventile fixed effects, we estimate variation in judge decisions according to defendant characteristics while flexibly controlling for such risk.[11]

Finally, given the imperfect nature of random assignment in our context, particularly concerning courts with fewer judges, we control for additional defendant characteristics, like age and gender ($\mathbf{x}_i'$). We also include the interaction between defendants' and judges' characteristics ($\mathbf{x}_i'\mathbf{w}_j'$) to account for potential correlations between judges' characteristics and defendants' or judges' race. For instance, we observe that there are no Black female judges. If male judges are, on average, harsher than female judges towards specific types of crimes and, at the same time, Black defendants are more likely to commit these crimes, we would observe in-group bias due to judges' sentiments toward specific crimes rather than race. Finally, we include judge fixed effects (α_j) to account for all time-invariant judge's characteristics. We cluster standard errors at the county×year level.

The coefficient of interest is β, which estimates the extent of in-group bias through a difference-in-difference estimate. Namely, β should be interpreted as how much Black judges are more likely than White judges to incarcerate Black defendants rather than White defendants. Still, the regression results should be interpreted with caution. A β different from zero does not necessarily indicate bias due to in-group preferences, since this might vary along with defendants' characteristics. In turn, a non-significant coefficient does not necessarily imply the absence of in-group bias. We will explore these issues in more detail

[11]For robustness, we create recidivism risk ventiles also within court-year-judge-race. Appendix Figure C.8 shows the across and within-race rankings showing are almost identical (the Pearson correlation is 0.9913). The results are robust to using either risk score ranking.

Table 3: In-Group Bias: Jail Decision

	Defendant is Incarcerated				
	(1)	(2)	(3)	(4)	(5)
Black Judge	-0.0300	-0.0175	-0.0168	0	0
	(0.0356)	(0.0326)	(0.0323)	(.)	(.)
Black Defendant	0.0414**	0.0630**	0.0506**	0.0512**	0.0519**
	(0.00268)	(0.00297)	(0.00292)	(0.00242)	(0.00231)
Black Judge × Black Defendant	0.0263	0.0205	0.0216	0.0599**	0.0518**
	(0.0208)	(0.0171)	(0.0175)	(0.0118)	(0.0105)
Obs.	894311	894311	894311	894291	883893
R^2	0.0403	0.0919	0.101	0.114	0.136
County-Year FE	X	X	X	X	X
Charge Severity FE		X	X	X	X
Risk Ventile FE			X	X	X
Judge FE				X	X
Additional Interactions & Controls					X

Notes: Estimated racial in-group bias in jail decision by judges. All specifications include county×year fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Risk ventiles are computed separately across all ethnicities using our machine-learning predicted recidivism risk. Additional interactions and controls include defendants' characteristics other than race and the interaction between these and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the county and at the year level (in parenthesis): + p < 0.1, * p < 0.05, ** p < 0.01.

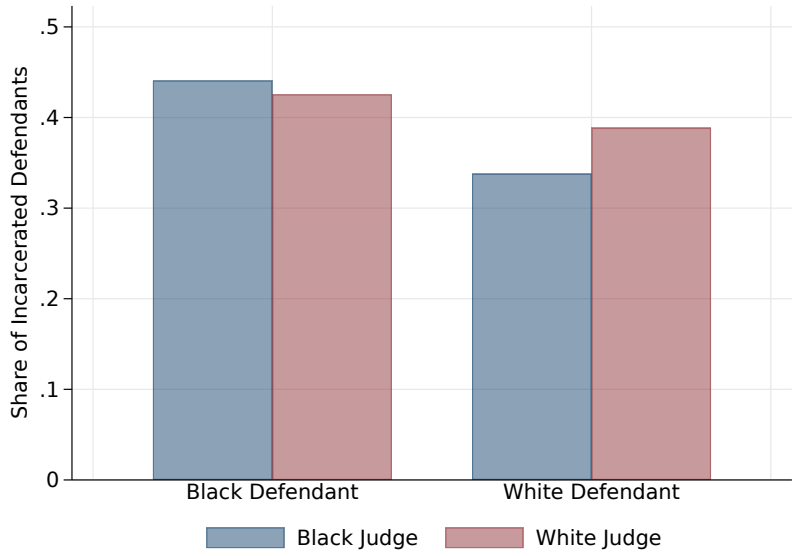
in Sections 5 and 6.

4.3 Results

We report the main results from Equation (2) in Table 3. Column 1 presents the baseline difference-in-difference analysis, including only county-year fixed effects to leverage the quasi-random assignment of judges and account for any selection issues. The first row says that Black judges do not differ from White judges in their overall incarceration rates. The second row says that Black defendants are about 4.1 percentage points more likely to be incarcerated than White defendants.

The interaction term, in the third row, shows mixed results on racial in-group bias. First, we can rule out pro-in-group bias – i.e., judges are not more lenient on same-race defendants. In Columns 1 through 3, there is a null estimate suggesting no in-group bias, which holds with the inclusion of charge severity and risk ventiles fixed effects (Columns 2

Figure 2: Share of Incarcerated Defendants by Judge and Defendant Race



Notes: Share of incarcerated defendants separately by judge and defendant race. The average share for each combination of judge-defendant race is calculated by adding the average share of incarcerated defendants to the residual variation from a regression of the outcome on county-year, charge severity, and judge-fixed effects.

and 3). However, when we include judge fixed effects in Columns 4 and 5, the estimate for in-group bias becomes statistically positive, suggesting *anti*-in-group bias – i.e., judges are *more* likely to incarcerate same-race defendants. Specifically, a defendant is 6 percentage points more likely to be incarcerated if his case is handled by a judge of the same race. In Column 5, we include controls for defendants' characteristics and interactions between these and judges' characteristics and observe that this increases explanatory power while keeping the estimate for in-group bias virtually unchanged.

The estimates from Column 4, including judge fixed effects, are visualized in Figure 2. The graph shows that, after the regression adjustments, Black judges (in blue) are slightly harsher on Black defendants and slightly more lenient on White defendants (relative to White judges, in red). These results are similar to the result on anti-in-group bias from Depew et al. (2017).

Next, we investigate the presence of in-group bias in sentence harshness, measured as the logarithm of days spent in jail. In Appendix Table E.11 we report our main results in Column 1 and the results for sentence length in Columns 2 and 3, using our preferred spec-

ification. Column 2 includes all defendants, both incarcerated and released, thus combining the intensive and extensive margin of jail decision-making. In this specification, we observe that Black defendants' jail sentences are about 25% longer when given by same-race judges, in line with our main result that Black judges are harsher towards same-race defendants. Column 3 presents the same sentence-length analysis while dropping released defendants, hence focusing only on the intensive margin. Differently from before, the coefficient on in-group bias is negative and non-significant. This suggests that there is no in-group bias in jail length, conditional on getting a jail sentence.

As mentioned in Section 4, it is possible that the in-group bias we observe is due to differences in characteristics between Black and White judges. Therefore, we investigate this in Appendix Table E.12, where we add interactions between the indicator for Black defendants and other judge's characteristics to our main specification. We observe that the coefficient on in-group bias is robust to the inclusion of these interactions and remains similar in magnitude. Moreover, there are no disparities by judges' gender or political affiliation, while we observe that more experienced judges are more likely to incarcerate Black defendants.

Finally, in Appendix Table E.13, we investigate the heterogeneity of in-group bias across additional judges' characteristics to investigate whether specific judges are driving the in-group bias we observe. We do this by interacting $BlackJudge \times BlackDefendants$ with judges' experience and political affiliation. The results suggest that there is heterogeneity in in-group bias, but adding all the interactions does not affect our estimate for in-group bias much.

5 A Model of Judging With Bias and Uncertainty

The previous section provided evidence of the presence of in-group bias against same-race defendants. We showed that this is not driven by differences in Black and White judges and it cannot be explained by judges' political affiliation and experience. Thus, we cannot say much yet about the mechanisms behind in-group bias in sentencing. This section provides a model of judges' sentencing decisions with recidivism risk that allows us to derive some empirically testable predictions for further exploring the mechanisms behind anti-in-group

bias.

There are two parts. Section 5.1 outlines a simple model of judges' decision-making when they receive a signal about defendants' recidivism risk. Section 5.2 focuses on how judges' decisions change with varying levels of recidivism risk.

5.1 A Model of Judicial Decision-Making

We model the decision of a judge j with characteristics $W_j \in \{A, B\}$, such as gender and race, about a defendant i with characteristics $X_i \in \{A, B\}$. Each defendant has true, unobserved recidivism risk $r_i \in \mathbb{R}$. The prior distribution on risk is $r_i \sim \mathcal{N}(\mu(X_i), 1)$, where the mean can vary according to defendant characteristics, and the variance (and precision) is assumed to always be one for simplicity.

The judge in case i observes an informative signal on recidivism risk, $\tilde{r}_i = r_i + \epsilon_i$, where $\epsilon_{ij} \sim \mathcal{N}(0, \frac{1}{\rho_{ij}})$. The precision of the signal $\rho_{ij} \geq 0$ depends on whether the defendants and the judge have similar characteristics and on the judges' experience ($e_j \geq 0$). In particular, if the judge is not from the same group as the defendant, the precision of her signal increases with the share of defendants from the out-group she observes, $\frac{n_{X_i}}{n_{X_i}+n_{-X_i}}$, where $n_{X_i} \geq 0$ is the number of defendants with characteristic X_i. On the other hand, precision does not depend on population shares for judges from the defendant's in-group. Moreover, precision increases with years of experience. That is,

$$\rho_{ij} = e_j \begin{cases} \rho & \text{if } X_i = W_j \\ \rho s_{X_i} & \text{if } X_i \neq W_j \end{cases} \tag{3}$$

where $s_{X_i} = \frac{n_{X_i}}{n_{X_i}+n_{-X_i}}$ is the share of defendants with characteristics X_i. The posterior belief of judge j about recidivism risk of i is $\hat{r}_{ij} \sim N\left(\bar{r}_{ij}, \frac{1}{1+\rho_{ij}}\right)$, where

$$\bar{r}_{ij} = \frac{\mu(X_i)}{1+\rho_{ij}} + \frac{\rho_{ij}}{1+\rho_{ij}}\tilde{r}_{ij} \tag{4}$$

that is, the average of the prior and the signal, weighted by the relative precision of the signal.

Next, we define judges' utility when incarcerating/releasing defendants of type i after observing a signal about recidivism risk:

$$U_j\left(Release_i|\tilde{r}_{ij}\right) = \phi\mathbb{1}\left\{X_i = W_j\right\} - \tau\mathbb{E}\left[r_i|\tilde{r}_{ij}\right]$$
$$U_j\left(Incarcerate_i|\tilde{r}_{ij}\right) = \gamma\mathbb{1}\left\{X_i = W_j\right\}\frac{n_{X_i}}{n_{-X_i}} - \bar{C}_j.$$

Here, $\phi\mathbb{1}\left\{X_i = W_j\right\}$ indicates the preference of judges to release in-group defendants, where $\phi > 0$ indicates *pro-in-group* bias and $\phi < 0$ *anti-in-group* bias. $\tau\mathbb{E}\left[r_i|\tilde{r}_{ij}\right]$ is the expected cost from releasing defendants with a given likelihood of re-offense. $\gamma\mathbb{1}\left\{X_i = W_j\right\}\frac{n_{X_i}}{n_{-X_i}}$ represents a benefit from incarcerating based on group-image concerns, arising if the judge is from the same group as the defendant and there is a sufficiently high number of defendants from that group. That is, judges want to punish same-group defendants, due to an image benefit that increases with the number of defendants from the same group, as the crime hurts the group's reputation (Guo et al., 2023).[12] Finally, $\bar{C}_j$ is a judge-specific cost of incarcerating. Judge j incarcerates defendant i if

$$\tau\mathbb{E}\left[r_i|\tilde{r}_{ij}\right] + \gamma\mathbb{1}\left\{X_i = W_j\right\}\frac{n_{X_i}}{n_{-X_i}} \geq \phi\mathbb{1}\left\{X_i = W_j\right\} + \bar{C}_j \tag{5}$$

that is, the cost of release is larger than the benefit.

Now, let us focus on the decision of a judge about a defendant from the same group, i.e., we assume that $X_i = W_j$. In this scenario, the decision rule becomes

$$\tau\left(\frac{\mu(X_i)}{1+e_j\rho} + \frac{e_j\rho}{1+e_j\rho}\tilde{r}_{ij}\right) + \gamma\frac{n_i}{n_{-i}} \geq \phi + \bar{C}_j. \tag{6}$$

For the same level of observed recidivism risk and judge-specific cost of incarceration, we can observe anti-in-group bias against same-race defendants on average for at least three reasons. First, there could be group-image concerns ($\gamma > 0$). Second, judges could have preferences for incarcerating defendants from the same group ($\phi < 0$). Third, if we allow

[12]This assumption combines the group-image theory from social psychology highlighting how misbehaving in-group members are judged more harshly than members of the out-group (Marques and Yzerbyt, 1988), and the literature on statistical stereotypes (Bordalo et al., 2016).

observed recidivism risk to vary, there could be statistical discrimination, where judges from different groups have different beliefs on defendants' riskiness – by either having a different prior mean for their own group, or responding differently to signals from their own group.

