Do Judges Vary in Their Treatment of Race?

David S. Abrams, Marianne Bertrand, and Sendhil Mullainathan[1]

Abstract

Are minorities treated differently by the legal system? Systematic racial differences in case characteristics, many unobservable, make this a difficult question to answer directly. In this paper, we estimate whether judges differ from each other in how they sentence minorities, avoiding potential bias from unobservable case characteristics by exploiting the random assignment of cases to judges. We measure the between-judge variation in the difference in incarceration rates and sentence lengths between African-American and White defendants. We perform a Monte Carlo simulation in order to explicitly construct the appropriate counterfactual, where race does not influence judicial sentencing. In our data set, which includes felony cases from Cook County, Illinois, we find statistically significant between-judge variation in incarceration rates, although not in sentence lengths.

[1] University of Pennsylvania, University of Chicago, and Harvard University, respectively. Corresponding author: dabrams@law.upenn.edu. The authors would like to thank Josh Fischman, Chris Hansen, Max Schanzenbach, an anonymous referee, and seminar participants at Chicago, Harvard, MIT, and the NBER Summer Institute. Many thanks to Chief Judge Timothy C. Evans, Presiding Judge Paul P. Biebel, Jr., Malcolm Rich and Karen Landon for providing the data and invaluable background information on the Cook County Courts. Excellent research assistance was provided by Rohit Gupta, Dhruva Kothari, Jessica Pan, Kathy Qian, Tommy Wong, and especially James Wang.

I. Introduction

In 2008, 38% of sentenced inmates in the U.S. were African-American, with African-American males incarcerated at six and a half times the rate of White males.[2] Do these differences in incarceration rates merely reflect racial differences in criminal behavior, or are they also partly an outcome of differential prosecution or sentencing practices? A long-standing principle embedded in our system of justice is that defendants should not be treated differently because of their race. This principle is codified in the "Equal Protection" clause of the 14th amendment to the Constitution.[3] Differential sentencing or conviction rates by race are presumably a violation of this clause, making this an important question to answer on legal grounds. Establishing whether or not courts treat minority defendants differently also has important social implications: such practices might further exacerbate social inequalities and might even lead to a self-confirming equilibrium where expectations of racial discrimination affect criminal behavior.

Numerous studies examine this question, and most encounter empirical hurdles, particularly small sample size and omitted variables bias. Although almost all proceedings in U.S. courts are public record, as a practical matter it is quite challenging to obtain a statistically significant sample size. The studies using small samples of archival data have produced mixed results.[4] Of equal concern is the fact that cross-

[2] From "Prisoners in 2008", Bureau of Justice Statistics.

[3] "No state shall make or enforce any law which shall abridge the privileges or immunities of citizens of the United States; nor shall any state deprive any person of life, liberty, or property, without due process of law; nor deny to any person within its jurisdiction the equal protection of the laws." Amendment XIV, US Constitution

[4] Given this difficulty, a number of studies (Devine, et al., 2000; Sommers and Ellsworth, 2000; MacCoun, 1989) have made use of experimental simulations of court cases, most often to understand the behavior of juries. While laboratory studies allow the careful manipulation of the variable of interest, defendant race,

sectional studies suffer from a potentially severe omitted variables bias. Apparently significant effects of defendant race may actually be due to omitted case characteristics that are correlated with race, such as criminal history or attorney quality[5]. Thus there are two potential reasons for finding a significant coefficient on race in a cross-sectional regression: discriminatory sentencing on the part of judges or juries, or unobservable characteristics that drive the sentencing gap. The central difficulty with the cross-sectional methodology is that race is not randomly assigned. Therefore, any regression and interpretation thereof is likely to suffer from omitted variables bias.

In this paper, we take a new approach to studying the impact of race in judicial sentencing, one that avoids some of the methodological pitfalls just discussed, and helps shed light on the central issue.[6] We attempt to determine whether there are systematic differences across judges in the racial gap in sentencing. At the heart of our research strategy is the ability to exploit the random assignment of cases to judges. This random assignment ensures that unobservable case and defendant characteristics are the same across judges. It allows us to distinguish between unobservable case and defendant variables on the one hand and judicial behavior on the other as explanations for a racial gap in sentencing.

Under the unobserved variables explanation, where no judge is discriminatory, we may see an overall difference in sentencing by race, but we do not expect systematic variation in that difference across judges, as random assignment ensures that each judge

---

they suffer from questionable external validity. Many studies simply involve having subjects read transcripts of cases, which removes potentially important non-verbal elements of a trial.

[5] Recent research by Abrams and Yoon (2007) has shown there is substantial variation in attorney ability, although they did not find an interaction with client race.

[6] Ayres and Waldfogel (1994) also take a novel approach to detecting discrimination in a different legal environment - bail setting. Consistent with the presence of racial prejudice, they show that courts set bail at much higher levels for minority defendants, "overdeterring" them from fleeing (compared to White defendants) after release on bail.

receives the same case and defendant mix. Under the discriminatory sentencing explanation, as long as there is some between-judge heterogeneity in the level of differential treatment, we have the opposite prediction; that is, some judges will systematically sentence African-Americans at a higher rate and some will sentence them at a lower rate. This logic underlies the examination in this paper of whether there is significant inter-judge disparity in the racial gap in sentencing.[7]

To proceed, we use data from felony cases to compute the racial gap in sentence length and incarceration rate for each judge. The main empirical challenge is to identify the correct counterfactual, in which inter-judge variation is due solely to sampling variability. The asymptotic F distribution is inappropriate for this data set because of the small number of observations at the level at which random assignment occurs. This is a problem that occurs frequently in datasets involving randomization procedures where data is collected over a long period of time.[8] We address this problem by employing a Monte Carlo methodology to explicitly construct the counterfactual where race has the same impact on sentencing for all judges. Besides its application to the current study, this technique could benefit a large array of empirical studies facing similar constraints without a great deal of learning costs.[9]

We find evidence of significant inter-judge disparity in the racial gap in incarceration rates, providing support for the model where at least some judges treat defendants differently based on their race. The magnitude of this effect is substantial.

---

[7] There have been several previous studies that have examined overall inter-judge heterogeneity in sentencing, but none that have looked at the effect of defendant race on this heterogeneity. See e.g. Gaudet, Harris, St. John (1933), Anderson, Kling, and Stith (1999), Payne (1997), Waldfogel (1991) and Waldfogel (1998).

[8] One example of a recent paper that might benefit from this technique is Cheng (2008). Josh Fischman employs the technique in his working paper *Estimating Preferences of Appellate Judges* (2010).

[9] The advantage of using simulations has been pointed out in other contexts, for example by Imbens and Rosenbaum (2005) in the case of weak instruments.

The gap in incarceration rates between White and African-American defendants increases by 18 percentage points (compared to a mean incarceration rate of 51% for African-Americans and 38% for Whites) when moving from the 10th to 90th percentile judge in the racial gap distribution. The corresponding sentence length gap increases by 10 months, but this cannot statistically be distinguished from a situation where race played no role in sentence length.

Although judges differ in the degree to which race influences their sentencing, we do not find evidence that observable characteristics such as judges' gender or age group significantly predict this differential treatment by race. Similarly, no systematic pattern emerges with respect to work history (such as whether the judge ever worked as a Public Defender). However, there is somewhat stronger evidence that the racial gap in sentencing is smaller among African-American judges. Further, judges who are harsher overall (as measured by incarceration rate) are more likely to sentence African Americans to jail than they are Whites. We also explore an important potential confound: that the heterogeneity we observe in the racial sentencing gap may actually be due to heterogeneity in treatment of type of crime. The results of this analysis indicate that there may be a difference in treatment of drug and non-drug crimes, but that there is still a heterogeneous treatment of race within non-drug crimes.

One limitation to our approach is that, while we can statistically establish that race matters in the courtroom, we cannot formally detect whether this is due to some judges discriminating against African-Americans, or some judges discriminating against Whites, or a mixture of both. In itself, though, the evidence we uncover on the importance of race in judicial decision-making should be of direct relevance to legal policy.

The rest of the paper proceeds as follows. Section II provides a brief overview of prior work on the role of race in judicial decisions. In Section III we describe the data from the courts of Cook County, Illinois. We discuss our econometric methodology, including the simulation procedure in Section IV. In Section V we report our basic results, and we discuss the influence of the crime category in Section VI. Section VII concludes.

II. Literature Review

There has been a great deal of scholarship investigating the role of race in the courtroom. Here we briefly summarize some of the previous research most relevant to this study. Many early studies were cross-sectional, and frequently used data sets that were not rich enough to include controls for important case and individual characteristics, such as criminal history, crime severity, and income. Thus it is unsurprising that an early review of the literature found a lack of consensus among these studies. Daly and Tonry (1997) note some of the shortcomings in some of the work between the 1960's and 1980's. Kleck (1981) finds that half of the 40 studies on non-capital cases that he reviews either support a finding of discrimination in sentencing or have mixed results, while the other half do not find evidence of judicial discrimination.

Written nearly two decades later, Spohn (2000) also reviews 40 recent studies on the role of race in sentencing, but splits outcomes into incarceration and sentence length. In her survey of the literature, a majority of studies find that race impacts the incarceration decision, but fewer than one-quarter report evidence that race affects sentence length. In one of the most sophisticated critiques of work on discrimination in

the criminal justice system, Klepper, Nagin and Tierney (1983) point out numerous methodological problems, including sample selection and omitted variables. Many of their insights are still often neglected in this field of research, almost three decades later.

Some of the earlier papers such as those by Thomson and Zingraff (1981) and Humphrey and Fogerty (1987) rely on relatively small data sets and are unable to distinguish a race effect from the impact of unobservables. Klein, Petersilia, and Turner (1990) use a dataset from California state courts with a large number of covariates to try to minimize the concern about unobservables. They find no impact of race on either the incarceration or sentencing decision, and little explanatory power. Albonetti (1997) uses federal data from the U.S. Sentencing Commission (USSC) on drug offenders. She finds that African-American and Hispanic defendants are more likely to be incarcerated and for longer duration. Steffensmeier and Demuth (2000) also use federal data collected by the USSC, and thus have a detailed and large data set with which to work. Their cross-sectional OLS and probit regressions indicate that African-Americans and Hispanics are jailed more frequently and receive longer sentences than White defendants. The same authors find similar results using state court data from Pennsylvania in their 2001 paper. This differs to some extent from the findings of Kramer and Steffensmeier (1993), which also used Pennsylvania state court data. This study found a small impact of race on the incarceration decision, but not on the length of imprisonment.

A more recent paper by Mustard (2001) improves on previous work by including additional controls in the regression analysis. Using federal data provided by the USSC, he examines the impact of race on the incarceration and sentencing decisions, as well as on departures from the sentencing guidelines. His cross-sectional regressions include

controls for income, as well as interaction terms between race and income, race and education, and race and criminal history. He finds that African-Americans are more likely to be incarcerated and receive longer sentences, although some of this appears to be due to more extensive criminal histories and more severe offenses.

Using state data from Maryland, Bushway and Piehl (2001) estimate a tobit model to isolate the impact of judicial discretion on sentence length. They find a greater impact of race than most prior work. A major strength of this paper is the use of guideline recommendations to instrument for potential unobservable case characteristics. Rachlinski and coauthors (2009) approach the question from an experimental psychological perspective. In a laboratory study of judges they find similar results on the implicit association test to that of the general population, which has been interpreted by some as evidence of bias. In studies with explicit racial identification, however, Rachlinski and co-authors do not find race effects.

