Institute for the Quantitative Study of Inclusion, Diversity, and Equity, Inc., 501(c)3
http://www.qsideinstitute.org
This report is available freely and without restriction at http://www.SocArXiv.org.

The Most Discriminatory Federal Judges Give Black and Hispanic Defendants At Least Double the Sentences of White Defendants

Christian Michael Smith[1], Nicholas Goldrosen[2], Maria-Veronica Ciocanel[3], Rebecca Santorella[4], Chad M. Topaz[5,6], Shilad Sen[7]

Abstract

In the aggregate, racial inequality in criminal sentencing is an empirically well-established social problem. Yet, data limitations have made it impossible to determine and name the most racially discriminatory federal judges. The authors use a new, large-scale database to determine and name the observed federal judges who impose the harshest sentence length penalties on Black and Hispanic defendants. Following the focal concerns framework, the authors (1) replicate previous findings that aggregate, conditional racial disparities in sentence lengths are large, (2) show that judges vary considerably in estimated racial discrimination, and (3) list the federal judges who exhibit the clearest evidence of racial discrimination. While acknowledging limitations of unobserved cases and variables, the authors find evidence that several judges give Black and Hispanic defendants double the sentences they give observationally equivalent white defendants. Accordingly, the results suggest that holding the most discriminatory judges accountable would yield meaningful reductions in racial inequality.

Institute for the Quantitative Study of Inclusion, Diversity, and Equity, Inc., 501(c)3
http://www.qsideinstitute.org
This report is available freely and without restriction at http://www.SocArXiv.org.

Keywords

punishment, racial discrimination, federal judges, criminal sentencing, inequality

Introduction

Existing research shows that the U.S. criminal justice system is rife with racial inequality and suggests that federal judges contribute to this inequality (Commission, 2018). With respect to criminal sentencing, the average level of apparent racial discrimination is high, with federal judges giving Black and Hispanic defendants harsher sentences than similar white defendants. This is especially true among men: according to the U.S. Sentencing Commission, federal judges give white men sentences 19% and 5% shorter, respectively, than the sentences they give Black and Hispanic men who commit similar crimes. The Code of Conduct for United States Judges expects judges to "perform the duties of the office fairly, impartially, and diligently" (Administrative Office of the United States Courts, b). Racial discrimination, whether implicit or overt, runs contrary to this norm and exacerbates racial inequality.

Still, average levels of discrimination do not express the variability in judges' discriminatory tendencies and they do not reveal which judges discriminate most egregiously. Presumably, not every judge applies the same sentencing penalties to Black defendants or to Hispanic defendants. Thus, judge-specific statistics can provide a public mechanism for accountability that aggregate statistics cannot.

---

[1]University of California, Merced, CA, USA
[2]University of Cambridge, Cambridge, Cambridgeshire, UK
[3]Duke University, Durham, NC, USA
[4]Brown University, Providence, RI, USA
[5]Institute for the Quantitative Study of Inclusion, Diversity, and Equity, Williamstown, MA, USA
[6]Williams College, Williamstown, MA, USA
[7]Macalester College, St. Paul, MN, USA

Corresponding author:
Christian Michael Smith, University of California, Merced, CA, 95343, USA.
Email: csmith97@ucmerced.edu

In this paper, we answer two primary questions: (1) Which federal judges give the harshest sentence length penalty to Black defendants? (2) Which federal judges give the harshest sentence length penalty to Hispanic defendants? To contextualize the answers to these questions, we first show that the aggregate level of sentencing discrimination against Black and Hispanic defendants is sizeable, and that the interjudge variability in this discrimination is wide. We then list the federal judges who show the most evidence of racial discrimination, finding that the most extreme of these judges give their Black and Hispanic defendants sentences that are twice as long as the sentences they give to observationally equivalent white defendants. We conclude by arguing for more accountability among judges with troubling sentencing patterns and for improved public reporting of judges' sentencing data.

Theory of Judicial Discrimination in Sentencing

Why might one expect some judges to discriminate against Black and Hispanic defendants in their sentencing decisions? The focal concerns framework offers one prevailing theory (Steffensmeier et al., 1998). This theory proposes that judges use extralegal characteristics like race as shorthands when determining appropriate sentences. In particular, judges follow three focal concerns: the defendant's blameworthiness, the extent to which a longer sentence would protect the community, and the practical constraints associated with a longer sentence. Race and ethnicity ought not affect how judges assess these concerns. Nevertheless, sentencing is a complex cognitive task that involves a good deal of ambiguity, and thus judges may consciously or subconsciously rely on stereotypes to assess the three concerns, in turn reducing the cognitive demand of sentencing.

According to the focal concerns framework, members of a social group will tend to receive unduly harsh sentences if the group faces stereotypes of being blameworthy, dangerous to the community, and in some way not very harmed by a long sentence (*e.g.*, being stereotyped as tough enough to handle prison, being stereotyped as already absent from one's dependents, etc.). Conversely, defendants are likely to

receive unduly lenient sentences if they benefit from stereotypes of being innocent, safe, and in some way harmed by a long sentence. There is good reason to believe that white defendants benefit from such stereotypes (Smith et al., 2014). Regarding blameworthiness, white individuals are underrepresented as perpetrators of crime in television news reports (Dixon and Linz, 2000), and this underrepresentation appears to provide white individuals an innocence premium in people's minds (Dixon, 2006). Regarding danger to the community, people implicitly view white individuals as safe and disinclined to commit crimes (Smith et al., 2014). Regarding practical constraints, qualitative evidence suggests that judges perceive white defendants and their families as being especially harmed by prison time (Kramer and Steffensmeir, 1993).

Judges may differ in their degree of racial discrimination for any number of reasons, including differential exposure to marginalized racial groups, differential exposure to stereotypical media, and differential effort to counteract implicit bias. The presence of aggregate discrimination does not imply that every judge discriminates and the absence of aggregate discrimination does not imply that every judge avoids discrimination. *

Literature on Aggregate Discrimination

Racial disparities are a common, and well-researched, issue in the sentencing practices of U.S. federal courts. Most research on aggregate discrimination in federal sentencing has focused on the disparity between Black and white defendants.† Black defendants

*While we forgo a thorough review of theories of discrimination in criminal sentencing, it behooves us to note the organizational perspective of this literature. Namely, racial disparities derive from a whole host of criminal justice actors and the stereotypes upon which they rely (Johnson and Betsinger, 2009). Legislatures can create disparity through the sentencing statutes they pass, such as the well-known disparity between crack and powder cocaine (Sklansky, 1995). Prosecutors, for example, might drive racial disparities through charging decisions (Rehavi and Starr, 2014). They can also drive disparities via the plea deals they offer to defendants, especially given the large "trial penalty" that federal defendants incur if they are convicted at trial (Ulmer et al., 2010). This prosecutorial effect is especially pronounced since the federal sentencing guidelines ceased to be binding on judges following *Booker* (Yang, 2014). At the tail end of this pipeline are federal judges, who exercise considerable discretion over sentencing post-*Booker* and who can create racial disparities through the focal concerns mechanisms detailed above.