The decision rule given by (6) does not allow us to directly disentangle these reasons why in-group bias could arise. To differentiate between the first two channels (image concerns and taste-based discrimination), we look at how the decision rule (6) changes when judges observe a low or a high share of defendants from the in-group, i.e., when $\frac{n_{X_i}}{n_{-X_i}} < 1$ and $\frac{n_{X_i}}{n_{-X_i}} \geq 1$, respectively.

Low Share of In-Group Defendants. We start by analyzing the scenario with a low share of defendants from the judge's in-group. As the number of in-group defendants goes to zero, group-image concerns become less relevant, i.e., $\frac{n_{X_i}}{n_{-X_i}} \to 0$, as the deviations from members of the in-group are not salient. Therefore, the decision rule becomes

$$\tau\left(\frac{\mu(X_i)}{1+e_j\rho} + \frac{e_j\rho}{1+e_j\rho}\tilde{r}_{ij}\right) \geq \phi + \bar{C}_j \tag{7}$$

leading to the following proposition.

Proposition 1. *Suppose that the judge observes a sufficiently low number of same-group defendants. Then, for the same level of observed recidivism and judge-specific cost, in-group bias in incarceration decisions against same-group defendants arises if and only if judges have preferences for incarcerating same-group defendants – $\phi < 0$.*

High Share of In-Group Defendants. Next, we move to the scenario where the judge observes that the majority of defendants are from her in-group. In this scenario, judges have stronger preferences for punishing members of their in-group who deviate, as they hurt the group's reputation. In other words, as the number of in-group defendants increases, $\frac{n_{X_i}}{n_{-X_i}} > 1$, group-image concerns become prevalent, and judges' decision rule is the same as in (6), leading to the following proposition.

Proposition 2. *Suppose that most defendants a judge observes are from her in-group. Then, for the same level of observed recidivism and judge-specific cost, in-group bias in incarceration*

decisions against same-group defendants arises if and only if at least one of the following happens:

1. *Judges have preferences for incarcerating same-group defendants:* $\phi < 0$*;*

2. *Judges have group-image concerns:* $\gamma > 0$*.*

5.2 In-Group Bias and Recidivism Risk

Now, we look at how the judge's decision rule changes for different levels of observed recidivism risk, i.e., for different levels of the signal. A change in the signal affects only the left-hand side of the decision rule (LHS_{ij}). Therefore, we calculate the derivative of the left-hand side with respect to the signal

$$\frac{\partial LHS_{ij}}{\partial \tilde{r}_{ij}} = \frac{\rho_{ij}}{1+\rho_{ij}} \geq 0. \tag{8}$$

The first-order derivative is a function of the signal's precision and is always larger than zero, leading to the following prediction.

Proposition 3. *On average, defendants with higher observed recidivism risk are more likely to be incarcerated.*

Additionally, we calculate the cross-derivative of the left-hand side with respect to the signal and judges' experience

$$\frac{\partial^2 LHS_{ij}}{\partial \tilde{r}_{ij} \partial e_j} = \frac{\rho_{ij}/e_j}{(1+\rho_{ij})^2} \geq 0 \tag{9}$$

As for the first-order derivative, the cross derivative is also always larger than zero, implying the following.

Proposition 4. *More experienced judges are more likely to incarcerate higher-risk defendants.*

Next, we are interested in comparing the change in incarceration rates when defendants face judges from their in-group versus those from the out-group. Given the definition of

the signal precision in (3), the derivative (8) changes depending on the group of the judge handling the case. Specifically,

$$\frac{\partial LHS_{ij}}{\partial \tilde{r}_{ij}} = \begin{cases} \frac{e_j \rho}{1+e_j \rho} & \text{if } X_i = W_j \\ \frac{e_j \rho s_{X_i}}{1+e_j \rho s_{X_i}} & \text{if } X_i \neq W_j. \end{cases} \tag{10}$$

Comparing the two derivatives in (10), we see that for $s_{X_i} < 1$, judges from the same group as the defendant are more responsive to changes in recidivism risk than judges from the out-group, since they receive a more precise signal about it.[13] Moreover, the difference in responsiveness between in-group and out-group judges increases when the share of in-group defendants decreases.

Note that this reasoning implies that in-group bias can occur due to information differences. When defendants from a minority group have a higher recidivism risk, and judges from that group get a more precise signal on the risk of that group, we might observe harsher treatment of the in-group. In that case, in-group bias arises due to statistical discrimination, rather than (or in addition to) taste-based discrimination or image concerns.

We can summarize these points as follows.

Proposition 5. *Suppose that we allow the observed level of recidivism risk to vary. Then, on average, Judges are more likely to incarcerate defendants from the same group with higher recidivism risk. Moreover, the lower the share of in-group defendants observed by the judge, the greater the in-group bias for higher-risk defendants.*

6 Empirical Analysis of Judge Bias Mechanisms

This section takes the insights from Section 5's model to the data. In particular, we look at how the results vary with predicted recidivism risk. First, we add recidivism risk among

[13] To see this, we show that the difference between the derivative when the judge is from the in-group versus when she is from the out-group is always larger than 0:

$$\frac{e_j \rho}{1+e_j \rho} - \frac{e_j \rho s_{X_i}}{1+e_j \rho s_{X_i}} = \frac{\rho(1-s_{X_i})}{(1+e_j \rho)(1+e_j \rho s_{X_i})} \geq 0 \tag{11}$$

Table 4: In-Group Bias and Recidivism Risk

	Defendant is Incarcerated		
	(1)	(2)	(3)
Exp. Judge	0.0179**	0.0174*	0.0125
	(0.00388)	(0.00688)	(0.00805)
Recid. Score	0.0340**	0.0346**	0.0186**
	(0.00148)	(0.00134)	(0.00264)
Exp. Judge × Recid. Score	0.00701**	0.00641**	0.00539*
	(0.00214)	(0.00206)	(0.00235)
Obs.	894311	894291	883893
R^2	0.102	0.115	0.137
County-Year FE	X	X	X
Charge Severity FE	X	X	X
Judge FE		X	X
Other Judge/Def Characteristics			X
Other Judge/Def. Characteristics-Risk Score Interactions			X

Notes: Estimated effect of judges' experience and recidivism risk on jail decisions. The recidivism risk score is standardized to have a mean of 0 and a standard deviation of 1. All specifications include county×year fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Additional interactions and controls include defendants' characteristics other than race and the interaction between these, the recidivism risk score, and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the court and at the year level (in parenthesis): + p < 0.1, * p < 0.05, ** p < 0.01.

the regressors, including interactions with judge experience, and judge and defendant race. Then we go further into the mechanisms for in-group bias by testing some key predictions from the model.

The estimating approach is the same as above, where we run regressions similar to Eq. (2). The main difference is that we will now add additional interactions between the recidivism risk score and judge/defendant characteristics. For this purpose, let $RecScore_{it}$ be the risk score (predicted probability of recidivism), residualized by judge-year, and then standardized to variance 1. Because risk is now included as a regressor, we will exclude the previously used controls for risk ventiles.

Judge Experience and Responses to Recidivism Risk. First, we analyze the effect of judge's experience, interacted with defendant recidivism risk. Table 4 shows the estimates from this specification, where we include an indicator for the judge having more than

the median experience in her court-year. In line with Proposition 3, the coefficient on recidivism risk score is positive and significant across all specifications. Defendants with 1 standard deviation higher observed recidivism risk are 1.9 percentage points more likely to be incarcerated.

Second, we assess the effect of experience on the responsiveness to recidivism risk. As articulated in Proposition 4, if more experienced judges obtain a more precise signal on defendant riskiness, they should be more responsive in their jailing decisions to our ML measure of recidivism probability. As shown in all Columns of Table 4, this prediction is borne out in the data. More experienced judges have a steeper slope for the relationship between recidivism risk and jailing frequency.

Interaction of Judge/Defendant Race and Recidivism Risk. Now we examine the interaction effects among judge/defendant race and recidivism risk. First, Table 5 Columns 1 and 2 provide estimates in the whole sample. These estimates are similar to those from Section 4.3, as the specification is the same except in how $RecScore_{it}$ is included (now as interacted regressor, rather than ventile FE). As before, the coefficient on the interaction between Black judges and Black defendants is significant when including judge fixed effects, where we see that Black judges are around 5.5 percentage points more likely to incarcerate same-race defendants.

Rows 4 through 6 provide the estimates for the interactions of judge/defendant race and recidivism risk. Row 4 shows there is no interaction with Black judge. In the fifth row, we see that judges are more responsive in their jail decisions to recidivism risk for Black defendants. In the sixth row, the triple interaction between the indicators for judges' and defendants' race and recidivism score is positive but not significant when judge fixed effects are included (Column 2).

Heterogeneity by Defendant Group Population Share. In the final part of this analysis, we explore the theoretical framework's predictions on the share of defendants from the minority group observed by judges. We start by splitting the sample into two parts, depending on the share of Black defendants seen by the judge. Specifically, we split the sample

between judges who, during their careers, handle cases involving mostly Black defendants – more than 50% of cases – and those who handle less than 50% of cases involving Black defendants.[14]

Propositions 1 and 2 provide insights on how to interpret heterogeneity according to the defendant group population shares. In the presence of taste-based anti-in-group bias against same-race defendants ($\phi < 0$), we would always observe a positive coefficient on the defendant-judge same-race interaction term, regardless of the number of in-group defendants observed by the judge. In the absence of taste-based anti-in-group bias ($\phi = 0$), but in the presence of image concerns ($\gamma > 0$), we would observe a positive coefficient only when there is a high Black-defendant population share, and not when there is a low share.

Table 5 Columns 3 and 4 report the results for the high-Black-share and low-Black-share samples, respectively. Note first that we do not see anti-in-group bias in Column 4 (low Black share). That goes against taste-based in-group bias as a central mechanism ($\phi = 0$). Instead, anti-in-group bias is concentrated when there is a high Black defendant share (Column 3). That is suggestive of group-image concerns ($\gamma > 0$), as articulated in Proposition 2. Namely, we observe higher incarceration rates of Black defendants by Black judges when they observe a high share of Black defendants due to group-image concerns.

Finally, we evaluate heterogeneity in the response to recidivism risk based on Proposition 5. That proposition says that judges should be more responsive to recidivism risk for same-race defendants, and that this higher responsiveness should be larger when the population share of same-race defendants is lower. We assess that empirically with the triple interaction of Black judge, Black defendant, and recidivism risk score, reported in the sixth and last row of coefficients in Table 5. Proposition 5 suggests that the term should be positive.

In Columns 1 through 3, there is not much of an effect of the triple interaction on the tendency of judges to assign jail time. However, the effect is positive and statistically significant for judges who see a low share of Black defendants (Column 4). That is consistent with the second part of Proposition 5: when there are relatively few Black defendants, White judges get a less precise signal than Black judges on Black defendants' riskiness. In that

[14] Appendix Tables D.9 and D.10 show balance checks on this sample split. In the sample with a high share of Black defendants, White judges handle cases with more Black defendants and cases where the defendants live in a zip code with a higher share of Black individuals.

Table 5: In-Group Disparities and Recidivism Risk for High/Low Share of Black Defendants

	Defendant is Incarcerated			
	All	All	High Black Shr.	Low Black Shr.
	(1)	(2)	(3)	(4)
Black Defendant	0.0420**	0.0405**	0.0341**	0.0452**
	(0.00491)	(0.00310)	(0.00646)	(0.00316)
Black Judge × Black Defendant	0.0126	0.0553**	0.0608**	0.00270
	(0.0193)	(0.0116)	(0.0132)	(0.0184)
Recid. Score	0.0213**	0.0226**	0.0158	0.0239**
	(0.00279)	(0.00253)	(0.0113)	(0.00245)
Black Judge × Recid. Score	0.0169	-0.0259	-0.0460+	-0.00912
	(0.0170)	(0.0164)	(0.0263)	(0.0218)
Black Defendant × Recid. Score	0.0193**	0.0252**	0.0340**	0.0217**
	(0.00394)	(0.00272)	(0.00607)	(0.00260)
Black Judge × Black Defendant × Recid. Score	0.0236+	0.0109	0.0103	0.0442**
	(0.0135)	(0.0108)	(0.0156)	(0.0115)
Obs.	883913	883893	122425	761461
R^2	0.124	0.137	0.144	0.120
County-Year FE	X	X	X	X
Charge Severity FE	X	X	X	X
Judge FE		X	X	X
Other Judge/Def Characteristics		X	X	X
Other Judge/Def. Characteristics-Risk Score Interactions		X	X	X

Notes: Estimates for in-group bias in jail decisions by judges, interacted with recidivism risk. Columns 1 and 2 show the results for the whole sample. Columns 3 and 4, respectively, show the results separately for the samples of judges seeing a share of Black defendants higher or lower than 0.5. Recidivism risk score is standardized to have a mean of 0 and a standard deviation of 1. All specifications include county×year fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Additional interactions and controls include defendants' characteristics other than race and the interaction between these, the recidivism risk score, and additional judges' characteristics. Standard errors are clustered at the court and at the year level (in parenthesis): + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

setting, the White judges are less responsive to the risk score. When there are relatively many Black defendants, White judges observe many cases and become familiar enough to form an accurate posterior about Black defendants. In that setting, White judges are equally responsive to recidivism risk among Black defendants as Black judges.