A recent contribution to the literature is from Schanzenbach (2005). This study focuses on understanding the impact of judicial characteristics on case outcomes, using variation in judicial characteristics at the federal district level.[10] While he finds that female judges reduce sex disparity in sentencing, results on racial disparity are mixed. He also finds no main effect of judges' race on average sentence length. Zussman and Shayo (2010) take a novel approach to understand the impact of ethnicity of various parties in the legal process. They exploits the random timing and location of terrorist attacks in Israel and shows that there is a short-lived local difference in case outcomes that is a function of defendant, plaintiff and judge ethnicity. Price and Wolfers (2010)

[10] Ashenfelter, et al (1995) is another study that focuses on the impact of judicial characteristics, using civil rights cases. They find no significant impact of the judges' race, sex, or political orientation on the case outcome.

also find evidence for race effects in a quasi-judicial context, that of NBA referees. In this paper, we focus primarily on defendant race effects in one large jurisdiction.

III. Data Description

Our data comes from the cases adjudicated in the Cook County Circuit of the Illinois state courts. Cook County is the largest unified court system in the country, with over 2.4 million cases processed per year in both civil and criminal courts.[11] It is also a racially mixed urban area, with a population that is 48% White, 26% African-American, and 20% Hispanic (see Table 1). The racial breakdown in our data is 12% White, 72% African-American, and 16% Hispanic, reflecting the substantially different rates of representation by race in the criminal justice system.

Illinois state courts are governed by sentencing guidelines, which provide suggested sentencing ranges by category of offense.[12] Previous studies, such as Anderson, et al. (1999), have found that guidelines mitigate interjudge sentencing variation, but not substantially. Judges in Cook County courts are initially appointed or elected, and subsequently subject to retention elections every six years.

While the original data set includes over 600,000 felony cases tried between 1985 and 2004, we use only a subset of the data. We discuss the primary restrictions used to obtain this subset here; further detail can be found in Appendix A. First, individual cases may have multiple defendants and multiple charges. In the data the number of charges per case ranges from 1 to 266 (see Table 2), but the median is 1. We retain one defendant and only the most severe charge for each case, since sentencing across charges for a

[11] See http://www.cookcountycourt.org/ for more detailed information about Cook County Courts.
[12] A rough description of Illinois sentencing guidelines is available at http://www.ilga.gov/commission/lru/2005PFC.pdf

given case will be highly correlated. Second, for the primary analysis, we restrict the data to defendants who are African-American or White (excluding the 16% of defendants classified as Hispanic).[13] Third, we retain only cases that were initiated between 1995 and 2001. The start date is used because it was impossible to verify random assignment of cases prior to 1995. The end date is used to allow sufficient time for completion of cases initiated towards the end of the time range (since some cases can take several years to adjudicate). Fourth, murder cases were excluded from the analysis because assignment of these cases often excluded certain judges.

We further limit the data to those cases adjudicated by a subset of the judges in the Cook County Criminal Courts Building, which handles the bulk of the criminal cases in Cook County. We included judges based on the following criteria: adjudicated at least 10 total cases throughout the time period of study; adjudicated cases only at the central courthouse location (in order to insure that all case randomization was performed on the same set of cases); did not preside over a special type of court (like drug court); did not have any unusual circumstances (such as lengthy capital trials) that would have resulted in non-random assignment of cases.

A full summary of the dataset we construct following the above criteria is provided in Tables 2A and 2B. Nearly all cases (92%) result in a guilty finding. The vast majority of defendants in the sample are African-American (86%), male (83%), and young (mean age is 29 and median age is 27). The mean length of incarceration is 20 months across all cases, and 42 months conditional on incarceration. Note that sentence length is top-coded at 60 years in our data. While the median case has only one charge

[13] Below and in the appendix we report the equivalent analysis on a dataset including only White or Hispanic defendants, and excluding African-Americans.

associated with it in the original data, the average number of charges per case is 2.4. As Table 2B shows, sentencing varies substantially by type of crime, with violent crimes receiving the most severe sentences. African-American defendants receive longer sentences on average and are over 30% more likely to be incarcerated than White defendants, not controlling for any case characteristics.[14]

Table 3 reports judicial characteristics collected from *Sullivan's Judicial Profiles*, *A Directory of State and Federal Judges in Chicago*, *The Directory of Minority Judges of the United States*, and several other sources listed in the references. The judiciary included in this study is largely White and male, with an average age of 49. Approximately half of the judges have some prior experience in private practice. Prior experience as a prosecutor is also a very common characteristic of these judges; over 70% have past experience as prosecutors, while 27% had previously served as public defenders or defense attorneys.

A crucial requirement for this analysis is that the court use random assignment of cases to judges. In the following section, we describe an econometric test for random assignment. But to establish even facial plausibility, one of the authors spent several days at the central Cook County Courthouse in Chicago, arranged by Presiding Judge Paul Biebel. Every morning in the courthouse, the clerks receive files for new cases and first remove those that have charges of murder or sex crimes. The remaining case numbers are typed individually into a monochromatic green-screen computer (almost certainly around since the 1980's) which then randomly chooses one of the judges

[14] Tables A1 and A2 report similar characteristics for the subset of the data containing Hispanic and White defendants.

currently hearing cases. The clerks verified that this procedure has been generally followed at least since the mid-1990s.

IV. Econometric Methodology

The focus of this paper is determining whether the impact of defendant race on sentencing varies across judges. There are two steps to testing this hypothesis. The first is to establish the random assignment of cases to judges, ensuring that sentencing outcomes can be fairly compared across judges. The second is to employ an appropriate method to evaluate whether there is excess heterogeneity in the racial gap in judicial sentencing beyond what would be expected due to sampling variability.

In theory, both steps may be accomplished using an ordinary least squares regression followed by an F-test. Under this approach, the random assignment of cases would be established by regressing a case characteristic, such as defendant age, on various controls and judge fixed effects, such as in Equation 1:

$$age_{ijt} = \alpha + \beta X_{ijt} + \sum \delta_j D_j + mo_t + \varepsilon_{ijt} \qquad (1)$$

where age is defendant age in years, X is an array of control variables, D are judge fixed effects, mo are month-year dummies, i is a defendant index, j is a judge index, and t a time index. An F-test on the equality of the judge fixed effects tests the hypothesis that cases are randomly assigned (with respect to defendant age). Similarly, in order to test the equality of the racial sentencing gap across judges, one would regress sentence length on a vector of control variables, defendant race, judge fixed effects, and interactions between the judge fixed effects and defendant race, such as in Equation 2:

$$sentence_{ijt} = \alpha + \beta X_{ijt} + race_{ijt} + \sum \delta_j D_j + \sum \gamma_j D_j * race_{ijt} + mo_t + \varepsilon_{ijt} \quad (2)$$

An F-test on the equality of the judge-race fixed effects γ_j would be a test of the equality of the racial gap in sentencing across judges.

In practice, rather than the asymptotic F-distribution, we rely instead on a Monte Carlo simulation to generate a correct finite-sample distribution. This methodology is analogous in spirit to that described above, but it addresses important shortcomings of using the standard F-test in this context.[15] Specifically, the methodology described above is likely to result in over-rejection of the null hypothesis (of random assignment, or no excess heterogeneity) for two reasons. First, although the overall sample is large, our regressions will suffer from finite sample bias because the sample cells are small within the short time periods that are of relevance. Indeed, it is necessary for the analysis to condition on short time periods because the random assignment of cases to judges occurs within these short periods, and there is substantial temporal variation in the judges available and the mix of case and defendant attributes. Our data structure will therefore not satisfy the large N assumption that the distribution of the F-statistic relies on. A second reason for not using the conventional F-statistic is that it will over-reject the null hypothesis when the errors are not normally distributed, as is the case where the dependant variable is Bernoulli with a mean substantially different from 0.5. This applies to several of the variables of interest here, such as race (test of random assignment) or incarceration (test of excess heterogeneity).[16]

[15] Methods analogous to the Bonferroni correction could also potentially be used to address some of the shortcomings of the asymptotic F-test. The advantage of the simulation approach is its simplicity, transparency, and ease of interpretation.
[16] See Kennedy (1998), chapter 4 for discussion of these issues.

The aforementioned reasons for empirically computing the finite-sample F-distribution are not unique to this paper, rather they are relatively frequent occurrences. In the law and economics literature, any study that compares judge effects without very high caseloads, like Cheng (2008) or Fischman (2010), is likely to suffer from the same problem. But this phenomenon is certainly not confined to judges; it applies to teacher studies, CEO's, leaders (see the discussion in Jones and Olken 2005), and numerous other contexts. Fortunately, the availability of cheap computing power makes the identification of the problem and the solution straightforward.

One way to test whether the small sample is a concern in this context is to simulate the F-distribution under the null for the given data set. Figure 1 illustrates the need for the simulation methodology in this context. In order to generate it, we ran 1000 tests similar to those we describe below, where by construction the null should not be rejected. Theoretically this should yield a uniform distribution. The dark bars are produced using the simulation methodology, and is nearly uniform. The light bars are produced using the standard F-test methodology. There is clearly an excess of p-values less than 0.05, which would lead to an over-rejection of the null.

For these reasons we instead use a Monte Carlo simulation methodology to both verify random assignment of cases to judges and to determine whether there is excess heterogeneity in the inter-judge racial gap in sentencing. Random assignment is verified by comparing the heterogeneity of the empirical distribution of case characteristics to that found in simulated data. The heterogeneity of the inter-judge racial gap is tested similarly. In both cases, statistical significance is determined by the dispersion of the empirical data relative to the distribution generated by the simulations. We now describe

the implementation of the simulation method, first for the random assignment test, and then for the test of excess heterogeneity across judges.

A. Testing for Random Assignment using a Monte Carlo Simulation

If cases are randomly assigned to judges, all observable case characteristics should have approximately the same moments for each judge. For example, the mean defendant age in the full data set is 29 years, and therefore if cases are randomly assigned, most judges should have a set of defendants with mean age around 29. Similarly, since 16% of cases are in the violent crime category, we expect a court that uses a random assignment procedure to produce a distribution of cases where most judges see violent crimes in about 16% of their cases. The difficulty in determining whether a data set results from random assignment is in quantifying exactly what it means for "most" judges to have a mean age "around 29." The question is – how much variation would there be in a randomly assigned data set, simply due to sampling variability? A straightforward way to establish whether the Cook County data does result from a random assignment process is by explicitly constructing a randomly assigned data set through simulation.

The procedure is as follows. Let X be a case characteristic of interest, such as defendant race, age, gender, or crime category. Denote a simulated observation by X_{ijcs} for observation i of judge j of simulation s ($i,j,s > 0$). X_{ijc0} refers to the empirical data set. The data is apportioned within cells (denoted by c) in order to approximate the actual

random assignment procedure done in the courthouse.[17] Create a simulated observation X_{ijcs} by assigning:

$$X_{ijcs} = X_{\alpha\beta c0}$$

where α is randomly chosen from the integers between 1 and N_c inclusive, and I_c is the number of observations in cell c (β is a function of α). This process is iterated for all i and j.