†What research does exist on other groups finds that Asian American defendants are sentenced similarly to white defendants (Johnson and Betsinger, 2009). Due to unique federal jurisdiction in Indian Country, Native American defendants face much harsher sentences for crimes prosecuted federally than defendants of other races who commit them elsewhere and who would only face state prosecution (Droske, 2008). On the whole, though, Native American defendants do not receive harsher sentences than

consistently receive harsher sentences than white defendants (Feldmeyer and Ulmer, 2011; Mustard, 2001; Rachlinski and Wistrich, 2017). Young Black men, in particular, are sentenced most harshly (Doerner and Demuth, 2010). Yang (2015) finds that Black-white disparities in sentencing increased after *United States v. Booker*, in which the Supreme Court held the U.S. Sentencing Guidelines to be advisory rather than mandatory. However, others find that no such increase occurred (Starr, 2013). While scholars debate whether *Booker* increased Black-white disparities in sentencing, there is little debate on whether such disparities have existed both before and after the decision. These disparities also show up in state sentencing (Abrams et al., 2012) and state bail decisions (Arnold et al., 2018).

The literature on Hispanic-white sentencing disparities is more variable. Young Hispanic men do receive particularly harsh sentences (Doerner and Demuth, 2010). Still, some studies argue that the Hispanic-white sentencing disparity can be explained as a function of noncitizens being sentenced more harshly and Hispanic defendants being disproportionately noncitizens (Light, 2014; Light et al., 2014).

Literature on Interjudge Differences in Discrimination

This paper focuses on *differences across* individual federal judges in their sentencing disparities based on race, particularly disparities at the extreme, and the field knows little about these differences. The literature on interjudge differences in racial disparities is much sparser than the literature on aggregate racial disparities. While interjudge differences in racial disparities are understudied among federal judges, some relevant evidence exists among judges at a lower level, namely, judges in the Circuit Court of Cook County: in terms of interjudge differences racial disparities, Abrams et al. (2012) find that Cook County judges differ substantially vis-à-vis incarceration rates but not vis-à-vis sentence lengths. The field also knows how much judges differ in

other defendants in the federal courts (Ulmer and Bradley, 2018). Young Native men, though, receive the harshest sentences of almost any group (Franklin, 2013).

terms of overall sentencing severity: both Scott (2010) and Yang (2014) find that interjudge differences in overall sentencing severity are wide and that the differences have generally grown Post-*Booker*. In sum, evidence is sparse on federal judges' variation in racial disparities, even though there is evidence on federal judges' variation in overall severity and on variation in racial disparities among judges in a lower court.

Policy has hindered research on federal judges' variation in racial disparities generally, and has especially hindered researchers from naming judges with the greatest racial disparities. Specifically, the U.S. Judicial Conference has a standing policy of not including judge names or information in federal databases such as the Integrated Database or U.S. Sentencing Commission database, a policy that the Judicial Conference claims is in place due to "the potential for judge-specific information taken out of context to be misinterpreted, the administrative burden of compiling information to satisfy outside requests, and the availability to researchers of the information from individual courts" (Judicial Conference of the United States, 1995). The federal court records system, Public Access to Court Electronic Records (PACER), is also not searchable by judge. Hence, all research on individual judge differences in sentencing has had to use hand-coded datasets (Schanzenbach and Tiller, 2008) or proprietary datasets (Cohen and Yang, 2019), which prevent the names of the judges from being publicly disclosed. This work fills a crucial gap in the literature by reporting interjudge racial sentencing disparities, identifiably linked to specific federal judges. We believe that doing so brings crucial transparency to what is ostensibly a public process.

Methods

Data

Source We analyze the JUSTFAIR (Judicial System Transparency through Federal Archive Inferred Records) database of criminal sentencing decisions in federal courts (Ciocanel et al., 2020). JUSTFAIR is compiled from five public sources, includes almost 600,000 records from fiscal years 2001-2018, and links information about

defendants, their federal crimes, their sentences, and the sentencing judges. Here we briefly summarize the JUSTFAIR data pipeline presented by Ciocanel et al. (2020). Ciocanel and colleagues obtained information about criminal cases, defendants, and sentences from the U.S. Sentencing Commission (USSC) and merged this dataset with the Federal Judicial Center (FJC) Integrated Database to obtain court docket numbers for each case. The docket numbers allowed them to access the PACER system and retrieve the initials of the sentencing judges. Finally, they connected these initials with name, demographic, and education information of the judges from Wikipedia as well as by linking with the Federal Judicial Center Biographical Directory of Article III Federal Judges. As a result, the records in JUSTFAIR contain variables pertaining to the demographic characteristics of the sentenced individual; sentence characteristics, date, and federal district location; and sentencing judge name, appointment, and education information. Crucially to the current work, JUSTFAIR also includes the sections of the law relevant to the conviction and factors influencing the recommended sentence. Because JUSTFAIR is a new database, we describe its limitations in detail in Appendix A.

We have extended the JUSTFAIR data pipeline to include the available 2018-2019 fiscal year federal sentencing data, which is updated yearly in the USSC datafile for individual offenders and quarterly in the FJC Integrated Database. Our merging of offender, sentence, and judge information proceeded largely as described above and by Ciocanel et al. (2020). Several USSC variables denoting the total prison sentence, the coding of the offense type, and the post-Booker reporting categories changed in the USSC database, therefore we adjusted the data processing approach of Ciocanel et al. (2020) to remain consistent with the variables in JUSTFAIR. This extended the dataset we analyze in this study by over 30,000 cases.

In our analysis, we include cases only from 2006 and later. We are principally concerned with interjudge differences in discrimination observed after the *Booker* decision, and so cases before 2006 lie outside of our target population. We further exclude immigration cases because of the unique use of fast-track sentencing in these

cases and the extreme concentration of these cases in several southwestern district courts (Hartley and Tillyer, 2012). We also do not consider cases for which we cannot infer the sentence length; specifically, in the JUSTFAIR database, there are about 13,000 cases that resulted in a sentence length of zero according to the continuous sentence total variable but resulted in prison time according to the categorical variable of imprisonment type. These two features contradict each other, hence we consider the 13,000 cases to be missing the outcome. We therefore remove these cases from the analysis. After these pre-processing steps, the analytic sample represents about 380,000 cases corresponding to 1092 judges. The median number of cases per judge is 263 (μ = 358.7, σ = 389.6).