7 Robustness Checks and Alternative Explanations

This section considers some alternative explanations that could be driving our results and provide evidence that our results are robust to account for these. Moreover, we conduct additional robustness tests and show that our results are robust to a series of alternative specifications.

Alternative Explanations. We start by ruling out alternative explanations to our information story. A first question is whether the observed in-group disparities across risk scores are due to the different political affiliations of judges. For this purpose, we run the heterogeneity analysis from above, replacing the indicator for African American judges with indicators for judges being more conservative, as measured by campaign donations. In Appendix Table E.12 and Appendix Figure E.9, we see no evidence of bias by conservative judges toward African American defendants across risk deciles. Thus, we can rule out that judge race is proxying for conservative ideology.

Another mechanism could be that cases with Black defendants are litigated by different types of prosecutors. We are able to test for this using the names of the prosecutors. In Appendix Table E.15, we run regressions with prosecutor fixed effects. We find no changes in the main results.

Moreover, it may be the case that judges' decisions are informed by the use of COMPAS, a commercial risk scoring algorithm that is sometimes used in Wisconsin courts. If judges indeed observe the suggestion by COMPAS, it may be that they are just (selectively) responding to some additional information rather than inferring this additional information by other characteristics, such as race. Wisconsin courts started to use COMPAS in 2012. Therefore, we re-run our main analysis, keeping only cases up until 2011. We report the

results in Appendix Table E.16 and find no differences in the main effects.

Finally, judges' decisions may be driven by additional information related to the socioeconomic status of defendants, for which observable defendants' characteristics are a proxy. We check this mechanism by including zip code fixed effects in the regression. Results are reported in Appendix Table E.17, showing that our main results are robust to the inclusion of zip code effects. That goes against defendants' socioeconomic characteristics being an important driver for the results.

Robustness Checks. In addition to excluding alternative explanations, we provide evidence that the findings are robust across a number of additional specifications. First, in Appendix Table E.18, we show that our results are robust to including county $\times$ year $\times$ charge-severity fixed effects. We also show that the results are robust when we use a logistic regression model instead of XGBoost to train the machine learning model to predict recidivism risk (Appendix Table E.19). Moreover, in Appendix Table E.20, we show that the results separately for the samples with high and low shares of Black defendants are robust to different ways of calculating the share. Specifically, we show results where the share of Black defendants is calculated within each county and year (Columns 1 and 2) and for each judge within each county and year (Columns 3 and 4).

Next, we show that our results are very similar if we consider only county-years with at least one Black defendant (Appendix Table E.21). Additionally, given the relatively few Black judges in our sample, it is possible that the estimates we obtain are downward or upward biased depending on whether it is Black or White judges who are driving the bias. For this reason, in Appendix Table E.22, we replicate the analysis keeping only county-years with at least one Black judge. The results are consistent with our main findings and even larger in magnitude, suggesting that our previous estimates are likely to be downward biased due to the low number of Black judges.

Additionally, we account for selective labeling bias in the recidivism risk score. The potential problem is that, since we observe recidivism episodes only for defendants whose sentence is lower than 24 months, the training data exclude some high-risk defendants who tend to be incarcerated. Therefore, it could be that the model is biased by the selected

training sample. To mitigate this problem, we adapt the approach from Lakkaraju et al. (2017) and re-train our machine learning model only on the data from the most lenient judges, namely, judges who release the most defendants. In particular, we perform two different exercises. First, we define judges as lenient if they are in the first tercile of the distribution of the share of released defendants (Appendix Table E.23). Second, we only include the two judges with the highest share of released defendants in each court and year (Appendix Table E.24). We produce results using the adjusted recidivism risk score and show that our results are once again robust, suggesting that selective labeling biases are not driving our results.

Finally, we run the same analysis restricting the sample to only first-time offenders (Appendix Table E.25). As mentioned in Section 2, repeat offenders within the same county are often assigned to the same judge if the judge is still in office. Therefore, there may be some violation of random assignment of judges when including repeat offenders. The results are robust to this restricted sample, even if the coefficient on in-group bias in Columns 1, 2, and 3 is smaller in magnitude. As an alternative robustness check, we include a control for whether the case involves a first-time offender as well as its interactions with judges' characteristics and the risk score. As shown in Appendix Table E.26, the results are robust to the inclusion of these variables.

8 Conclusion

This paper provides new evidence and analysis of bias and uncertainty in judge decisions about sentencing. The Circuit Courts of Wisconsin are a desirable setting, given the rich dataset on criminal cases, 2005-2017. Our empirical strategy, grounded in the procedural rules of case assignment within the courts, allows us to investigate the presence of in-group bias in sentencing decisions while accounting for potential confounders like judge preferences and case characteristics. We further explored the mechanisms behind these disparities through a model of judicial decision-making, which suggests that judges' in-group bias and the perceived recidivism risk of defendants play critical roles in sentencing decisions. The empirical evidence supports the presence of in-group bias among judges, particularly when

the share of in-group defendants is high, pointing towards group-image concerns as a driving factor. Additionally, our findings suggest that judges' responsiveness to recidivism risk varies with the racial composition of the defendants they handle, indicating a nuanced relationship between judge characteristics, defendant characteristics, and sentencing outcomes.

This research contributes to the broader literature on racial bias in the criminal justice system. These results might help reconcile divergent findings in the previous literature. For example, the mixed evidence on bias discussed in Ash et al. (2022) might come from differences across jurisdictions in the experiences of judges, or how much identity features are correlated with recidivism risk.

This research informs policy choices aimed at mitigating these biases. For example, there could be increased transparency in sentencing decisions, training programs to raise awareness of unconscious biases among judges, and implementation of checks and balances to ensure fairness in the judicial process. Future research could further explore the implications of these findings for the broader criminal justice system and examine interventions to reduce racial disparities in sentencing.

References

Arnold, D., Dobbie, W., and Hull, P. (2022). Measuring Racial Discrimination in Bail Decisions. *American Economic Review*, 112(9):2992–3038.

Arnold, D., Dobbie, W., and Yang, C. S. (2018). Racial bias in bail decisions. *The Quarterly Journal of Economics*, 133(4):1885–1932.

Ash, E., Asher, S., Bhowmick, A., Bhupatiraju, S., Chen, D., Devi, T., Goessmann, C., Novosad, P., and Siddiqi, B. (2022). In-group bias in the Indian judiciary. page 59.

Ash, E., Goel, N., Li, N., Marangon, C., and Sun, P. (2023). WCLD: Curated Large Dataset of Criminal Cases from Wisconsin Circuit Courts.

Becker, G. (1957). *The Economics of Discrimination*. University of Chicago Press, 2 edition.

Berdejó, C. (2018). Criminalizing race: Racial disparities in plea-bargaining. *BCL Rev.*, 59:1187.

Berdejó, C. (2019). Gender disparities in plea bargaining. *Ind. LJ*, 94:1247.

Bonica, A. (2016). Database on Ideology, Money in Politics, and Elections (DIME).

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes*. *The Quarterly Journal of Economics*, 131(4):1753–1794.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. Association for Computing Machinery.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Depew, B., Eren, O., and Mocan, N. (2017). Judges, Juveniles, and In-Group Bias. *The Journal of Law and Economics*, 60(2):209–239.

Fagan, J. and Ash, E. (2017). New policing, new segregation: from ferguson to new york. *Geo. LJ Online*, 106:33.

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., and Hutter, F. (2018). Practical automated machine learning for the automl challenge 2018. In *International Workshop on Automatic Machine Learning at ICML*, pages 1189–1232.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? *arXiv preprint arXiv:2207.08815*.

Guo, F., Li, J., Ma, M., and Zha, D. (2023). Identity on the bench: Gender in-group bias in the judiciary. Available at SSRN: https://ssrn.com/abstract=4625401 or http://dx.doi.org/10.2139/ssrn.4625401.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hunt, K. S. and Dumville, R. (2016). *Recidivism Among Federal Offenders: A Comprehensive Overview*. United States Sentencing Commission. Google-Books-ID: 7l7MjwEACAAJ.

Jung, J., Corbett-Davies, S., Gaebler, J. D., Shroff, R., and Goel, S. (2024). Mitigating included-and omitted-variable bias in estimates of disparate impact. *arXiv preprint arXiv:1809.05651*.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining, KDD '17, pages 275–284, New York, NY, USA. Association for Computing Machinery.

Larson, J., Angwin, J., Kirchner, L., and Mattu, S. (2016). How We Analyzed the COMPAS Recidivism Algorithm.

Li, N., Goel, N., and Ash, E. (2022). Data-Centric Factors in Algorithmic Fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 396–410, Oxford United Kingdom. ACM.

Lim, C. S., Silveira, B. S., and Snyder, J. M. (2016). Do judge characteristics matter? ethnicity, gender, and partisanship in texas state trial courts. *American Law and Economics Review*, 18(2):302–357.

Marques, J. M. and Yzerbyt, V. Y. (1988). The black sheep effect: Judgmental extremity towards ingroup members in inter-and intra-group situations. *European Journal of Social Psychology*, 18(3):287–292. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.2420180308.

Nellis, A. (2021). The Color of Justice: Racial and Ethnic Disparity in State Prisons.

Phelps, E. S. (1972). The Statistical Theory of Racism and Sexism. *The American Economic Review*, 62(4):659–661. Publisher: American Economic Association.

Shayo, M. and Zussman, A. (2011). Judicial Ingroup Bias in the Shadow of Terrorism. *The Quarterly Journal of Economics*, 126(3):1447–1484.

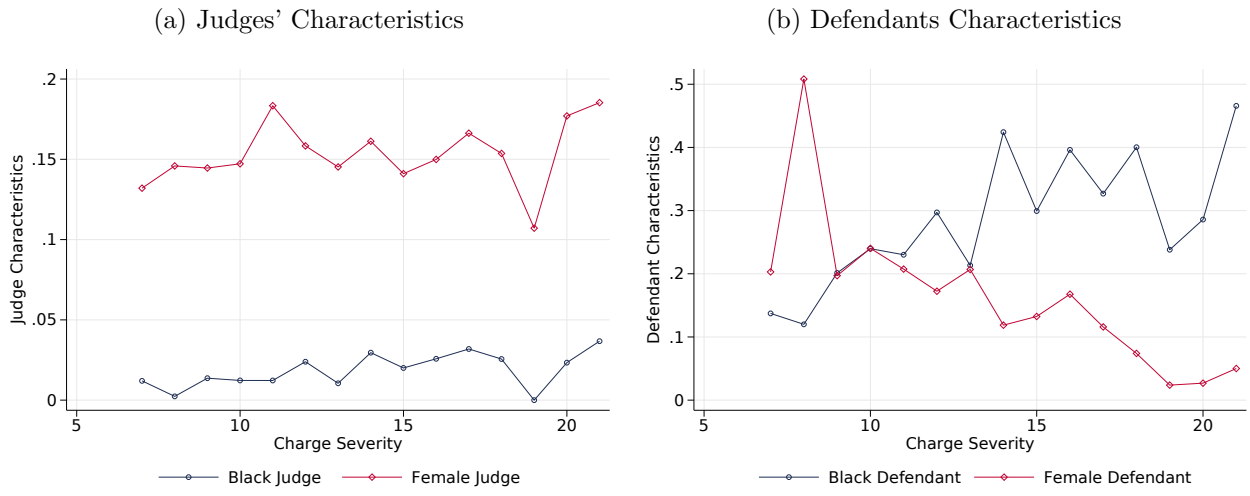
A Data Limitations

Following our agreement with the WCCA, we acknowledge the following data limitations:

- WCCA Information includes only court records open to public view under Wisconsin's Open Records Law, Wis. Stat. 19.31-19.39.
- WCCA Information does not comprise the complete court record. For instance, it does not include information that may be confidential, sealed, or redacted in accordance with all applicable statutes, court orders, and rules related to confidentiality, sealing, and redaction.
- WCCA Information is not the Judgment and Lien Docket under Wis. Stat. 806.10. The Judgment and Lien Docket is available from the Clerk of Circuit Court.
- In criminal cases, any designation in any race field contains subjective information generally provided by the agency that filed the case.

B Summary Statistics

Figure B.1: Judges/Defendants' Characteristics Across Charge Severity Risk



Notes: Average share of Female and Black judges/defendants plotted by charge severity. Panel (a) shows the average share of Female and Black judges, while Panel (b) shows that of Female and Black defendants.