For each simulated data set, judge means may be computed:

$\overline{X}_{js} = \frac{1}{N_j}\sum_{i \in J} X_{ijcs}$, where J is the set of cases of judge j and has size N_j. We similarly compute a measure of inter-judge disparity (such as inter-quartile range, $D_s^{25\text{-}75}$) for each simulated dataset .[18] These measures may then be ranked across simulations, and a p-value found for the empirical distribution ($D_0^{25\text{-}75}$) based on where it falls in the $D_s^{25\text{-}75}$ distribution.

We refer to Table 4 as an illustration of the simulation for the random assignment test. For the purpose of this illustration, the outcome variable used to test random assignment is race.[19] The null hypothesis is that each judge has the same fraction of African-American defendants. If the case mix and eligible judge mix were time invariant, we would not need to restrict ourselves in time. But given that there is substantial variation in both, we choose the cell size to be one month.

---

[17] Since random assignment is done on a daily basis in the courthouse, this is the ideal cell size to use. Because there is unlikely to be substantial variation in case mix and judge mix within a month, we use one month as the cell size for computational simplicity.

[18] We use 3 different inter-percentile ranges, 25-75, 10-90, and 5-95. Other measures, such as standard deviation or absolute mean deviation could be used as well. We choose inter-percentile ranges because we are interested in the central tendencies of the distribution. These will not be substantially impacted by a small number of outliers.

[19] Race is a dummy that is zero if the defendant race is White and one if African-American.

In this abridged data set there are six total cases, four of which were assigned to judges in January. Thus the observation in simulation 1, case #1001 will be randomly chosen from cases 1001, 1414, 3141, and 2718. Since three of the four defendants in those cases are African-American, there is a 75% chance that the simulated data point will be African-American. In fact, in simulation 1, the simulated defendant race is indeed African-American.

This procedure is repeated for each observation in Table 4 to produce a full simulated data set. The process is then repeated 1000 times to produce 1000 simulated data sets. For each simulated data set, the mean of the race variable is then computed by judge, producing a distribution similar to the empirical distribution shown in Figure 2. We then calculate a measure of dispersion of this simulated distribution, for example, the interquartile range, which is denoted by the vertical lines in Figure 2. This measure is computed for each of the 1000 simulations. The data is then reduced to a distribution of these simulated interquartile ranges. We then compare the empirical interquartile range to the distribution produced from the simulations to obtain an estimate of how likely it is that the empirical distribution occurred due to chance. Figure 3 shows the 1000 simulated interquartile ranges along with the empirical interquartile range.

It is worth noting that the random draw in the procedure may be either with or without replacement (which would be akin to a permutation). Both procedures may be used, but have slightly different interpretations. Drawing with replacement is correct if the data is assumed to be one manifestation of a larger universe of potential empirical realizations. The permutation approach is correct if the data is assumed to be the only relevant realization. The main results presented in the paper were produced using

random draws with replacement; however, as a check we reproduced Figure 7 using a draw without replacement (Fig A1 in the Appendix). Given the size of the data set, it is unsurprising that there is no apparent difference between the two approaches.

B. Testing for Heterogeneous Sentencing by Race using a Monte Carlo Simulation

Once random case assignment has been established, we can infer that any differences in judicial decisions are due to differences across judges, and not to differences in case or defendant characteristics. We may then test the hypothesis that all judges have identical sentencing propensities with respect to race through a simulation procedure similar to the one described above.[20] The only difference is replacing a case characteristic with a case outcome measure, like incarceration rate or sentence length. The simulation procedure is as follows.

First, we compute the outcome of interest for each judge. For example, we compute the difference in average sentence length between African-American defendants and White defendants. If race has no impact on judicial decision-making, this difference should be very similar across judges.[21] We can test whether there is excess inter-judge disparity in this outcome by comparing the empirical dispersion with that from simulated data in which there is no excess disparity by construction. In order to construct the distribution under the null of no disparity, we simulate new data as above, replacing the original case data with that from a randomly chosen case from the same cell. The only difference is that now the cells are restricted further – the simulated case must be from

[20] We implicitly assume that cases do not affect each other. In particular we assume that the racial composition of a judge's other previous cases, do not affect future decisions.
[21] Alternatively, we would find the same result if race impacted all judges' decisions the same way.

the same month *and* have the same defendant race as in the original case. In this way, we compute a simulated distribution of racial gaps by judge. We then calculate a measure of the inter-judge dispersion of the difference in average sentence length by race for each simulation as the test statistic. Finally, we compare the empirical measure of the test statistic to its distribution from all the simulations. This allows us to determine, for example, what proportion of the simulated distributions have a larger 5-95 spread than the empirical distribution. This proportion is the probability that the empirical distribution would have a dispersion of the magnitude observed or larger by chance, when there was in fact no inter-judge difference in the racial gap in sentencing.

This procedure has three benefits. First, it allows us to simulate the sentencing gap for each judge.[22] Second, it allows us to address the small sample problem. The simulated data produces an unbiased distribution of the inter-judge disparity measure which is not reliant on a large N assumption. Finally, this distribution allows us to compute a traditional p-value. Using it, we can determine the probability of observing the empirical inter-judge disparity measure if cases are randomly assigned to judges and race has no impact on judicial decision-making. All of the above-described procedures focus on the racial gap, but could of course also be used to identify the impact of any case characteristics on judicial decision-making.

V. Results

Because random case assignment is crucial to determine whether judges vary in their treatment of race, we examine it first, using the Monte Carlo methodology discussed

[22] Because judges may vary in the time periods they serve, the expected racial gap may be different across judges.

in the previous section. Figure 3 displays the results of the simulation using defendant race as a check for random assignment of cases. Since the empirical interquartile range falls squarely in the middle of the simulated distribution, with a p-value of .26, we conclude that there was no systematic bias in the distribution of defendant race among judges in our sample. Figure 4 reports the results of the random assignment check using defendant gender as the case characteristic of interest. We find a p-value of .57 and therefore cannot reject the null hypothesis that cases were also randomly assigned to judges with respect to defendant gender.

We find similar results when we perform the same Monte Carlo simulations using other specifications. In particular, we test case type and defendant age as case characteristics, and we also test defendant characteristics by subset of case types. These test results are presented in Table 5, where we report, for each defendant or case characteristic, the empirical interquartile range (IQR), mean and standard deviation of the simulated IQRs, as well as the associated p-value.

Additional measures of the spread of the distribution of observable case characteristics, including 10-90 percentile range and 5-95 percentile range, all support the basic hypothesis that cases were randomly assigned to judges. Based on the random assignment of all observables we can test, we conclude that judges will receive the same distribution of unobservable case characteristics as well. Thus differences in sentencing between judges are attributable solely to their characteristics and preferences, and not to differences in case types.

Having established the random assignment of cases to judges, we first examine inter-judge variation in sentence length and incarceration rates. While not the focus of

our inquiry, this is a useful baseline measure before examining differential sentencing by race. Even independent of defendant characteristics, judges in our sample demonstrate substantial heterogeneity in their sentencing decisions. In Table 6, we report results comparing actual heterogeneity to the null hypothesis of no mean differences in sentencing and incarceration rates, using the Monte Carlo methodology detailed above. All measures of dispersion are at least 20% lower than that in a federal district court evaluated in Waldfogel (1998). This is not particularly concerning, given that federal and state courts differ in numerous ways.

In comparison with a simulated dispersion, judges' decisions show excess heterogeneity in all measures including incarceration ("jail"), average sentence length ("sentence"), and average sentence length conditional on receiving a non-zero jail sentence ("sentence2"). This is true not only in the inter-quartile range but also in the 10-90 gap and the 5-95 gap. Figure 5 shows the inter-judge variability in incarceration rate. We can reject the null hypothesis that the average incarceration rate does not vary across judges with a p-value of less than .001. There appears to be substantial heterogeneity in judicial sentencing in our dataset. This finding of inter-judge sentencing disparity is consistent with previous research focusing on other courts. In particular, Anderson et al. (1999) found significant inter-judge sentencing variation in federal courts. They further found that this disparity was reduced only modestly by federal sentencing guidelines.

We now turn to the main objective of this paper, which is to study whether there is excess heterogeneity across judges with regard to the racial gap in sentencing. Table 7 summarizes the results of the Monte Carlo simulations. Figure 6 shows that the inter-quartile range of the empirical distribution of the racial difference in incarceration rates is

significantly larger (with a p-value of .01) than if judges were sentencing without regard to race. That is, we find significant judge-race interactions in the incarceration rate. This result indicates that there is variation in judicial behavior in our sample when it comes to the decision of whether or not to incarcerate defendants of different races.

We next examine whether there is an analogous impact of defendant race on sentence length. In Table 7 and Figure 7 we present the empirical interquartile range and simulated interquartile ranges for the racial gap in sentence length. Unlike incarceration, there is no evidence of excess inter-judge variation in the racial sentencing gap beyond what we would expect from sampling variation alone. Thus it appears there are substantial differences in behavior across the judges when it comes to the decision of whether or not to incarcerate defendants of different races, but not to the same extent when it comes to the decision of setting sentence length. Table 7 also shows that the lack of excess inter-judge heterogeneity in the racial gap in sentence length extends to conditioning on strictly positive sentences.[23]

These findings are consistent with recent criminology literature describing attempts to measure the direct effect of race on sentence length. For example, Spohn (2000) notes that the evidence is more compelling for a racial impact in the incarceration decision rather than the sentence length. While none of the studies reviewed avoid the omitted variables bias difficulty, it is worth noting that these earlier findings are consistent with those in this study. This begs the question, why do we find excess

[23] We conduct the same analysis that is reported in Table 7 for a Hispanic subset of data (that is the original data restricted to Hispanic and White defendants). We follow the same criteria in constructing this subset as we did for the African-American subset (see Section III and Appendix A for detail). The main characteristics of the Hispanic subset are reported in Tables A1 and A2. Like African-American defendants, the Hispanic defendants also have higher raw incarceration rates than White defendants. However, the difference is much smaller, and not statistically significant. The main finding reported in Table A3 is that, unlike for the African-American sample, we find no evidence of excess inter-judge heterogeneity in the Hispanic-White gap in incarceration rate or sentence length.

heterogeneity in incarceration rate, but not sentence length. One possible explanation is that Illinois sentencing guidelines reduce the latitude of individual judges to tailor sentences.[24]

It is important to gain an idea of the magnitude of the inter-judge racial gap in incarceration rate. Table 8 reports the effect of a shift from a judge at the 25th percentile of the racial sentencing gap to the 75th percentile judge to be an increase of 11 percentage points in probability of incarceration and nearly 3 months in sentence length. This compares with a mean incarceration rate of 49% and racial gap of 13 percentage points, and mean sentence length of 20 months and racial gap of 5 months. The difference between a defendant who is randomly assigned to the 10th percentile judge versus one assigned to the 90th percentile judge is (not surprisingly) even more striking. There, the racial gap in incarceration rate rises by a full 18 percentage points while expected sentence length increases by 10 months. While the sentencing gap is large in magnitude, this gap cannot, as we established above, statistically be distinguished from that which would arise simply due to sampling variability (See Figure 7 and Table 7).