Measures The outcome of interest is the length of the sentence assigned to the defendant. We cap the sentence length at 470 months because life sentences are generally coded as 470 months. We log-transform the outcome because its distribution has a positive skew (Bushway and Piehl, 2001). Accordingly, we interpret the results in terms of percentage changes rather than in terms of linear changes in months of prison time.

The case-level characteristic of principal interest is the defendant's race and ethnicity, which we measure with the categories 'Hispanic,' 'Non-Hispanic Black,' 'Non-Hispanic White,' and 'Other Race/Ethnicity.' We collapse the groups in this way because, unfortunately, the sample sizes of the various groups contained in the final category are not large enough for us to perform cross-judge analyses, which are the main focus of our study. Using these definitions of race and ethnicity, the median judge imposed sentences in 41 cases with Hispanic defendants (μ = 97.3, σ = 210.6) and 69 cases with Black defendants (μ = 114.1, σ = 137.7).

We include an extensive set of control variables:

- the guideline minimum sentence, with any statutory minimum sentences taken into account,
- the defendant's criminal history points,

- whether the judge departed from the guideline sentence range for a government-sponsored reason,
- whether the judge departed from the guideline sentence range for a family-related reason,
- crime type, namely 'violent crime,' 'drug-related crime,' 'embezzlement, fraud, theft,' or 'other,'
- whether the case was settled by plea agreement or trial,
- sentencing year, and
- defendant demographics, namely sex, age, U.S. citizenship status, and educational attainment.

Arguably the most important of these control variables is the guideline minimum sentence, which is standardized based on factors like the severity of the crime. Thus, at least theoretically, we expect a non-discriminatory judge to assign roughly equal sentences to two defendants who have equal guideline minimum sentences.

We do not control for every reason a judge could depart from the guidelines, as these are quite numerous; see *e.g.*, Kaiser and Spohn (2018). In our analysis, amongst the most common reasons for departure, government-sponsored departures and family-related reasons are the only two that were not 18 USC 3553(a) considerations for sentencing. The remaining most common departure reasons in our data *are* 18 USC 3553(a) considerations, such as the nature of offense, need for rehabilitation, and so on. We do not control for these as they themselves are mechanisms through which racial discrimination plausibly occurs.

While the defendant's demographics are extralegal factors that judges ought not consider in their sentencing, we nevertheless control for these demographics because they may capture variation in legal factors that confound the race-sentence length relationship. For example, if a judge hears cases where many Hispanic defendants have a unique situation not captured by observed legal variables, and if this unique situation also is associated with educational attainment, then controlling for educational attainment reduces confounding due to the unobserved prevalence of this situation.

Analytic Strategy

Estimand: What We Mean by 'Discrimination' Before explaining our estimation procedures, it is important to define precisely what we wish to estimate (Lundberg et al., 2021). Racial discrimination is differential treatment due directly to one's race. We wish to estimate the extent to which each judge discriminates against Black and Hispanic defendants when imposing sentences. For precision, it is useful to translate this wish into the potential outcomes framework (Winship and Morgan, 1999). We do so below, focusing on discrimination against Black defendants for simplicity. Note that our estimand for discrimination against Hispanic defendants is defined analogously.

Across all federal criminal court cases i after 2005, we define judge *j*'s degree of racial discrimination against Black defendants as

$$\mathbf{E}[Y_{ijb} - Y_{ijw}] \tag{1}$$

where Y_{ijb} is the potential sentence that judge j imposes in case i if the case has a Black defendant and Y_{ijw} is the potential sentence that judge j imposes in the same case i if the case has a white defendant. It is helpful to imagine a theoretical experiment, even if this experiment would not be realistic or ethical in practice. An experiment to estimate the estimand in Equation (1) would have each judge j hear two cases that, while being legally identical, differ extralegally in that one has a Black defendant and one has a white defendant. The estimate of $\mathbf{E}[Y_{ijb} - Y_{ijw}]$ would then be the difference in the sentences judge j imposed in the two legally identical cases. Note that, while unit-specific quantities usually vary across persons in uses of the potential outcomes framework, unit-specific quantities vary across legal cases in our application because race is resistant to person-level manipulation (Sen and Wasow, 2016). Our approach is analogous to that of racial discrimination studies where the unit-specific quantities vary by job applications rather than persons (Lundberg et al., 2021; Pager, 2003).

By defining our causal estimand, we are not claiming that our results perfectly estimate this quantity. In fact, as with all causal estimands, it is impossible to calculate

$\mathbf{E}[Y_{ijb} - Y_{ijw}]$ exactly. This is impossible because at most one of the potential outcomes can be observed, in our case, the potential outcome corresponding to the actual race of the defendant in case i. Instead of forgoing investigation altogether, the researcher should *estimate* the causal estimand as closely as possible given limitations, as Blank et al. (2004) explain in their work on causally assessing racial discrimination. To do so, the researcher "exploits knowledge of population averages of outcomes among aggregates of members of a racial group" (pg. 81). In the next section, we describe our strategy for this estimation.

Empirical Model We estimate a hierarchical linear model of log transformed sentence lengths where cases (Level 1) are nested within federal judges (Level 2). In particular, we estimate the following model:

$$\text{(Level 1)} \quad \log(Y_{ijk}) = \beta_{0j} + \beta_{\mathbf{100}}\mathbf{X_{ij}} + \beta_{\mathbf{1j}}\mathbf{C_{ij}} + \epsilon_{ij} \tag{2}$$

$$\text{(Level 2, Intercepts)} \quad \beta_{0j} = \gamma_{000} + \zeta_{0j} \tag{3}$$

$$\text{(Level 2, Slopes)} \quad \beta_{1jk} = \delta_{100k} + \eta_{1jk} \tag{4}$$

where Y_{ijk} is the length of the sentence corresponding to case i by judge j for a defendant of race k, β_{0j} is the intercept for cases heard by judge j, $\mathbf{X_{ij}}$ is a vector of defendant control characteristics corresponding to case i by judge j, $\beta_{\mathbf{100}}$ is a vector of slopes between control characteristics and log transformed sentence length, $\mathbf{C_{ij}}$ is a vector of racial group binary indicators, $\beta_{\mathbf{1j}}$ is a vector of random slopes between logged sentence length and each racial group indicator among cases heard by judge j, ϵ_{ij} is an idiosyncratic error term, γ_{000} is the overall intercept, ζ_{0j} is the deviation of the intercept for cases heard by judge j from the overall intercept, β_{1jk} is the random slope corresponding to racial group k represented in $\beta_{\mathbf{1jk}}$, δ_{100k} is the overall slope between

log transformed sentence length and racial group indicator k, and η_{1jk} is the deviation of the slope for cases heard by judge j from the overall slope corresponding to racial group k. Our model assumes that ϵ_{ij}, ζ_{0j}, and η_{1jk} each are normally distributed with mean zero.