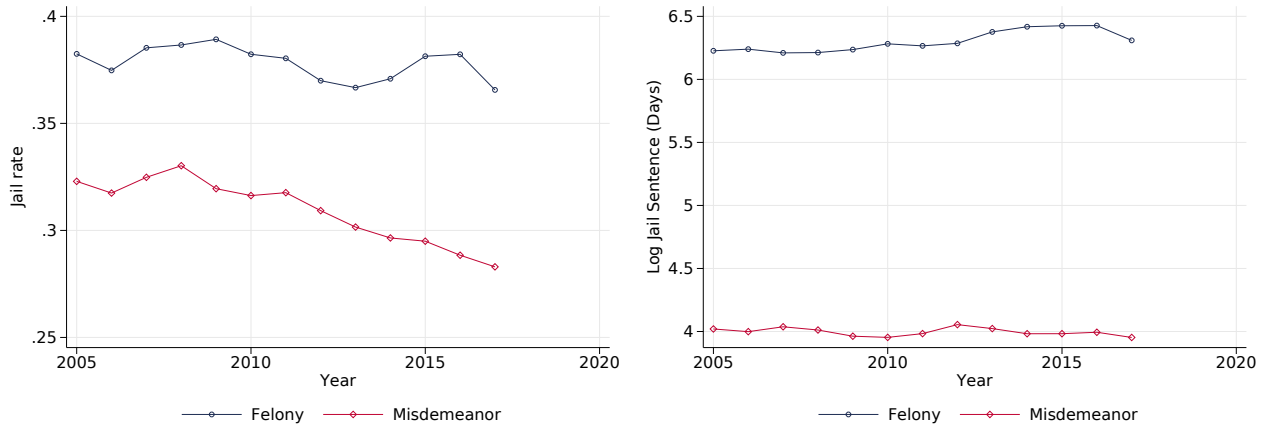
Table B.1: Summary Statistics on Judge Characteristics

	Mean		
	White Judges	Black Judges	Difference
	(1)	(2)	(3)
Female Judge	0.179	0.000	-0.178***
	(0.384)	(0.000)	(0.016)
Harshness (Terciles)	2.813	2.667	-0.147
	(0.483)	(0.651)	(0.189)
Experience	29.055	26.924	-2.193
	(8.142)	(8.309)	(2.424)
Political Contributions	-0.035	-0.522	-0.487
	(1.098)	(0.968)	(0.283)

Notes: Summary statistics of judges in the sample calculated using judges as units of observation. *Experience* indicates the years of judicial activity of judges, and *Political Contributions* indicates a political contributions score (where positive values indicate contribution by Republicans). We define leniency by looking at the number of incarcerated defendants by judges in each county and year. Lenient judges are those for whom the share of incarcerated defendants is in the first tercile of the distribution of incarceration rates, while moderate and harsh ones are those in the second and third terciles, respectively.

Figure B.2: Summary Statistics: Judge Decisions

(a) Incarceration Rate by Year and Charge Type (b) Log Jail Sentence (Days) by Year and Charge Type



Notes: Average summary statistics of judges' decisions plotted by year separately for misdemeanors and felonies. Panel (a) shows the incarceration rates by year, while Panel (b) shows the average sentence length (log days) by year.

Table B.2: Summary Statistics by Judge Harshness

	Lenient	Moderate	Harsh	Total
Female Judge	0.184	0.133	0.239	0.145
	(0.388)	(0.339)	(0.426)	(0.352)
Black Judge	0.0194	0.0128	0.0207	0.0142
	(0.138)	(0.112)	(0.142)	(0.118)
Experience	33.00	31.69	26.75	31.87
	(7.343)	(6.855)	(8.479)	(7.035)
Political Contributions	-0.0733	0.00694	-0.0311	-0.00967
	(1.011)	(1.049)	(0.954)	(1.040)
Black Defendant	0.380	0.172	0.271	0.215
	(0.485)	(0.377)	(0.444)	(0.411)
Female Defendant	0.182	0.211	0.215	0.205
	(0.385)	(0.408)	(0.411)	(0.404)
Charge Severity	11.32	9.458	8.992	9.821
	(3.092)	(2.461)	(2.490)	(2.705)
Defendant Age	30.96	31.45	31.66	31.36
	(11.09)	(11.05)	(11.05)	(11.06)

Notes: Summary statistics of judges' and defendants' characteristics. *Experience* indicates the years of judicial activity of judges, and *Political Contributions* indicates a political contributions score (where positive values indicate contribution by Republicans). We define leniency by looking at the number of incarcerated defendants by judges in each county and year. Lenient judges are those for whom the share of incarcerated defendants is in the first tercile of the distribution of incarceration rates, while moderate and harsh ones are those in the second and third terciles, respectively.

Table B.3: Summary Tabulations: Misdemeanors

	Freq.	Percent
Disorderly Conduct	68409	15.47
Battery	62636	14.17
Resisting Officer	60500	13.68
Bail Jumping	42566	9.628
Drug Possession	38238	8.649
Theft	28317	6.405
Retail Theft (Shoplifting)	25086	5.674
Criminal Damage	24451	5.531
Drug Paraphernalia	12313	2.785
Operating While Intoxicated	9980	2.257
Worthless Checks	9541	2.158
Other Misdemeanor	8966	2.028
OAR/OAS	8912	2.016
Weapons/Explosives	8786	1.987
Violation of TRO	4950	1.120
Crimes Against Children	4517	1.022
Sex Crimes	3902	0.883
Criminal Trespass	3286	0.743
Operate Without License	2924	0.661
Unidentified Misdemeanor	2798	0.633
Other Fraud	2363	0.534
Intimidate Witness/Victim	2112	0.478
Receiving Stolen Property	2073	0.469
Fourth Degree Sexual Assau	1609	0.364
Operate Vehicle Without Consent	924	0.209
Hit and Run	405	0.0916
Escape	363	0.0821

	Freq.	Percent
Entering Locked Vehicle	319	0.0722
Contempt of Court	276	0.0624
Other Public Safety Crimes	134	0.0303
Operating while intoxicated	65	0.0147
BAC	59	0.0133
Non-Traffic Forfeiture	42	0.00950
Other Crimes Against Children	42	0.00950
Forgery	33	0.00746
Reckless Driving	30	0.00679
Other Drug Offenses	29	0.00656
Other Felony	27	0.00611
Operate Vehicle w/out Consent	20	0.00452
Substantial/Aggravated Battery	19	0.00430
Public Assistance Fraud	15	0.00339
Arson	13	0.00294
Drug Manufacture/Deliver	13	0.00294
Burglary	10	0.00226
Unidentified Felony	7	0.00158
Local or Unidentified Forfeiture	6	0.00136
Gambling	5	0.00113
Child Abuse	3	0.000679
2nd Deg. Sex. Assault of Child	1	0.000226
Extradition	1	0.000226
Other Bodily Security	1	0.000226
Stalking	1	0.000226
Unidentified Traffic Forfeiture	1	0.000226
Total	442099	100

Table B.4: Summary Tabulations: Felonies

	Freq.	Percent
Drug Possession	58210	16.82
Bail Jumping	38973	11.26
Burglary	25254	7.296
Theft	23482	6.784
Drug Manufacture/Deliver	23007	6.647
Operating while intoxicated	21953	6.342
Other Felony	16672	4.817
Battery	11853	3.424
Forgery	10202	2.947
Substantial/Aggravated Battery	9714	2.806
Other Bodily Security	9252	2.673
Other Public Safety Crimes	9093	2.627
Other Crimes Against Children	7946	2.296
Weapons/Explosives	7493	2.165
Child Abuse	7233	2.090
Operate Vehicle w/out Consent	7004	2.024
Armed Robbery	5985	1.729
2nd Deg. Sex. Assault of Child	4814	1.391
Other Drug Offenses	4142	1.197
Intimidate Witness/Victim	3789	1.095
Escape	3648	1.054
1st Deg. Sex. Assault of Child	3644	1.053
Unarmed Robbery	3552	1.026
Sexual Assault	3522	1.018
Kidnap/Hostage/False Imprisonment	3203	0.925
Unidentified Felony	3028	0.875
Disorderly Conduct	1746	0.504
Resisting Officer	1740	0.503
Other Fraud	1544	0.446
Criminal Damage	1336	0.386
Stalking	1240	0.358
First Degree Intentional Homicide	1158	0.335
Hit and Run	1132	0.327
Other Homicide	1121	0.324
Other Misdemeanor	1092	0.315
First Degree Reckless Homicide	901	0.260
Arson	785	0.227
Receiving Stolen Property	735	0.212
Operating While Intoxicated	721	0.208
Fourth Degree Sexual Assau	650	0.188
Worthless Checks	603	0.174
Retail Theft (Shoplifting)	596	0.172
BAC	476	0.138
Crimes Against Children	462	0.133
Operate Vehicle Without Consent	398	0.115
Drug Paraphernalia	206	0.0595
Perjury	206	0.0595
Criminal Trespass	163	0.0471
Public Assistance Fraud	98	0.0283
OAR/OAS	76	0.0220
Sex Crimes	73	0.0211
Violation of TRO	59	0.0170
Unidentified Misdemeanor	53	0.0153
Reckless Driving	33	0.00953
Contempt of Court	20	0.00578
Operate Without License	16	0.00462
Gambling	13	0.00376
Entering Locked Vehicle	5	0.00144
Unidentified Felony Traffic	4	0.00116
Extradition	3	0.000867
Total	346132	100

Table B.5: Summary Tabulations: Criminal Traffic

	Freq.	Percent
Operating While Intoxicated	109771	45.56
OAR/OAS	92189	38.26
Operate Without License	24148	10.02
Hit and Run	4816	1.999
Bail Jumping	2660	1.104
Unidentified Misdemeanor Traffic	2041	0.847
Other Misdemeanor	1828	0.759
Resisting Officer	1118	0.464
Drug Possession	596	0.247
Disorderly Conduct	534	0.222
BAC	412	0.171
Drug Paraphernalia	297	0.123
Reckless Driving	177	0.0735
Operating while intoxicated	144	0.0598
Weapons/Explosives	79	0.0328
Criminal Damage	46	0.0191
Battery	22	0.00913
Operate Vehicle Without Consent	21	0.00872
Theft	10	0.00415
Retail Theft (Shoplifting)	9	0.00374
Crimes Against Children	7	0.00291
Criminal Trespass	6	0.00249
Receiving Stolen Property	6	0.00249
Violation of TRO	3	0.00125
Other Felony	2	0.000830
Contempt of Court	1	0.000415
Escape	1	0.000415
Intimidate Witness/Victim	1	0.000415
Other Fraud	1	0.000415
Sex Crimes	1	0.000415
Unidentified Felony Traffic	1	0.000415
Worthless Checks	1	0.000415
Total	240949	100

Table B.6: Summary Tabulations: Charge Severity

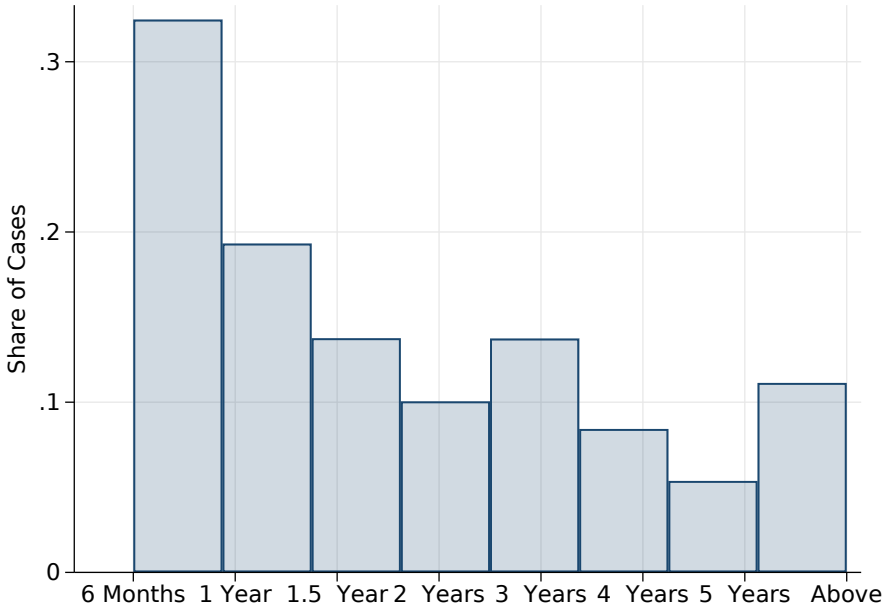
	Freq.	Percent
7	347046	33.72
8	425	0.0413
9	99657	9.682
10	322108	31.29
11	2203	0.214
12	63918	6.210
13	100638	9.777
14	25855	2.512
15	32731	3.180
16	12112	1.177
17	5461	0.531
18	13353	1.297
19	84	0.00816
20	3124	0.304
21	599	0.0582
Total	1029314	100

Table B.7: Charge Severity: Felony and Misdemeanor Classifications

Misdemeanor	Felony
Misdemeanor A	Felony A
Misdemeanor B	Felony B
Misdemeanor C	Felony BC
Misdemeanor U	Felony C
	Felony D
	Felony E
	Felony F
	Felony G
	Felony H
	Felony I
	Felony U

C Recidivism Risk Prediction

Figure C.3: Share of Defendants/Cases by Intervals between Offenses



Notes: Share of defendant/cases with recidivism episodes by time intervals between current and subsequent offense.