To make these results a bit more concrete, consider the expectations of incarceration for two pairs of otherwise identically situated defendants, who differ only by race. William L., who is White, and Bob L., who is African-American, have their cases heard before Judge Lenient, who is at the 10th percentile in the racial gap in incarceration rate. Bill H., who is African-American and Walter H,, who is White, appear before Judge Harsh, whose mean racial gap in incarceration rate puts him at the 90th percentile. Besides their race and (random) judicial assignment, all four defendants

[24] Waldfogel (1998) shows that under some realistic assumptions that guidelines are not an effective way to reduce interjudge sentencing disparity. Pfaff (2006) points out that Illinois guidelines are relatively broad, compared to other states.

and their crimes are otherwise identical. The difference between Bill H. and Walter H.'s likelihood of incarceration is 18 percentage points greater than that for Bob L. and William L. So while William L. may expect a 35% chance of incarceration and Bob L. a 45% likelihood, Walter H. may face a 40% probability of incarceration and Bill H. a 68% chance.

Given the significant heterogeneity between judges, a further question suggests itself: are any observable judge characteristics predictive of where judges fall in the empirical distribution of the racial gap in sentencing? We examine this question in Table 9. To perform this analysis, we construct a dataset of judge fixed effects and regress these fixed effects on judge-level characteristics such as those reported in Table 3. We estimate the judge fixed effects γ_j in Equation (2) above for both incarceration rate and sentence length. We use the inverse of the square of the estimated standard error to weight each observation in the judge-level regressions. For the sake of completeness, we also estimate judge fixed effects in average incarceration rate and average sentence length and relate those to observable judge characteristics. We do this by estimating the judge fixed effects δ_j in Equation (1) above using both incarceration rate and sentence length as dependent variables. Estimated standard errors are again used for weighting in the judge-level regressions.

As the first two columns of Table 9 indicate, there is no systematic relationship between judges' characteristics such as their race, gender, age or prior experience in public defense and how harsh judges are on average. For example, while the point estimates indicate that male judges give sentences that are on average about 50 days longer (column 1) and that they incarcerate about 3 percentage points more (column 2),

these differences are not statistically significant. The point estimates in columns 1 and 2 are of different signs for African-American judges; they are associated with longer sentences on average but incarcerate at a lower rate, though again, neither of these is statistically significant.

The remaining columns of Table 9 relate judge fixed effects in the racial gap in sentencing (columns 3 and 4) and in the racial gap in incarceration rate (columns 5 and 6) to judge characteristics. A few somewhat more robust patterns emerge from these regressions. First, and most interestingly, it appears that African-American judges are associated with a smaller racial gap in sentence length. This effect is substantial (about 150 days) and statistically significant. The point estimates indicate that African-American judges are also associated with smaller racial differences in incarceration rate (about 3 percentage points) but this effect is not statistically significant. The point estimates indicate that older male judges might be associated with larger racial differences but these effects are statistically insignificant and smaller in magnitude than the "African-American judge" effect. Also, no clear pattern emerges based on whether the judge has prior experience in public defense.

In columns 4 and 6, we include additional regressors, judge fixed effects for average sentence length (column 4) and for average incarceration rate (column 6). Both are positively correlated with the fixed effects on racial differences in sentencing. Hence, judges that are tougher on average are also relatively tougher on African Americans.

VI. Potential Confounds and Analysis by Crime Category

Our results are consistent with differential judicial treatment of minority defendants, at least with respect to the decision to incarcerate. Some judges show a much larger racial gap in incarceration rates than other judges, even when facing the same types of defendants and cases. There are several potential concerns regarding the interpretation of these findings, which we now discuss in detail.

African-Americans may commit different crimes than Whites and judges may have different sentencing policies for different crimes. For example, suppose some judges are stricter on violent crimes than others. Suppose also that African-Americans commit more violent crimes. This correlation would then lead to the appearance of heterogeneity in racial gaps in sentencing even if judges were race blind. One strategy for accounting for these differences in crime categories is to look separately in different categories of crime. The difficulty with this approach is that once divided this way, each category contains a relatively small number of observations. In performing this analysis (not reported in tables) we find no evidence for excess heterogeneity in racial gap in any crime category. This result is almost certainly due to lack of power.

In order to address the problem of diminishing the sample size, we run our central analysis while controlling for the category of crime committed. We implement this by subtracting out judge-specific means by crime category for both incarceration and sentence length. The results are reported in Panel A of Table 10. We find very similar results to those in the main specification. There is evidence of excess heterogeneity in the racial gap in incarceration rates, but not sentence lengths.

In order to get some purchase on whether there is any variation in judicial decisions by case type while maintaining sufficient observations to ensure a meaningful test, we subdivide the data into drug and non-drug cases. The results from this analysis are reported in Panels B and C of Table 10. Focusing on the incarceration racial gap, we find excess dispersion for non-drug cases ($p = .043$), but not drug cases ($p = .868$). Although there are fewer drug cases than non-drug cases, the disparity is only 35% and thus a lack of power is unlikely to be the cause of the difference. One plausible explanation is that the Illinois sentencing guidelines provide less judicial discretion in the incarceration decision for drug offenses than with non-drug offenses.

While correlation between race and crime type is the most obvious potential confound, this is an example of a more general concern. Suppose there are unobservable (to us) features of the case, which some judges care more about than others. For example, there may be details of the crime that are not captured by the statute the person is being charged under. Alternatively, there may be details of the evidence (such as use of DNA tests), which are not in our data set. These unobservable case features could in principle generate the type of variation we observe if these unobserved features vary systematically across racial groups and judges differ in their treatment of these characteristics. This would happen in the above example if DNA evidence was used more against one racial group than another. It seems unlikely that under this model, a characteristic such as judge's race would systematically predict the racial gap in sentencing (Table 9). While still potentially a concern, the approach in this paper advances the field over previous work, because now the unobservables would have to be correlated with defendant race and elicit differential treatment across judges.

VII. Conclusion

In this paper we have sought to shed light on the influence of race in judicial sentencing practices. Previous research has largely made use of OLS regressions in addressing this topic. This approach may suffer from an omitted variables problem, which could substantially bias any estimate of the influence of race on sentencing.

We make use of the random assignment of cases to judges in order to address omitted variables bias. With random assignment of cases, judges will receive the same distribution of case characteristics, both observed and unobserved. Thus if all judges are unbiased, one would expect the racial gap in sentencing to be the same across judges, to within sampling error. The core of our analysis is establishing what the gap would be for unbiased judges, and comparing this with the actual data.

This is accomplished using a Monte Carlo simulation, sampling from the actual data, but mechanically breaking the judge-defendant race link. We find that there is substantial excess heterogeneity in the empirical distribution of the racial gap in incarceration rate. The quantitative impact of this gap on sentencing disparity is of considerable magnitude. In moving from a defendant that was assigned to the 10th percentile judge to the 90th percentile judge, the racial gap in incarceration rate rises by a full 18 percentage points.

It is also useful to consider potential legal policy implications in light of these findings. One goal of policy changes could be to try to reduce or eliminate the excess inter-judge heterogeneity in the racial gap. This analysis can inform how big an impact that sort of policy change would make. If the excess inter-judge racial gap in

incarceration was eliminated, the interquartile range of the racial gap in incarceration would drop from .106 to .073 (Table 7). This represents a 31% reduction in the variability of the black-white racial gap in incarceration due *just to judicial assignment*. The magnitude of this potential effect would decrease one element of the randomness in the judicial process and surely increase confidence in the fairness of the courts.

One important limitation of our work is that while we show that race appears to play a role in judicial decision-making, we cannot make statements about its optimality. That is, we can say that judges vary in their treatment of race, but not whether this is evidence of discrimination or reverse discrimination. It is theoretically possible that the heterogeneity in the racial gap in incarceration reflects favoritism by some judges towards African-American defendants. For example, suppose unobservable case characteristics dictated that an unbiased racial gap in sentencing would be 50%. In this case heterogeneity in the race gap between 20% and 50% would indicate a great deal of favoritism towards African-Americans, not discrimination. In future work, information on inter-judge differences in the racial gap in recidivism may further guide the interpretation of our findings. In particular, one may relate the variation we observe in the racial gap in sentencing to variation in the racial gap in recidivism. Additionally, information on the success rate of appeals may provide another method of evaluating the optimality of the racial gap. The theoretical ideal would be to evaluate a social welfare function with terms that included both recidivism and appeals, and all other relevant factors.

Despite this interpretational limitation, our findings nevertheless raise important legal questions. Heterogeneity across judges in sentencing by race suggests that

courtroom outcomes may not be race blind. This may be one source of the substantial overrepresentation of African-Americans in the prison population. Understanding the sources of variation in the criminal justice system is an important first step toward reducing disparities of various kinds.

Appendix A. Data Cleaning Procedure

The data for this study comes from the Cook County Circuit of the Illinois state courts. For each felony case that is prosecuted, a record is made of key case details including defendant characteristics (race, sex, age, etc.), case traits (crime type, assigned judge, court location), and outcomes (sentence length, plea, finding of guilt). A substantial amount of data cleaning was necessary to prepare the data for analysis. This appendix details that process.

The initial data processing removed observations with erroneous data. For example, observations where the sentence length was inaccurate or unintelligible, such as "2 months 400 days" were excluded. Other dropped observations include those with erroneous dates (too far in the past or in the future), negative sentences, duplicate observations based on case number, and missing race.

Sentences were top coded to 60 years under the assumption that defendants were unlikely to serve longer, based on the median defendant age. Life sentences were also coded as 60 years. The guilty binary indicator was set to equal guilty when sentences were nonzero and the guilty variable was missing. We dropped any observation where the guilty and sentence variables both were non-missing and contradicted each other (i.e. defendant found not guilty but with non-zero sentence length).

Defendants with cases already pending in the courts are sometimes assigned to the same judge, thus we keep only the first time a defendant appears in the data, because only these cases are likely to be truly random. Establishing unique defendant identities is difficult due to frequent miscoding, which we attempt to address with several procedures.

A unique defendant ID is defined by last name, race, and sex. Last name is defined as the last word in the defendant's name. The identification is further refined by a fuzzy match on date of birth. Due to miscoding of this variable, we count two observations as having the same defendant if they match on last name, race, and sex, and have at most one digit different in their dates of birth. For example, Kev Marshall with birthday 124278 (with the tens digit in day miscoded) would be the same individual as Kevin Marshall with birthday 120278.

Once the dataset is winnowed to a single observation per defendant, there are still a number of other data cleaning procedures we undertake, due to further idiosyncrasies of the dataset and coding errors. Homicide cases are not allocated using the standard random assignment method, (their assignment takes into account judicial caseload) and thus we exclude them from our sample. The variable indicating the courthouse location is often miscoded. This poses a serious problem because cases arising in Rolling Meadows, Skokie, and other suburban courthouses have vastly different characteristics than cases from Chicago.

We use two procedures to attempt to exclude cases actually originating from suburban locations. First, we drop all of the cases in a given year for a judge who has any cases outside the main Chicago courthouse (located at 26th & California) in that year. For example, Judge Roberts may have 100 cases at 26th & California every year from 1994 to 2003, but in 1996, he took on a case at Rolling Meadows. This would drop all of his cases for 1996. Second, we compute a measure of the dispersion of defendant home zip codes for each judge. We drop all cases for a judge in a year in which this measure deviates from the mean by over 10%.

For certain years in our range, the Cook County courts had judges who adjudicated only drug cases. The cases assigned to these judges were clearly non-random along the case type dimension. In order to exclude them, we drop cases heard by judges for whom drug cases comprise more than 70% of their caseload for the year.

