Judicial Scarring

Karthik Srinivasan*

First Draft: January 5, 2023
Last Revision: February 7, 2023

I document that experienced decision makers can be influenced by irrelevant events in a high stakes setting, felony sentencing in Cook County. Using a stacked difference-in-differences design, I estimate that judges hand down sentences that are 13% longer after sentencing a first degree murder. The effect is twice as large for defendants who resemble the murderer along the dimensions of race and charge severity. The bias affects 6% of defendants on an ongoing basis and temporarily increases the Black sentencing penalty by 91%.

Keywords: Anchoring, Behavioral Bias, Judges, Salience, Sentencing, Sequential Decisions

JEL Codes: D91, J15, K14

Karthik Srinivasan
University of Chicago
Booth School of Business
5807 S Woodlawn Ave,
Chicago, IL 60637
ks@chicagobooth.edu

*I thank Alex Frankel, Eric Zwick, Devin Pope, Walter Zhang, Michael Galperin, Benedict Guttman-Kenney, Claire Bergey, Derek Neal, Andrew Jordan, Aurelie Ouss, Alex Imas, Marianne Bertrand and participants at the CEG American Politics Conference, the Booth Student Research in Economics Seminar, and the Booth Behavioral Economics Lab for valuable comments. I thank Era Laudermilk, Hayley Hopkins, Alexi Stocker, and Reuben Bauer for providing important legal context. Thanks to Arjun Srinivasan, Anand Srinivasan, Shanthi Srinivasan, Muthayyah Srinivasan, Gabi Hirsch, Alex Duner, Jaclyn Zhou, James Kiselik, Sam Osburn, Anna Cormack, Janani Nathan and Mochi.

Sequential judgments are a key feature of many of the most important decision making problems we face such as searching for a job, buying a home, or finding a partner. Yet despite their importance and ubiquity, sequential judgments may be subject to systematic psychological biases such as anchoring and contrast effects.

Anticipating how these biases will effect judgment is challenging because behavioral theories can make opposite predictions in the same setting. This is the case in the context I study, judicial sentencing decisions, where anchoring and contrast effects are both plausible and suggest opposite biases. Consider a judge who has recently handed down a long sentence for a particularly gruesome crime and now must sentence a typical defendant. If the long sentence serves as an anchor, the judge may adjust down insufficiently resulting in over-sentencing. Alternatively, after handling the gruesome crime, the judge may find a run-of-the-mill crime to appear mild in contrast resulting in under-sentencing. These opposite predictions are not purely theoretical: lab evidence documents anchoring and contrast effects in the context of judicial sentencing (Pepitone and DiNubile, 1976; Guthrie et al., 2000; Enough and Mussweiler, 2001). Field evidence can help adjudicate whether one bias dominates in practice.

I study felony sentencing decisions in Cook County. This is an attractive setting to test behavioral theories in the field because institutional features push against finding bias. The stakes are high: the average defendant is sentenced to over two years in prison. The decision makers are educated and experienced: judges are required to have a post-graduate degree and typically serve on the bench for multiple six-year terms. Moreover, judges have a legal obligation to hand down fair sentences.

Despite these factors, I find that Cook County judges exhibit a persistent behavioral bias. Using administrative data on the universe of Cook County felony sentencing decisions from March 2011 to December 2022, I estimate a stacked difference-in-differences design around the date of a first-degree murder sentencing. Judges hand down sentences that are 97 days longer in the 10 days after a first-degree murder sentencing, an effect which I refer to as scarring. This is a sizable effect: it represents 13% of the mean and 26% of the median sentence length. Because judges handle first-degree murders regularly, scarring distorts sentencing for 6% of defendants on an ongoing basis. The scarring effect is consistent with anchoring, but I cannot rule out some alternative behavioral explanations (see Subsection 3.4).

