Online Appendix:
Measuring Racial Discrimination in Bail Decisions

By DAVID ARNOLD, WILL DOBBIE, AND PETER HULL[*]

[*] Arnold: Department of Economics, University of California-San Diego (daarnold@ucsd.edu); Dobbie: Harvard Kennedy School and NBER (will_dobbie@hks.harvard.edu); Hull: Department of Economics, Brown University and NBER (peter_hull@brown.edu).

Table of Contents

A Appendix Figures and Tables 3

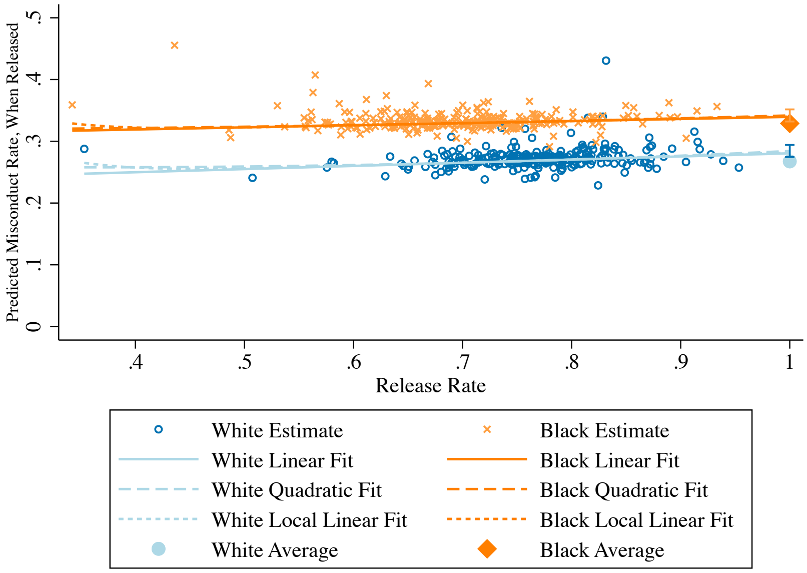
Figure A1: Placebo Mean Risk Extrapolation 3
Figure A2: Judge-Specific Release Rates and Conditional Misconduct Rates, With Covariate Adjustment 4
Figure A3: Disparate Impact Estimates, Model-Based Mean Risk Estimates 5
Figure A4: Predictiveness of Observational Release Rate Disparities 6
Table A1: Judge Leniency and Sample Attrition 7
Table A2: Descriptive Statistics by Sample 8
Table A3: Tests of Quasi-Random Judge Assignment 9
Table A4: First Stage Effects of Judge Leniency 10
Table A5: Simple Numerical Example of Disparate Impact Estimation 11
Table A6: Disparate Impact Estimation for NYC Release Decisions 12
Table A7: Mean Risk and Disparate Impact Estimates, Shrunk Leniency Estimates 13
Table A8: Mean Risk and Disparate Impact Estimates, With Covariate Adjustment 14
Table A9: Disparate Impact and Judge Characteristics 15
Table A10: Mean Risk and Disparate Impact Estimates by Defendant Characteristics 16
Table A11: Mean Risk and Disparate Impact Bounds 17
Table A12: Decomposition of Disparate Impact by Misconduct Potential 18
Table A13: Mean Risk and Disparate Impact Estimates, Borough-Specific Estimates 19
Table A14: Mean Risk and Disparate Impact Estimates, Borough-Specific Estimates with Judge-Specific Time Effects 20
Table A15: Mean Risk and Disparate Impact Estimates, Alternative Misconduct Outcome 21
Table A16: Mean Risk and Disparate Impact Estimates, Alternative Judge Decisions 22
Table A17: Mean Risk and Disparate Impact Estimates, Alternative Race Definition 23
Table A18: Hierarchical MTE Model Hyperparameter Estimates 24
Table A19: Tests of Conventional MTE Monotonicity 25
Table A20: Disparate Impact and Judge Characteristics, Model-Based Mean Risk . . . 26
Table A21: Racial Bias and Judge Characteristics 27
Table A22: Signal Quality Differences and Judge Characteristics 28

B Econometric Appendix 29

Section B.B1: Defining and Measuring Disparate Impact with Multi-Valued Y_i^* 29
Section B.B2: Included Variables Bias 30
Section B.B3: Empirical Bayes Methods 30
Section B.B4: Rescaled Benchmarking Regressions: Numerical Example and in NYC . 32
Section B.B5: Bounding Mean Risk and Racial Discrimination 34
Section B.B6: Judge Decision-Making Model and Extensions 34
Section B.B7: Conventional Monotonicity Violations and Judge Signal Quality 36
Section B.B8: SMD Estimation of the Hierarchical MTE Model 37

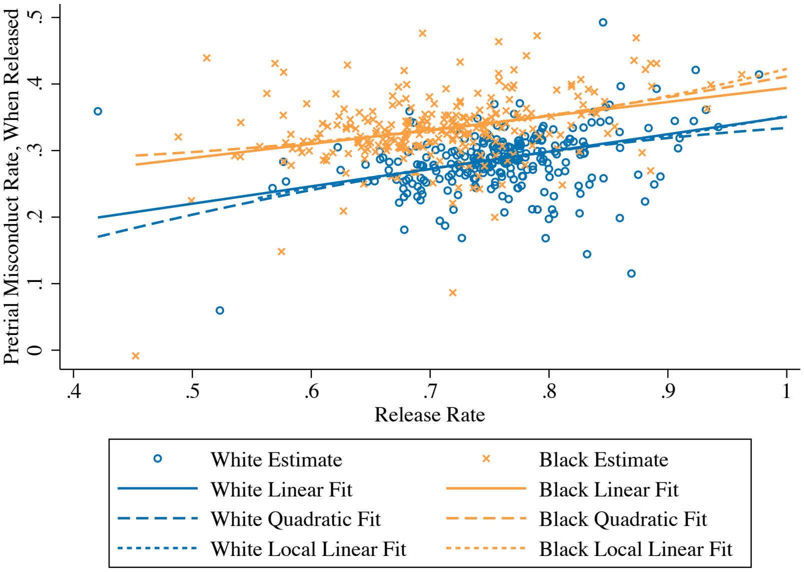
Appendix Figures and Tables

Appendix Figure A1. Placebo Mean Risk Extrapolation



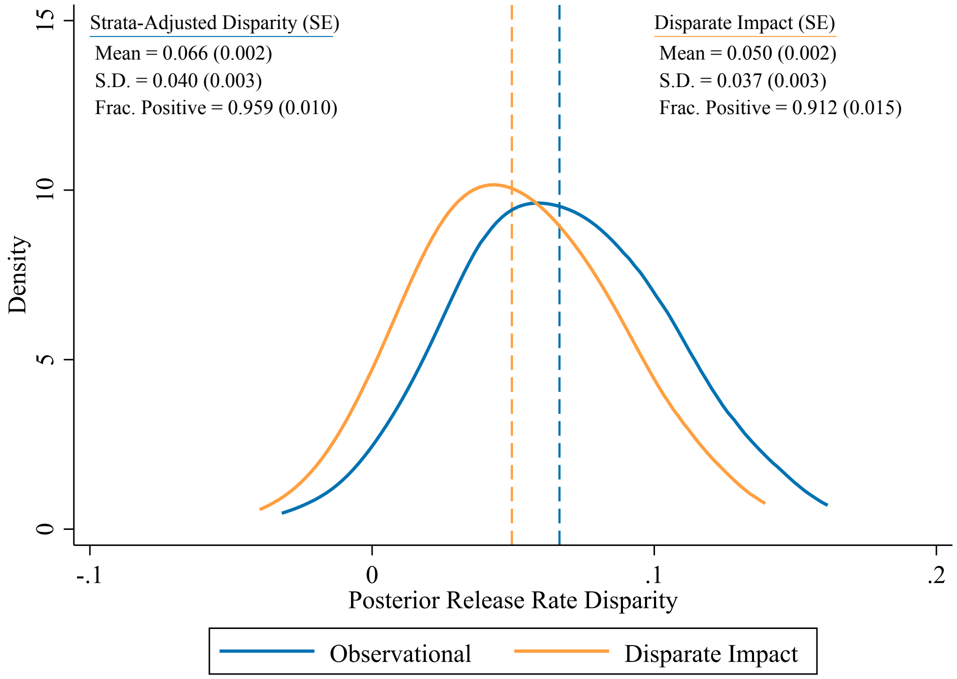
Notes. This figure plots race-specific release rates for the 268 judges in our sample against rates of predicted pretrial misconduct among the set of released defendants. Predicted misconduct is given by the fitted values of an OLS regression of misconduct on the regressors in column 3 of Table 2, estimated in the set of released defendants. Average predicted misconduct rates in the full sample of white and Black defendants are indicated with solid markers at the maximal release rate of one. All estimates adjust for court-by-time fixed effects. The figure also plots race-specific linear, quadratic, and local linear curves of best fit, obtained from judge-level regressions that inverse-weight by the variance of the estimated predicted misconduct rate among released defendants. The local linear regression uses a Gaussian kernel with a race-specific rule-of-thumb bandwidth. 95 percent confidence intervals for the local linear extrapolations' intercept estimates at one, obtained from robust standard errors two-way clustered at the individual and judge level, are indicated with brackets.

APPENDIX FIGURE A2. JUDGE-SPECIFIC RELEASE RATES AND CONDITIONAL MISCONDUCT RATES, WITH COVARIATE ADJUSTMENT



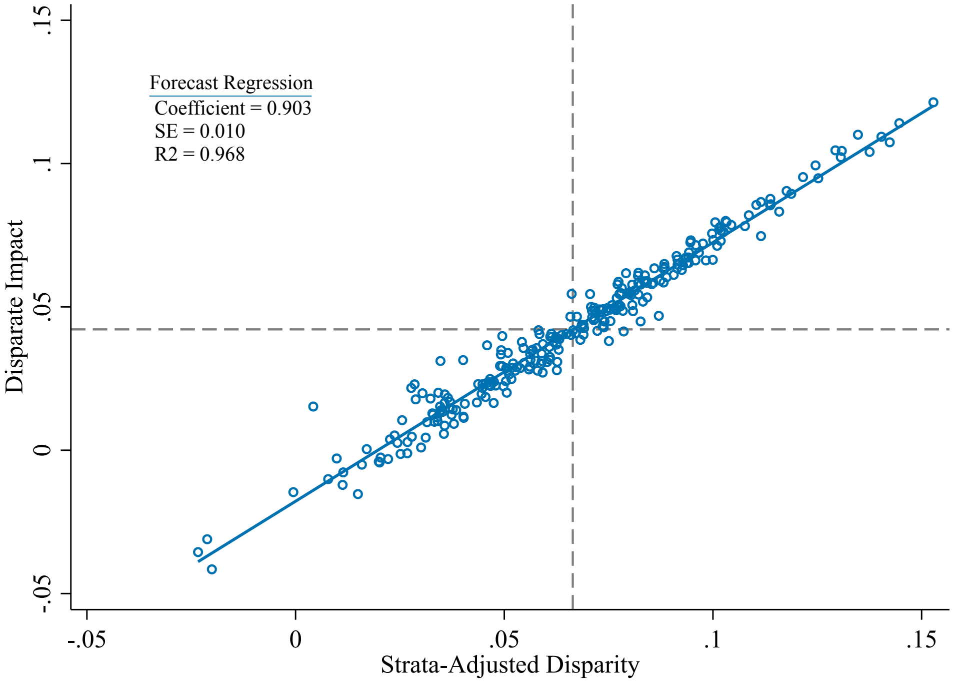
Notes. This figure plots race-specific release rates for the 268 judges in our sample against rates of pretrial misconduct among the set of released defendants. All estimates adjust for court-by-time fixed effects and the case and defendant observables in Table 2. The figure also plots race-specific linear, quadratic, and local linear curves of best fit, obtained from judge-level regressions that inverse-weight by the variance of the estimated misconduct rate among released defendants. The local linear regressions use a Gaussian kernel with a race-specific rule-of-thumb bandwidth.