We are primarily interested in the values of η_{1jk} for each judge. Taking the example where $k = Black$ and the reference category is white defendants, the value of $\eta_{1j,Black}$ will answer the following question: how much greater is the Black-white sentencing disparity when judge j hears cases compared to when the average judge hears cases, controlling for defendant and case characteristics? Interpreting this value as reflecting discrimination and therefore causation requires that one assume ignorability conditional on the observed defendant and case characteristics. We note that, because our model estimation shrinks random slopes and intercepts for judges with a small number of cases, such judges do not cause random slopes and intercepts to appear unduly variable.

The model includes judge random effects instead of judge fixed effects because the former accommodate random slopes, which are the primary parameters of interest in this study. We forgo a hybrid model because it is less parsimonious and does not influence the estimated random slopes (Schunck, 2013), our primary interest. The model excludes district fixed effects because judges' levels of discrimination should be evaluated based on all other judges' levels, rather than just those in the same district. Still, we estimated an alternative model with district fixed effects and the results did not differ in any substantial way (model output available in the online supplement). In particular, 14 of the 15 judges we identify as the top 2% most discriminatory against Black defendants and 12 of the 16 identified as the top 2% most discriminatory against Hispanic defendants also appear on those lists with district fixed effects. Thus, even if the two specification options were equally compelling theoretically, our preferred specification is more parsimonious and does not affect the main results.

Identification of Especially Discriminatory Judges Identifying racially discriminatory judges is challenging for two main reasons. First, above and beyond observed

characteristics, cases with defendants of different races might differ on average with respect to unobserved characteristics that are correlated with sentence length, and these correlations may be systematically more prevalent for Judge A's cases than for Judge B's cases. Such a systematic difference could arise, for example, if Judge A specializes in a certain type of case and our observed covariates like case type do not fully account for this specialization. Structural factors, for example federal prosecutors' charging decisions and sentencing arguments, present another potential source of confounding—although we suspect that our control variables like guideline sentence, reasons for departure, and sentencing year account for most of this confounding. While work from a state judicial circuit has demonstrated random assignment of cases to judges in order to quell concerns of residual confounding (Abrams et al., 2012), we do not find that cases are randomly assigned within federal districts and hence we cannot rely on that possibility (see Appendix A for more details). As a second challenge, even if Judge A is not latently more discriminatory than Judge B, and even if Judge A's cases do not systematically differ from Judge B's, Judge A might by random chance have cases where defendants of certain races have unobserved, justifiable reasons for greater sentencing.

To combat the foregoing issues, we identify the judges that seem to be most discriminatory against Black defendants and Hispanic defendants in two steps. First, to limit this part of our investigation to those who have a recurring pattern of unequal sentencing, we include only judges with twenty-five or more sentences issued to non-Black defendants in the dataset and twenty-five or more to Black defendants (or twenty-five or more to Hispanic defendants instead, when we investigate Hispanic-white discrimination). There are 760 observed judges who imposed sentences on at least 25 Black defendants and at least non-Black defendants, and there are 702 who imposed sentences on at least 25 Hispanic defendants and at least non-Hispanic defendants. We still include the other judges in our model estimation because they provide meaningful variation for that purpose, but we exclude them when identifying especially discriminatory judges because racial differences in these judges' sentencing

decisions are particularly likely to be flukes. Next, from the narrowed set of judges, we identify the 2% of judges with the greatest $\eta_{1j,Black}$ values and the 2% with the greatest $\eta_{1j,Hispanic}$ values. Because these judges have such dramatic conditional disparities in sentencing, it is less likely that systematic or random-chance differences across judges in their defendants can explain the disparities completely. Accordingly, especially given the extensive control variables, this leaves racial discrimination as a likely explanation for the disparities. In sum, (1) restricting judges based on race-specific sample sizes reduces the risk of mislabeling as discriminatory a judge whose racial differences in sentencing arose only by random chance, and (2) labeling judges as discriminatory only if they have the most extreme conditional disparities reduces the risk of mislabeling a judge due to *either* systematic *or* random-chance differences in unobserved defendant characteristics.

Results

Aggregate Discrimination

Conditional on observed case and defendant characteristics, judges assign white defendants sentences that are about 9% lower than Black defendants' and 19% lower than Hispanic defendants', on average. In contrast, defendants in other racial groups receive sentences about 15% conditionally lower than white defendants' sentences, on average (Table 1). Under the assumption that defendants from different racial groups differ trivially with respect to relevant unobserved characteristics, these estimates suggest that judges tend to discriminate against Black and Hispanic defendants compared to white defendants. Judges may also discriminate in favor of defendants in other racial groups. Given the extremely small standard errors and *p*-values of all these estimates, sampling error is a very weak candidate for explaining these patterns.

Table 1. The average estimated effect that race has on logged sentencing length. Each percent change in the first column represents what percent greater/lesser the conditional average sentence length is for defendants from the corresponding racial group, relative to white defendants. This quantity comes from the following formula, where δ_{100k} represents the estimated effect of being in racial group k, as shown in the "Estimate" column: Percent Change $= 100 \times (e^{\delta_{100k}} - 1)$.

	% Change	Estimate	Std. Error	p-value
Intercept		2.48	0.02	$<2 \times 10^{-16}$
Black	9.41	0.09	0.01	$<2 \times 10^{-16}$
Hispanic	18.53	0.17	0.01	$<2 \times 10^{-16}$
Other Race	-14.79	-0.16	0.02	$<2 \times 10^{-16}$

Our aggregate results are in line with previous studies, though they perhaps point to even more aggregate discrimination than found in earlier timespans. In an investigation of a 2000–2002 pre-*Booker* USSC dataset, Feldmeyer and Ulmer (2011) find that, after controlling for a considerable set of legal and extralegal factors, Black defendant sentences are 6% longer and Hispanic defendant sentences are 1% longer on average than White sentences. In an analysis of a 2006–2008 multi-agency-based dataset, Rehavi and Starr (2014) similarly find that, in the aggregate, Black defendants receive sentences that are 9% longer than those of similarly situated white defendants who commit the same crimes. Finally, using a federal sentencing dataset similar to ours but covering cases from 2000 to 2010, Yang (2015) finds that Black offenders receive 1.9 months longer sentences, and Hispanic offenders over 1.9 months longer sentences, than those of similar white offenders post-*Booker*.