To get at mechanisms, I estimate triple-difference designs comparing treatment effects between defendants who are similar to and different from the murderer along the dimensions of race and charge severity. Treatment effects are 186 days longer for same-race defendants,

and 211 days longer for high class felonies. These heterogeneous treatment effects are suggestive evidence that cases that call to mind the murderer are most affected by scarring. These findings are consistent with recent theoretical models that posit that salience underlies many behavioral biases (Bordalo et al., 2022).

While a salience-driven mechanism could be race-neutral in principle, the fact that 76% of defendants convicted of first-degree murders in Cook County are Black means that scarring compounds existing inequalities in practice. Using a triple-difference design, I find that treatment effects are 131 days longer for Black defendants. This disparity is 91% of the size of the observed sentencing penalty faced by Black defendants in Cook County after conditioning on observables.

Identifying each of these effects presents a non-standard causal inference challenge because each judge can sentence multiple first-degree murders. I estimate a stacked difference-in-differences design in order to leverage these multiple treatments for identification. I construct an event-by-event dataset where each event is a first-degree murder sentencing. Treated sentencing decisions are those made by the judge who sentenced the murder in a window of time around the sentencing date. Control sentencing decisions are taken from judges in the same courthouse at the same point in calendar time who do not sentence a first-degree murder within the event time window.

This stacked difference-in-differences design requires a parallel trends assumption: I assume that residual sentence length would evolve similarly for treated and control judges in the absence of treatment. Across specifications, I find that sentence length is similar not only in trend but also in level to treated judges in the pre-period after controlling for observables which lends credibility to this design.

This paper sits at the intersection of two large interdisciplinary literatures. A literature that spans economics, political science, and empirical legal studies considers the factors that influence judicial decision making. Much of this literature documents the importance of judge characteristics, defendant characteristics, and institutional features.[1] This paper relates most closely to a newer strand of the literature which tests whether normatively irrelevant factors influence judicial decisions. Hunger (Danziger et al., 2011), sleep deprivation (Cho et al.,

[1]Harris and Sen (2019) provides a recent review emphasizing the role of a judge's characteristics such as race, gender, and political ideology. Defendant characteristics such as race, gender, and education have been shown to influence incarceration rates and sentence lengths (Mustard, 2001; Abrams et al., 2012; Rehavi and Starr, 2014). A few papers emphasize the relationship between judge and defendant characteristics (Shayo and Zussman, 2011; Depew et al., 2017). A smaller literature in law and economics highlights the role of institutional features with a particular emphasis on caseload (Engel and Weinshall, 2020; Shumway and Wilson, 2022).

2016), newspaper coverage (Lim et al., 2015; Philippe and Ouss, 2018), football results (Eren and Mocan, 2018), temperature (Heyes and Saberian, 2019), and terrorist attacks (Shayo and Zussman, 2011; Brodeur and Wright, 2019; McConnell and Rasul, 2021; Asadi, 2022) have all been linked to judicial decisions, though some of these findings are contested (Weinshall-Margel and Shapard, 2011; Spamann, 2018, 2022).

A nearly century-old literature[2] in psychology and economics theorizes and estimates biases that arise in the context of sequential decision making (Beebe-Center, 1929; Sherif et al., 1958; Tversky and Kahneman, 1974). This paper joins a recent strand documenting bias in the field. Evidence of anchoring in the field has been found in art auctions, house sales, sports betting, game shows, and procurement auctions (Beggs and Graddy, 2009; Bucchianeri and Minson, 2013; McAlvanah and Moul, 2013; Jetter and Walker, 2017; Ferraro et al., 2022). Evidence of contrast effects in the field has been found in commuting patterns, speed dating, and financial markets (Simonsohn, 2006; Bhargava and Fisman, 2014; Hartzmark and Shue, 2018).

I am most closely related to a small literature investigating the role of behavioral biases in judicial decision making in the field. Chen et al. (2016) provides evidence of the gambler's fallacy in sequential decision making for baseball umpires, loan officers, and asylum judges. Using data from the Pennsylvania courts, Leibovitch (2016) argues that judges make decisions relative to the historical severity of their caseload.