APPENDIX FIGURE A3. DISPARATE IMPACT ESTIMATES, MODEL-BASED MEAN RISK ESTIMATES



Notes. This figure plots the posterior distribution of observational disparities and disparate impact for the 268 judges in our sample. Strata-adjusted disparities are estimated by the coefficients of an OLS regression of an indicator for pretrial release on white×judge fixed effects, controlling for judge main effects and court-by-time fixed effects. Disparate impact is estimated as described in Section IV, using the hierarchical MTE model estimates of mean risk for each race. The distribution of judge disparities and disparate impact estimates, and fractions of positive disparities and disparate impact estimates, are computed from these estimates as posterior average effects; see Appendix B.B3 for details. Means and standard deviations refer to the estimated prior distribution.

Appendix Figure A4. Predictiveness of Observational Release Rate Disparities



Notes. This figure plots disparate impact estimates against the corresponding strata-adjusted release rate disparity posteriors for the 268 judges in our sample. Observational disparities are estimated by the coefficients of an OLS regression of an indicator for pretrial release on white×judge fixed effects, controlling for judge main effects and court-by-time fixed effects. Disparate impact is estimated as described in Section IV, using the local linear extrapolation from Figure 2 to estimate the mean risk of each race. Empirical Bayes posteriors are computed using a standard shrinkage procedure, as described in Appendix B.B3. The slope of the solid line indicates the forecast coefficient.

APPENDIX TABLE A1—JUDGE LENIENCY AND SAMPLE ATTRITION

	All Defendants	White Defendants	Black Defendants
	(1)	(2)	(3)
Dropped from Sample	0.00007	0.00003	0.00012
	(0.00012)	(0.00013)	(0.00014)
Court x Time FE	Yes	Yes	Yes
Mean Sample Attrition	0.416	0.409	0.424
Cases	1,425,652	726,284	697,597

Notes. This table reports OLS estimates of regressions of judge leniency on an indicator for leaving the sample due to case adjournment or case disposal and court-by-time fixed effects. The regressions are estimated on the sample of all arraignments made in NYC between November 1, 2008 and November 1, 2013. Judge leniency is estimated using data from other cases assigned to a given bail judge, following the procedure described in Section III.A. Robust standard errors, two-way clustered at the individual and the judge level, are reported in parentheses.

Appendix Table A2—Descriptive Statistics by Sample

	All Defendants		White Defendants		Black Defendants	
	Full	Estimation	Full	Estimation	Full	Estimation
	Sample	Sample	Sample	Sample	Sample	Sample
Panel A: Pretrial Release	(1)	(2)	(3)	(4)	(5)	(6)
Released Before Trial	0.852	0.730	0.872	0.767	0.832	0.695
Share ROR	0.601	0.852	0.616	0.852	0.586	0.851
Share Disposed	0.301	0.000	0.274	0.000	0.327	0.000
Share Adjourned	0.191	0.000	0.199	0.000	0.183	0.000
Share Money Bail	0.068	0.144	0.070	0.144	0.066	0.145
Share Other Bail Type	0.332	0.004	0.314	0.004	0.348	0.004
Share Remanded	0.000	0.000	0.000	0.000	0.000	0.000
Panel B: Defendant Characteristics						
White	0.483	0.478	1.000	1.000	0.000	0.000
Male	0.822	0.821	0.831	0.839	0.813	0.804
Age at Arrest	31.819	31.969	31.540	32.055	32.080	31.890
Prior Rearrest	0.192	0.229	0.168	0.204	0.214	0.253
Prior FTA	0.085	0.103	0.071	0.087	0.099	0.117
Panel C: Charge Characteristics						
Number of Charges	1.094	1.150	1.111	1.184	1.078	1.118
Felony Charge	0.184	0.362	0.181	0.355	0.188	0.368
Misdemeanor Charge	0.816	0.638	0.819	0.645	0.812	0.632
Any Drug Charge	0.347	0.256	0.342	0.257	0.352	0.256
Any DUI Charge	0.031	0.046	0.046	0.067	0.017	0.027
Any Violent Charge	0.072	0.143	0.062	0.124	0.081	0.160
Any Property Charge	0.217	0.136	0.209	0.127	0.226	0.144
Cases	1,358,278	595,186	656,711	284,598	701,567	310,588

Notes. This table summarizes the difference between the NYC analysis sample and the full sample of NYC arraignments. The full sample consists of all bail hearings between November 1, 2008 and November 1, 2013. The analysis sample consists of bail hearings that were quasi-randomly assigned to judges between November 1, 2008 and November 1, 2013, as described in the text. Information on demographics and criminal outcomes is derived from court records as described in the text. Pretrial release is defined as meeting the bail conditions set by the first assigned bail judge. ROR (released on recognizance) is defined as being released without any conditions. FTA (failure to appear) is defined as failing to appear at a mandated court date.

Appendix Table A3—Tests of Quasi-Random Judge Assignment

	All Defendants	White Defendants	Black Defendants
	(1)	(2)	(3)
White	0.00013		
	(0.00009)		
Male	0.00003	0.00003	0.00004
	(0.00014)	(0.00019)	(0.00018)
Age at Arrest	-0.00011	-0.00015	-0.00008
	(0.00004)	(0.00006)	(0.00005)
Prior Rearrest	-0.00021	0.00006	-0.00042
	(0.00011)	(0.00018)	(0.00015)
Prior FTA	0.00016	-0.00011	0.00036
	(0.00016)	(0.00024)	(0.00023)
Number of Charges	-0.00001	-0.00001	-0.00001
	(0.00001)	(0.00001)	(0.00003)
Felony Charge	0.00025	0.00011	0.00039
	(0.00020)	(0.00023)	(0.00025)
Any Drug Charge	-0.00022	-0.00017	-0.00027
	(0.00016)	(0.00021)	(0.00018)
Any DUI Charge	0.00045	0.00051	0.00008
	(0.00027)	(0.00032)	(0.00045)
Any Violent Charge	-0.00008	-0.00023	0.00001
	(0.00023)	(0.00033)	(0.00025)
Any Property Charge	-0.00033	-0.00028	-0.00036
	(0.00018)	(0.00019)	(0.00027)
Joint p-value	[0.10689]	[0.29792]	[0.10136]
Court x Time FE	Yes	Yes	Yes
Cases	595,186	284,598	310,588

Notes. This table reports OLS estimates of regressions of judge leniency on defendant characteristics. The regressions are estimated on the sample described in Table 1. Judge leniency is estimated using data from other cases assigned to a given bail judge, following the procedure described in Section III.A. All regressions control for court-by-time fixed effects. The p-values reported at the bottom of each column are from F-tests of the joint significance of the variables listed in the rows. Robust standard errors, two-way clustered at the individual and the judge level, are reported in parentheses.

APPENDIX TABLE A4—FIRST STAGE EFFECTS OF JUDGE LENIENCY

	All Defendants	White Defendants	Black Defendants
	(1)	(2)	(3)
Judge Leniency	0.960	0.788	1.104
	(0.025)	(0.029)	(0.033)
Court x Time FE	Yes	Yes	Yes
Mean Release Rate	0.730	0.767	0.695
Cases	595,186	284,598	310,588

Notes. This table reports OLS estimates of regressions of an indicator for pretrial release on judge leniency. The regressions are estimated on the sample described in Table 1. Judge leniency is estimated using data from other cases assigned to a bail judge, following the procedure described in Section III.A. All regressions control for court-by-time fixed effects. Robust standard errors, two-way clustered at the individual and the judge level, are reported in parentheses.

Appendix Table A5—Simple Numerical Example of Disparate Impact Estimation

		Number of Defendants	Number Released	Scaling Factor	Rescaled Released	Release Rate	Release Disparity
Panel A: Observational Estimates		(1)	(2)	(3)	(4)	(5)	(6)
L Defendants	$Y_i^* = 0$	75	60	1	60	0.65	0.30
	$Y_i^* = 1$	25	5	1	5		
H Defendants	$Y_i^* = 0$	25	20	1	20	0.35	
	$Y_i^* = 1$	75	15	1	15		
Panel B: Rescaled Estimates							
L Defendants	$Y_i^* = 0$	75	60	2/3	40	0.50	0.00
	$Y_i^* = 1$	25	5	2	10		
H Defendants	$Y_i^* = 0$	25	20	2	40	0.50	
	$Y_i^* = 1$	75	15	2/3	10		

Notes: This table uses a simple numerical example to illustrate how disparate impact can be measured with observational release rate comparisons that are rescaled using average group-specific misconduct risk. We assume there is one type-neutral judge who releases 80 percent of defendants with $Y_i^* = 0$ and 20 percent of defendants with $Y_i^* = 1$. The judge observes the type of the defendant, which is either High-risk or Low-risk. There are 100 High-risk defendants where 75 have $Y_i^* = 1$, and 100 Low-risk defendants where 25 have $Y_i^* = 1$. Panel A shows that the judge has a Low-risk release rate of 0.65 but a High-risk release rate of 0.35, meaning that an observational comparison would find that Low-risk defendants have a 30 percentage point higher release rate than High-risk defendants despite the judge being type-neutral. Panel B shows that the true disparate impact of zero can be measured by rescaling this observational release rate comparison with the scaling factor described in the text. Column 3 of Panel B shows the scaling factor (Ω_i) in this example, and column 6 shows the resulting disparate impact estimate.

Appendix Table A6—Disparate Impact Estimation for NYC Release Decisions

		Number of Defendants	Number Released	Scaling Factor	Rescaled Released	Release Rate	Release Disparity
Panel A: Observational Estimates		(1)	(2)	(3)	(4)	(5)	(6)
White Defendants	$Y_i^* = 0$	186,250	159,296	1.000	159,296	0.765	0.068
	$Y_i^* = 1$	98,348	58,425	1.000	58,425		
Black Defendants	$Y_i^* = 0$	175,120	145,528	1.000	145,528	0.697	
	$Y_i^* = 1$	135,468	70,952	1.000	70,952		
Panel B: Rescaled Estimates							
White Defendants	$Y_i^* = 0$	186,250	159,296	0.928	147,788	0.753	0.042
	$Y_i^* = 1$	98,348	58,425	1.137	66,418		
Black Defendants	$Y_i^* = 0$	175,120	145,528	1.077	156,709	0.710	
	$Y_i^* = 1$	135,468	70,952	0.901	63,905		

Notes: This table calculates system-wide disparate impact in NYC by rescaling observational release rate comparisons using estimates of average white and Black misconduct risk. In Panel A we use the local linear estimates of mean risk in Table 3 to estimate the number of defendants with and without misconduct potential (column 1) as well as the number of such defendants that are released (column 2). In Panel A, column 6 we display the observational release rate disparity between white and Black defendants. In Panel B we use the same mean risk estimates to rescale this observational release rate comparison with the scaling factor described in the text. Column 3 of Panel B shows the scaling factor (Ω_i) given by these estimates, and column 6 shows the resulting disparate impact estimate.