Interjudge Variability in Discrimination

How much do judges vary in their estimated degree of racial discrimination? The estimates in the previous section show only the average extent of estimated discrimination. We turn now to a discussion of how much judges are spread around this average. Table 2 shows that the spread is considerable for each racial group, with random slope standard deviations ranging from 0.24 to 0.33 log transformed years. The

random intercepts, representing judges' sentencing severity against white defendants, also vary to a similar extent.

Table 2. Dispersion of random effects. Random intercepts—ζ_{0j} in Equation (3)—represent judges' sentencing severity toward white defendants. Random slopes—η_{1jk} in Equation (4)—represent judges' estimated racial discrimination compared to the average judge's level of discrimination.

	Std. Dev.
Random Intercepts	0.38
Random Black Slopes	0.24
Random Hispanic Slopes	0.26
Random Other Race Slopes	0.33

Because interpreting the foregoing standard deviations is challenging given the log transformed dependent variable, we present a more intuitive account of the interjudge variability in Table 3. As noted before, the average judge assigns Black defendants sentences that are conditionally 9% greater than white defendants'. However, a judge that is one standard deviation above average in terms of estimated anti-Black discrimination assigns Black defendants sentences that are conditionally 39% greater than white defendants', and a judge that is two standard deviations above average assigns Black defendants sentences that are conditionally 77% greater than white defendants'. Similarly, while the average judge assigns Hispanic defendants sentences that are conditionally 19% greater than white defendants', a judge that is one standard deviation above average in terms of estimated anti-Hispanic discrimination assigns Hispanic defendants sentences that are 54% greater than white defendants'. A judge that is two standard deviations above average assigns Hispanic defendants sentences that are conditionally a full 99% greater than white defendants'.

Table 3. The percent change in sentence length conditionally associated with the defendant being in different racial groups relative to being white by the average judge (column 1), by a judge one standard deviation above average (column 2), and by a judge two standard deviations above average (column 3), in terms of estimated discrimination against the group.

	% Change (Avg.)	% Change (1 SD)	% Change (2 SD)
Black	9.41	39.10	76.83
Hispanic	18.53	53.73	99.37
Other Race	-14.79	18.53	64.87

Especially Discriminatory Judges

Table 4 shows the top 2% of judges in the sample who show the greatest Black-white conditional sentencing disparities, and Table 5 shows the top 2% of judges in the sample who show the greatest Hispanic-white conditional sentencing disparities. In both cases, the top two constitute the same pair of judges, and we estimate that each of them imposes more than a two-fold penalty on Black and Hispanic defendants relative to white defendants. Even the judges with the 15th greatest conditional sentencing disparities impose more than an estimated 75% penalty on Black and Hispanic defendants relative to white defendants. Each of the judges on these lists presents very troubling patterns that warrant further scrutiny.

Our estimates cannot perfectly capture each judge's degree of discrimination, but nevertheless, recall that three main features of our analysis support the notion that these judges are discriminating against Black and Hispanic defendants. First, the estimates are based on *conditional* disparities, so that we compare the sentence lengths of defendants who are in different racial groups but are similar with respect to many other characteristics, including their recommended minimum sentence lengths. Second, each list contains only those judges with extremely large conditional disparities. Thus, if one of the judges' estimated degree of discrimination were due entirely to confounding factors, and not at all due to actual discrimination, then the extent of confounding would have to be extraordinary. Finally, for the purpose of Table 4 and Table 5, we restrict the list to judges who imposed sentences on twenty-five or more defendants in each of

Table 4. Top 2% of judges with greatest Black-white conditional disparities in sentencing. η represents the random slope associated with each judge. % Increase represents the percent by which each judge's sentences given to Black defendants are conditionally greater than the sentences the judge gives to white defendants. % Data represents the percent of cases matched for a judge's district during the years they were on the bench (or the entire database, if their career spanned that timeframe).

Rank	Judge	District	Status	η	% Increase	% Data
1	C. Darnell Jones II	Eastern PA	Senior	0.7	126	49
2	Timothy J. Savage	Eastern PA	Senior	0.7	117	46
3	Joseph P. Stadtmueller	Eastern WI	Judge	0.6	106	48
4	Frederick Scullin	Northern NY	Senior	0.6	98	65
5	Stanley R. Chesler	NJ	Senior	0.6	98	73
6	James C. Turk	Western VA	Deceased, 2014	0.6	91	68
7	Robin L. Rosenberg	Southern FL	Judge	0.5	90	70
8	John T. Elfvin	Western NY	Deceased, 2009	0.5	88	43
9	Kenneth Ryskamp	Southern FL	Deceased, 2017	0.5	84	70
10	John G. Murtha	VT	Senior	0.5	81	75
11	Gary L. Lancaster	Western PA	Deceased, 2013	0.5	80	74
12	Richard J. Arcara	Western NY	Senior	0.5	80	56
13	Patrick M. Duffy	SC	Retired, 2019	0.5	77	69
14	Dora Irizarry	Eastern NY	Senior	0.5	77	61
15	William K. Sessions	VT	Senior	0.5	76	75

Table 5. Top 2% of judges with greatest Hispanic-white conditional disparities in sentencing. η represents the random slope associated with each judge. % Increase represents the percent by which each judge's sentences given to Hispanic defendants are conditionally greater than the sentences the judge gives to non-Hispanic white defendants. % Data represents the percent of cases matched for a judge's district during the years they were on the bench (or the entire database, if their career spanned that timeframe).