After the preceding case culling we ran the random assignment check across multiple dimensions on the remaining data at the month level. We were unable to verify random assignment prior to 1995, so we exclude this data. We further restrict ourselves to cases begun before 2002, in order to prevent truncation bias from impacting the results, as cases can often stretch on for several years.

Figure A1

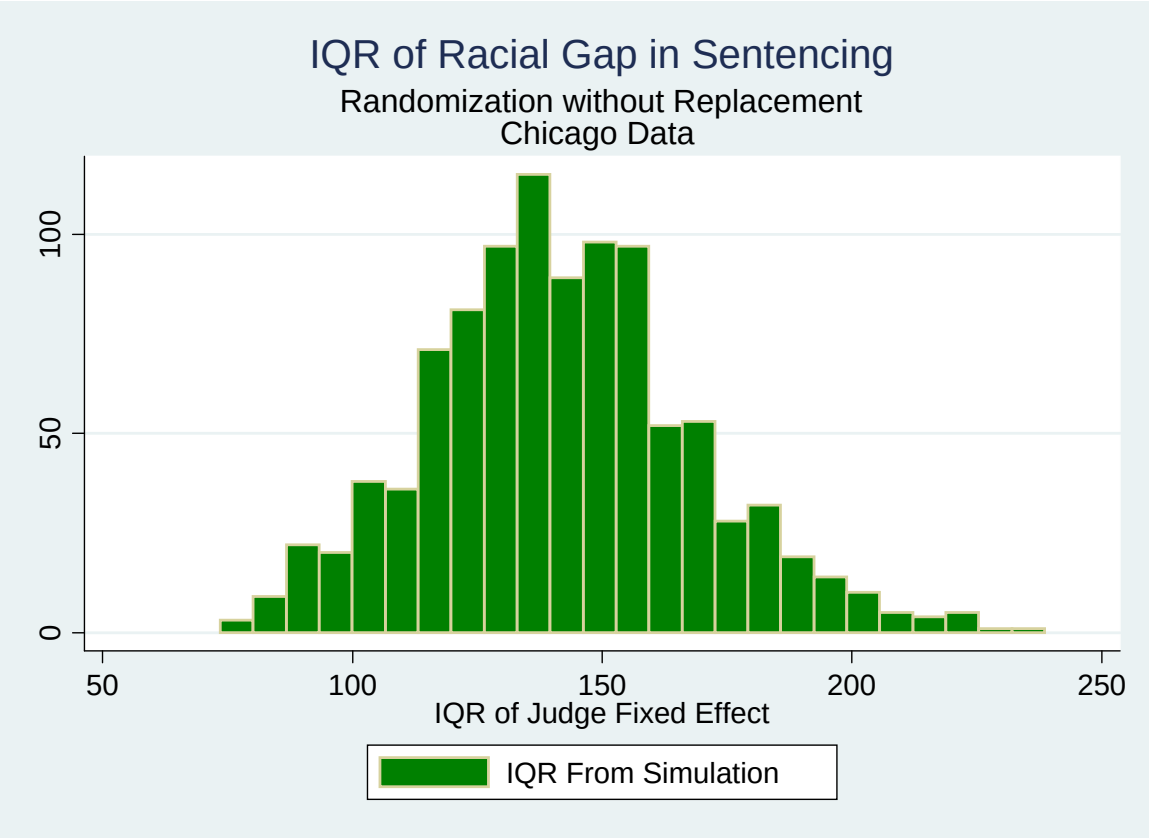
IQR of Racial Gap in Sentencing
Randomization without Replacement
Chicago Data
0
50
100
50
100
150
200
250
IQR of Judge Fixed Effect
IQR From Simulation

Table A1: Summary Statistics
Hispanic Subset

	Mean	Standard Deviation
Fraction Hispanic	0.56	0.5
Fraction Male	0.88	0.32
Age	29	10
Cases Per Judge	174	133
Charges per Case	2.4	4.2
Plea	0.76	0.43
Guilty Verdict	0.92	0.27
Probation	0.29	0.46
Incarceration	0.41	0.49
Sentence Length (months)	18	37
Sentence length (non-zero)	43	46
Judges	75	
Total Cases	11946	

Table reports means and standard deviations of case characterstics. Cases involve felony offenses in Cook County District Court initiated from 1995-2001 in which the defendant was Hispanic or White (see appendix for further detail on dataset).

Table A2: Sentencing Breakdown
Hispanic Subset

	Incarceration Rate		Sentence Length		Sentence Length Conditional on non-zero	
	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev
Total:	0.41	0.49	18	37	43	46
...by Type of Charge						
Drugs	0.34	0.48	7.1	16	20	22
Violent Crime	0.41	0.49	21	40	50	49
EFT	0.48	0.5	19	29	40	30
Other	0.41	0.49	22	46	55	59
...by Race						
Hispanic	0.44	0.5	21	39	47	49
White	0.38	0.49	15	32	39	42
Judges	75					
Total Cases	11946					

Table reports means and standard deviations of case characterstics by charge category and race. Cases involve felony offenses in Cook County District Court initiated from 1995-2001 in which the defendant was African-American or White (see appendix for further detail on dataset). Sentence length measured in months.

Table A3: Dispersion of Racial Gap in Sentencing and Incarceration Rates, Hispanic Subset

Variable Name	Empirical IQR	Simulation Mean	Simulation St Dev	P Value	Observations
jail	0.06	0.09	0.02	0.97	11946
sentence	172.58	193.31	32.52	0.75	11946
sentence2	288.84	383.91	66.68	0.93	4888

The Empirical IQR column reports the interquartile range of the distribution of the racial gap judge fixed effect for the given variable. Simulation mean reports the mean of the interquartile range from 1000 simulations; St Dev reports the standard deviation from the simulations. The p-value indicates the percentile of the simulated data to which the empirical data corresponds. Simulations randomly choose an outcome chosen from cases initiated in the same month and with the same defendant race as the original case. jail is a binary variable indicating whether the defendant was incarcerated. sentence2 is sentence length conditional on receiving a non-zero sentence. sentence and sentence2 measured in days. Cases involve felony offenses in Cook County District Court initiated from 1995-2001 in which the defendant was Hispanic or White. See additional explanation in the text.

References

Abrams, D. S., & Yoon, A. H. (2007). The Luck of the Draw: Using Random Case Assignment to Investigate Attorney Ability. *University of Chicago Law Review*, *74*(4), 1145-1177.

Albonetti, Celesta A. "Sentencing under the Federal Sentencing guidelines: Effects of defendant characteristics, guilty pleas, and departures on sentence outcomes for drug offenses, 1991-1992." *Law and Society Review*, pp. 789-822, 1997.

Anderson, James M; Kling, Jeffrey R and Stith, Kate. "Measuring Interjudge Sentencing Disparity: Before and after the Federal Sentencing Guidelines." *Journal of Law and Economics*, 1999, 42(1; 2), pp. 271-307 38 pages.

Ashenfelter, Orley, Theodore Eisenberg, and Stewart J. Schwab. "Politics and the Judiciary: The Influence of Judicial Background on Case Outcomes." *Journal of Legal Studies* 24:257–81, 1995.

Ayres, Ian and Joel Waldfogel. "A Market Test for Race Discrimination in Bail Setting." Stanford Law Review, Vol. 46 (5), pp. 987-1047, 1994.

Bushway, Shawn D. and Anne M. Piehl. "Judging Judicial Discretion: Legal Factors and Racial Discrimination in Sentencing." *Law and Society Review*, Vol. 35 (4), pp. 733-764, 2001.

Cheng, Edward K. "The Myth of the Generalist Judge." *Stanford Law Review*, Vol 61(3), pp. 519 – 572, 2008.

Daly, Kathleen and Michael Tonry. "Gender, Race, and Sentencing." *Crime & Justice*, Vol. 22, pp. 201-252, 1997.

Devine, Dennis J., et al. "Jury Decision Making: 45 Years of Empirical Research on Deliberating Groups." *Psychology, Public Policy, and Law*, Vol. 7(3), pp. 622-727, 2001.

Gaudet F, Harris GS, St. John CW. "Individual Differences in the Sentencing Tendencies of Judges." *Journal of Criminal Law and Criminology*. 23(5):811-818. 1933.

Fischman Josh B. "Estimating Preferences of Appellate Judges: A Model of 'Consensus Voting'" University of Virginia School of Law School Law and Economics Research Paper Series No. 2009-142010. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1361348. 2010.

Humphrey, John A. and Timothy Fogarty. "Race and Plea Bargained Outcomes: A Research Note." *Social Forces*, Vol. 66 (1), pp. 176-182, 1987.

Imbens, Guido W. and Paul R. Rosenbaum. "Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Vol. 168(1), pp. 109-126, 2005.

Kennedy, Peter. A Guide to Econometrics. MIT Press. Cambridge, MA. 1998.

Klein, Stephen, Joan Petersilia, and Susan Turner. "Race and Imprisonment Decisions in California." *Science*, Vol. 247 (4944), pp. 812-816, 1990.

Klepper, Steven, Daniel Nagin, and Luke-Jon Tierney. "Discrimination in the criminal justice system: A critical appraisal of the literature." In *Research on sentencing: The search for reform*, edited by Alfred Blumstein, Jacqueline Cohen, Susan E. Martin, and Michael H. Tonry. Vol. 2. Washington, D.C.: National Academy Press. 1983.

Kramer, John H. and Darrell Steffensmeier. "Race and Imprisonment Decisions." *Sociological Quarterly*, Vol. 342 (2), pp. 357-376, 1993.

MacCoun, Robert J. "Experimental Research on Jury Decision-Making." *Science*, Vol. 244, pp. 1046-1050, 1989.

Mustard, David B. "Racial, ethnic, and gender disparities in sentencing: Evidence from the U.S. federal courts." *Journal of Law and Economics*, Vol. 44, pp. 285-314, 2001.

Payne, A. Abigail. "Does inter-judge disparity really matter? An analysis of the effects of sentencing reforms in three federal district courts." *International Review of Law and Economics*, Vol. 17, pp. 337-366, 1997.

Pfaff, John F. "Continued Vitality of Structured Sentencing Following Blakely: The Effectiveness of Voluntary Guidelines." *UCLA Law Review*, Vol. 54, pp. 235-308. 2006.

Price, Joseph and Justin Wolfers, "Racial Discrimination among NBA Referees," *Quarterly Journal of Economics*, Vol. 125 (2010): forthcoming.

Rachlinski, Jeffrey J., Sheri Johnson, Andrew J. Wistrich, and Chris Guthrie, "Does Unconscious Racial Bias Affect Trial Judges?" *Notre Dame Law Review*, Vol. 84, pp/ 1195-1246. 2008.

Sabol, William J., West, Heather C., & Cooper, Matthew. "Prisoners in 2008." US Department of Justice, Bureau of Justice Statistics. (Washington, DC: US Department of Justice, Apr. 2010)
Available at http://bjs.ojp.usdoj.gov/index.cfm?ty=pbdetail&iid=1763.

Schanzenbach, Max. "Racial and Sex Disparities in Prison Sentences: The Effect of District-Level Judicial Demographics." *Journal of Legal Studies*, Vol. 34, pp. 57-92, 2005.

Sommers, Samuel R. and Phoebe C. Ellsworth. "Race in the Courtroom: Perceptions of Guilt and Dispositional Attributions." *Personality and Social Psychology Bulletin*, Vol 26 (11), pp. 1367-1379, 2000.