Appendix Table A7—Mean Risk and Disparate Impact Estimates, Shrunk Leniency Estimates

	Linear Extrapolation	Quadratic Extrapolation	Local Linear Extrapolation
Panel A: Mean Risk by Race	(1)	(2)	(3)
White Defendants	0.342	0.368	0.358
	(0.008)	(0.036)	(0.014)
Black Defendants	0.403	0.436	0.441
	(0.007)	(0.026)	(0.014)
Panel B: System-Wide Disparate Impact			
Mean Across Cases	0.054	0.046	0.042
	(0.003)	(0.014)	(0.006)
Panel C: Judge-Level Disparate Impact			
Mean Across Judges	0.053	0.046	0.042
	(0.003)	(0.013)	(0.006)
Std. Dev. Across Judges	0.029	0.029	0.029
	(0.002)	(0.002)	(0.002)
Fraction Positive	0.963	0.938	0.920
	(0.011)	(0.075)	(0.037)
Judges	268	268	268

Notes. This table summarizes estimates of mean risk and disparate impact from different extrapolations of the variation in Figure 2, after applying conventional empirical Bayes shrinkage to the judge- and race-specific leniency estimates. Panel A reports estimates of race-specific average misconduct risk, Panel B reports estimates of system-wide (case-weighted) disparate impact, and Panel C reports empirical Bayes estimates of summary statistics for the judge-level disparate impact prior distribution. To estimate mean risk, column 1 uses a linear extrapolation of the variation in Figure 2, while column 2 uses a quadratic extrapolation and column 3 uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

APPENDIX TABLE A8—MEAN RISK AND DISPARATE IMPACT ESTIMATES, WITH COVARIATE ADJUSTMENT

	Linear Extrapolation	Quadratic Extrapolation	Local Linear Extrapolation
Panel A: Mean Risk by Race	(1)	(2)	(3)
White Defendants	0.351	0.334	0.352
	(0.007)	(0.018)	(0.013)
Black Defendants	0.394	0.412	0.423
	(0.006)	(0.021)	(0.016)
Panel B: System-Wide Disparate Impact			
Mean Across Cases	0.043	0.037	0.035
	(0.002)	(0.006)	(0.005)
Panel C: Judge-Level Disparate Impact			
Mean Across Judges	0.043	0.036	0.035
	(0.002)	(0.006)	(0.005)
Std. Dev. Across Judges	0.031	0.030	0.031
	(0.003)	(0.003)	(0.003)
Fraction Positive	0.923	0.891	0.878
	(0.017)	(0.042)	(0.036)
Judges	268	268	268

Notes. This table summarizes estimates of mean risk and disparate impact from different extrapolations of the variation in Figure 2, where release and misconduct rates adjust for both the court-by-time effects and the case and defendant observables in Table 2. Panel A reports estimates of race-specific average misconduct risk, Panel B reports estimates of system-wide (case-weighted) disparate impact, and Panel C reports empirical Bayes estimates of summary statistics for the judge-level disparate impact prior distribution. To estimate mean risk, column 1 uses a linear extrapolation, while column 2 uses a quadratic extrapolation and column 3 uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

Appendix Table A9—Disparate Impact and Judge Characteristics

	Full-Sample Disparate Impact					Split-Sample Disparate Impact	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
New Judge	-0.022				-0.011		-0.004
	(0.007)				(0.005)		(0.007)
Lenient Judge		-0.014			-0.018		-0.008
		(0.007)			(0.005)		(0.006)
Above-Median Black Share			-0.022		-0.007		0.003
			(0.007)		(0.008)		(0.009)
Manhattan Courtroom				0.062	0.058		0.053
				(0.008)	(0.007)		(0.011)
Bronx Courtroom				-0.003	-0.005		0.005
				(0.005)	(0.009)		(0.010)
Queens Courtroom				0.047	0.041		0.045
				(0.008)	(0.011)		(0.010)
Richmond Courtroom				0.028	0.021		0.047
				(0.011)	(0.008)		(0.017)
Lagged Disparate Impact						0.860	0.385
						(0.093)	(0.132)
Mean Disparate Impact	0.044	0.044	0.044	0.044	0.044	0.061	0.061
R2	0.059	0.027	0.066	0.452	0.508	0.294	0.428

Notes. This table reports OLS estimates of regressions of disparate impact estimates on judge characteristics. Disparate impact is estimated as described in Section IV, using the benchmark local linear estimate of mean risk. New judges are defined as judges appointed during our estimation period. Lenient judges are defined as judges with above-average leniency, controlling for court-by-time fixed effects. Courtroom locations are defined using the location of the modal case heard by each judge. Split-sample disparate impact estimates are computed by splitting each judge's sample of cases at the median case and constructing two samples, a before-median case sample and an after-median case sample. Disparate impact is then re-estimated within each subsample. The estimation procedure conditions on court-by-time effects, which causes a small number of judge effects to become collinear with the court-by-time effects and dropped. All specifications are weighted by the inverse variance of the disparate impact estimates. Columns 6 and 7 include empirical Bayes posteriors of lagged disparate impact, computed using a standard shrinkage procedure (Morris, 1983). Robust standard errors are reported in parentheses.

APPENDIX TABLE A10—MEAN RISK AND DISPARATE IMPACT ESTIMATES BY DEFENDANT CHARACTERISTICS

	Criminal History		Type of Arraignment Charge					
	Prior	No Prior	Felony	Misdemeanor	Drug	DUI	Property	Violent
Panel A: Mean Risk by Race	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
White Defendants	0.389	0.264	0.358	0.303	0.347	0.140	0.343	0.165
	(0.064)	(0.007)	(0.062)	(0.007)	(0.032)	(0.007)	(0.058)	(0.079)
Black Defendants	0.506	0.317	0.501	0.386	0.462	0.175	0.464	0.339
	(0.066)	(0.008)	(0.106)	(0.008)	(0.035)	(0.010)	(0.042)	(0.091)
Panel B: System-Wide Disparate Impact								
Mean Across Cases	0.026	0.024	0.030	0.046	0.055	0.024	0.010	0.107
	(0.020)	(0.002)	(0.035)	(0.003)	(0.009)	(0.005)	(0.022)	(0.397)
Panel C: Judge-Level Disparate Impact								
Mean Across Judges	0.026	0.024	0.030	0.046	0.058	0.024	0.006	0.105
	(0.020)	(0.002)	(0.035)	(0.003)	(0.009)	(0.005)	(0.022)	(0.389)
Std. Dev. Across Judges	0.038	0.014	0.034	0.035	0.052	0.000	0.036	0.031
	(0.009)	(0.004)	(0.016)	(0.003)	(0.007)	(0.011)	(0.015)	(0.237)
Fraction Positive	0.752	0.960	0.821	0.915	0.870	1.000	0.569	1.000
	(0.106)	(0.027)	(0.138)	(0.019)	(0.036)	(0.038)	(0.072)	(0.084)
Judges	263	264	261	264	258	174	222	219

Notes. This table summarizes estimates of mean risk and disparate impact by defendant characteristics separately. For each subgroup, we require that a judge observe at least 25 cases in order to be included in the sample. Therefore, the number of judges does vary across the columns depending on how many judges in the sample meet the requirement. Information on demographics and criminal outcomes is derived from court records as described in the text. Prior is an indicator equal to one if the defendant has a prior conviction. Estimates come from a local linear extrapolation of the variation in Figure 2, although unlike Figure 2, the extrapolations are done within the given characteristic. Panel A reports estimates of race-specific average misconduct risk, Panel B reports estimates of system-wide (case-weighted) disparate impact, and Panel C reports empirical Bayes estimates of summary statistics for the judge-level disparate impact prior distribution. To estimate mean risk, this table uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

Appendix Table A11—Mean Risk and Disparate Impact Bounds

	From 0.80 Leniency	From 0.85 Leniency	From 0.90 Leniency
Panel A: Mean Risk by Race	(1)	(2)	(3)
White Defendants	[0.221,0.421]	[0.248,0.398]	[0.277,0.377]
	(0.001,0.001)	(0.002,0.002)	(0.004,0.004)
Black Defendants	[0.280,0.480]	[0.313,0.463]	[0.349,0.449]
	(0.002,0.002)	(0.003,0.003)	(0.006,0.006)
Panel B: System-Wide Disparate Impact			
Mean Across Cases	[0.021,0.092]	[0.029,0.083]	[0.035,0.073]
	(0.003,0.002)	(0.002,0.001)	(0.002,0.001)
Panel C: Judge-Level Disparate Impact			
Mean Across Judges	[0.021,0.091]	[0.029,0.083]	[0.035,0.073]
	(0.003,0.002)	(0.003,0.002)	(0.002,0.002)
Std. Dev. Across Judges	[0.036,0.046]	[0.037,0.042]	[0.037,0.039]
	(0.003,0.004)	(0.003,0.004)	(0.003,0.005)
Fraction Positive	[0.694,0.989]	[0.770,0.982]	[0.821,0.975]
	(0.021,0.011)	(0.021,0.011)	(0.017,0.008)
Judges	268	268	268

Notes. This table summarizes bounds on mean risk and disparate impact estimated from the variation in Figure 2. Panel A reports bounds on race-specific average misconduct risk, Panel B reports bounds on system-wide (case-weighted) disparate impact, and Panel C reports bounds on empirical Bayes estimates of summary statistics for the judge-level disparate impact prior distribution. To estimate bounds on mean risk, column 1 uses a local linear fit of released misconduct rates among judges releasing 80% of white and Black defendants. Columns 2 and 3 form bounds from judges releasing 85% and 90% of white and Black defendants, respectively. The local linear regressions use a Gaussian kernel and a rule-of-thumb bandwidth. Bounds are formed under the assumption that either none or all of the detained defendants in each column have pretrial misconduct potential. Panels B and C search within these bounds to find the combination of white and Black mean risk that minimize or maximize each disparate impact statistic. Robust standard errors on the endpoints of each set of bounds, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

APPENDIX TABLE A12—DECOMPOSITION OF DISPARATE IMPACT BY MISCONDUCT POTENTIAL

	Linear Extrapolation	Quadratic Extrapolation	Local Linear Extrapolation
Panel A: Mean Risk by Race	(1)	(2)	(3)
White Defendants	0.338	0.319	0.346
	(0.007)	(0.021)	(0.016)
Black Defendants	0.400	0.394	0.436
	(0.006)	(0.022)	(0.016)
Panel B: Racial Disparity in Conditional on Misconduct Potential			
Δ_{j1}	0.033	0.060	0.066
	(0.016)	(0.054)	(0.037)
Δ_{j0}	0.066	0.050	0.027
	(0.011)	(0.038)	(0.030)
Judges	268	268	268

Notes. This table summarizes estimates of mean risk and racial disparities in true/false negative rates from different extrapolations of the variation in Figure 2. Panel A reports estimates of race-specific average misconduct risk and Panel B reports estimates of true/false negative rates. Δ_{j0} corresponds to defendants with $Y_i^* = 0$ while Δ_{j1} corresponds to defendants with $Y_i^* = 1$. To estimate mean risk, column 1 uses a linear extrapolation of the variation in Figure 2, while column 2 uses a quadratic extrapolation and column 3 uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

Appendix Table A13—Mean Risk and Disparate Impact Estimates, Borough-Specific Estimates

	Linear Extrapolation	Quadratic Extrapolation	Local Linear Extrapolation
Panel A: Mean Risk by Race	(1)	(2)	(3)
White Defendants	0.337	0.342	0.337
	(0.014)	(0.037)	(0.025)
Black Defendants	0.415	0.399	0.420
	(0.009)	(0.023)	(0.021)
Panel B: System-Wide Disparate Impact			
Mean Across Cases	0.050	0.052	0.046
	(0.002)	(0.008)	(0.007)
Panel C: Judge-Level Disparate Impact			
Mean Across Judges	0.042	0.048	0.040
	(0.003)	(0.008)	(0.007)
Std. Dev. Across Judges	0.032	0.040	0.039
	(0.003)	(0.008)	(0.007)
Fraction Positive	0.902	0.885	0.846
	(0.019)	(0.047)	(0.046)
Judges	267	267	267