Relative to the literature, I make two contributions. I am the first to document an anchoring-like effect in judicial decision making in the field. The scarring effect I find is large and causes ongoing distortions in sentencing. This effect is surprising given previous work. I find a directionally opposite effect to the lab experiment that most closely matches my setting (Pepitone and DiNubile, 1976). My finding also contrasts with the gambler's fallacy mechanism studied in the most closely related empirical paper Chen et al. (2016). Chen et al. (2016) note that the gambler's fallacy makes identical predictions to sequential contrast effects in their binary asylum decision setting, and sequential contrast effects are directionally opposed to the scarring effect that I document.

My second contribution is to provide empirical evidence of the role that salience plays as a mediating factor in behavioral biases. While salience has been proposed as a theoretical mechanism underlying anchoring and contrast effects (Bordalo et al., 2015), my heterogeneity results provide evidence from the field showing that the size of a behavioral bias varies systematically with a measure of salience.

[2]See Furnham and Boo (2011) for a literature review.

The remainder of the paper proceeds as follows. Section 1 provides institutional details of the Cook County judicial system. Section 2 outlines the econometric framework and identification strategy of the paper. Section 3 presents the main results and discusses limitations.

1 Data and Institutional Details

1.1 Data

The data come from the Cook County State's Attorney's Office, and represent the universe of (non-juvenile) Cook County felony sentencing decisions. In March 2018, State's Attorney Kim Foxx began releasing data on the county's handling of felony cases with historical records going back to 2011.[3]

1.2 Felony Classes in Cook County

There are six felony classes in Cook County. In descending order of severity, they are Class M, X, 1, 2, 3, and 4. The nature of crimes across felony classes varies widely. Examples of Class 4 felonies include domestic battery, obstruction of justice, and theft of less than $300. Examples of Class X felonies include armed robbery, sexual assault, and home invasion.

Sentencing guidelines are pegged to felony class. For example, Class 4 felonies carry sentences between 1 and 3 years while Class 1 felonies carry sentences between 5 and 15 years. Judges have additional discretion in applying aggravating factors, such as the defendant's criminal history or motivations, which may double the maximum sentence length.

The Class M felony charge is reserved for first-degree murders which are murders that are intentional and premeditated. Class M felonies carry a sentence of 20–60 years, though defendants can face a life sentence if there are aggravating circumstances.

1.3 Defining Prison Sentence Length

My primary outcome variable is prison sentence length. I standardize the length of all prison sentences to days. Out of the set of defendants initially charged with a non-M felony who are ultimately sentenced, 44% of sentences are not prison sentences (nearly all non-prison sentences are probation sentences).[4] I assign a length of 0 to non-prison sentences.

[3]To download the data, visit https://datacatalog.cookcountyil.gov/Courts/Sentencing/tg8v-tm6u.

[4]I define sentences at the Illinois Department of Corrections, Cook County Department of Corrections, Cook County Boot Camp, Cook County Impact Incarceration Program, and Juvenile IDOC as prison sen-

I cap sentence length at 75 years and assign life sentences this length. This cap affects a small number of observations (20 sentences) and is intended to reflect the fact that years beyond this point are very unlikely to be served.

1.4 Random Assignment

While not necessary for identification, random assignment of cases to judges (conditional on court and calendar time) improves identification because it ensures that control judges are presiding over cases that are balanced on unobservables in expectation. Random assignment of cases to judges is common in Cook County. According to Loeffler (2013), case assignment is handled by the presiding justice who uses a program called "the randomizer" to assign new cases to available judges in view of the Cook County State's Attorneys Office and Cook County Public Defender's Office.