Spohn, Cassia C. "Thirty Years of Sentencing Reform: The Quest for a Racially Neutral Sentencing Process." in *Policies, Processes, and the Decisions of the Criminal Justice System*, vol. 3. Washington, DC: Department of Justice, 2000.

Steffensmeier, Darrell and Stephen Demuth. "Ethnicity and Sentencing Outcomes in U.S. Federal Courts: Who is Punished More Harshly?" *American Sociological Review*, Vol. 65 (5), pp. 705-729, 2000.

Steffensmeier, Darrell and Stephen Demuth. "Ethnicity and Judges' Sentencing Decisions: Hispanic-Black-White Comparisons" *Criminology*, Vol. 39, pp. 154-178, 2001.

Thomson, Randall J. and Matthew T. Zingraff. "Detecting Sentencing Disparity" Some Problems and Evidence." *The American Journal of Sociology*, Vol. 86 (4), pp. 869-880, 1981.

Waldfogel, Joel. "Aggregate Inter-Judge Disparity in Sentencing: Evidence from Three Districts." *Federal Sentencing Reporter*, Vol. 4, pp. 151-154, 1991.

Waldfogel, Joel. "Does Inter-Judge Disparity Justify Empirically Based Sentencing Guidelines?" *International Review of Law and Economics*. Vol. 18:293-304, 1998.

Zussman, Asaf and Moses Shayo. "Judicial Ingroup Bias in the Shadow of Terrorism." Working Paper. 2010.