Notes. This table summarizes estimates of mean risk and disparate impact. We estimate conditional regression models for each borough and averages the resulting estimates by borough share. Panel A reports estimates of race-specific average misconduct risk, Panel B reports estimates of system-wide (case-weighted) disparate impact, and Panel C reports empirical Bayes estimates of summary statistics for the judge-level disparate impact prior distribution. To estimate mean risk, column 1 uses a linear extrapolation of the variation in Figure 2, while column 2 uses a quadratic extrapolation and column 3 uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

APPENDIX TABLE A14—MEAN RISK AND DISPARATE IMPACT ESTIMATES, BOROUGH-SPECIFIC ESTIMATES WITH JUDGE-SPECIFIC TIME EFFECTS

Panel A: Mean Risk by Race	(1)	(2)	(3)	(4)	(5)	(6)
White Defendants	0.338	0.370	0.337	0.314	0.323	0.316
	(0.033)	(0.035)	(0.036)	(0.033)	(0.033)	(0.039)
Black Defendants	0.421	0.483	0.422	0.443	0.458	0.416
	(0.038)	(0.032)	(0.044)	(0.040)	(0.038)	(0.047)
Panel B: System-Wide Disparate Impact						
Mean Across Cases	0.046	0.027	0.045	0.027	0.037	0.033
	(0.031)	(0.021)	(0.045)	(0.082)	(0.039)	(0.053)
Judges	262	159	244	262	159	244
Judge x Year-Month	Yes	Yes	No	Yes	Yes	No
Judge x Year-Month Squared	No	Yes	No	No	Yes	No
Judge x Year, Judge x Month	No	No	Yes	No	No	Yes
With Race Interactions	No	No	No	Yes	Yes	Yes

Notes. This table summarizes estimates of mean risk and disparate impact. We estimate conditional regression models for each borough and averages the resulting estimates by borough share. The columns add different levels of judge-specific time effects as well as judge-specific time effects interacted with race. Panel A reports estimates of race-specific average misconduct risk, and Panel B reports estimates of system-wide (case-weighted) disparate impact. To estimate mean risk, each column uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth of the variation in Figure 2 which is estimated for each borough separately. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

APPENDIX TABLE A15—MEAN RISK AND DISPARATE IMPACT ESTIMATES, ALTERNATIVE MISCONDUCT OUTCOME

	Any Misconduct	Case FTA	Any Rearrest	Violent Rearrest
Panel A: Mean Risk by Race	(1)	(2)	(3)	(4)
White Defendants	0.346	0.176	0.233	0.014
	(0.014)	(0.011)	(0.019)	(0.004)
Black Defendants	0.436	0.242	0.314	0.014
	(0.017)	(0.014)	(0.019)	(0.006)
Panel B: System-Wide Disparate Impact				
Mean Across Cases	0.042	0.051	0.050	0.068
	(0.006)	(0.005)	(0.005)	(0.141)
Panel C: Judge-Level Disparate Impact				
Mean Across Judges	0.042	0.051	0.050	0.068
	(0.006)	(0.005)	(0.005)	(0.130)
Std. Dev. Across Judges	0.037	0.039	0.039	0.045
	(0.003)	(0.003)	(0.004)	(0.099)
Fraction Positive	0.873	0.913	0.910	0.948
	(0.036)	(0.025)	(0.027)	(0.089)
Judges	268	268	268	268

Notes. This table summarizes estimates of mean risk and disparate impact for different outcome variables. Panel A reports estimates of race-specific average misconduct risk, Panel B reports estimates of system-wide (case-weighted) disparate impact, and Panel C reports empirical Bayes estimates of summary statistics for the judge-level disparate impact prior distribution. Column 1 adjusts for differences by race in the mean risk of any misconduct (either rearrest or FTA). Column 2 adjusts for differences by race in the mean risk of FTA. Column 3 adjusts for differences by race in the mean risk of rearrest. Column 4 adjusts for differences by race in the mean risk of rearrest for a violent crime. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

APPENDIX TABLE A16—MEAN RISK AND DISPARATE IMPACT ESTIMATES, ALTERNATIVE JUDGE DECISIONS

	Linear Extrapolation	Quadratic Extrapolation	Local Linear Extrapolation
Panel A: Mean Risk by Race	(1)	(2)	(3)
White Defendants	0.343	0.341	0.345
	(0.007)	(0.026)	(0.031)
Black Defendants	0.405	0.415	0.447
	(0.006)	(0.022)	(0.039)
Panel B: System-Wide Disparate Impact			
Mean Across Cases	0.045	0.042	0.032
	(0.002)	(0.007)	(0.013)
Panel C: Judge-Level Disparate Impact			
Mean Across Judges	0.044	0.042	0.032
	(0.003)	(0.007)	(0.012)
Std. Dev. Across Judges	0.043	0.043	0.043
	(0.004)	(0.004)	(0.004)
Fraction Positive	0.855	0.838	0.769
	(0.017)	(0.041)	(0.082)
Judges	268	268	268

Notes. This table summarizes estimates of mean risk and disparate impact from different extrapolations of the variation in Figure 2. The judge's decision variable in this table is release on recognizance (ROR) versus the assignment of any monetary bail, where there is a 5.8 percentage point disparity in the assignment of ROR between white and Black defendants after controlling for court-by-time effects. Panel A reports estimates of race-specific average misconduct risk, Panel B reports estimates of system-wide (case-weighted) disparate impact, and Panel C reports empirical Bayes estimates of summary statistics for the judge-level disparate impact prior distribution. To estimate mean risk, column 1 uses a linear extrapolation of the variation in Figure 2, while column 2 uses a quadratic extrapolation and column 3 uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

Appendix Table A17—Mean Risk and Disparate Impact Estimates, Alternative Race Definition

	Linear Extrapolation	Quadratic Extrapolation	Local Linear Extrapolation
Panel A: Mean Risk by Race	(1)	(2)	(3)
White Defendants	0.208	0.138	0.187
	(0.009)	(0.018)	(0.014)
Black or Hispanic Defendants	0.393	0.419	0.415
	(0.006)	(0.019)	(0.012)
Panel B: System-Wide Disparate Impact			
Mean Across Cases	0.089	0.213	0.112
	(0.007)	(0.031)	(0.017)
Panel C: Judge-Level Disparate Impact			
Mean Across Judges	0.090	0.211	0.112
	(0.007)	(0.030)	(0.016)
Std. Dev. Across Judges	0.000	0.000	0.000
	(0.007)	(0.020)	(0.016)
Fraction Positive	1.000	1.000	1.000
	(0.018)	(0.004)	(0.016)
Judges	250	250	250

Notes. This table summarizes estimates of mean risk and disparate impact from different extrapolations of the variation in Figure 2. The racial comparison in this table is between Black or Hispanic defendants to non-Hispanic white defendants, where there is a 8.4 percentage point release rate disparity after adjusting for court-by-time effects. Panel A reports estimates of race-specific average misconduct risk, Panel B reports estimates of system-wide (case-weighted) disparate impact, and Panel C reports empirical Bayes estimates of summary statistics for the judge-level disparate impact prior distribution. To estimate mean risk, column 1 uses a linear extrapolation of the variation in Figure 2, while column 2 uses a quadratic extrapolation and column 3 uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

Appendix Table A18—Hierarchical MTE Model Hyperparameter Estimates

	White Defendants			Black Defendants		
	(1)	(2)	(3)	(4)	(5)	(6)
Mean Misconduct Risk (μ)	0.346	0.391	0.371	0.423	0.441	0.437
	(0.008)	(0.007)	(0.014)	(0.009)	(0.007)	(0.016)
Mean ln(Signal Quality) (α)	0.538	0.316	0.523	-0.038	-0.044	-0.080
	(0.128)	(0.074)	(0.125)	(0.146)	(0.075)	(0.104)
Mean Release Threshold (γ)	0.912	1.055	1.144	0.893	1.072	1.089
	(0.045)	(0.023)	(0.080)	(0.051)	(0.034)	(0.079)
Release Threshold Std. Dev. (δ)	0.369	0.109	0.149	0.417	0.194	0.203
	(0.039)	(0.011)	(0.037)	(0.052)	(0.021)	(0.049)
ln(Signal Quality) Std. Dev. (ψ)		0.140	0.134		0.166	0.151
		(0.019)	(0.016)		(0.014)	(0.013)
Regression of ln(Signal Quality) on Release Threshold (β)			-0.376			-0.007
			(0.153)			(0.212)
Judges	268	268	268	268	268	268

Notes. This table reports simulated minimum distance estimates of the MTE model described in the text. 500 simulation draws are used. Columns 3 and 6 estimate the full model with all hyperparameters. Columns 2 and 5 restrict $\beta = 0$, while columns 1 and 4 also restrict $\psi = 0$. The baseline model used in the text and summarized in Table 4 comes from columns 2 and 5 of this table. Robust standard errors, two-way clustered at the individual and the judge level, are reported in parentheses.

Appendix Table A19—Tests of Conventional MTE Monotonicity

	Number of Spline Knots			
	5	10	15	20
Panel A: White Defendants	(1)	(2)	(3)	(4)
Test Statistic	303.8	303.5	303.4	303.3
Deg. of Freedom	260	255	250	245
p-value	[0.032]	[0.020]	[0.012]	[0.007]
Cases	284,598	284,598	284,598	284,598
Panel B: Black Defendants				
Test Statistic	403.8	402.9	402.8	402.3
Deg. of Freedom	260	255	250	245
p-value	[<0.001]	[<0.001]	[<0.001]	[<0.001]
Cases	310,588	310,588	310,588	310,588

Notes. This table reports the results of the tests of conventional MTE monotonicity proposed by Frandsen et al. (2019), computed separately by defendant race. Test statistics are based on quadratic b-spline estimates of the relationship between misconduct outcomes and judge leniency, with the number of knots specified in each column, controlling for court-by-time fixed effects.

Appendix Table A20—Disparate Impact and Judge Characteristics, Model-Based Mean Risk

	Full-Sample Disparate Impact					Split-Sample Disparate Impact	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
New Judge	-0.023				-0.011		-0.002
	(0.007)				(0.005)		(0.007)
Lenient Judge		-0.015			-0.019		-0.011
		(0.008)			(0.005)		(0.006)
Above-Median Black Share			-0.021		-0.007		0.003
			(0.007)		(0.008)		(0.009)
Manhattan Courtroom				0.060	0.056		0.046
				(0.009)	(0.008)		(0.011)
Bronx Courtroom				-0.004	-0.005		-0.003
				(0.006)	(0.009)		(0.011)
Queens Courtroom				0.045	0.040		0.036
				(0.008)	(0.011)		(0.011)
Richmond Courtroom				0.025	0.018		0.039
				(0.010)	(0.009)		(0.014)
Lagged Disparate Impact						0.733	0.395
						(0.087)	(0.126)
Mean Disparate Impact	0.050	0.050	0.050	0.050	0.050	0.050	0.050
R2	0.061	0.032	0.063	0.435	0.499	0.308	0.420
Judges	268	268	268	268	268	252	252

Notes. This table reports OLS estimates of regressions of disparate impact estimates on judge characteristics. Disparate impact is estimated as described in Section IV, using the hierarchical MTE model estimate of mean risk. New judges are defined as judges appointed during our estimation period. Lenient judges are defined as judges with above-average leniency, controlling for court-by-time fixed effects. Courtroom locations are defined using the location of the modal case heard by each judge. Split-sample disparate impact estimates are computed by splitting each judge's sample of cases at the median case and constructing two samples, a before-median case sample and an after-median case sample. Disparate impact is then re-estimated within each subsample. The estimation procedure conditions on court-by-time effects, which causes a small number of judge effects to become collinear with the court-by-time effects and dropped. All specifications are weighted by the inverse variance of the disparate impact estimates. Columns 6 and 7 include empirical Bayes posteriors of lagged disparate impact, computed using a standard shrinkage procedure (Morris, 1983). Robust standard errors are reported in parentheses.