However, no law in Cook County requires random assignment of cases to judges, and in practice random assignment may be circumvented due to a variety of exceptions. Based on a conversation I had with the authors of Jordan et al. (2021), there are several exceptions to random assignment. First, some cases are siphoned off to "problem-solving courts" which are designed to handle sentencing related to mental health, drug addiction, and veteran's issues. Second, if a defendant violates their probation, they return to the judge who originally sentenced them. Third, some state benefits fraud cases are not randomly assigned due to involvement by the governor's office. Fourth, some high profile cases, colloquially referred to as "heaters," are assigned to judges who are comfortable with increased attention and scrutiny. Fifth, some judges are floaters who do not have a permanent courtroom but instead pick up slack. For cases that avoid these exceptions, randomization occurs at the "call" level, which essentially represents a fixed courtroom. I am able to exclude cases that are associated with the first three issues. I do not observe if a case is a "heater," so I cannot remove these cases directly. Additionally, I do not observe call numbers directly so I cannot guarantee random assignment.

In an attempt to exclude judges who may violate random assignment, I place three restrictions on the inclusion of judges into the sample.[5] I require that the each judge has

tences. The vast majority of prison sentences are served at the Illinois Department of Corrections and the Cook County Department of Corrections.

[5]There are three reasons why a judge could not be randomly assigned cases as far as I know. First, they could be a floater who picks up slack. Second, they could be an associate judge; associate judge's need permission from the head judge to preside over felonies. In practice, many associate judges do preside over felonies and are given calls, but not all. Third, they might be assigned to both a regular call as well as a problem solving court call which could result in the regular call being non-random due to the partial nature of their availability.

presided over at least one Class M felony. Following Jordan et al. (2021), I require that each judge has handed down at least 500 felony sentences. Finally, I exclude judges who appear to violate random assignment directly. I predict the average expected sentence length of cases seen by a judge on the basis of observables and exclude judges whose predicted average sentence length would occur with a less than 1% chance if they were being assigned cases randomly.[6]

1.5 Sample Restrictions

Beyond the restrictions placed on the set of judges included in my sample, I also place some restrictions on sentences directly. After imposing all restrictions, I am left with a sample of 117,915 sentencing decisions made by 56 judges from March 2011 to December 2022.

To exclude cases that are not randomly assigned to judges, I exclude sentences that are not an original sentence, sentences that were diverted to problem solving courts, and sentences for state benefits fraud charges.

Since treatment is defined using Class M felony sentences, I restrict attention to defendants initially charged with non-M felonies. I also exclude any sentences associated with either the charge ID or the defendant ID of a Class M felony sentence to prevent effects from being driven by longer sentences for co-conspirators or longer sentences among a Class M felon's other charges.

Following Abrams et al. (2012), I define an observation as a judge-date-defendant tuple, keeping the longest sentence in cases where there are two or more sentences handed down by the same judge to the same defendant on a given day. The reason to do this is that such sentences are unlikely to be independent. I do not observe whether multiple sentences are concurrent or consecutive, so I cannot combine sentences. This restriction results in dropping about 10% of sentences.

Finally, I exclude a small number of sentences where judge information is missing or date information is mis-entered in an ambiguous way.

[6]The estimating equation for this restriction is a poisson regression of sentence length on race, gender, five year age bins, and class. I compare the predicted sentence length of cases seen by a judge to the predicted sentence length of cases seen by other judges at the same courtroom during the same stretch of calendar time.

2 Econometric Framework

2.1 Identification Challenges

I want to estimate how many additional days the typical defendant faces in prison if they are sentenced by a judge who has recently sentenced a first-degree murder. Estimating this effect presents a causal inference challenge for three reasons. First, each treatment unit (a judge) can be treated multiple times (sentence multiple first-degree murders), so treatment status is not an absorbing state. Second, time between treatments varies and the duration of the treatment effect is unknown ex ante. Third, only a small number of units (typically one) are treated at any point in calendar time.

Observing multiple treatments is potentially helpful because it increases the sample of events, but it simultaneously complicates inference because treatment from an earlier event could pollute the control period of a later event, biasing the estimated coefficient downwards. This bias is particularly concerning if the murder sentencing hearings happen in quick succession: for example, if a judge sentences two murders in a month, we might be worried that the sentences handed down between the murders are not a good control for the sentences handed down after the second murder. This problem is further complicated by the fact that the duration of the treatment effect is not known ex ante. In the extreme case, if the treatment effect were to last forever (i.e. sentencing a murder changes a judge's judicial philosophy permanently), an event study estimate of the treatment effect for second and subsequent murders would be zero.