Rank	Judge	District	Status	η	% Increase	% Data
1	C. Darnell Jones II	Eastern PA	Senior	0.7	156	49
2	Timothy J. Savage	Eastern PA	Senior	0.7	137	46
3	Elizabeth A. Kovacevich	Middle FL	Senior	0.7	137	68
4	Joseph F. Bianco	Eastern NY	2nd Circuit, 2019	0.7	131	61
5	James S. Moody	Middle FL	Senior	0.7	130	68
6	Stanley R. Chesler	NJ	Senior	0.6	125	73
7	Richard J. Arcara	Western NY	Senior	0.6	124	56
8	D. Lowell Jensen	Northern CA	Retired, 2014	0.6	123	56
9	Lawrence L. Piersol	SD	Senior	0.6	122	72
10	Charles J. Siragusa	Western NY	Senior	0.6	119	56
11	James L. Holmes	Eastern AR	Retired, 2020	0.6	117	64
12	Richard A. Lazzara	Middle FL	Senior	0.6	117	68
13	Ronald E. Longstaff	Southern IA	Senior	0.6	115	63
14	Edward R. Korman	Eastern NY	Senior	0.6	115	59
15	Alvin Hellerstein	Southern NY	Senior	0.6	112	59
16	Daniel D. Crabtree	KS	Judge	0.6	112	64

the relevant racial groups. Therefore, random chance is not a viable explanation for the large disparities corresponding to each judge on the lists. One unknown potential

source of error is that we cannot determine what percentage of each judge's cases were matched in the JUSTFAIR database. If this missingness is as-if random with respect to sentencing variables of interest, that should not bias our results, but we have little way of determining this. This limitation is an inevitable effect of the Judicial Conference policy prohibiting judge-identifiable data; since one cannot search PACER, FJC, or USSC records to determine how many cases a judge heard during the years in the JUSTFAIR database, one also cannot determine what proportion of cases are captured. As a rough proxy, we have included in the table the percent of cases matched for a judge's district during the years they were on the bench (or the entire database, if their career spanned that timeframe). Further transparency from the federal judiciary would greatly improve the accuracy of these results.

In sum, we have good reason to believe that the racially unequal sentencing patterns of these judges reflects some level of discrimination, despite the methodological limitations, and therefore, we encourage examination of these judges, especially those still on the bench.

Replication Package

The replication package in the online supplement of this article contains code for replicating all of the results above. The supplement also contains a full list of available judges, their estimated random slopes, and transformations of these slopes expressed intuitively in percent changes. This list allows readers to search for judges of interest, for example, judges in their own district. However, we emphatically encourage caution: For all the reasons detailed in the *Identification of Especially Discriminatory Judges* section, we urge readers not to infer a true difference between two judges when (1) the estimated random slopes are close to one another, or (2) one or both of the judges has only a few observed cases with defendants from a racial group of interest. We also urge readers not to infer much about a judge who is estimated to have a very high proportion of missing cases. We hope this full list will be useful, but we equally hope that all uses of it will be judicious.

Discussion

Using the JUSTFAIR database of district court criminal sentences, we have carried out the first judge-level analysis of racial disparities in federal district court sentencing. Specifically, we computed a hierarchical linear model of log transformed sentence length with cases nested within federal judges. We controlled for factors such as recommended sentence, sentence departure reasons, crime type, presence of a plea agreement, sentencing year, defendant demographics, and more. From our model, we calculated the percentage by which each judge's sentences given to Black and Hispanic defendants are conditionally greater than the sentences the judge gives to white defendants.

We find large disparities. For example, the three judges with the largest Black-white sentencing disparities give Black defendants sentences that are more than twice as long as those given to White defendants, controlling for factors as mentioned previously. Perhaps even more striking, the 2% of judges—16 judges in all—displaying the largest Hispanic-White disparities all give Hispanic defendants sentences that are more than twice as long as those given to white defendants. There are four judges who have especially large sentencing disparities for both Black and Hispanic defendants, namely Hons. C. Darnel Jones II, Timothy J. Savage, Stanley R. Chesler, and Richard J. Arcara. We note that all four have Senior status and are situated in the Northeast (and more specifically, in the Second and Third Circuits). The first three are George W. Bush appointees and the fourth is a Reagan appointee.

Writing on troubling immigration judges in particular, Alexander (2006) advocates for a campaign to publicly name these judges. Alexander argues that this campaign would improve policymakers' awareness of judicial problems as well as deter all judges away from troubling practice, even judges not initially sanctioned. We argue that a campaign analogous to Alexander's proposal may be in order: to the extent that our estimates validly capture these judges' discriminatory practices, these practices are a

significant social problem that amplifies racial inequality, runs counter to federal judges' Code of Conduct, and warrants rectification.

Accordingly, we propose several potential uses for our results. First, federal judges interested in assessing their own behavior could use them. We do not know the degree to which federal district judges are aware of their own sentencing patterns, but for inclined judges, our results can be used for self-assessment and reflection. Wistrich and Rachlinski (2017) offer practical advice for judges on how to diminish their own bias. Second, our results provide transparency to the public, who have a constitutional right to knowledge about what happens in criminal courtrooms. Third, for a president concerned with racial equity, our analysis could inform appointments of federal district judges to higher courts, especially if the results allow the public to lobby against judges with troubling records.

We reiterate some limitations of our study. First, because cases are not randomly assigned to judges, unobserved factors that systematically differ across judges may cause some judges to appear more discriminatory than they actually are. We would be surprised if the judges in our top 2% lists were not racially discriminatory, given three previously mentioned analytic choices: controlling for case characteristics, restricting to judges who have sentenced many people from each relevant racial group, and listing only the judges with the most extreme conditional disparities. Nevertheless, residual confounding could cloud the exact ranking of judges with respect to discrimination. A second limitation is our inability to observe every case, or even every district. Like most social science datasets, the JUSTFAIR database contains a sample rather than population. The database is the most thorough curation of federal sentencing data that includes judges' names, but because the judicial branch obfuscates judge identities in the data it releases, JUSTFAIR can only include a portion of cases from the last two decades. If sample inclusion is unrelated to our measures of interest, then lacking the full population leaves our estimates more susceptible to random sampling error than they would be otherwise. If sample inclusion is related to our measures of interest, then

lacking the full population leaves our estimates more susceptible both to systematic error and random sampling error than they would be otherwise.

The solution to these data limitations is straightforward: the Judicial Conference of the United States should repeal its current policy of redacting judge names in sentencing data. It should instead include those names in the databases released by the Federal Judicial Center and the United States Sentencing Commission, as well as in searches of the PACER online court records system. If we had this complete data, we could analyze whether the *Booker* decision (which made sentencing guidelines advisory) made individual judges more discriminatory or less discriminatory, and perhaps propose solutions to mitigate disparities.

Future research might extend analyses to state courts. While our work has examined criminal sentencing in federal district courts, we are cognizant that most sentencing in the United States is carried out in state courts. To understand the role individual judges play in race-based sentencing disparities, it will be necessary to gather and analyze data from all 50 state court systems, as well as districts, territories, and other possessions. This will be challenging given that each of these systems has its own sentencing frameworks, data keeping practices, and levels of transparency, but will be worthwhile in order to help move the justice system towards greater equity.