APPENDIX TABLE A21—RACIAL BIAS AND JUDGE CHARACTERISTICS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
New Judge	-0.021				-0.023		-0.007
	(0.009)				(0.007)		(0.005)
Lenient Judge		0.023			0.017		0.033
		(0.007)			(0.005)		(0.003)
Above-Median Black Share			-0.008		-0.013		-0.002
			(0.007)		(0.008)		(0.005)
Manhattan Courtroom				0.052	0.044		-0.006
				(0.008)	(0.008)		(0.006)
Bronx Courtroom				-0.016	-0.027		-0.015
				(0.007)	(0.010)		(0.006)
Queens Courtroom				0.038	0.023		-0.007
				(0.009)	(0.011)		(0.008)
Richmond Courtroom				0.037	0.019		-0.010
				(0.007)	(0.009)		(0.014)
Disparate Impact						1.369	1.403
						(0.086)	(0.085)
Mean Bias	0.072	0.072	0.072	0.072	0.072	0.072	0.072
R2	0.026	0.053	0.007	0.332	0.397	0.646	0.770
Judges	268	268	268	268	268	268	268

Notes. This table reports OLS estimates of regressions of racial bias estimates on judge characteristics. Bias estimates are obtained from the heirarchical MTE model as described in Section V. New judges are defined as judges appointed during our estimation period. Lenient judges are defined as judges with above-average leniency, controlling for court-by-time fixed effects. Courtroom locations are defined using the location of the modal case heard by each judge. All specifications are weighted by the inverse variance of the racial bias posteriors. Robust standard errors are reported in parentheses.

Appendix Table A22—Signal Quality Differences and Judge Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
New Judge	-0.092				-0.073		-0.014
	(0.029)				(0.022)		(0.012)
Lenient Judge		0.031			0.016		0.074
		(0.020)			(0.016)		(0.009)
Above-Median Black Share			-0.031		-0.029		-0.006
			(0.020)		(0.024)		(0.013)
Manhattan Courtroom				0.172	0.153		-0.001
				(0.023)	(0.025)		(0.016)
Bronx Courtroom				-0.042	-0.062		-0.044
				(0.022)	(0.030)		(0.016)
Queens Courtroom				0.120	0.090		-0.018
				(0.028)	(0.034)		(0.021)
Richmond Courtroom				0.117	0.081		-0.050
				(0.023)	(0.029)		(0.037)
Disparate Impact						4.575	4.584
						(0.197)	(0.215)
Mean Difference	0.412	0.412	0.412	0.412	0.412	0.412	0.412
R2	0.055	0.011	0.010	0.338	0.379	0.738	0.812
Judges	268	268	268	268	268	268	268

Notes. This table reports OLS estimates of regressions of differences in signal quality estimates on judge characteristics. Signal quality estimates are obtained from the heirarchical MTE model as described in Section V. New judges are defined as judges appointed during our estimation period. Lenient judges are defined as judges with above-average leniency, controlling for court-by-time fixed effects. Courtroom locations are defined using the location of the modal case heard by each judge. All specifications are weighted by the inverse variance of the signal quality difference posteriors. Robust standard errors are reported in parentheses.

Econometric Appendix

B1. Defining and Measuring Disparate Impact with Multi-Valued Y_i^*

This appendix first generalizes our definition of disparate impact and derivation of OVB in observational comparisons to settings where the decision-maker's objective is non-binary. We then discuss how our quasi-experimental framework for measuring disparate impact extends to this case.

Natural generalizations of Equation (3) are given by

$$\Delta_j = \sum_{y \in Supp(Y_i^*)} \left(\delta_{jw}^y - \delta_{jb}^y\right) p_y \tag{B1}$$

in the multi-valued Y_i^* case, where $p_y = Pr(Y_i^* = y)$, and:

$$\Delta_j = \int_{Supp(Y_i^*)} \left(\delta_{jw}^y - \delta_{jb}^y\right) dF(y) \tag{B2}$$

in the case of continuous Y_i^*, where $F(\cdot)$ is the cumulative distribution function of Y_i^*. In both cases, $\delta_{jr}^y = E[D_{ij} \mid Y_i^* = y, R_i = r]$ gives conditional release rates for each race r and each $y \in Supp(Y_i^*)$.

As in Section II.B, the bias of observational benchmarking regressions relative to these parameters, when judges are as-good-as-randomly assigned, is given by

$$\begin{aligned} \xi_j &= \sum_{y \in Supp(Y_i^*)} \delta_{jw}^y p_{yw} - \sum_{y \in Supp(Y_i^*)} \delta_{jb}^y p_{yb} - \sum_{y \in Supp(Y_i^*)} \left(\delta_{jw}^y - \delta_{jb}^y\right)(p_{yw} p_w + p_{yb} p_b) \\ &= \sum_{y \in Supp(Y_i^*)} \left(\delta_{jw}^y p_b + \delta_{jb}^y p_w\right)(p_{yw} - p_{yb}) \end{aligned} \tag{B3}$$

in the multi-valued Y_i^* case, where $p_{yr} = Pr(Y_i^* = y \mid R_i = r)$ and again $p_r = Pr(R_i = r)$, and:

$$\begin{aligned} \xi_j &= \int_{Supp(Y_i^*)} \delta_{jw}^y dF_w(y) - \int_{Supp(Y_i^*)} \delta_{jb}^y dF_b(y) - \int_{Supp(Y_i^*)} \left(\delta_{jw}^y - \delta_{jb}^y\right) d(F_w(y) p_w + F_b(y) p_b) \\ &= \int_{Supp(Y_i^*)} \left(\delta_{jw}^y p_b + \delta_{jb}^y p_w\right) d(F_w(y) - F_b(y)) \end{aligned} \tag{B4}$$

in the case of continuous Y_i^*, where $F_r(\cdot)$ is the cumulative distribution function of Y_i^* given $R_i = r$.

As in Section IV, disparate impact is identified by the distribution of misconduct outcomes Y_i^* within each race when judges are quasi-randomly assigned. By Bayes' law:

$$\delta_{jr}^y = Pr(Y_i^* = y \mid D_{ij} = 1, R_i = r) \frac{E[D_{ij} \mid R_i = r]}{Pr(Y_i^* = y \mid R_i = r)} \tag{B5}$$

for multi-valued Y_i^* and similarly for continuous Y_i^*. The first two terms, $Pr(Y_i^* = y \mid D_{ij} = 1, R_i = r)$ and $E[D_{ij} \mid R_i = r]$, are identified by $Pr(Y_i = y \mid D_i = 1, Z_{ij} = 1, R_i = r)$ and $E[D_i \mid Z_{ij} = 1, R_i = r]$ under quasi-random judge assignment as before. In the continuous Y_i^* case, the first term is given by the conditional density of Y_i^* given $D_i = 1$, $Z_{ij} = 1$, and $R_i = r$. Estimates of the race-specific misconduct distribution corresponding to the third $Pr(Y_i^* = y \mid R_i = r)$ term (which might be obtained from similar extrapolations of quasi-experimental data as in the binary Y_i^* case) thus yield a plug-in estimator of each δ_{jr}^y, which can be combined to estimate Δ_j.

B2. *Included Variables Bias*

This appendix derives the included variables bias (IVB) formula (9) in a conditional release rate comparison that adjusts for a binary characteristic X_i:

$$\tilde{\Delta}_j = \tilde{\Delta}_{j,X=0}(1-\bar{\mu}^X) + \tilde{\Delta}_{j,X=1}\bar{\mu}^X, \tag{B6}$$

where $\tilde{\Delta}_{j,X=x} = \delta_{jw,X=x} - \delta_{jb,X=x}$ and $\bar{\mu}^X = E[X_i] = \mu_b^X p_b + \mu_w^X p_w$. Since we assume here that white and Black misconduct risk are equal, $\mu_w = \mu_b$, we have no OVB, and:

$$\Delta_j = \alpha_j = \left(\delta_{jw,X=0}(1-\mu_w^X) + \delta_{jw,X=1}\mu_w^X\right) - \left(\delta_{jb,X=0}(1-\mu_b^X) + \delta_{jb,X=1}\mu_b^X\right). \tag{B7}$$

It thus follows similarly to Equation (8) that:

$$\begin{aligned}\tilde{\Delta}_j - \Delta_j &= \left(\delta_{jw,X=0}(\bar{\mu}^X - \mu_w^X) + \delta_{jw,X=1}(\mu_w^X - \bar{\mu}^X)\right) - \left(\delta_{jb,X=0}(\bar{\mu}^X - \mu_b^X) + \delta_{jb,X=1}(\mu_b^X - \bar{\mu}^X)\right) \\ &= \left[\left(\delta_{jw,X=0} - \delta_{jw,X=1}\right)p_b + \left(\delta_{jb,X=0} - \delta_{jb,X=1}\right)p_w\right] \times (\mu_b^X - \mu_w^X). \end{aligned} \tag{B8}$$

B3. *Empirical Bayes Methods*

This appendix summarizes the two conventional empirical Bayes approaches used in this paper: the posterior mean calculation of Morris (1983) and the posterior average effect calculation of Bonhomme and Weidner (2020). We use the former to gauge sensitivity of our main extrapolations in Appendix Table A7 (see footnote 19), and to compute the prior means and standard deviations in Figures 1, 3, and A3. We use the latter to compute the posterior distribution and fraction of judges with positive disparities in these figures, and to interpret the coefficient estimates in Tables A9, A20, A21, and A22.

Let $\hat{\theta}_j$ be an estimate of an unknown judge-specific parameter θ_j, such as an observational benchmarking coefficient or our rescaled disparate impact measure. Applying a usual asymptotic approximation, we write $\hat{\theta}_j = \theta_j + \varepsilon_j$ where $\varepsilon_j \sim N(0, \Sigma_j)$ for known Σ_j. Conventional empirical Bayes methods further assume $\theta_j \sim N(\bar{\theta}, \Lambda)$, where $\bar{\theta}$ and Λ are unknown hyperparameters. Given this prior distribution, the posterior mean of θ_j after observing the estimate $\hat{\theta}_j$ is given by

$$\theta_j^* \equiv E[\theta_j \mid \hat{\theta}_j] = \frac{\Sigma_j}{\Lambda + \Sigma_j}\bar{\theta} + \frac{\Lambda}{\Lambda + \Sigma_j}\hat{\theta}_j \tag{B9}$$

More generally, Equation (B9) gives the minimum mean-squared error prediction of θ_j given $\hat{\theta}_j$ when the normality of θ_j is relaxed, provided $\bar{\theta}$ and Λ continue to parameterize the mean and variance of the prior distribution.

Empirical Bayes posteriors estimate $\bar{\theta}$ and Λ and plug these hyperparameter estimates into Equation (B9). We estimate $\bar{\theta}$ and Λ by the weighted iterative procedure studied by (Morris, 1983), which is equivalent to a maximum likelihood procedure. At iteration k the hyperparameter estimates are:

$$\hat{\hat{\theta}}_k = \sum_j \frac{\omega_{jk}}{\sum_{j'} \omega_{j'k}} \hat{\theta}_j \tag{B10}$$

$$\hat{\Lambda}_k = \sum_j \frac{\omega_{jk}}{\sum_{j'} \omega_{j'k}} \left((\hat{\theta}_j - \hat{\hat{\theta}}_k)^2 - \Sigma_j \right) \tag{B11}$$

with inverse-variance weights that are proportional to $\omega_{jk} = (\hat{\Lambda}_{k-1} + \Sigma_j)^{-1}$ and where $\omega_{j0} = 1$. We iterate this procedure to convergence.