C.1 Target

The target of the prediction model is a binary variable taking value 1 if the defendant commits another crime within two years from the previous one and 0 otherwise (See Figure C.3). One issue we had to deal with when computing this variable was how to decide which cases to include, because whether we observe recidivism or not is affected by the decisions of judges. This is is not trivial because defendants serve different sentence lengths. To see this, consider three defendants who stay in jail for one year, for 22 months, and for 2 years after judgment. The first defendant has 1 year left to re-offend in the follow-up period, the second has only 2 months, and for the third, we can not observe recidivism in the follow-up period of 2 years at all. Yet extending the follow-up period for two years after the assigned sentence period instead of the judgment date is also problematic because defendants often serve more or less than the assigned sentence. Since there is not a comparable data source that has the exact jail record of every defendant in Wisconsin, we don't observe the actual

length of the sentence served. Moreover, the sentence itself could affect the probability of recidivism. Further, the defendants who receive sentences are a selected group, so there is the issue of selective labeling (Lakkaraju et al., 2017).

There is no consensus in the literature about how to deal with this problem. We use a cutoff for sentence length of 180 days, such that we don't throw away a lot of useful data and still leave enough time in the follow-up period for the defendant to reveal crime potential. We drop observations above this sentence length cutoff.

C.2 XGBoost Implementation and Performance

The XGBoost algorithm is an ensemble method that produces forests of decision trees, a state-of-the-art binary classification model for tabular datasets is gradient boosted trees (Friedman, 2001; Hastie et al., 2009).[15] Gradient boosting models consist of an ensemble of decision trees that "vote" on the predicted outcome. Each decision tree iteratively selects informative variables (in our case, e.g., number of previous arrests), splits on a value of that variable (e.g., $x > 2$) to better predict the outcome, branches off for additional splitting, and so on, until reaching a terminal node and an associated prediction ($\hat{Y} = 0$ or $\hat{Y} = 1$). With gradient boosting, additional layers of trees are gradually added during the training process to fit residuals and fix errors in the initial layers. A number of hyperparameters, such as the number of trees, and their depth, can be selected to calibrate the level of regularization.

A popular implementation of gradient boosting is XGBoost ("eXtreme Gradient Boosting"; Chen and Guestrin 2016). Besides being optimized for fast training, XGBoost has a number of computational adjustments to improve out-of-sample fit. Feurer et al. (2018) systematically compared XGBoost to many other classifiers, including a sophisticated automated ML system, and found that XGBoost consistently performed best on classification tasks with tabular data (see also Grinsztajn et al., 2022). Hence, we take XGBoost as our preferred model for predicting recidivism. In a robustness check, we replicate our results using a Logit model to predict the recidivism risk score, and show that the results are robust to this change in prediction model.

[15] For non-tabular data, such as images, text, or audio, neural nets are often preferred (Goodfellow et al., 2016).

C.3 Predicting Recidivism Risk

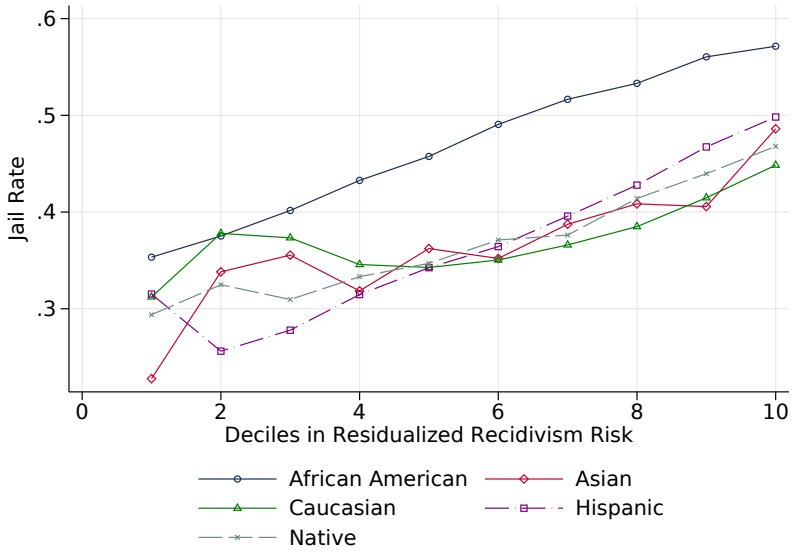
We start by training and fine-tuning the model to predict the target variable as described above on 70% of the whole sample, using gender, type of offense, prior criminal count (separately by type), and age at judgment and first offense as features. We select the hyperparameters using grid search and 5-fold cross validation.

After obtaining the best-performing model we compute the recidivism risk score by forming cross-predictions. First, we split the sample in two at the defendant level, such that the same defendant does not appear in both the training and the test set. Then, we train a new model on each of the two samples and use the model to predict recidivism risk on the other sample.

Table C.8: Performance of Classifier

Metric	Caucasian	African American
Accuracy	0.6648	0.6459
AUC	0.7044	0.7033
FPR	0.2159	0.2454
FNR	0.5113	0.4792

Figure C.4: Incarceration Rate by Recidivism Risk and Race

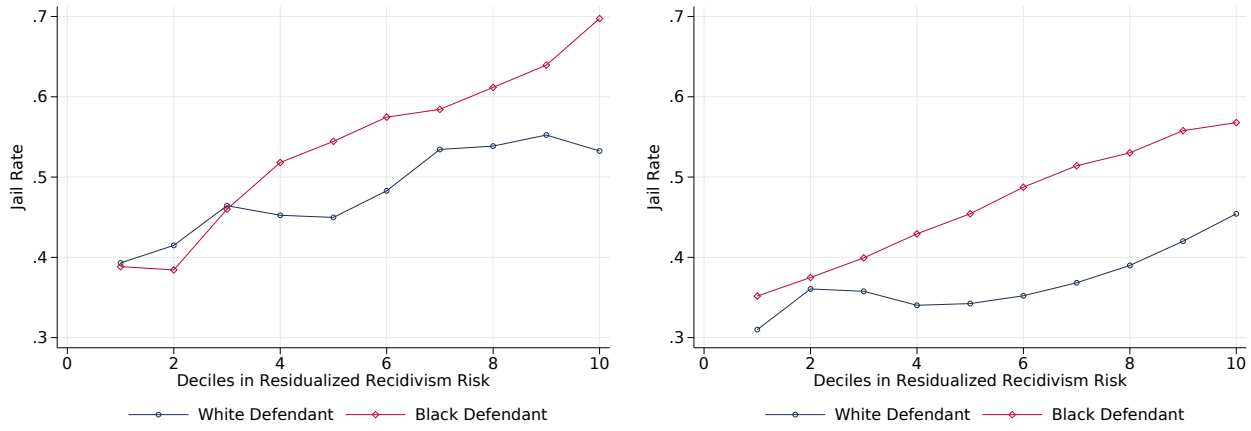


Notes: Average share of incarcerated defendants plotted by recidivism risk deciles and separately by ethnicity. Recidivism risk deciles are computed across all ethnicities and are centered on court and year and charge severity.

Figure C.5: Incarceration Rate by Recidivism Risk and Race

(a) Black Judges

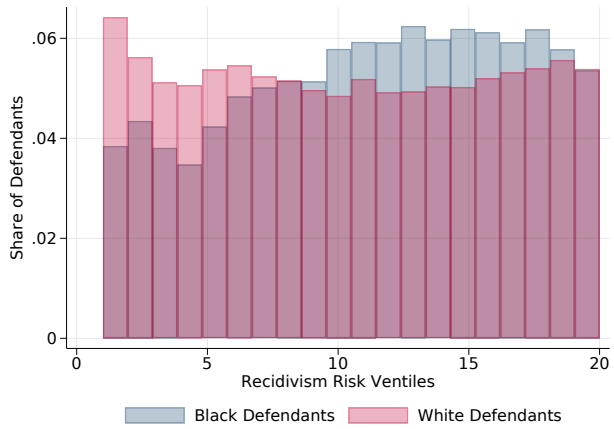
(b) White Judges



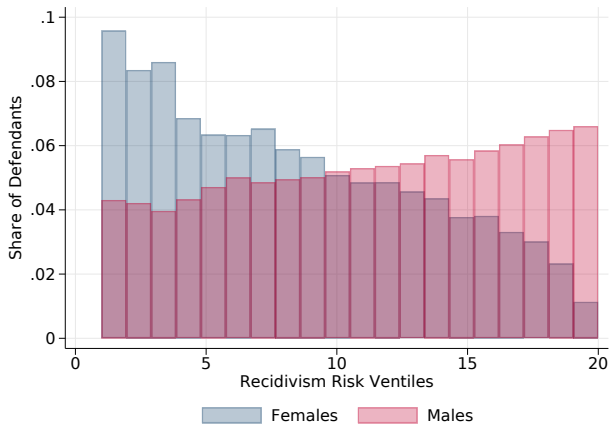
Notes: Average share of incarcerated defendants plotted by recidivism risk deciles for Black and White defendants separately. Recidivism risk deciles are computed across all ethnicities and are centered by court-year and charge severity. Panel (a) shows incarceration rates by recidivism risk decile for Black judges, while Panel (b) for White judges.

Figure C.6: Distribution of Defendants by Recidivism Risk

(a) Defendants by Race



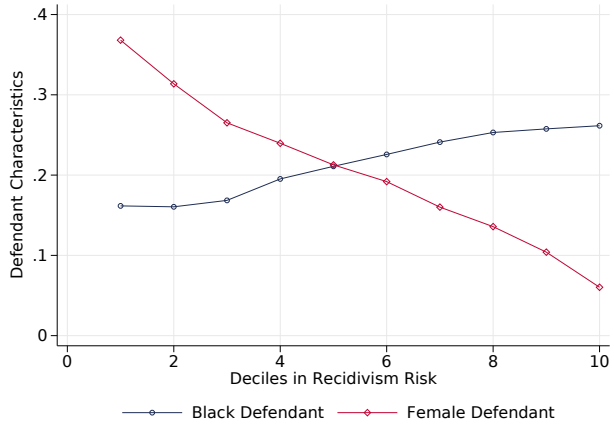
(b) Defendants by Gender



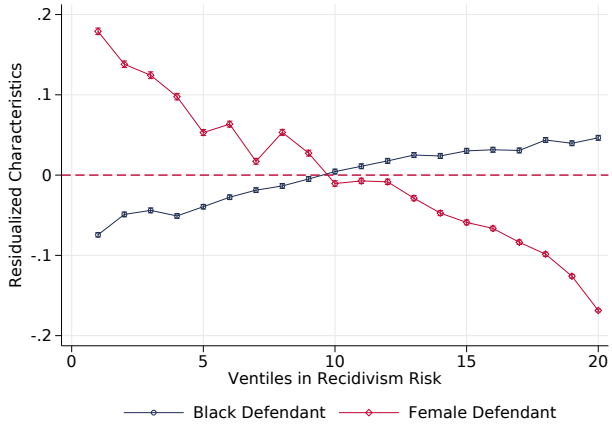
Notes: Share of defendants plotted by recidivism risk ventiles. Panel (a) shows the distribution of defendants separately by black and white defendants, while Panel (b) shows the distribution of defendants separately by genders.

Figure C.7: Defendant Characteristics Across Recidivism Risk

(a) Raw Characteristics



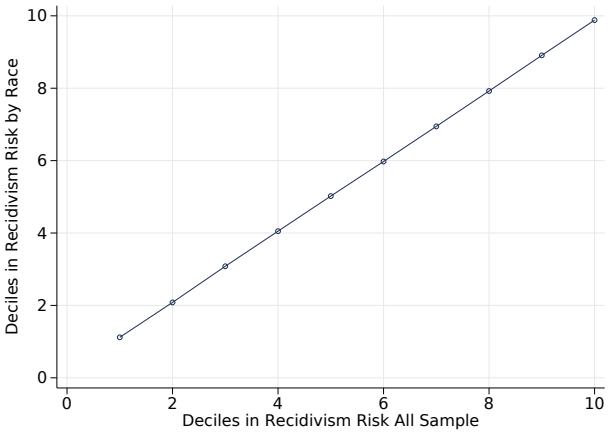
(b) Residualized Characteristics



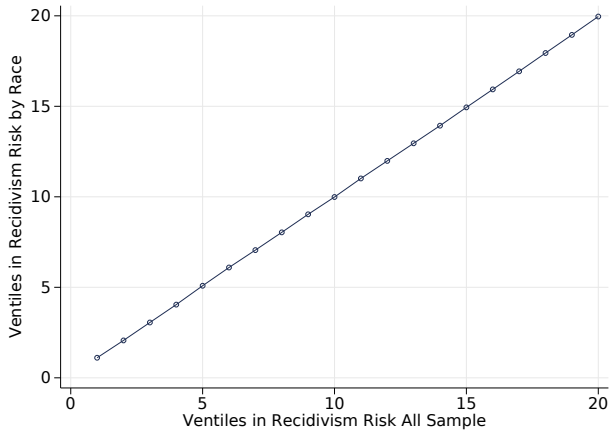
Notes: Panel (a) shows the average share of Female and Black defendants plotted by recidivism risk deciles. Recidivism risk deciles are computed across all ethnicities and are centered by court-year and charge severity. Panel (b) shows the residuals from regressions of defendants' characteristics on court-year and charge severity fixed effects plotted by recidivism risk ventiles, with 95% confidence intervals.