Author Biographies

CHRISTIAN MICHAEL SMITH is a Postdoctoral Scholar in the Department of Sociology at the University of California, Merced. He conducts interdisciplinary, quantitative research on social stratification, mobility, higher education, and criminal justice.

NICHOLAS GOLDROSEN is a graduate student in the Institute of Criminology at the University of Cambridge. His research interests include discretion and policing, progressive prosecution, and discrimination in sentencing.

MARIA-VERONICA CIOCANEL is an Assistant Professor of Mathematics and Biology at Duke University. She uses dynamical systems modeling and data analysis to study protein dynamics and transport in cells as well as criminal justice.

REBECCA SANTORELLA is a Ph.D. candidate in the Division of Applied Mathematics at Brown University. She uses machine learning and mathematical modeling to study algorithmic fairness, single cell biology, and criminal justice.

CHAD M. TOPAZ is Co-Founder and Executive Director of Research at the Institute for the Quantitative Study of Inclusion, Diversity, and Equity (QSIDE), as well as Professor of Mathematics at Williams College. His research uses applied mathematical modeling and data science to study criminal justice, education equity, diversity and inclusion in arts/media, and health care equity.

SHILAD SEN is a Professor of Computer Science at Macalester College. He studies the relationship between algorithms, software, and people and focuses on biases and inequalities along dimensions such as race, gender, and geography.

References

Abrams, D. S., M. Bertrand, and S. Mullainathan (2012). Do judges vary in their treatment of race? *J. Legal Stud. 41*(2), 347–383.

Administrative Office of the United States Courts. U.S. Courts FAQs: Filing a case. https://www.uscourts.gov/faqs-filing-case#faq-How-are-judges-assigned-to-cases? Accessed: 2020-12-22.

Administrative Office of the United States Courts. U.S. courts judges & judgeships: Code of conduct for United States judges. https://www.uscourts.gov/judges-judgeships/code-conduct-united-states-judges. Accessed: 2021-4-26.

Alexander, III, S. B. (2006). A political response to crisis in the immigration courts. *Georgetown Immigration Law Journal 21*, 1.

Anderson, J. M., J. R. Kling, and K. Stith (1999). Measuring interjudge sentencing disparity: Before and after the federal sentencing guidelines. *The Journal of Law and Economics 42*(S1), 271–308.

Arnold, D., W. Dobbie, and C. S. Yang (2018). Racial bias in bail decisions. *Q. J. Econ. 133*(4), 1885–1932.

Blank, R. M., M. Dabady, and C. F. Citro (2004). *Measuring racial discrimination*. National Academies Press.

Bushway, S. D. and A. M. Piehl (2001). Judging judicial discretion: Legal factors and racial discrimination in sentencing. *Law and Society Review*, 733–764.

Ciocanel, M.-V., C. M. Topaz, R. Santorella, S. Sen, C. M. Smith, and A. Hufstetler (2020). Justfair: Judicial system transparency through federal archive inferred records. *PLOS One 15*(10), e0241381.

Cohen, A. and C. S. Yang (2019). Judicial politics and sentencing decisions. *Am. Econ. J. Econ. Pol. 11*(1), 160–191.

Commission, U. S. S. (2018). Demographic differences in sentencing: An update to the 2012 booker report. *Federal Sentencing Reporter 30*(3), 212–229.

Dixon, T. L. (2006). Schemas as average conceptions: Skin tone, television news exposure, and culpability judgments. *Journalism & Mass Communication Quarterly 83*(1), 131–149.

Dixon, T. L. and D. Linz (2000). Overrepresentation and underrepresentation of African Americans and Latinos as lawbreakers on television news. *Journal of communication 50*(2), 131–154.

Doerner, J. K. and S. Demuth (2010). The independent and joint effects of race/ethnicity, gender, and age on sentencing outcomes in U.S. federal courts. *Justice Quarterly 27*(1), 1–27.

Droske, T. (2008). Correcting Native American sentencing disparity post-booker. *Marquette Law Review 91*(3), 723–813.

Feldmeyer, B. and J. T. Ulmer (2011). Racial/ethnic threat and federal sentencing. *J. Res. Crime Delinq. 48*(2), 238–270.

Franklin, T. W. (2013). Sentencing native americans in us federal courts: An examination of disparity. *Justice Quarterly 30*(2), 310–339.

Hartley, R. D. and R. Tillyer (2012). Defending the homeland: Judicial sentencing practices for federal immigration offenses. *Justice Quarterly 29*(1), 76–104.

Johnson, B. D. and S. Betsinger (2009). Punishing the "model minority": Asian-american criminal sentencing outcomes in federal district courts. *Criminology 47*(4), 1045–1090.

Judicial Conference of the United States (1995). Report of the proceedings of the Judicial Conference of the United States, March 14, 1995. Technical report.

Kaiser, K. and C. Spohn (2018). Why do judges depart: A review of reasons for judicial departures in federal sentencing. *Criminol. Crim. Justic. Law. Soc. 19*(2), 44–62.

Kramer, J. and D. Steffensmeir (1993). Race and imprisonment decisions. *The Sociological Quarterly 34*(2), 357–376.

Light, M. T. (2014). The new face of legal inequality: Noncitizens and the long-term trends in sentencing disparities across US district courts, 1992–2009. *Law Soc. Rev. 48*(2), 447–478.

Light, M. T., M. Massoglia, and R. D. King (2014). Citizenship and punishment: The salience of national membership in US criminal courts. *Am. Sociol. Rev. 79*(5), 825–847.

Lundberg, I., R. Johnson, and B. M. Stewart (2021). What is your estimand? defining the target quantity connects statistical evidence to theory. *American Sociological Review 86*(3), 532–565.

Mustard, D. B. (2001). Racial, ethnic, and gender disparities in sentencing: Evidence from the US federal courts. *J. Law Econ. 44*(1), 285–314.

Pager, D. (2003). The mark of a criminal record. *American journal of sociology 108*(5), 937–975.

Rachlinski, J. J. and A. J. Wistrich (2017). Judging the judiciary by the numbers: Empirical research on judges. *Ann. Rev. Law Soc. Sci. 13*, 203–229.

Rehavi, M. M. and S. B. Starr (2014). Racial disparity in federal criminal sentences. *J. Pol. Econ. 122*(6), 1320–1354.

Schanzenbach, M. M. and E. H. Tiller (2008). Reviewing the sentencing guidelines: Judicial politics, empirical evidence, and reform. *U. Chi. Law Rev. 75*(2), 715–760.

Schunck, R. (2013). Within and between estimates in random-effects models: Advantages and drawbacks of correlated random effects and hybrid models. *The Stata Journal 13*(1), 65–76.