Figure 1

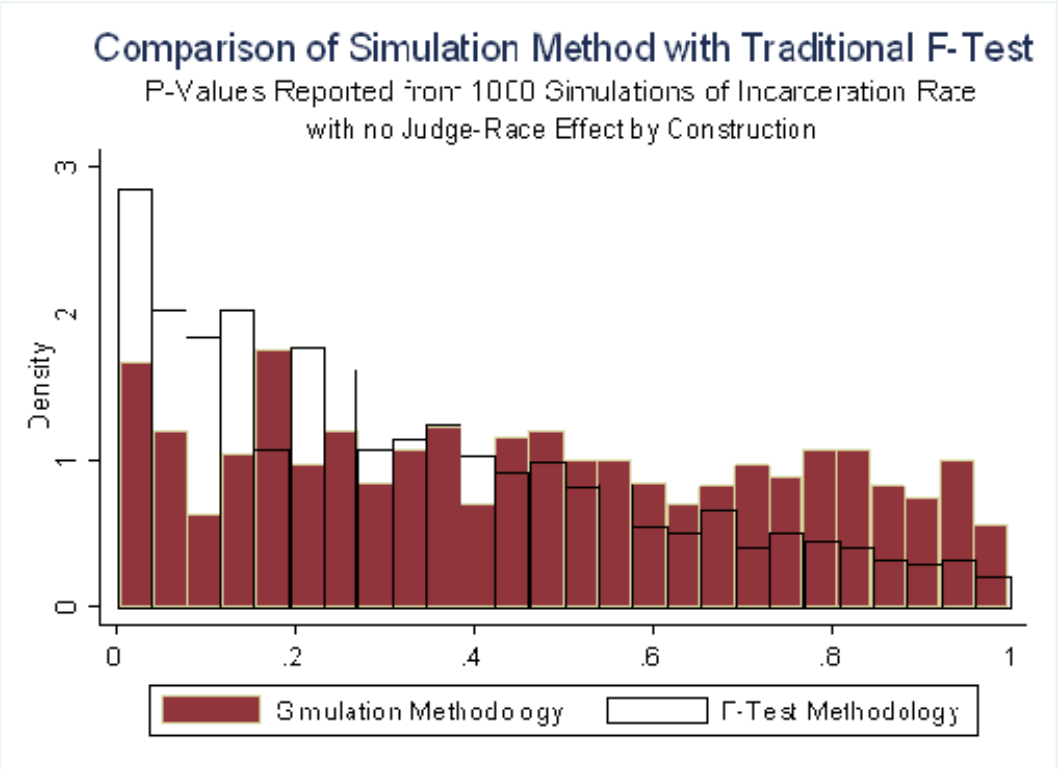

Figure 2

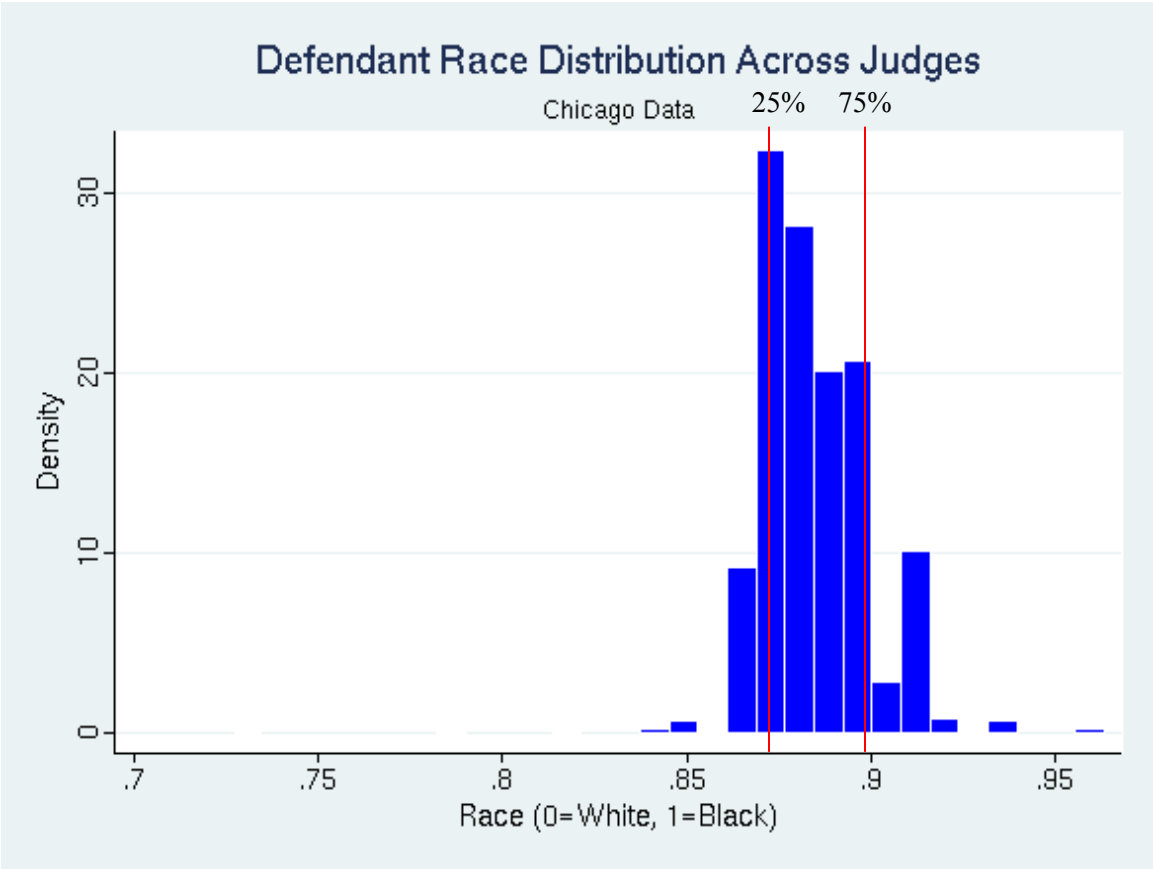

Figure 3

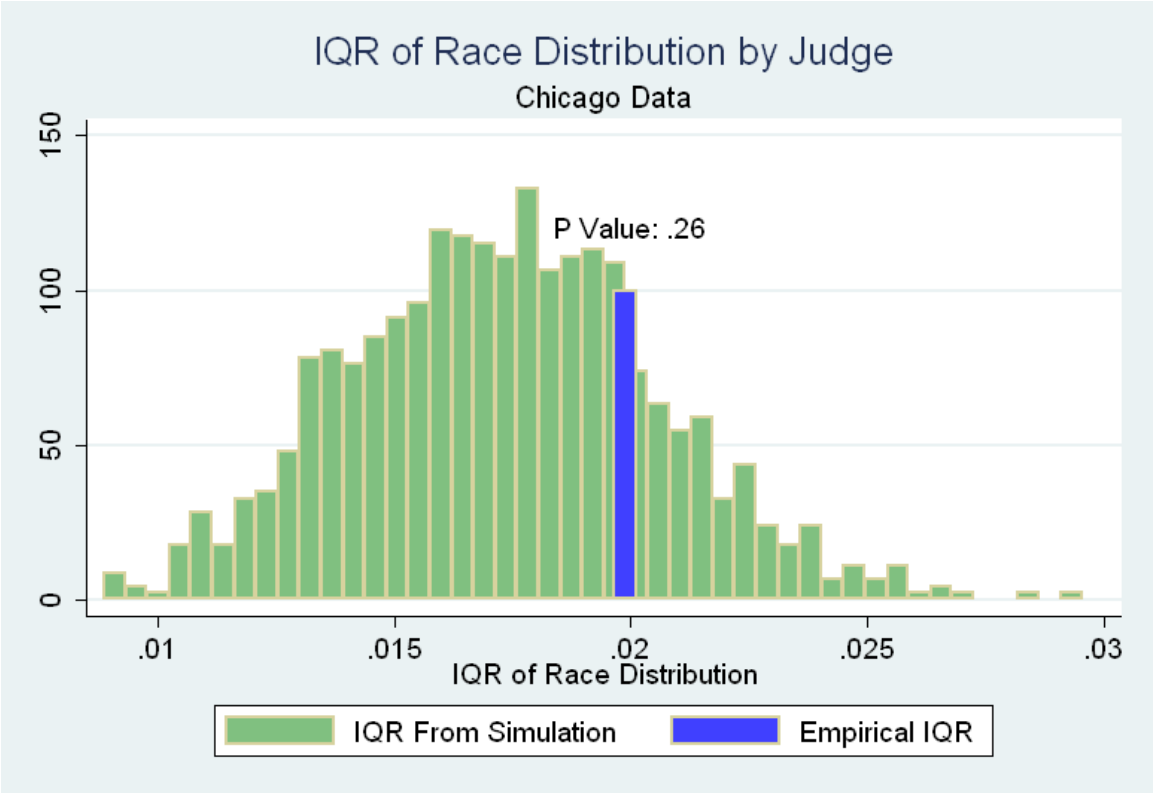

Figure 4

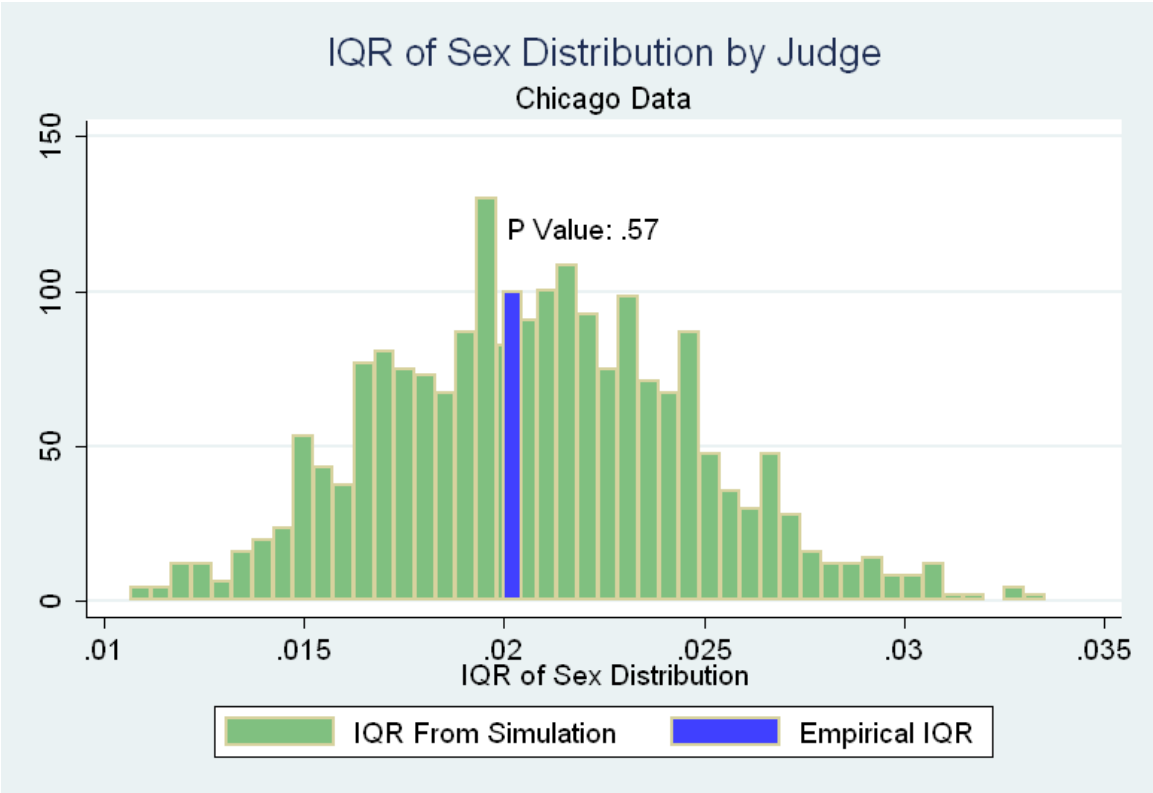

Figure 5

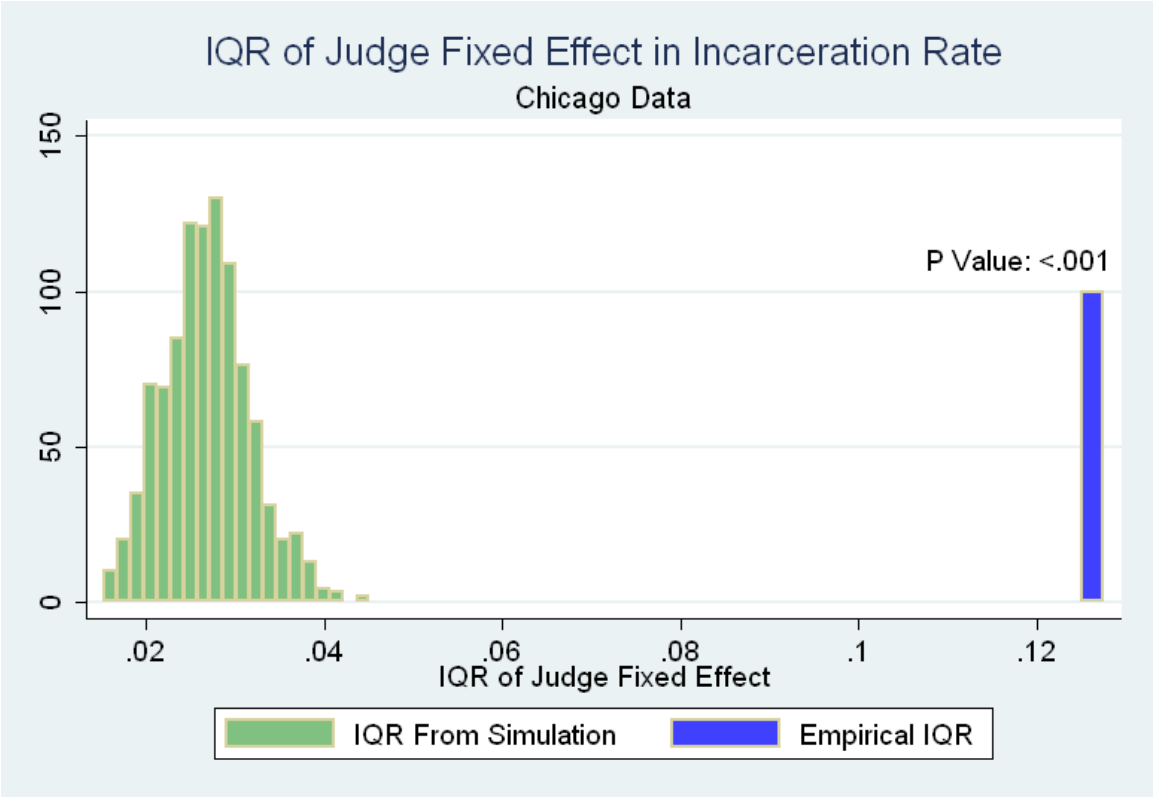

Figure 6

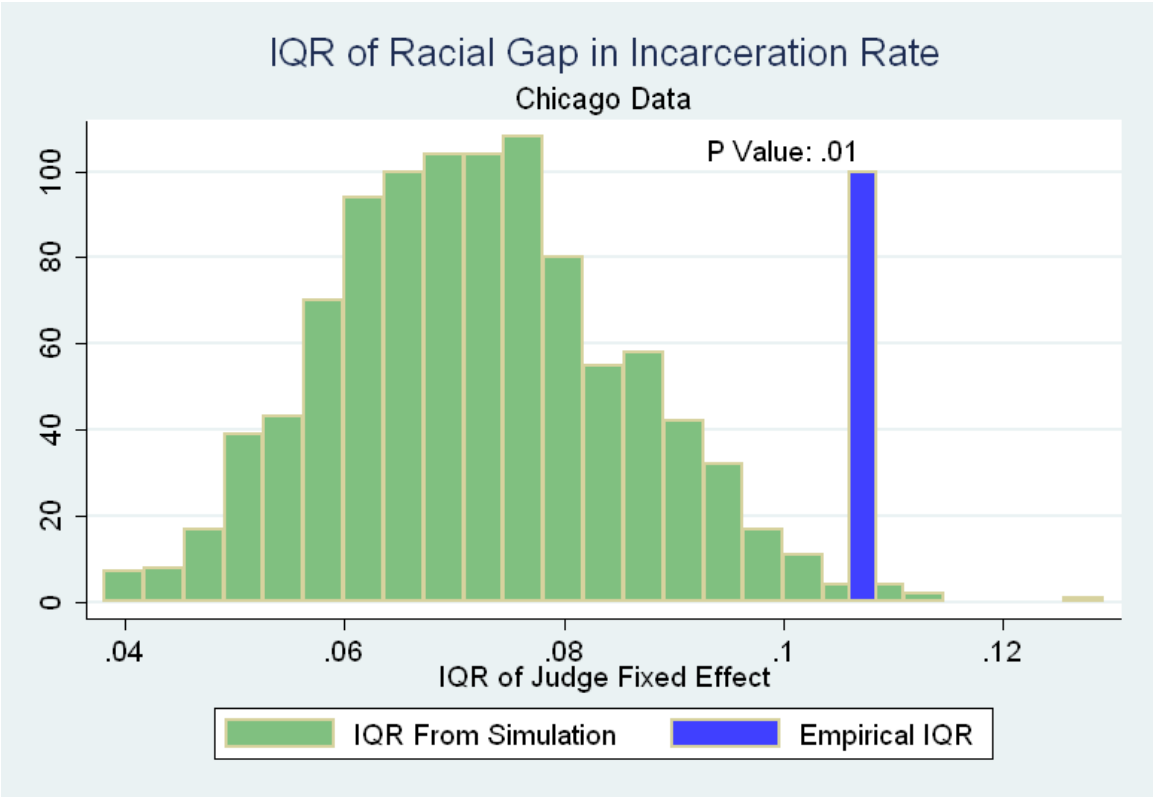

Figure 7

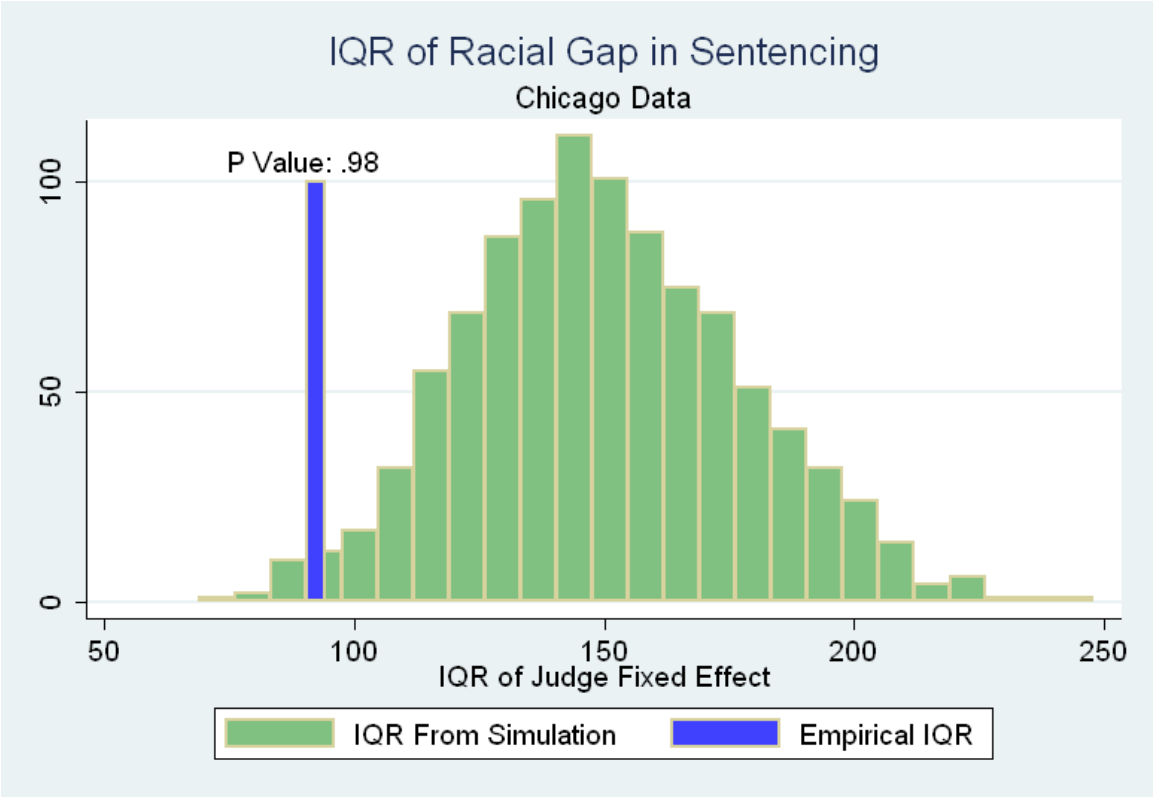
IQR of Racial Gap in Sentencing
Chicago Data
P Value: .98
100
50
0
50
100
150
200
250
IQR of Judge Fixed Effect
IQR From Simulation
Empirical IQR

Table 1: Summary Statistics for Cook County and Chicago, IL

	Cook County		Chicago		Court Data	
	Population	Percent	Population	Percent	Population	Percent
White (Non-Hispanic)	2,558,709	47.6%	907,166	31.3%	120,389	18.0%
Black (Non-Hispanic)	1,390,448	25.9%	1,053,739	36.4%	487,732	73.1%
Other	355,844	6.6%	181,467	6.3%	3,031	0.5%
Hispanic	1,071,740	19.9%	753,644	26.0%	56,328	8.4%
Total	5,376,741		2,896,016		667,480	

Source: U.S. Census Bureau, Census 2000
Cook County District Court felony cases 1985-2005

Table 2A: Summary Statistics
African-American Subset

	Mean	Standard Deviation
African American	0.86	0.35
Male	0.83	0.38
Age	29	10
Cases Per Judge	489	417
Charges per case	2.4	5.1
Plea	0.69	0.46
Guilty verdict	0.92	0.27
Probation	0.25	0.44
Incarcertation	0.49	0.5
Sentence Length (months)	20	36
Sentence length (non-zero)	42	42
Judges	70	
Total Cases	34227	

Table reports means and standard deviations of case characterstics. Cases involve felony offenses in Cook County District Court initiated from 1995-2001 in which the defendant was African-American or White (see appendix for further detail on dataset).