Figure C.8: Recidivism Risk Deciles are Identical Across Race Categories

(a) Recidivism Risk Deciles



(b) Recidivism Risk Ventiles



Notes: Average recidivism risk percentiles computed separately for each ethnicity plotted against average recidivism risk percentiles computed across all ethnicities. Panel (a) reports deciles, centered by court-year and charge severity, while Panel (b) reports ventiles.

D Additional Checks for Random Assignment of Judges

Table D.9: Randomization Check for Judge Assignment – High Share of Black Def.

	Mean		Difference in Means	
	White Judges	Black Judges	Without FE	With FE
	(1)	(2)	(3)	(4)
Charge Severity	10.745	10.943	-0.198**	-0.333
	(3.140)	(3.133)	(0.034)	(0.292)
Recid. Risk	0.391	0.378	0.013**	0.0136**
	(0.151)	(0.150)	(0.002)	(0.00295)
Black Defendant	0.679	0.656	0.023**	0.0335*
	(0.467)	(0.475)	(0.005)	(0.0158)
Female Defendant	0.157	0.156	0.000	-0.00510
	(0.363)	(0.363)	(0.004)	(0.0136)
Defendant Age	30.915	30.387	0.528**	0.288
	(11.126)	(10.822)	(0.121)	(0.469)
Prior Offense	0.728	0.696	0.033**	0.0313*
	(0.445)	(0.460)	(0.005)	(0.0142)
Misdemeanor	0.362	0.370	-0.008	-0.0186
	(0.481)	(0.483)	(0.005)	(0.0239)
Felony	0.472	0.551	-0.079**	-0.0526+
	(0.499)	(0.497)	(0.005)	(0.0304)
Criminal Traffic	0.166	0.079	0.087**	0.0713**
	(0.372)	(0.270)	(0.004)	(0.0117)
Zip Shr. Black	0.442	0.431	0.011**	0.0212**
	(0.314)	(0.302)	(0.003)	(0.00779)
Zip Shr. Male	0.478	0.479	-0.001+	-0.00104*
	(0.028)	(0.026)	(0.000)	(0.000415)
Zip Shr. Urban	0.970	0.977	-0.008**	-0.00280+
	(0.146)	(0.120)	(0.002)	(0.00149)
Zip Shr. College	0.198	0.199	-0.000	-0.00216
	(0.101)	(0.100)	(0.001)	(0.00165)
Zip Shr. Food Stamps	0.227	0.227	-0.000	0.00441+
	(0.107)	(0.102)	(0.001)	(0.00261)
Zip Median Income	10.504	10.504	0.001	-0.011
	(0.321)	(0.307)	(0.004)	(0.008)

Notes: Balance test of defendants' characteristics in the sample with only White and Black defendants when the share of Black defendants is larger than 0.5 (N=132,952). Recid. Risk is the predicted probability of recidivism. Columns 1 and 2 report the mean and standard deviation of defendants' characteristics separately for White and Black judges. Column 3 reports the simple difference in means between White and Black judges. Column 4 reports the difference in means after taking out county-year and charge severity fixed effects. Standard errors in Column 4 are clustered at the county×year level (in parenthesis): + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table D.10: Randomization Check for Judge Assignment – Low Share of Black Def.

	Mean		Difference in Means	
	White Judges	Black Judges	Without FE	With FE
	(1)	(2)	(3)	(4)
Charge Severity	9.769	9.345	0.424**	0.0606
	(2.612)	(2.454)	(0.040)	(0.0573)
Recid. Risk	0.432	0.467	-0.035**	-0.00986**
	(0.174)	(0.171)	(0.003)	(0.00367)
Black Defendant	0.176	0.417	-0.241**	0.00383
	(0.381)	(0.493)	(0.006)	(0.00837)
Female Defendant	0.215	0.198	0.018**	0.00800+
	(0.411)	(0.398)	(0.006)	(0.00482)
Defendant Age	31.720	31.268	0.452**	0.151
	(11.255)	(10.762)	(0.173)	(0.154)
Prior Offense	0.778	0.811	-0.032**	-0.00779
	(0.415)	(0.392)	(0.006)	(0.00947)
Misdemeanor	0.448	0.413	0.035**	-0.0153+
	(0.497)	(0.492)	(0.008)	(0.00792)
Felony	0.328	0.249	0.079**	0.00415**
	(0.470)	(0.433)	(0.007)	(0.00106)
Criminal Traffic	0.223	0.338	-0.114**	0.0112
	(0.416)	(0.473)	(0.006)	(0.00843)
Zip Shr. Black	0.054	0.076	-0.021**	0.000669
	(0.126)	(0.087)	(0.002)	(0.000830)
Zip Shr. Male	0.502	0.503	-0.001*	0.00147*
	(0.035)	(0.032)	(0.001)	(0.000673)
Zip Shr. Urban	0.520	0.773	-0.253**	0.0156*
	(0.460)	(0.354)	(0.007)	(0.00782)
Zip Shr. College	0.235	0.393	-0.159**	0.00968**
	(0.110)	(0.151)	(0.002)	(0.00180)
Zip Shr. Food Stamps	0.106	0.095	0.011**	0.0000361
	(0.058)	(0.055)	(0.001)	(0.000687)
Zip Median Income (Log)	10.805	10.886	-0.081**	-.040**
	(0.239)	(0.317)	(0.004)	(0.006)

Notes: Balance test of defendants' characteristics in the sample with only White and Black defendants when the share of Black defendants is lower than 0.5 (N=761,461). Recid. Risk is the predicted probability of recidivism. Columns 1 and 2 report the mean and standard deviation of defendants' characteristics separately for White and Black judges. Column 3 reports the simple difference in means between White and Black judges. Column 4 reports the difference in means after taking out county-year and charge severity fixed effects. Standard errors in Column 4 are clustered at the county×year level (in parenthesis): + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

E Additional Results

E.1 Tables

Table E.11: In-Group Bias: Alternative Outcomes

	Defendant is Incarcerated	Jail Time (Log Days)	
	Main	Including Zeroes	Excluding Zeroes
	(1)	(2)	(3)
Black Defendant	0.0517**	0.158**	-0.184**
	(0.00253)	(0.0185)	(0.0221)
Black Judge × Black Defendant	0.0544**	0.250**	-0.0149
	(0.0110)	(0.0595)	(0.0679)
Obs.	883893	883893	348541
R^2	0.136	0.195	0.566
County-Year FE	X	X	X
Charge Severity FE	X	X	X
Risk Ventile FE	X	X	X
Judge FE	X	X	X
Additional Interactions & Controls	X	X	X

Notes: Estimated racial in-group bias in jail decision by judges and harshness of the sentence, measured in jail time. All specifications include county×year, charge severity, risk ventiles, and judge fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Risk ventiles are computed separately across all ethnicities using our machine-learning predicted recidivism risk. Additional interactions and controls include defendants' characteristics other than race and the interaction between these and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the county and year level (in parenthesis): + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table E.12: Judicial Bias: Jail Decision

	Defendant is Incarcerated			
	(1)	(2)	(3)	(4)
Black Defendant	0.0514**	0.0516**	0.0463**	0.0450**
	(0.00313)	(0.00254)	(0.00303)	(0.00394)
Black Judge × Black Defendant	0.0523**	0.0521**	0.0510**	0.0521**
	(0.0111)	(0.0105)	(0.0105)	(0.0110)
Black Defendant × Republican Judge	0.00108			0.00167
	(0.00426)			(0.00422)
Black Defendant × Female Judge		0.00145		0.00220
		(0.00531)		(0.00523)
Black Defendant × Exp. Judge			0.0162**	0.0163**
			(0.00461)	(0.00461)
Obs.	883893	883893	883893	883893
R^2	0.136	0.136	0.136	0.136
County-Year FE	X	X	X	X
Charge Severity FE	X	X	X	X
Risk Ventile FE	X	X	X	X
Judge FE	X	X	X	X
Additional Interactions & Controls	X	X	X	X

Notes: Estimated in-group bias and judicial bias along judges' characteristics other than race. All specifications include county×year, charge severity, risk ventiles, and judge fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Risk ventiles are computed separately across all ethnicities using our machine-learning predicted recidivism risk. Additional interactions and controls include defendants' characteristics other than race and the interaction between these and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the county and year level (in parenthesis): + p < 0.1, * p < 0.05, ** p < 0.01.

Table E.13: In-Group Bias: Jail Decision – Judge Heterogeneity

	Defendant is Incarcerated			
	(1)	(2)	(3)	(4)
Black Defendant	0.0458**	0.0459**	0.0462**	0.0462**
	(0.00315)	(0.00315)	(0.00314)	(0.00314)
Black Judge × Black Defendant	0.0529**	0.0529**	0.0360*	0.0429**
	(0.0108)	(0.0102)	(0.0160)	(0.0155)
Black Judge × Black Defendant × Republican Judge		-0.0922**		-0.0766**
		(0.0245)		(0.0265)
Black Judge × Black Defendant × Exp. Judge			0.0418*	0.0358+
			(0.0200)	(0.0202)
Obs.	883893	883893	883893	883893
R^2	0.136	0.136	0.136	0.136
County-Year FE	X	X	X	X
Charge Severity FE	X	X	X	X
Risk Ventile FE	X	X	X	X
Judge FE	X	X	X	X
Additional Interactions & Controls	X	X	X	X

Notes: Heterogeneity of racial in-group bias in jail decision by judges along judges characteristics. All specifications include county×year, charge severity, risk ventiles, and judge fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Risk ventiles are computed separately across all ethnicities using our machine-learning predicted recidivism risk. Additional interactions and controls include defendants' characteristics other than race and the interaction between these and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the county and year level (in parenthesis): + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

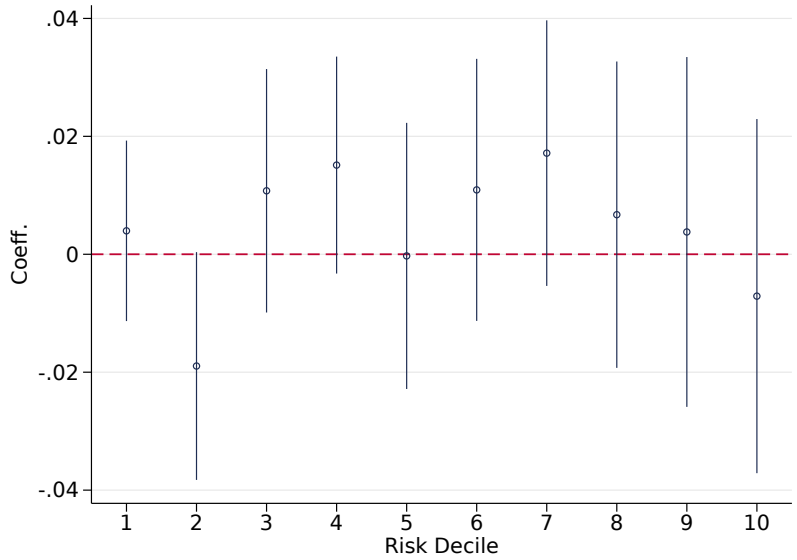
Table E.14: In-Group Bias: Alternative Outcomes

	Defendant is Incarcerated	Jail Time (Log Days)	
	Main	Including Zeroes	Excluding Zeroes
	(1)	(2)	(3)
Black Defendant	0.0399**	0.121**	-0.223**
	(0.00309)	(0.0175)	(0.0212)
Black Judge × Black Defendant	0.0550**	0.253**	-0.0322
	(0.0116)	(0.0632)	(0.0643)
Recid. Score	0.132**	0.712**	0.651**
	(0.0148)	(0.0776)	(0.0731)
Black Judge × Recid. Score	-0.152	-0.572	-0.250
	(0.0957)	(0.474)	(0.419)
Black Defendant × Recid. Score	0.147**	0.941**	0.742**
	(0.0159)	(0.0938)	(0.0678)
Black Judge × Black Defendant × Recid. Score	0.0636	0.294	0.0544
	(0.0632)	(0.284)	(0.246)
Obs.	883893	883893	348541
R^2	0.137	0.196	0.567
County-Year FE	X	X	X
Charge Severity FE	X	X	X
Judge FE	X	X	X
Other Judge/Def. Characteristics	X	X	X
Other Judge/Def. Characteristics-Risk Score Interactions	X	X	X

Notes: Estimated racial in-group bias in jail decision by judges and harshness of the sentence, measured in jail time. The recidivism risk score is standardized to have a mean of 0 and a standard deviation of 1. All specifications include county×year, charge severity, and judge fixed effects. Risk ventiles are computed separately across all ethnicities using our machine-learning predicted recidivism risk. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Additional interactions and controls include defendants' characteristics other than race and the interaction between these, the recidivism risk score, and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the county and year level (in parenthesis): + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

E.2 Robustness

Figure E.9: Black Def./Conservative Judge Effect on Incarceration Rate, by Risk Decile



Notes: Effect of Black defendant-Conservative judge pairs on incarceration decision by judges. Regressions include court×year and charge severity fixed effects. We also add recidivism risk deciles fixed effects interacted with indicators for defendant/judge characteristics. Recidivism risk deciles are computed across all ethnicities and are centered by court-year and charge severity. Errors are clusters at the court-year level.