Scott, R. W. (2010). Inter-judge sentencing disparity after Booker: A first look. *Stanford Law Rev. 63*(1), 1–66.

Sen, M. and O. Wasow (2016). Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science 19*, 499–522.

Sklansky, D. A. (1995). Cocaine, race, and equal protection. *Stanford Law Review 47*, 1287–1322.

Smith, R. J., J. D. Levinson, and Z. Robinson (2014). Implicit white favoritism in the criminal justice system. *Ala. L. Rev. 66*, 871.

Starr, S. (2013). Did Booker increase disparity? Why the evidence is unpersuasive. *Fed. Sent. Rep. 25*(5), 323–326.

Steffensmeier, D., J. Ulmer, and J. Kramer (1998). The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, black, and male. *Criminology 36*(4), 763–798.

Ulmer, J. T. and M. S. Bradley (2018). Punishment in indian country: Ironies of federal punishment of native americans. *Justice Quarterly 35*(5), 751–781.

Ulmer, J. T., J. Eisenstein, and B. Johnson (2010). Trial penalties in federal sentencing: Extra-guidelines factors and district variation. *Justice Quarterly 27*(4), 560–592.

Winship, C. and S. L. Morgan (1999). The estimation of causal effects from observational data. *Annual review of sociology 25*(1), 659–706.

Wistrich, A. J. and J. J. Rachlinski (2017). Implicit bias in judicial decision making: How it affects judgment and what judges can do about it. LawArXiv, 2017 Dec 22.

Yang, C. S. (2014). Have interjudge sentencing disparities increased in an advisory guidelines regime? evidence from Booker. *NYUL Rev. 89*, 1268–1342.

Yang, C. S. (2015). Free at last? Judicial discretion and racial disparities in federal sentencing. *J. Legal. Stud. 44*(1), 75–111.

Appendix A: Data Limitations

Ciocanel et al. (2020) detail several measures they took to ensure the quality of the JUSTFAIR database. They validated merged records by checking for additional common variables (such as offense codes) in the original datasets, used two independent sources to identify judge names from judge initials, and excluded records where sentencing dates fall outside the judges' activity periods. In addition, they carried out a manual data validation procedure for randomly sampled cases from the assembled dataset (Ciocanel et al., 2020). The validation set confirms the accuracy of the merging procedure: very few cases violate the assumption that the judge involved in a proceeding

is the same as the sentencing judge. Nevertheless, while they corrected these cases in the validation set, there could be rare additional instances where a judge other than the sentencing judge became involved in proceedings post-sentencing, leading to an incorrect inference in the database.

While JUSTFAIR is, to our knowledge, the largest publicly-available database of federal sentences currently available, it nevertheless remains an incomplete database of the cases sentenced in 2001–2018 (extended in this work to 2019) due to issues pertaining to data quality and merging challenges. For example, when merging USSC and FJC sentence and defendant information, JUSTFAIR only retains cases where there is a unique matching. Similarly, when retrieving sentencing judge initials from PACER, criminal cases may include more than one defendant, and therefore JUSTFAIR only keeps records where all defendants with the same court docket number are associated with the same judge. JUSTFAIR also cannot include information on sealed federal cases. Other factors that prevent JUSTFAIR from being complete include the inability to distinguish between judges who have both the same initials and district, and the complete or partial lack of judge initials in certain districts or records. In particular, due to such data quality issues, JUSTFAIR contains no sentencing data from the Eastern District of North Carolina, the Southern District of West Virginia, the Southern District of Texas, the Middle District of Tennessee, the Northern District of Illinois, the District of Guam, and the District of the Northern Mariana Islands. The database also includes limited sentencing data (less than 33% of the starting USSC cases) from the Northern District of Texas, the Southern District of California, the District of Oregon, the District of New Mexico, the Western District of Oklahoma, and the Northern District of Florida.

Non-random assignment of cases to judges presents another limitation, not just in the case of the JUSTFAIR database but rather whenever federal judge effects are of interest. The Administrative Office of the US Courts claims that "The majority of courts use some variation of a random drawing" (Administrative Office of the United States Courts, a). This means that, "to assure equitable distribution of caseloads and avoid judge shopping" (Administrative Office of the United States Courts, a), the district

courts have a rotation plan for cases assigned to judges; however, the US Courts Office also mentions that special expertise and geography are also considered in case assignment. Previous studies have exploited the random assignment of cases to judges in their analyses of sentencing equity (Anderson et al., 1999; Abrams et al., 2012; Cohen and Yang, 2019). Establishing random assignment has been considered key for these studies since it ensures that each judge receives a similar combination of cases and defendants (in terms of observable characteristics), so that unobservable characteristics can also be assumed to be similar across judges.

For instance, Abrams et al. (2012) analyze racial sentencing disparities in felony cases from Cook County, Illinois. To verify random assignment, they use a Monte Carlo simulation methodology for felony data from 1995 to 2001 to construct a randomly assigned dataset across characteristics such as defendant race, age, gender, and crime category, and use it to establish random assignment for their dataset (Abrams et al., 2012). Similarly, the study of the proprietary federal sentence dataset in Cohen and Yang (2019) relies on the assumption that cases are randomly assigned to judges in the same district court, focusing on observed case and defendant characteristics across Republican-appointed vs Democrat-appointed judges. Motivated by whether the political affiliation of a judge's appointing president influences racial and gender gaps in sentencing decisions, the authors use a joint F-test to test whether there are significant differences in defendant characteristics by judge political affiliation (as well as by judge tenure and gender). Conditioning on sentencing year and district court fixed effects, they find that cases are randomly assigned to sentencing judges from each political affiliation (Cohen and Yang, 2019).

Our setting is different from these studies, given that we are 1) analyzing the large, publicly-available JUSTFAIR dataset of federal sentences from 2001-2018 (Ciocanel et al., 2020) and 2) interested in whether the cases in this dataset are assigned randomly to individual judges within each district (rather than to judges with different characteristics). We find that, when testing for random assignment using an F-test method similar to Cohen and Yang (2019) or using a Monte Carlo simulation method

similar to Abrams et al. (2012), nearly every district shows evidence of nonrandom judge assignment. Specifically, nearly every district shows statistically significant between-judge differences in at least one case characteristic, such as defendant race, defendant sex, and offense type. In the model described here, we expect that control variables, especially the recommended sentence, adjust for most confounding that could erroneously make one judge appear more racially discriminatory than another, but we acknowledge that residual confounding may still be at play. Thus, one should exercise caution when interpreting the difference between two judges' estimated degree of racial discrimination, especially when the difference is small.