Additionally, estimating treatment effects in the context of multiple individual-level treatments precludes the use of recently proposed estimators from the rapidly growing literature exploring biases in two-way fixed effects estimation (e.g., Callaway and Sant'Anna (2021); Goodman-Bacon (2021); Sun and Abraham (2021)) as these estimators rely on large cohorts being treated at the same time and assume that treatment is an absorbing state.

2.2 Overview of Identification Strategy

My identification strategy proceeds in three steps.

First, I construct a stacked difference-in-differences dataset by iterating through a set of events (first-degree murder sentences) and keeping a set of sentences issued by treated and control judges in a window around the date of each event.

Second, I compute residual sentence length by estimating a poisson regression of sentence

length on controls for defendant gender, race, and age as well as fixed effects for felony class, sentencing judge, and event.

Third, I estimate results by taking simple differences in mean residual sentence length.

This identification strategy requires a parallel trends assumption. Specifically, I assume that, in the absence of treatment, the mean residual sentence length for treated judges would have followed the same path as mean residual sentence length for control judges.

In the next three subsections, I'll discuss each of the three steps in detail. In Subsection 3.2 I discuss potential violations of the identifying assumption.

2.3 Constructing a Stacked Difference-in-Differences Dataset

I construct the stacked difference-in-differences dataset using the following procedure. First, I define an event as a Class M felony sentence with a clean pre-period. I say that a Class M felony has a clean pre-period if the sentencing judge has not sentenced another Class M felony in the prior 100 days. For each event, a treated judge spell is made up of sentences issued by the sentencing judge in a 100 day window around the event. For each event, I identify a set of control judges who issue sentences in the same courthouse as the treated judge and who do not sentence a Class M felony during the 100 day window around the event. Because this dataset is constructed event-by-event, decisions from the same judges appear in both treatment spells and control spells.

2.4 Estimating Residual Sentence Length

Residual sentence length is the difference between observed sentence length and predicted sentence length. I compute predicted sentence length by estimating Equation 1 on the stacked difference-in-differences dataset. I estimate Equation 1 using a poisson regression because sentence length is strictly positive and has a right skewed distribution:

$$y_{dje} = \alpha + \beta X_d + FE_j + FE_e + \epsilon. \tag{1}$$

In this equation, d indexes the defendant, j indexes the judge, and e indexes the event; Y_{dje} is sentence length measured in days; and X_d is a set of defendant-level characteristics. The characteristics are race (Black, Hispanic, White, other), gender (male, female), defendant age, and initial felony charge class. Defendant age is included as a set of five year age bin fixed effects to allow for nonlinearities in how age affects sentencing. A small number of

observations are missing age; these observations are assigned a missing age fixed effect. I choose to control for initial felony charge class because previous work argues that final charge class is an endogenous outcome of the interaction between the judge and the prosecutor and so could be affected by treatment. Further, FE_j is a fixed effect for the sentencing judge j and FE_e is a fixed effect for event e. Since each event identifies a treatment judge spell and a set of control judge spells where all judges come from the same courtroom, FE_e can be thought of as a courtroom–calendar time fixed effect.

2.5 Estimating Treatment Effects

All treatment effects I report are computed by taking simple differences in mean residual sentence length. Difference-in-differences figures report the difference in mean residual sentence length between treatment and control judges for each 10 day event time bin. Triple difference figures report the difference between differenced residual sentence lengths across two groups within each 10 day event time bin. Figures plot the data in 10 day event time bins in order to visualize how treatment effects evolve dynamically. In order to provide a single top line statistic summarizing each treatment effect, I compare differenced estimates in a treatment period (defined as event time between 0 and 10) to differenced estimates in the pre-period (defined as event time < 0).