Table E.15: Main Results: Incarceration Decision - Prosecutor Fixed Effects

	Defendant is Incarcerated			
	All	All	High Black Shr.	Low Black Shr.
	(1)	(2)	(3)	(4)
Black Defendant	0.0507**	0.0453**	0.0571**	0.0467**
	(0.00283)	(0.00281)	(0.00677)	(0.00298)
Black Judge × Black Defendant	0.0419**	0.0435**	0.0575**	0.00495
	(0.0109)	(0.0131)	(0.0108)	(0.0175)
Recid. Score		0.0236**	0.0238	0.0236**
		(0.00241)	(0.0162)	(0.00243)
Black Judge × Recid. Score		-0.0396*	-0.111**	-0.00752
		(0.0190)	(0.0310)	(0.0213)
Black Defendant × Recid. Score		0.0239**	0.0215*	0.0202**
		(0.00257)	(0.00820)	(0.00257)
Black Judge × Black Defendant × Recid. Score		0.0192+	0.0457**	0.0481**
		(0.0117)	(0.0165)	(0.0138)
Obs.	819746	819746	68824	750857
R^2	0.146	0.146	0.180	0.134
County-Year FE	X	X	X	X
Charge Severity FE	X	X	X	X
Risk Ventile FE	X			
Judge FE	X	X	X	X
Other Judge/Def Characteristics	X	X	X	X
Other Judge/Def. Characteristics-Risk Score Interactions		X	X	X
Prosecutor FE	X	X	X	X

Notes: Estimated racial in-group bias in incarceration decisions by judges. Column 1 shows the baseline in-group bias; Column 2 adds the interaction with recidivism risk; Columns 3 and 4 report the results from Column 2 separately for the samples of judges seeing a share of Black defendants higher or lower than 0.5, respectively. The recidivism risk score is standardized to have a mean of 0 and a standard deviation of 1. All specifications include county×year, charge severity, judge fixed effects, and prosecutor fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Risk ventiles are computed separately across all ethnicities using our machine-learning predicted recidivism risk. Additional interactions and controls include defendants' characteristics other than race and the interaction between these, the recidivism risk score, and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the court and year level (in parenthesis): + p < 0.1, * p < 0.05, ** p < 0.01.

Table E.16: Main Results: Jail Decision - Before COMPAS

	Defendant is Incarcerated			
	All	All	High Black Shr.	Low Black Shr.
	(1)	(2)	(3)	(4)
Black Defendant	0.0458**	0.0405**	0.0341**	0.0452**
	(0.00315)	(0.00310)	(0.00646)	(0.00316)
Black Judge × Black Defendant	0.0529**	0.0553**	0.0608**	0.00270
	(0.0108)	(0.0116)	(0.0132)	(0.0184)
Recid. Score		0.0226**	0.0158	0.0239**
		(0.00253)	(0.0113)	(0.00245)
Black Judge × Recid. Score		-0.0259	-0.0460+	-0.00912
		(0.0164)	(0.0263)	(0.0218)
Black Defendant × Recid. Score		0.0252**	0.0340**	0.0217**
		(0.00272)	(0.00607)	(0.00260)
Black Judge × Black Defendant × Recid. Score		0.0109	0.0103	0.0442**
		(0.0108)	(0.0156)	(0.0115)
Obs.	883893	883893	122425	761461
R^2	0.136	0.137	0.144	0.120
County-Year FE	X	X	X	X
Charge Severity FE	X	X	X	X
Risk Ventile FE	X			
Judge FE	X	X	X	X
Other Judge/Def Characteristics	X	X	X	X
Other Judge/Def. Characteristics-Risk Score Interactions		X	X	X

Notes: Estimated racial in-group bias in incarceration decisions by judges before COMPAS was introduced. Column 1 shows the baseline in-group bias; Column 2 adds the interaction with recidivism risk; Columns 3 and 4 report the results from Column 2 separately for the samples of judges seeing a share of Black defendants higher or lower than 0.5, respectively. The recidivism risk score is standardized to have a mean of 0 and a standard deviation of 1. All specifications include county×year, charge severity, and judge fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Risk ventiles are computed separately across all ethnicities using our machine-learning predicted recidivism risk. Additional interactions and controls include defendants' characteristics other than race and the interaction between these, the recidivism risk score, and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the court and year level (in parenthesis): + p < 0.1, * p < 0.05, ** p < 0.01.

Table E.17: Main Results: Incarceration Decision - Zip Fixed Effects

	Defendant is Incarcerated			
	All	All	High Black Shr.	Low Black Shr.
	(1)	(2)	(3)	(4)
Black Defendant	0.0414**	0.0342**	0.0303**	0.0395**
	(0.00334)	(0.00337)	(0.00882)	(0.00312)
Black Judge × Black Defendant	0.0515**	0.0560**	0.0619**	0.00848
	(0.0124)	(0.0126)	(0.0191)	(0.0187)
Recid. Score		0.0228**	0.0233+	0.0235**
		(0.00249)	(0.0118)	(0.00248)
Black Judge × Recid. Score		-0.0317+	-0.0627*	-0.0106
		(0.0173)	(0.0256)	(0.0171)
Black Defendant × Recid. Score		0.0274**	0.0379**	0.0235**
		(0.00256)	(0.00623)	(0.00245)
Black Judge × Black Defendant × Recid. Score		0.0165	0.0223	0.0440**
		(0.0107)	(0.0154)	(0.0123)
Obs.	852228	852228	108340	742898
R^2	0.153	0.153	0.168	0.137
County-Year FE	X	X	X	X
Charge Severity FE	X	X	X	X
Risk Ventile FE	X			
Judge FE	X	X	X	X
Other Judge/Def Characteristics	X	X	X	X
Other Judge/Def. Characteristics-Risk Score Interactions		X	X	X
Zip FE	X	X	X	X

Notes: Estimated racial in-group bias in incarceration decisions by judges. Column 1 shows the baseline in-group bias; Column 2 adds the interaction with recidivism risk; Columns 3 and 4 report the results from Column 2 separately for the samples of judges seeing a share of Black defendants higher or lower than 0.5, respectively. The recidivism risk score is standardized to have a mean of 0 and a standard deviation of 1. All specifications include county×year, charge severity, judge, and zipcode fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Risk ventiles are computed separately across all ethnicities using our machine-learning predicted recidivism risk. Additional interactions and controls include defendants' characteristics other than race and the interaction between these, the recidivism risk score, and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the court and year level (in parenthesis): + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table E.18: Main Results: Incarceration Decision - County-Year-Severity Fixed Effects

	Defendant is Incarcerated			
	All	All	High Black Shr.	Low Black Shr.
	(1)	(2)	(3)	(4)
Black Defendant	0.0436**	0.0383**	0.0333**	0.0433**
	(0.00308)	(0.00299)	(0.00605)	(0.00301)
Black Judge × Black Defendant	0.0461**	0.0486**	0.0584**	0.00157
	(0.00923)	(0.00987)	(0.0111)	(0.0150)
Recid. Score		0.0235**	0.0209+	0.0243**
		(0.00242)	(0.0117)	(0.00242)
Black Judge × Recid. Score		-0.0310+	-0.0493+	-0.0118
		(0.0160)	(0.0251)	(0.0211)
Black Defendant × Recid. Score		0.0265**	0.0328**	0.0221**
		(0.00264)	(0.00608)	(0.00254)
Black Judge × Black Defendant × Recid. Score		0.0114	0.0115	0.0462**
		(0.0103)	(0.0147)	(0.0124)
Obs.	882642	882642	122352	760191
R^2	0.162	0.163	0.155	0.148
County-Year-C. Severity FE	X	X	X	X
Risk Ventile FE	X			
Judge FE	X	X	X	X
Other Judge/Def Characteristics	X	X	X	X
Other Judge/Def. Characteristics-Risk Score Interactions		X	X	X

Notes: Estimated racial in-group bias in incarceration decisions by judges. Column 1 shows the baseline in-group bias; Column 2 adds the interaction with recidivism risk; Columns 3 and 4 report the results from Column 2 separately for the samples of judges seeing a share of Black defendants higher or lower than 0.5, respectively. The recidivism risk score is standardized to have a mean of 0 and a standard deviation of 1. All specifications include county×year×charge severity and judge fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Risk ventiles are computed separately across all ethnicities using our machine-learning predicted recidivism risk. Additional interactions and controls include defendants' characteristics other than race and the interaction between these, the recidivism risk score, and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the court and year level (in parenthesis): + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table E.19: Main Results: Incarceration Decision - Logistic Regression Model

	Defendant is Incarcerated			
	All	All	High Black Shr.	Low Black Shr.
	(1)	(2)	(3)	(4)
Black Defendant	0.0449**	0.0443**	0.0385**	0.0485**
	(0.00307)	(0.00305)	(0.00696)	(0.00319)
Black Judge × Black Defendant	0.0542**	0.0567**	0.0616**	0.0185
	(0.0110)	(0.0123)	(0.0161)	(0.0185)
Recid. Score		0.0658**	0.0815**	0.0658**
		(0.00306)	(0.0147)	(0.00308)
Black Judge × Recid. Score		-0.0227	-0.0986*	0.0535+
		(0.0297)	(0.0453)	(0.0296)
Black Defendant × Recid. Score		0.0146**	0.0189*	0.0111**
		(0.00288)	(0.00726)	(0.00265)
Black Judge × Black Defendant × Recid. Score		0.00939	0.0201	0.0139
		(0.0135)	(0.0225)	(0.0163)
Obs.	883893	883893	122425	761461
R^2	0.140	0.139	0.140	0.124
County-Year FE	X	X	X	X
Charge Severity FE	X	X	X	X
Risk Ventile FE	X			
Judge FE	X	X	X	X
Other Judge/Def Characteristics	X	X	X	X
Other Judge/Def. Characteristics-Risk Score Interactions		X	X	X

Notes: Estimated racial in-group bias in incarceration decisions by judges. Column 1 shows the baseline in-group bias; Column 2 adds the interaction with recidivism risk; Columns 3 and 4 report the results from Column 2 separately for the samples of judges seeing a share of Black defendants higher or lower than 0.5, respectively. The recidivism risk score is obtained by training a logistic regression model and it is standardized to have a mean of 0 and a standard deviation of 1. All specifications include county×year, charge severity, and judge fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Risk ventiles are computed separately across all ethnicities using our machine-learning predicted recidivism risk. Additional interactions and controls include defendants' characteristics other than race and the interaction between these, the recidivism risk score, and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the court and year level (in parenthesis): + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table E.20: Main Results: Incarceration Decision - Different Black Shares

	Defendant is Incarcerated			
	Share within county-year		Share within county-year-judge	
	High Black Shr.	Low Black Shr.	High Black Shr.	Low Black Shr
	(1)	(2)	(3)	(4)
Black Defendant	0.0356**	0.0454**	0.0436**	0.0438**
	(0.00672)	(0.00316)	(0.00628)	(0.00311)
Black Judge × Black Defendant	0.0605**	0.00136	0.0436**	0.0195
	(0.0140)	(0.0184)	(0.0151)	(0.0172)
Recid. Score	0.0115	0.0240**	0.0226*	0.0235**
	(0.0115)	(0.00246)	(0.0113)	(0.00245)
Black Judge × Recid. Score	-0.0429	-0.0111	-0.0382	-0.0207
	(0.0270)	(0.0216)	(0.0319)	(0.0222)
Black Defendant × Recid. Score	0.0367**	0.0216**	0.0289**	0.0218**
	(0.00593)	(0.00257)	(0.00590)	(0.00253)
Black Judge × Black Defendant × Recid. Score	0.00827	0.0452**	-0.00183	0.0353*
	(0.0160)	(0.0114)	(0.0162)	(0.0139)
Obs.	122260	761625	119692	764101
R^2	0.143	0.119	0.151	0.121
County-Year FE	X	X	X	X
Charge Severity FE	X	X	X	X
Risk Ventile FE	X			
Judge FE	X	X	X	X
Other Judge/Def Characteristics	X	X	X	X
Other Judge/Def. Characteristics-Risk Score Interactions		X	X	X

*Notes:*Estimated in-group basis in jail decisions by judges, and recidivism riskeparately for the samples of judges seeing a share of Black defendants higher or lower than 0.5. Columns 1 and 2 report the results using the share of Black defendants within each county-year, while Columns 3 and 4 the share of Black defendants calculated for each judge within each county-year. The recidivism risk score is standardized to have a mean of 0 and a standard deviation of 1. All specifications include county×year, charge severity, and judge fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Risk ventiles are computed separately across all ethnicities using our machine-learning predicted recidivism risk. Additional interactions and controls include defendants' characteristics other than race and the interaction between these, the recidivism risk score, and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the court and year level (in parenthesis): + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table E.21: Main Results: Incarceration Decision - Court-Years With Black Defendants