Bonhomme and Weidner (2020) discuss posterior average effect estimators of the cumulative distribution function for θ_j, given by

$$\hat{F}_\theta(t) = \frac{1}{J} \sum_j E[\mathbf{1}[\theta_j \leq t] \mid \hat{\theta}_j] \tag{B12}$$

for each t in the support of θ_j. Note that $1 - \hat{F}_\theta(0)$ is a posterior average effect estimate of the fraction of θ_j in the population that is positive. Under the normality assumption:

$$E[\mathbf{1}[\theta_j \leq t] \mid \hat{\theta}_j] = \Phi\left(-\frac{\theta_j^*}{\sqrt{\frac{\Lambda \Sigma_j}{\Lambda + \Sigma_j}}} \right) \tag{B13}$$

which can, as with Equation (B9), be estimated by plugging in the estimates of the mean and variance hyperparameters. Just as with the empirical Bayes posterior estimator, Bonhomme and Weidner (2020) show that this posterior average effect estimator has certain robustness properties: it is optimal in terms of local worst-case bias, and its global bias is bounded by the minimum worst-case bias within a large class of estimators. They further show how regressions of the empirical Bayes posterior means on judge characteristics also have a posterior average effect interpretation and thus the same robustness properties for estimating conditional mean functions.

To estimate the density of θ_j as posterior average effects, we consider

$$\hat{f}_\theta(t) = \frac{1}{J} \sum_j E\left[\frac{1}{h} K\left(\frac{t - \theta_j}{h} \right) \mid \hat{\theta}_j \right] \tag{B14}$$

where $K(\cdot)$ is a kernel function and h is a bandwidth. For the posterior densities in Figures 1, 3, and A3 we use an Epanechnikov kernel, $K(u) = \frac{3}{4}(1 - u^2)\mathbf{1}[|u| \leq 1]$, and a rule-of-thumb bandwidth. To compute $\hat{f}_\theta(t)$, we note that under the reference model (i.e., normality)

$$Pr\left(\left| \frac{t - \theta_j}{h} \right| \leq 1 \mid \hat{\theta}_j \right) = \Phi\left(\frac{t + h - \theta_j^*}{\sqrt{\frac{\Lambda \Sigma_j}{\Lambda + \Sigma_j}}} \right) - \Phi\left(\frac{t - h - \theta_j^*}{\sqrt{\frac{\Lambda \Sigma_j}{\Lambda + \Sigma_j}}} \right) \tag{B15}$$

$$E\left[\theta_j \mid \left|\frac{t-\theta_j}{h}\right| \leq 1, \hat{\theta}_j\right] = \theta_j^* + \sqrt{\frac{\Lambda\Sigma_j}{\Lambda+\Sigma_j}} \frac{\phi\left(\frac{t-h-\theta_j^*}{\sqrt{\frac{\Lambda\Sigma_j}{\Lambda+\Sigma_j}}}\right) - \phi\left(\frac{t+h-\theta_j^*}{\sqrt{\frac{\Lambda\Sigma_j}{\Lambda+\Sigma_j}}}\right)}{\Phi\left(\frac{t+h-\theta_j^*}{\sqrt{\frac{\Lambda\Sigma_j}{\Lambda+\Sigma_j}}}\right) - \Phi\left(\frac{t-h-\theta_j^*}{\sqrt{\frac{\Lambda\Sigma_j}{\Lambda+\Sigma_j}}}\right)} \tag{B16}$$

and

$$\begin{aligned} Var\left(\theta_j \mid \left|\frac{t-\theta_j}{h}\right| \leq 1, \hat{\mu}_j\right) =& \frac{\Lambda\Sigma_j}{\Lambda+\Sigma_j}\left(1 + \frac{\left(\frac{t-h-\theta_j^*}{\sqrt{\frac{\Lambda\Sigma_j}{\Lambda+\Sigma_j}}}\right)\phi\left(\frac{t-h-\theta_j^*}{\sqrt{\frac{\Lambda\Sigma_j}{\Lambda+\Sigma_j}}}\right) - \left(\frac{t+h-\theta_j^*}{\sqrt{\frac{\Lambda\Sigma_j}{\Lambda+\Sigma_j}}}\right)\phi\left(\frac{t+h-\theta_j^*}{\sqrt{\frac{\Lambda\Sigma_j}{\Lambda+\Sigma_j}}}\right)}{\Phi\left(\frac{t+h-\theta_j^*}{\sqrt{\frac{\Lambda\Sigma_j}{\Lambda+\Sigma_j}}}\right) - \Phi\left(\frac{t-h-\theta_j^*}{\sqrt{\frac{\Lambda\Sigma_j}{\Lambda+\Sigma_j}}}\right)}\right) \\ & - \frac{\Lambda\Sigma_j}{\Lambda+\Sigma_j}\left(\frac{\phi\left(\frac{t-h-\theta_j^*}{\sqrt{\frac{\Lambda\Sigma_j}{\Lambda+\Sigma_j}}}\right) - \phi\left(\frac{t+h-\theta_j^*}{\sqrt{\frac{\Lambda\Sigma_j}{\Lambda+\Sigma_j}}}\right)}{\Phi\left(\frac{t-h-\theta_j^*}{\sqrt{\frac{\Lambda\Sigma_j}{\Lambda+\Sigma_j}}}\right) - \Phi\left(\frac{t+h-\theta_j^*}{\sqrt{\frac{\Lambda\Sigma_j}{\Lambda+\Sigma_j}}}\right)}\right)^2 . \end{aligned} \tag{B17}$$

We compute these by again plugging in estimates of the mean and variance hyperparameters, and use

$$\begin{aligned} E\left[\frac{1}{h}K\left(\frac{t-\theta_j}{h}\right) \mid \hat{\theta}_j\right] =& \frac{3}{4h}Pr\left(\left|\frac{t-\theta_j}{h}\right| \leq 1 \mid \hat{\theta}_j\right)\left(1 - \frac{t^2 - E\left[\theta_j \mid \left|\frac{t-\theta_j}{h}\right| \leq 1, \hat{\theta}_j\right]^2 - Var\left(\theta_j \mid \left|\frac{t-\theta_j}{h}\right| \leq 1, \hat{\theta}_j\right)}{h^2}\right) \\ & + \frac{3t}{2h^3}Pr\left(\left|\frac{t-\theta_j}{h}\right| \leq 1 \mid \hat{\theta}_j\right) E\left[\theta_j \mid \left|\frac{t-\theta_j}{h}\right| \leq 1, \hat{\theta}_j\right] \end{aligned} \tag{B18}$$

to compute the posterior density.

B4. Rescaled Benchmarking Regressions: Numerical Example and in NYC

This appendix illustrates how our rescaling approach allows us to measure disparate impact in bail decisions, even though misconduct potential is unobserved and cannot be directly conditioned on. We first consider a simple numerical example. Suppose that there are two types of hypothetical defendants assigned to a single bail judge: high-risk H types and low-risk L types. 75 of the 100 H-type defendants have misconduct potential ($Y_i^* = 1$) but only 25 of the 100 L-type defendants have misconduct potential, such that $\mu_H = 0.75$, $\mu_L = 0.25$, and $p_H = p_L = 0.5$. The judge is type-neutral when making release decisions: if the defendant has $Y_i^* = 1$ there is an 80 percent chance the defendant is released regardless of type, and if the defendant has $Y_i^* = 0$ there is a 20 percent chance the defendant is released regardless of type. Thus, while the judge receives a signal of the defendant's unobserved misconduct potential, this signal is not perfectly predictive, implying the judge will release some defendants with pretrial misconduct potential and detain some defendants without pretrial misconduct potential.

Appendix Table A5 summarizes the setup. Panel A shows that this judge has a release rate of 0.65 for L-type defendants but a release rate of 0.35 for H-type defendants. This means that a simple benchmarking regression would suffer from OVB: it finds that L-type defendants have a 30 percentage point higher release rate than H-type defendants ($\alpha_j = 0.3$), despite the judge being type-neutral.

Panel B of Appendix Table A5 shows how discrimination can be measured in this simple numerical example with observational release rate comparisons that are rescaled using average misconduct risk. Following Equations (14) and (15), we compute $\Omega_i = \frac{0.50}{0.75} = 2/3$ for released H-type defendants with $Y_i = 0$ and released L-type defendants with $Y_i = 1$, and $\Omega_i = \frac{0.50}{0.25} = 2$ for released L-type defendants with $Y_i = 1$ and released H-type defendants with $Y_i = 0$. The rescaling factor thus up-weights the release rates of individuals who are relatively less common in each type (risky L-type defendants and non-risky H-type defendants), while down-weighting the release rates of individuals who are relatively more common (non-risky L-type defendants and risky H-type defendants). This pattern of up- and down-weighting generally arises when H-type defendants have higher misconduct risk: i.e., when $\mu_H > \bar{\mu} > \mu_L$. In such cases, observations of released L-type defendants who subsequently offend are up-weighted ($Y_i - \mu_L > 0$ and $\bar{\mu} - \mu_L > 0$ so $\Omega_i > 1$), as are observations of released H-type defendants who do not subsequently offend ($Y_i - \mu_H < 0$ and $\bar{\mu} - \mu_H < 0$, so again $\Omega_i > 1$.

The rescaling factor removes OVB by implicitly equalizing the proportion of risky and non-risky defendants by type. This means that a rescaled benchmarking regression correctly find that H- and L-type defendants with the same misconduct potential have identical release rates ($\Delta_j = 0$). This is shown in the final column of Appendix Table A5, Panel B.

Appendix Table A6 similarly illustrates our finding of significant disparate impact in NYC bail decisions. We use the benchmark local linear estimates of mean risk in Table 3 to estimate the number of white and Black defendants with and without misconduct potential in column 1 of Panel A. In column 2, we combine these estimates with estimates of release and released misconduct rates adjusted by court-by-time fixed effects to compute the number of released defendants in each race and misconduct category, as in Equation (19). This calculation yields the case-weighted average observational disparity of 6.8 percentage points in column 6. In Panel B, we use the local linear estimates of mean risk to compute and apply the appropriate rescaling factor Ω_i. Our baseline estimates of average misconduct risk are $\mu_w = 0.346$ for white defendants and $\mu_b = 0.436$ for Black defendants. Combining these estimates with the share of white and Black defendants in our sample yields an overall average misconduct risk of $\bar{\mu} = 0.392$. Following Equations (14) and (15), these estimates yield a rescaling factor of $\Omega_i = \frac{1-0.392}{1-0.346} = 0.928$ for released white defendants with $Y_i = 0$, $\Omega_i = \frac{0.392}{0.436} = 0.901$ for released Black defendants with $Y_i = 1$, $\Omega_i = \frac{0.392}{0.346} = 1.137$ for released white defendants with $Y_i = 1$, and $\Omega_i = \frac{1-0.392}{1-0.436} = 1.077$ for released Black defendants with $Y_i = 0$. Thus the rescaling factor up-weights the release rates of risky white defendants and non-risky Black defendants (who are relatively less common) while down-weighting the release rates of non-risky white defendants and risky Black defendants (who are relatively more common). Applying these rescaling factors to the observational release rates yields a system-wide disparate impact estimate of 4.2 percentage points, matching the estimate in Panel B of Table 3.

B5. Bounding Mean Risk and Racial Discrimination

This appendix details the construction of mean risk and disparate impact bounds in Appendix Table A11. As in the baseline analysis, the procedure uses estimates of race- and judge-specific release rates $\rho_{jr} = E[D_{ij} \mid R_i = r]$ and released misconduct rates $\lambda_{jr} = E[Y_i^* \mid D_{ij} = 1, R_i = r]$. Instead of extrapolating the latter estimates to estimate the mean risk parameters μ_{jr}, and the corresponding estimates of disparate impact Δ_j, we bound the range of logically possible μ_{jr} given typical misconduct rates of highly lenient judges and search within these ranges to bound statistics of the prior distribution of disparate impact.