I compute standard errors assuming that sentence length is independent and identically distributed after conditioning on observables and treatment status. This assumption is justified by the plausibly random assignment of cases to judges. Abadie et al. (2017) argue that when treatment status is randomly assigned, clustering standard errors is excessively conservative.[7]

The choice to estimate effects by residualizing sentence length and then differencing means has a few advantages. First, coefficients derived from differences in residual sentence length have a straightforward interpretation: effects are measured in terms of a number of additional days in prison. Second, by computing each mean separately, I can report estimates of the standard error for every event time period (a single-step procedure would required dropping one time period fixed effect to avoid perfect collinearity). Third, the fact

[7]Quoting Abadie et al. (2017), "If one has a random sample of units from a large population with randomized treatment assignment at the unit level, there is no reason to cluster the standard errors of the least squares estimator. ... Similarly, in a judge-leniency design—where defendants are randomly assigned to judges—standard errors should not be clustered at the level of the judge." Note that it is not a problem if some murders are "heaters" which are not randomly assigned as long as the other cases that a judge sentences after the first-degree murder were randomly assigned, as this would ensure that treatment status is randomly assigned.

that pre-period 95% confidence intervals typically include zero in figures is not a mechanical consequence of normalization: rather, the interpretation of these graphs is that control judges offer sentences that are similar in level after controlling for observables which, though it is not a key identifying assumption, is still reassuring.

The two-step procedure of residualizing and then differencing means is conceptually similar to estimating a single poisson regression and backing out coefficients in terms of days or estimating a poisson regression to residualize sentence length followed by an OLS regression with event time $\times$ treated dummies.

3 Results

3.1 Discussion of Treatment Effect Estimates

Figure 1 visualizes the primary difference-in-differences design. After controlling for observables, mean residual sentence length for treatment and control judge spells is less than 5 days apart in the pre-period (defined as event time < 0). In the treatment period (defined as event time $\in [0, 10]$), residual sentence length spikes for treated judges: the difference-in-differences point estimate of the treatment effect is 97 days ($p < 0.001$). The effect is large relative to the length of typical sentences: 97 days is 26% of the median (365 days) and 13% of the mean (756 days) length of non-M felony sentences in Cook County.

To test whether this effect persists over a longer time horizon, I compute a difference-in-differences coefficient comparing sentencing in the post-treatment period (defined as event time $\in (10, 100]$) to sentencing in the pre-period: residual sentence length remains elevated with the average sentence being 37 days longer ($p < 0.001$) in the 90 day post-treatment period. Note that this estimate requires a stronger identifying assumption: control judge sentencing must capture how treatment judge sentencing would have evolved over the 90 days following the treatment period, so I cannot appeal to the sharp timing of treatment.

In order to get at mechanisms, I compute heterogeneous treatment effects by estimating triple-difference designs comparing defendants who are similar to and different from the murderer along the dimensions of race and charge severity. Panel A of Figure 2 compares same-race defendants to different-race defendants. The triple-difference point estimate is that the scarring effect is 186 days longer for same-race defendants than different-race defendants ($p < 0.001$). This result is not a mechanical consequence of differential sentencing by race as residual sentence length includes controls for defendant race.

Panel B of Figure 2 splits defendants into high- and low-class felonies. I define high-

class felonies as Class X and 1 felonies and low-class felonies as Class 2, 3 and 4 felonies. The scarring effect is 211 days longer for high-class felonies compared to low-class felonies ($p = 0.053$). Note that high-class felonies differ from low-class felonies in two ways: high-class felonies are more similar to the murder event because they are more severe, but they also offer judges more discretion as the sentencing guidelines for these classes are looser. Nevertheless, the fact that I cannot reject zero difference in residual sentencing between high and low-class felonies for the vast majority of the pre-period shows that this result is not solely due to additional discretion. Again, this result is not a mechanical consequence of the fact that high-class felonies receive longer sentences on average because residual sentence length controls for felony class.