	Defendant is Incarcerated			
	All	All	High Black Shr.	Low Black Shr.
	(1)	(2)	(3)	(4)
Black Defendant	0.0458**	0.0405**	0.0341**	0.0452**
	(0.00315)	(0.00309)	(0.00646)	(0.00316)
Black Judge × Black Defendant	0.0529**	0.0553**	0.0608**	0.00269
	(0.0108)	(0.0116)	(0.0132)	(0.0184)
Recid. Score		0.0228**	0.0158	0.0241**
		(0.00254)	(0.0113)	(0.00246)
Black Judge × Recid. Score		-0.0262	-0.0460+	-0.00943
		(0.0164)	(0.0263)	(0.0218)
Black Defendant × Recid. Score		0.0252**	0.0340**	0.0217**
		(0.00272)	(0.00607)	(0.00260)
Black Judge × Black Defendant × Recid. Score		0.0109	0.0103	0.0442**
		(0.0108)	(0.0156)	(0.0115)
Obs.	882408	882408	122425	759976
R^2	0.136	0.136	0.144	0.120
County-Year FE	X	X	X	X
Charge Severity FE	X	X	X	X
Risk Ventile FE	X			
Judge FE	X	X	X	X
Other Judge/Def Characteristics	X	X	X	X
Other Judge/Def. Characteristics-Risk Score Interactions		X	X	X

Notes: Estimated racial in-group bias in incarceration decisions by judges for court-years with at least one Black defendant. Column 1 shows the baseline in-group bias; Column 2 adds the interaction with recidivism risk; Columns 3 and 4 report the results from Column 2 separately for the samples of judges seeing a share of Black defendants higher or lower than 0.5, respectively. The recidivism risk score is standardized to have a mean of 0 and a standard deviation of 1. All specifications include county×year, charge severity, and judge fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Risk ventiles are computed separately across all ethnicities using our machine-learning predicted recidivism risk. Additional interactions and controls include defendants' characteristics other than race and the interaction between these, the recidivism risk score, and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the court and year level (in parenthesis): + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table E.22: Main Results: Incarceration Decision - Court-Years With Black Judge

	Defendant is Incarcerated			
	All	All	High Black Shr.	Low Black Shr.
	(1)	(2)	(3)	(4)
Black Defendant	0.0292**	0.0318**	0.0341**	0.0325**
	(0.00620)	(0.00671)	(0.00669)	(0.0109)
Black Judge × Black Defendant	0.0607**	0.0598**	0.0615**	0.0221
	(0.0124)	(0.0125)	(0.0137)	(0.0204)
Recid. Score		0.00539	0.0137	-0.00203
		(0.00755)	(0.0115)	(0.00883)
Black Judge × Recid. Score		-0.00791	-0.0436	0.0146
		(0.0175)	(0.0263)	(0.0235)
Black Defendant × Recid. Score		0.0336**	0.0359**	0.0271**
		(0.00504)	(0.00602)	(0.00779)
Black Judge × Black Defendant × Recid. Score		0.00739	0.00796	0.0393*
		(0.0124)	(0.0156)	(0.0173)
Obs.	186716	186716	118937	67777
R^2	0.139	0.140	0.143	0.141
County-Year FE	X	X	X	X
Charge Severity FE	X	X	X	X
Risk Ventile FE	X			
Judge FE	X	X	X	X
Other Judge/Def Characteristics	X	X	X	X
Other Judge/Def. Characteristics-Risk Score Interactions		X	X	X

Notes: Estimated racial in-group bias in incarceration decisions by judges for court-years with at least one Black judge. Column 1 shows the baseline in-group bias; Column 2 adds the interaction with recidivism risk; Columns 3 and 4 report the results from Column 2 separately for the samples of judges seeing a share of Black defendants higher or lower than 0.5, respectively. The recidivism risk score is standardized to have a mean of 0 and a standard deviation of 1. All specifications include county×year, charge severity, and judge fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Risk ventiles are computed separately across all ethnicities using our machine-learning predicted recidivism risk. Additional interactions and controls include defendants' characteristics other than race and the interaction between these, the recidivism risk score, and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the court and year level (in parenthesis): + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table E.23: Main Results: Incarceration Decision - Lenient Judges Risk Score

	Defendant is Incarcerated			
	All	All	High Black Shr.	Low Black Shr.
	(1)	(2)	(3)	(4)
Black Defendant	0.0461**	0.0404**	0.0341**	0.0453**
	(0.00318)	(0.00310)	(0.00656)	(0.00317)
Black Judge × Black Defendant	0.0527**	0.0556**	0.0605**	0.000901
	(0.0109)	(0.0120)	(0.0141)	(0.0188)
Recid. Score		0.0187**	0.0161	0.0197**
		(0.00242)	(0.0109)	(0.00238)
Black Judge × Recid. Score		-0.0259	-0.0508+	-0.00714
		(0.0169)	(0.0272)	(0.0195)
Black Defendant × Recid. Score		0.0280**	0.0382**	0.0235**
		(0.00278)	(0.00669)	(0.00260)
Black Judge × Black Defendant × Recid. Score		0.0115	0.00803	0.0505**
		(0.0117)	(0.0159)	(0.0156)
Obs.	883893	883893	122425	761461
R^2	0.136	0.136	0.144	0.119
County-Year FE	X	X	X	X
Charge Severity FE	X	X	X	X
Risk Ventile FE	X			
Judge FE	X	X	X	X
Other Judge/Def Characteristics	X	X	X	X
Other Judge/Def. Characteristics-Risk Score Interactions		X	X	X

Notes: Estimated racial in-group bias in incarceration decisions by judges. Column 1 shows the baseline in-group bias; Column 2 adds the interaction with recidivism risk; Columns 3 and 4 report the results from Column 2 separately for the samples of judges seeing a share of Black defendants higher or lower than 0.5, respectively. The recidivism risk score is obtained by training the model on the sample of lenient judges, defined as judges who release the most defendants. The risk score is standardized to have a mean of 0 and a standard deviation of 1. All specifications include county×year, charge severity, and judge fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Risk ventiles are computed separately across all ethnicities using our machine-learning predicted recidivism risk. Additional interactions and controls include defendants' characteristics other than race and the interaction between these, the recidivism risk score, and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the court and year level (in parenthesis): + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table E.24: Main Results: Incarceration Decision - Most Lenient Judges Risk Score

	Defendant is Incarcerated			
	All	All	High Black Shr.	Low Black Shr.
	(1)	(2)	(3)	(4)
Black Defendant	0.0458**	0.0406**	0.0344**	0.0453**
	(0.00315)	(0.00309)	(0.00651)	(0.00317)
Black Judge × Black Defendant	0.0530**	0.0556**	0.0602**	0.00152
	(0.0107)	(0.0119)	(0.0138)	(0.0187)
Recid. Score		0.0220**	0.0201+	0.0229**
		(0.00246)	(0.0108)	(0.00243)
Black Judge × Recid. Score		-0.0311+	-0.0573*	-0.0102
		(0.0170)	(0.0265)	(0.0224)
Black Defendant × Recid. Score		0.0257**	0.0341**	0.0218**
		(0.00276)	(0.00644)	(0.00264)
Black Judge × Black Defendant × Recid. Score		0.0119	0.00999	0.0494**
		(0.0114)	(0.0153)	(0.0147)
Obs.	883893	883893	122425	761461
R^2	0.136	0.136	0.144	0.120
County-Year FE	X	X	X	X
Charge Severity FE	X	X	X	X
Risk Ventile FE	X			
Judge FE	X	X	X	X
Other Judge/Def Characteristics	X	X	X	X
Other Judge/Def. Characteristics-Risk Score Interactions		X	X	X

Notes: Estimated racial in-group bias in incarceration decisions by judges. Column 1 shows the baseline in-group bias; Column 2 adds the interaction with recidivism risk; Columns 3 and 4 report the results from Column 2 separately for the samples of judges seeing a share of Black defendants higher or lower than 0.5, respectively. The recidivism risk score is obtained by training the model on the sample of the two most lenient judges for each county and year, defined as judges who release the most defendants. The risk score is standardized to have a mean of 0 and a standard deviation of 1. All specifications include county×year, charge severity, and judge fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Risk ventiles are computed separately across all ethnicities using our machine-learning predicted recidivism risk. Additional interactions and controls include defendants' characteristics other than race and the interaction between these, the recidivism risk score, and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the court and year level (in parenthesis): + p < 0.1, * p < 0.05, ** p < 0.01.

Table E.25: Main Results: Incarceration Decision - First Offense

	Defendant is Incarcerated			
	All	All	High Black Shr.	Low Black Shr.
	(1)	(2)	(3)	(4)
Black Defendant	0.0330**	0.0258**	0.0101	0.0300**
	(0.00447)	(0.00447)	(0.00868)	(0.00479)
Black Judge × Black Defendant	0.0104	0.0138	0.0116	0.000197
	(0.0158)	(0.0141)	(0.0146)	(0.0413)
Recid. Score		-0.102**	-0.0854**	-0.106**
		(0.00450)	(0.0191)	(0.00456)
Black Judge × Recid. Score		-0.0264	-0.0435	0.0103
		(0.0362)	(0.0543)	(0.0275)
Black Defendant × Recid. Score		0.0386**	0.0305**	0.0365**
		(0.00346)	(0.0101)	(0.00452)
Black Judge × Black Defendant × Recid. Score		-0.0218	-0.0195	0.0117
		(0.0231)	(0.0300)	(0.0293)
Obs.	279555	279555	42881	236671
R^2	0.211	0.212	0.189	0.210
County-Year FE	X	X	X	X
Charge Severity FE	X	X	X	X
Risk Ventile FE	X			
Judge FE	X	X	X	X
Other Judge/Def Characteristics	X	X	X	X
Other Judge/Def. Characteristics-Risk Score Interactions		X	X	X

Notes: Estimated racial in-group bias in incarceration decisions by judges on the sample of first-time offenders. Column 1 shows the baseline in-group bias; Column 2 adds the interaction with recidivism risk; Columns 3 and 4 report the results from Column 2 separately for the samples of judges seeing a share of Black defendants higher or lower than 0.5, respectively. The recidivism risk score is standardized to have a mean of 0 and a standard deviation of 1. All specifications include county×year, charge severity, and judge fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Risk ventiles are computed separately across all ethnicities using our machine-learning predicted recidivism risk. Additional interactions and controls include defendants' characteristics other than race and the interaction between these, the recidivism risk score, and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the court and year level (in parenthesis): + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

Table E.26: Main Results: Incarceration Decision - Controlling for First Offense

	Defendant is Incarcerated			
	All	All	High Black Shr.	Low Black Shr.
	(1)	(2)	(3)	(4)
Black Defendant	0.0462**	0.0411**	0.0319**	0.0459**
	(0.00314)	(0.00310)	(0.00641)	(0.00315)
Black Judge × Black Defendant	0.0481**	0.0538**	0.0624**	-0.000922
	(0.0105)	(0.0116)	(0.0128)	(0.0171)
Recid. Score		0.0266**	0.00321	0.0291**
		(0.00256)	(0.0121)	(0.00239)
Black Judge × Recid. Score		-0.0325*	-0.0342	-0.0262
		(0.0163)	(0.0271)	(0.0259)
Black Defendant × Recid. Score		0.0248**	0.0351**	0.0206**
		(0.00279)	(0.00604)	(0.00262)
Black Judge × Black Defendant × Recid. Score		0.0112	0.00923	0.0450**
		(0.0109)	(0.0156)	(0.0138)
Obs.	883893	883893	122425	761461
R^2	0.137	0.138	0.145	0.121
County-Year FE	X	X	X	X
Charge Severity FE	X	X	X	X
Risk Ventile FE	X			
Judge FE	X	X	X	X
Other Judge/Def Characteristics	X	X	X	X
Other Judge/Def. Characteristics-Risk Score Interactions		X	X	X

Notes: Estimated racial in-group bias in incarceration decisions by judges. Column 1 shows the baseline in-group bias; Column 2 adds the interaction with recidivism risk; Columns 3 and 4 report the results from Column 2 separately for the samples of judges seeing a share of Black defendants higher or lower than 0.5, respectively. The recidivism risk score is standardized to have a mean of 0 and a standard deviation of 1. All specifications include county×year, charge severity, and judge fixed effects. Charge severity indicates the severity of the case, defined as the severity of the highest charge in the case. Risk ventiles are computed separately across all ethnicities using our machine-learning predicted recidivism risk. Additional interactions and controls include defendants' characteristics other than race, including whether the case is concerns a first-time offender, and the interaction between these, the recidivism risk score, and judges' characteristics, like gender or political affiliation, measured with party contributions. Standard errors are clustered at the court and year level (in parenthesis): + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.