Each column of Appendix Table A11 forms bounds from a different leniency threshold $\overline{\rho} \in \{0.8, 0.85, 0.9\}$. For each race r, we first use a local linear regression of the estimated λ_{jr} on the estimated ρ_{jr} to estimate the average λ_{jr} for judges with $\rho_{jr} = \overline{\rho}$, parameters we denote by $\overline{\lambda}_r$. By definition, each $\overline{\lambda}_r$ bounds the mean risk of race r as

$$\mu_r \in [\overline{\lambda}_r\overline{\rho}, \overline{\lambda}_r\overline{\rho} + (1 - \overline{\rho})]. \tag{B19}$$

The lower bound $\overline{\lambda}_r\overline{\rho}$ is obtained from assuming all detained defendants for a judge with a leniency of $\overline{\rho}$ have $Y_i^* = 0$ while the upper bound is obtained from assuming the $(1 - \overline{\rho})$ share of detained defendants have pretrial misconduct potential ($Y_i^* = 1$). Panel A of Appendix Table A11 reports estimates of these bounds for each race, along with their associated standard errors in parentheses. Note that by construction the width of each interval is equal to $1 - \overline{\rho}$.

To obtain bounds on the statistics in Panels B and C of Appendix Table A11, we perform grid searches within the mean risk bounds in Panel A. For example, to bound the system-wide level of discrimination we search within the mean risk bounds to find the (μ_w, μ_b) pair that minimizes and maximizes the case-weighted average of judge-specific disparate impact Δ_j. We report these bounds and their associated standard errors in parentheses. Note that the width of each statistic's interval is weakly increasing in $1 - \overline{\rho}$, reflecting the increase in the range of mean risk parameters.

B6. Judge Decision-Making Model and Extensions

This appendix first derives the specific form of the posterior function $p_j(\cdot)$ in the model discussed in Section V.A. We then show how equivalent models are obtained when judges have inaccurate beliefs over the risk of white and Black defendants, and when judges minimize race-specific costs of misconduct classification errors. Finally, we show how disparate impact manifests in this model.

The initial model assumes judges form accurate posteriors of defendant misconduct potential Y_i^* after observing noisy signals $\nu_{ij} = Y_i^* + \eta_{ij}$ with normally distributed noise: $\eta_{ij} \mid Y_i^*, (R_i = r) \sim N(0, 1/\tau_{jr}^2)$. The distribution of these posteriors is given by Bayes' rule as:

$$\begin{aligned} p_j(\nu; r) &\equiv Pr(Y_i^* = 1 \mid \nu_{ij} = \nu, R_i = r) \\ &\frac{Pr(\nu_{ij} = \nu \mid Y_i^* = 1, R_i = r)Pr(Y_i^* = 1, R_i = r)}{Pr(\nu_{ij} = \nu, R_i = r)} \\ &= \frac{\phi(\tau_{jr}(\nu - 1))\tau_{jr}\mu_r}{\phi(\tau_{jr}(\nu - 1))\tau_{jr}\mu_r + \phi(\tau_{jr}\nu)\tau_{jr}(1 - \mu_r)} \end{aligned} \tag{B20}$$

where $\phi(x) \propto \exp(-x^2/2)$ is the standard normal density and $\mu_r = E[Y_i^* \mid R_i = r]$ is the mean risk of race

r. Simplifying this expression yields:

$$p_j(\nu;r) = \left(1+\exp(\tau_{jr}^2(1-2\nu)/2)\frac{1-\mu_r}{\mu_r}\right)^{-1} \tag{B21}$$

With π_{jr} giving the private benefits of releasing defendants of race r, the judge's release rule is then given by $D_{ij} = \mathbf{1}[\pi_{jR_i} \geq p_j(\nu_{ij};R_i)]$.

Equation (B21) shows that risk posteriors are strictly increasing in ν, such that they can be inverted to write the judge's release decision as a cutoff rule for her observed signals ν_{ij}:

$$D_{ij} = \mathbf{1}\left[\frac{1}{2} - \ln\left(\frac{\mu_{R_i}(1-\pi_{jR_i})}{(1-\mu_{R_i})\pi_{jR_i}}\right)/\tau_{jR_i}^2 \geq \nu_{ij}\right] \tag{B22}$$

We use this fact to parameterize the hierarchical model, as discussed in Section V.A. Here

$$\kappa_{jr} = \frac{1}{2} - \ln\left(\frac{\mu_r(1-\pi_{jr})}{(1-\mu_r)\pi_{jr}}\right)/\tau_{jr}^2. \tag{B23}$$

It follows from Equation (B23) that if judges form posteriors with inaccurate priors $\tilde{\mu}_{jr} \neq \mu_r$, this bias in beliefs cannot be distinguished from bias in the preference parameters π_{jr}. Only the index $I_{jr} = \frac{\tilde{\mu}_{jr}(1-\pi_{jr})}{\pi_{jr}(1-\tilde{\mu}_{jr})}$, which combines beliefs and preferences, is relevant to the judge's decision-making process. Consequently, the judge's marginal released outcomes

$$E[Y_i^* \mid p_j(\nu_{ij};r) = \pi_r, R_i = r] = \left(1 + I_{jr}\left(\frac{1-\mu_r}{\mu_r}\right)\right)^{-1} \tag{B24}$$

will generally differ by race when either $\tilde{\mu}_{jr} \neq \mu_r$ for one or both races (indicating inaccurate beliefs) or when $\pi_{jw} \neq \pi_{jb}$ (indicating racial animus).

An equivalent model is derived by assuming the judge minimizes the cost of making "false positive" decisions (detaining an individual with no pretrial misconduct risk) and "false negative" decisions (releasing an individual with pretrial misconduct risk), rather than having explicit benefits of releasing white and Black defendants. Denote these judge- and race-specific type-I and type-II error costs by $c_{jr}^I, c_{jr}^{II} > 0$. A judge's ex-post utility for a given release decision $D_{ij} \in \{0,1\}$ is then:

$$U_{ij} = -c_{jR_i}^{II} D_{ij} Y_i^* - c_{jR_i}^I (1-D_{ij})(1-Y_i^*) \tag{B25}$$

and her expected utility over her posterior risk beliefs is

$$E[U_{ij} \mid \nu_{ij}, R_i] = -c_{jR_i}^{II} D_{ij} p_j(\nu_{ij}, R_i) - c_{jR_i}^I (1-D_{ij})(1-p_j(\nu_{ij}, R_i)) \tag{B26}$$

The judge's expected utility is thus maximized by cutoff rule:

$$D_{ij} = \mathbf{1}[\pi_{jR_i} \geq p_j(\nu_{ij}, R_i)] \tag{B27}$$

where $\pi_{jr} = \frac{c^{II}_{jR_i}}{c^{I}_{jR_i}+c^{II}_{jR_i}} \in (0,1)$ gives the judge's relative cost of type-II error.

To characterize discrimination in this model, note that Equation (B22) and the conditional normality of ν_{ij} implies that the judge's conditional release rates can be written

$$\delta_{jr0} = Pr(D_{ij} = 1 \mid Y_i^* = 0, R_i = r) = \Phi\left(\frac{1}{2}\tau_{jr} - \frac{1}{\tau_{jr}}\ln I_{jr}\right) \tag{B28}$$

$$\delta_{jr1} = Pr(D_{ij} = 1 \mid Y_i^* = 1, R_i = r) = 1 - \Phi\left(\frac{1}{2}\tau_{jr} + \frac{1}{\tau_{jr}}\ln I_{jr}\right) \tag{B29}$$

When signal quality is the same by race, $\tau_{jw} = \tau_{jb}$, these expressions show that disparate impact $\Delta_j = (\delta_{jw0} - \delta_{jb0})(1-\bar{\mu}) + (\delta_{jw1} - \delta_{jb1})\bar{\mu}$ is only zero when $I_{jw} = I_{jb}$. By comparison with Equation (B24), this scenario will generally lead to bias at the margin unless white and Black average misconduct risk are also equal ($\mu_w = \mu_b$). Furthermore, the fact that Δ_j is strictly decreasing (to zero) in the white index I_{jw} and strictly increasing (to one) in the Black index I_{jb} implies that there exist a set of thresholds (I_{jw}, I_{jb}) resulting in no disparate impact on average, even when signal quality differs. Again, this will typically yield racial bias, per Equation (B24), to the extent mean risk differs by race.

B7. *Conventional Monotonicity Violations and Judge Signal Quality*

This appendix shows how differences in the way judges consider defendant and case characteristics, which lead to violations of conventional MTE monotonicity, can be viewed as differences in judge signal quality within models like the one we develop in Section V.A. In doing so we show that such models are without observational loss, provided judge release decisions are better-than-random.

Consider a setting with a binary potential misconduct outcome Y_i^* and a set of binary judge release decisions D_{ij}. The distribution of these random variables is fully specified by the mean risk $\mu = E[Y_i^*]$ and the conditional release rates $\delta_{j0} = E[D_{ij} \mid Y_i^* = 0]$ and $\delta_{j1} = E[D_{ij} \mid Y_i^* = 1]$. With mean risk fixed, any restriction on judicial decision-making —such as conventional MTE monotonicity or alternative parameterizations—can thus be understood as restricting the set of $(\delta_{j0}, \delta_{j1})$.

We first show that when judges are making better-than-random release decisions, in the sense of $0 < \delta_{j0} < \delta_{j1} < 1$ for each j, it is without observational loss to assume a decision-making model of $D_{ij} = \mathbf{1}[\kappa_j \geq Y_i^* + \eta_i/\tau_j]$, with $\eta_i \mid Y_i^*$ continuously distributed and $\tau_j > 0$. This follows since then $\tau_j = G_\eta^{-1}(\delta_{j0}) - G_\eta^{-1}(\delta_{j1}) > 0$ and $\kappa_j = G_\eta^{-1}(\delta_{j0})/\tau_j$ rationalize each $(\delta_{j0}, \delta_{j1})$, where $G_\eta(\cdot)$ specifies the cumulative distribution of $\eta_i \mid Y_i^*$:

$$\begin{aligned} E[D_{ij} \mid Y_i^* = y] &= Pr(\kappa_j \geq y + \eta_i/\tau_j) \\ &= G_\eta((\kappa_j - y)\tau_j) \\ &= G_\eta(G_\eta^{-1}(\delta_{j0})) + y(G_\eta^{-1}(\delta_{j1}) - G_\eta^{-1}(\delta_{j0})) \\ &= \delta_{j0} + y(\delta_{j1} - \delta_{j0}) \end{aligned} \tag{B30}$$

In particular, Equation (B30) shows that our risk signal threshold decision rule (23), in which $\eta_i \mid Y_i^* \sim N(0,1)$, is without loss in this case. In general, we may think of τ_j as capturing judge j's signal quality: how less likely she is to release defendants with $Y_i^* = 1$ than those with $Y_i^* = 0$.

We next relate differences in such signal quality to conventional monotonicity violations in a simple behavioral model of judicial decision-making. Suppose judges observe a vector of defendant and case characteristics $\mathbf{X_i}^*$ which are, without loss, mean zero and positively correlated with misconduct potential: $\boldsymbol{\mu_X}(1) \equiv E[\mathbf{X_i}^* \mid Y_i^* = 1] > E[\mathbf{X_i}^* \mid Y_i^* = 0] \equiv \boldsymbol{\mu_X}(0)$. Judges place different weights $\boldsymbol{\beta_j}$ on the elements of this vector and also vary in their overall leniency π_j, such that:

$$D_{ij} = \mathbf{1}[\pi_j \geq \mathbf{X_i}^{*\prime}\boldsymbol{\beta_j} + U_i] \tag{B31}$$

where we assume $U_i \mid \mathbf{X_i}^*, Y_i^*$ is uniformly distributed. In this model $E[D_{ij} \mid Y_i^* = y] = \pi_j - \boldsymbol{\mu_X}(y)'\boldsymbol{\beta_j}$, assuming the parameters are such that these are all between zero and one.