Table 2B: Sentencing Breakdown
African-American Subset

	Incarceration Rate		Sentence Length		Sentence Length Conditional on non-zero	
	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev
Total:	0.49	0.5	20	36	42	42
...by Type of Charge						
Drugs	0.5	0.5	15	22	30	23
Violent Crime	0.47	0.5	24	43	52	50
EFT	0.56	0.5	23	31	41	31
Other	0.46	0.5	24	48	53	31
...by Race						
African American	0.51	0.5	21	36	42	41
White	0.38	0.48	16	33	42	43
Judges	70					
Total Cases	34227					

Table reports means and standard deviations of case characterstics by charge category and race. Cases involve felony offenses in Cook County District Court initiated from 1995-2001 in which the defendant was African-American or White (see appendix for further detail on dataset). Sentence length measured in months. EFT is a case category that stands for "embezzlement, fraud, theft".

Table 3: Judge Characteristics

	Mean
Male	0.82
White	0.86
Age	49
Private Practice	0.49
Defense attorney	0.27
Prosecutor	0.70
Judges	70

Table reports judge characteristics for cases involve felony offenses in Cook County District Court initiated from 1995-2001 in which the defendant was African-American or White (see appendix for further detail on dataset).

Source: Sullivans Judicial Profiles
Directory of State and Federal Judges in Chicago
The Directory of Minority Judges in the United States

Table 4: Monte Carlo Race Simulation Example

			Real Data	Simulation 1	Simulation …
Judge	Case #	Date	Race	Race	Race
Wapner	1001	1/1/2000	Black	Black	White
	1414	1/15/2000	White	Black	Black
	…				
Judy	3141	1/5/2000	Black	Black	Black
	6789	3/12/2000	White	White	Black
	…				
Dredd	2718	1/20/2000	Black	White	Black
	8765	2/29/2000	Black	Black	White
	…				

Table 5: Random Assignment Simulation Results

Subset	Variable Name	IQR	Simulation Mean	Simulation St Dev	P Value	Observations
ALL	race	0.02	0.02	0.00	0.26	34298
	age	0.03	0.02	0.00	0.11	34298
	sex	0.02	0.02	0.00	0.57	34298
	violent	0.03	0.03	0.00	0.12	34298
	drugs	0.02	0.03	0.00	0.53	34298
	eft	0.02	0.02	0.00	0.53	34298
	other	0.03	0.03	0.00	0.45	34298
Violent	race	0.04	0.04	0.01	0.30	5482
	age	0.06	0.06	0.01	0.60	5482
	sex	0.04	0.03	0.01	0.09	5482
Drugs	race	0.01	0.02	0.00	0.97	13322
	age	0.05	0.04	0.01	0.15	13322
	sex	0.03	0.03	0.01	0.37	13322
EFT	race	0.07	0.05	0.01	0.04	6484
	age	0.06	0.06	0.01	0.50	6484
	sex	0.06	0.05	0.01	0.10	6484
Other	race	0.03	0.04	0.01	0.96	9010
	age	0.05	0.05	0.01	0.62	9010
	sex	0.04	0.04	0.01	0.25	9010

The IQR column reports the interquartile range of the distribution of judge fixed effects for a given variable. Simulation mean reports the mean of the interquartile range from 1000 simulations; St Dev reports the standard deviation from the simulations. The p-value indicates the percentile of the simulated data to which the empirical data corresponds. Simulations randomly choose an outcome chosen from cases initiated in the same month as the original case. Cases involve felony offenses in Cook County District Court initiated from 1995-2001 in which the defendant was African-American or White See additional explanation in the text. EFT is a case category that stands for "embezzlement, fraud, theft".

Table 6: Dispersion of Judicial Sentencing and Incarceration Rates

		jail	sentence	sentence2
25-75 Percentile	Empirical Value	0.13	148.28	257.14
	Simulation Mean	0.03	68.24	110.52
	Simulation St Dev	0.00	13.17	19.25
	P Value	<.001	<.001	<.001
10-90 Percentile	Empirical Value	0.20	251.19	527.25
	Simulation Mean	0.05	143.69	231.50
	Simulation St Dev	0.01	19.27	30.98
	P Value	<.001	<.001	<.001
5-95 Percentile	Empirical Value	0.25	390.72	684.25
	Simulation Mean	0.07	200.40	323.26
	Simulation St Dev	0.01	24.50	41.88
	P Value	<.001	<.001	<.001
Observations		34298	34298	16825

Each panel reports analogous measures of the empirical and simulated distributions of judge fixed effects for a given variable, using either IQR, 10-90 range, or 5-95 range. Empirical value reports the empirical measure. Simulation mean reports the mean of the measure from 1000 simulations; St Dev reports the standard deviation from the simulations. The p-value indicates the percentile of the simulated data to which the empirical data corresponds. Simulations randomly choose an outcome chosen from cases initiated in the same month as the original case. jail is a binary variable indicating whether the defendant was incarcerated. sentence2 is sentence length conditional on receiving a non-zero sentence. sentence and sentence2 measured in days. Cases involve felony offenses in Cook County District Court initiated from 1995-2001 in which the defendant was African-American or White. See additional explanation in the text.

Table 7: Dispersion of Racial Gap in Sentencing and Incarceration Rate

Variable Name	Empirical IQR	Simulation Mean	Simulation St Dev	P Value	Observations
jail	0.11	0.07	0.01	0.01	34298
sentence	90.50	150.35	29.17	0.98	34298
sentence2	238.36	295.21	53.51	0.85	16825

The Empirical IQR column reports the interquartile range of the distribution of the racial gap judge fixed effect for the given variable. Simulation mean reports the mean of the interquartile range from 1000 simulations; St Dev reports the standard deviation from the simulations. The p-value indicates the percentile of the simulated data to which the empirical data corresponds. Simulations randomly choose an outcome chosen from cases initiated in the same month and with the same defendant race as the original case. jail is a binary variable indicating whether the defendant was incarcerated. sentence2 is sentence length conditional on receiving a non-zero sentence. sentence and sentence2 measured in days. Cases involve felony offenses in Cook County District Court initiated from 1995-2001 in which the defendant was African-American or White. See additional explanation in the text.

Table 8: Impact of Judicial Heterogeneity in Sentencing by Race

	Change in Black-White Incarceration Rate Gap		Change in Black-White Sentencing Gap (months)	
Judge Percentile Shift	Simulation mean (sd)	Empirical	Simulation mean (sd)	Empirical
25%-75%	0.07 (0.01)	0.11	4.85 (0.94)	2.92
10%-90%	0.14 (0.02)	0.18	9.52 (1.38)	10.47

Table compares the empirical shift in the racial gap in sentencing with the counterfactual of no interjudge variation in racial gap, as produced by simulation. Second and fourth columns report empirical impact on incarceration and sentencing, respectively, of moving from the 25th (10th) percentile judge to the 75th (90th) percentile judge in the 1st (2nd) row. Analogous simulation means are reported in the first and third columns, along with the standard deviation. Cases involve felony offenses in Cook County District Court initiated from 1995-2001 in which the defendant was African-American or White. See additional explanation in the text.

Table 9: Correlation with Judge Characteristics

	Dependent Variable: Judge Fixed Effects in…					
	Sentence length	Incarceration rate	Black-White difference in sentence length		Black-White difference in incarceration rate	
Black judge? (Y=1)	45.03	-0.02	-152.69	-156.71	-0.03	-0.03
	(60.20)	(0.04)	(80.14)	(81.34)	(0.04)	(0.04)
Male judge? (Y=1)	54.02	0.03	61.14	57.6	0.02	0.02
	(56.50)	(0.03)	(74.22)	(75.28)	(0.04)	(0.04)
Older judge? (Y=1)	-11.03	-0.03	48.80	48.79	0.01	0.01
	(42.78)	(0.03)	(57.19)	(57.59)	(0.03)	(0.03)
Judge was public defender? (Y=1)	-0.56	0.02	30.77	31.39	-0.04	-0.05
	(49.19)	(0.03)	(65.04)	(65.50)	(0.03)	(0.03)
Judge F.E. in sentence length				0.07		
				(0.17)		
Judge F.E. in incarceration rate						0.3
						(0.15)
R^2	0.02	0.03	0.10	0.16	0.04	0.11
Observations:	67	67	67	67	67	67

Standard errors in parentheses. Each column correspond to a different regression. In each regression, each observation is weighted by the inverse of the square of the estimated standard error for the fixed effect used a dependent variable in that column. See text for additional detail.

Table 10: Crime Category Analysis

Variable Name	Empirical IQR	Simulation Mean	Simulation St Dev	P Value	Observations
Panel A - All cases with Crime Controls					
jail	0.090	0.069	0.012	0.046	34227
sentence	141.57	150.49	27.68	0.599	34227
sentence2	283.06	279.24	47.91	0.457	16807
Panel B - Drug Cases					
jail	0.112	0.143	0.028	0.868	13317
sentence	114.50	145.61	26.63	0.891	13317
sentence2	175.55	330.76	66.25	0.997	6588
Panel C - Non-Drug Cases					
jail	0.108	0.083	0.015	0.043	20910
sentence	175.11	192.08	36.22	0.632	20910
sentence2	350.91	352.24	71.67	0.487	10219

The Empirical IQR column reports the interquartile range of the distribution of the racial gap judge fixed effect for the given variable. Simulation mean reports the mean of the interquartile range from 1000 simulations; St Dev reports the standard deviation from the simulations. The p-value indicates the percentile of the simulated data to which the empirical data corresponds. Simulations randomly choose an outcome chosen from cases initiated in the same month and with the same defendant race as the original case. jail is a binary variable indicating whether the defendant was incarcerated. sentence2 is sentence length conditional on receiving a non-zero sentence. sentence and sentence2 measured in days. Cases involve felony offenses in Cook County District Court initiated from 1995-2001 in which the defendant was African-American or White. See additional explanation in the text.