Conventional monotonicity in this model requires $Pr(D_{ij} \geq D_{ik} = 1)$ or $Pr(D_{ik} \geq D_{ij} = 1)$ for each (j,k), which generally restricts the weights $\boldsymbol{\beta_j}$ to be the same across judges. If some elements of $\mathbf{X_i}^*$ were observed to the econometrician, one could relax this assumption by a conditional analysis within sets of defendants with identical observables (e.g., Mueller-Smith, 2015). Conditional monotonicity would then generally constrain the weights corresponding to unobserved characteristics to be constant.

Judge decision-making is here better-than-random when $\delta_{j0} - \delta_{j1} = (\boldsymbol{\mu_X}(1) - \boldsymbol{\mu_X}(0))'\boldsymbol{\beta_j} > 0$ or when the weights in each $\boldsymbol{\beta_j}$ are non-negative with at least one element strictly positive. In this case we have from the above result an equivalent representation of:

$$D_{ij} = \mathbf{1}[\kappa_j \geq Y_i^* + V_i/\tau_j] \tag{B32}$$

where $V_i \mid Y_i^* \sim U(0,1)$. Here judge signal quality is given by $\tau_j = (\boldsymbol{\mu_X}(1) - \boldsymbol{\mu_X}(0))'\boldsymbol{\beta_j}$ and has a straightforward interpretation: with only one element in $\mathbf{X_i}^*$, for example, differences in τ_j are proportional to differences in the behavioral weights $\boldsymbol{\beta_j}$. More generally, this discussion shows how parameterizations of the distribution of signal quality across judges can be thought to structure differences in how judges weigh defendant and case characteristics when making release decisions.

B8. SMD Estimation of the Hierarchical MTE Model

We estimate the hierarchical model described in Sections V.A and V.B by a simulated minimum distance (SMD) procedure that targets moments of the distribution of race- and judge-specific release rates $\rho_{jr} = E[D_{ij} \mid R_i = r]$ and released misconduct rates $\lambda_{jr} = E[Y_i^* \mid D_{ij} = 1, R_i = r]$, estimated from quasi-experimental judge assignments. This appendix formally specifies this procedure.

We first obtain estimates of ρ_{jr} and λ_{jr} from OLS regressions of pretrial release D_i and pretrial misconduct Y_i on judge-by-race interactions, adjusting for the quasi-experimental court-by-time effects) and defendant and case observables as discussed in Section IV.B. Subject to the usual asymptotic approximation, the resulting estimates $\hat{\rho}_{jr}$ and $\hat{\lambda}_{jr}$ can be modeled as noisy measures of the true parameters, with a known distribution of sampling error. Specifically:

$$\hat{\rho}_{jr} = \rho_{jr} + \varepsilon_{jr}^{\rho} \tag{B33}$$

$$\hat{\lambda}_{jr} = \lambda_{jr} + \varepsilon_{jr}^{\lambda} \tag{B34}$$

where $\boldsymbol{\varepsilon} \mid \boldsymbol{\rho}, \boldsymbol{\lambda} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ for a variance-covariance matrix $\boldsymbol{\Sigma}$ that is given by conventional asymptotics. Let $\mathscr{X} = ((\hat{\rho}_{jr}, \hat{\lambda}_{jr})_{j=1,\ldots,268, r\in\{w,b\}})$ collect these estimates across the 268 judges in our sample and both races w and b.

The model in Appendix B.B6 specifies ρ_{jr} and λ_{jr} as functions of mean misconduct risk μ_r, judge signal quality τ_{jr}, and risk thresholds π_{jr}:

$$\rho_{jr} = \Phi((f(\pi_{jr}, \mu_r, \tau_{jr}) - 1)\tau_{jr}))\mu_r + \Phi(f(\pi_{jr}, \mu_r, \tau_{jr})\tau_{jr}))(1 - \mu_r) \tag{B35}$$

$$\lambda_{jr} = \Phi((f(\pi_{jr}, \mu_r, \tau_{jr}) - 1)\tau_{jr}))\mu_r / \rho_{jr} \tag{B36}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function and:

$$f(\pi, \mu, \tau) = \frac{1}{2} - \ln\left(\frac{\mu(1-\pi)}{\pi(1-\mu)}\right) / \tau^2. \tag{B37}$$

We further model signal thresholds $\kappa_{jr} = f(\pi_{jr}, \mu_r, \tau_{jr})$ and log signal quality $\ln \tau_{jr}$ as being joint-normally distributed across judges, with reisdual correlation across races. That is, we specify:

$$\ln \tau_{jr} = \alpha_r + \beta_r \kappa_{jr} + \varepsilon_{jr} \tag{B38}$$

for each race r, with $(\kappa_{jw}, \kappa_{jb})' \sim N(\boldsymbol{\mu_\kappa}, \boldsymbol{\Lambda_\kappa})$ and $(\varepsilon_{jw}, \varepsilon_{jb})' \mid \boldsymbol{\kappa} \sim N(\mathbf{0}, \boldsymbol{\Lambda_\tau})$.

Equations (B33)–(B38) specify a complete distribution for the observed quasi-experimental estimates $\mathscr{X}$ in terms of a hyperparameter vector $\boldsymbol{\Theta} = (\mu_w, \mu_b, \alpha_w, \alpha_b, \beta_w, \beta_b, \boldsymbol{\mu}'_{\boldsymbol{\kappa}}, vec(\boldsymbol{\Lambda}_{\boldsymbol{\kappa}}^{1/2})', vec(\boldsymbol{\Lambda}_{\tilde{\tau}}^{1/2})')'$. We estimate $\boldsymbol{\Theta}$ by SMD, targeting moments of $\mathscr{X}$ as discussed in Section V.A. Specifically, let $\hat{\mathbf{M}}$ be a vector with the first two race-specific elements of:

$$\hat{M}_{1r} = \sum_{j=1}^{268} \omega_{jr}^{\rho} \hat{\rho}_{jr} \tag{B39}$$

$$\hat{M}_{2r} = \sum_{j=1}^{268} \omega_{jr}^{\rho} (\hat{\rho}_{jr} - \hat{M}_{1r})^2 \tag{B40}$$

the next three race-specific elements corresponding to coefficient estimates from the ω_{jr}^{λ}-weighted quadratic OLS regression of:

$$\hat{\lambda}_{jr} = \hat{M}_{3r} + \hat{M}_{4r}\hat{\rho}_{jr} + \hat{M}_{5r}\hat{\rho}_{jr}^2 + \hat{\upsilon}_{jr} \tag{B41}$$

and the sixth race-specific element corresponding to the ω_{jr}^{λ}-weighted residual variance estimate:

$$\hat{M}_{6r} = \sum_{j=1}^{268} \omega_{jr}^{\lambda} \hat{\upsilon}_{jr}^2 \tag{B42}$$

The weights are derived from the estimation error matrix $\boldsymbol{\Sigma}$: ω_{jr}^{ρ} is proportional to the inverse variance of $\hat{\rho}_{jr} - \rho_{jr}$ while ω_{jr}^{λ} is proportional to the inverse variance of $\hat{\lambda}_{jr} - \lambda_{jr}$, with both weights rescaled to sum to one in the population of judges. We further include in $\hat{M}$ the $\sqrt{\omega_{jw}^{\rho}\omega_{jb}^{\rho}}$-weighted covariance of $\hat{\rho}_{jw}$ and $\hat{\rho}_{jw}$ as well as the $\sqrt{\omega_{jw}^{\lambda}\omega_{jb}^{\lambda}}$-weighted covariance of $\hat{\lambda}_{jw}$ and $\hat{\lambda}_{jw}$. Together this gives 14 elements in

$\hat{\mathbf{M}}$, the same number of hyperparameters in $\mathbf{\Theta}$.

The SMD procedure matches the empirical moments in $\hat{\mathbf{M}}$ with the corresponding model-implied moments averaged across 500 simulated draws of the above data-generating process. That is, we estimate:

$$\hat{\mathbf{\Theta}} = \arg\min_{\tilde{\mathbf{\Theta}}} \sum_{m=1}^{14} \left(\hat{M}_m - \frac{1}{500} \sum_{s=1}^{500} M_{ms}(\tilde{\mathbf{\Theta}}) \right)^2 \tag{B43}$$

where the functions $M_{ms}(\cdot)$ of candidate hyperparameters $\tilde{\mathbf{\Theta}}$ are given by applying the previous moment calculations to data generated from 500 fixed simulation draws s. Conventional asymptotic theory for $\hat{\mathbf{\Theta}}$ applies under appropriate regularity conditions (e.g., Pakes and Pollard, 1989).

Columns 3 and 6 of Appendix Table A18 report SMD estimates and standard errors for the full model. As discussed in the main text, our baseline model estimates set $\beta_r = 0$. Per the intuition in Section V.A and to keep the model just-identified, we correspondingly drop the quadratic term from the moment regression in Equation (B41). The resulting estimates are reported in columns 2 and 5 of Appendix Table A18. To impose conventional MTE monotonicity, we further set the variance of τ_{jr} to zero. The resulting estimates are reported in columns 1 and 4 of Appendix Table A18.

Lastly, given $\hat{\Theta}$, we compute maximum *a posteriori* probability estimates (also known as posterior modes) of the judge-specific parameters $\boldsymbol{\theta_j} = (\kappa_{jw}, \ln\tau_{jw}, \kappa_{jb}, \ln\tau_{jb})'$, following an approach similar to that which Angrist et al. (2017) apply for a similar hierarchical model. Note that the log-likelihood of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1' \ldots, \boldsymbol{\theta}_{268}')'$ and quasi-experimental estimates $\mathscr{X}$ can be written:

$$\mathscr{L}(\boldsymbol{\theta}, \mathscr{X}) = \ln\phi_m\left(\mathscr{X} - \bar{\mathbf{X}}(\boldsymbol{\theta}); \mathbf{\Sigma}\right) + \ln\phi_m\left(\boldsymbol{\theta} - \boldsymbol{\mu_\theta}; \mathbf{\Lambda_\theta}\right) \tag{B44}$$

where $\phi_m(\cdot;\mathbf{V})$ gives the density of a mean-zero multivariate normal vector with variance-covariance matrix $\mathbf{V}$; $\bar{\mathbf{X}}(\cdot)$ collects the formulas from Equations (B35) and (B36), for ρ_{jr} and λ_{jr} in terms of μ_w, μ_b, and $\boldsymbol{\theta}$; and both $\boldsymbol{\mu_\theta}$ and $\mathbf{\Lambda_\theta}$ are derived from the α_r and β_r, $\boldsymbol{\mu_\kappa}$, $\mathbf{\Lambda}_\kappa$, and $\mathbf{\Lambda_\tau}$. Our estimates of $\boldsymbol{\theta}$ are given by maximizing this likelihood, plugging in our baseline hyperparameter estimates $\hat{\mathbf{\Theta}}$.

*

REFERENCES

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

Bonhomme, Stephane, and Martin Weidner. 2020. "Posterior Average Effects." *Unpublished Working Paper.*

Morris, Carl. 1983. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association*, 78(381): 47–55.

Mueller-Smith, Michael. 2015. "The Criminal and Labor Market Impacts of Incarceration." *Unpublished Working Paper.*

Pakes, Ariel, and David Pollard. 1989. "Simulation and the Asymptotics of Optimization Estimators." *Econometrica*, 57(5): 1027–1057.