These heterogeneous treatment effects are consistent with a model of salience-driven behavioral bias as described in Bordalo et al. (2015). Both panels of Figure 2 document ways in which defendants who are similar to the murder face a larger scarring effect. These results suggest that it is the cases which call to mind the murderer that are penalized. The fact that the scarring effect fades out over time bolsters this interpretation because fading out points towards the importance of memory. In this interpretation, as the first-degree murder fades further into the past, it exerts less of an effect because it is no longer salient. In this sense, time can be construed as a third dimension of similarity that mediates the scarring effect.

Next, I show that scarring exacerbates existing inequalities because it differentially affects Black defendants. Figure 3 plots a triple-difference between Black and non-Black defendants. Non-Black pools White, Hispanic, and Asian defendants, who together make up 29% of defendants. The triple-difference point estimate is 131 days ($p = 0.014$). While the scarring effect is short-lived, it represents a large increase in inequality as 131 days is 91% of the 142 day average sentencing penalty that Black defendants face conditional on observables.[8]

One reason these results are concerning is that judges sentence first-degree murders regularly. As a back-of-the-envelope calculation, I find that 6% of defendants in my sample are sentenced by a judge who sentenced a first-degree murder in the 10 days prior.

[8]My estimate of a 142 day sentencing penalty between Black and non-Black defendants in Cook County comes from regressing sentence length on a dummy for Black as well as fixed effects for gender, age (5 year bins), initial charge class and sentencing judge. This specification does not distinguish between two potential explanations for a sentencing gap. The average Black defendant within each class may have committed an unobservably more severe crime resulting in a longer sentence. Alternatively, judges may be prejudiced against Black defendants, handing down longer sentences for crimes of the same severity. Decomposing these two explanations is beyond the scope of this paper.

3.2 Threats to Identification

The identifying assumption of the stacked difference-in-differences design is that, in the absence of treatment, residual sentence length for treated judges would have evolved in the same way that residual sentence length evolves for control judges. Because the treatment period is defined from event time 0 to event time 10 for the top line estimates, this assumption only needs to hold for this 10 day period. The primary evidence for this assumption is that, across all figures, pre-period 95% confidence intervals almost always contain zero and appear consistent with a flat trend. In the context of Figure 1, this means that sentence length is similar in both level and trend for treated and control judge spells after controlling for observables.

One intuitive concern is that the timing of a first-degree murder sentencing may be correlated with a broader rise in the severity of crime biasing my estimates upwards. Changes in the severity of the felonies that treated judges sentence alone are not an issue for my identification strategy because I control for felony class. Since felony class determines the sentencing guidelines in Cook County, results are driven by increases in the length of sentences for cases that were determined to be deserving of similar sentences ex ante.

A related concern is that felonies could be becoming unobservably more severe within felony class around the time of the first-degree murder sentencing. To the extent that I succeed in restricting attention to a sample where cases are randomly assigned to judges, this ensures that cases are balanced along unobservables between treatment and control judge spells. However, even if random assignment fails, the fact that pre-period trends are parallel suggests that control judge spells do a good job predicting treated judge sentence length after controlling for observables.

The most serious identification challenge is the potential for sorting on unobservables around the time of the event. For example, if a judge anticipates that sentencing a first-degree murder will involve a large workload, they could try to push off unobservably more severe cases until after the sentencing which would bias my estimates upwards. This kind of sorting would not be picked up by control judge spells since control judges do not have a first-degree murder to sentence by construction and so do not have an incentive to sort.

Institutional features push against this concern. According to a conversation I had with an attorney from the Cook County Public Defender's Office, felony sentencing hearings are typically scheduled weeks or months in advance by mutual agreement of the prosecutor, the defense, and the judge. Mutual scheduling constraints and advanced scheduling make precise sorting around the date of the first-degree murder sentencing less likely.

Figure 4 directly tests sorting on observables. Each panel in Figure 4 is estimated analogously to the primary difference-in-differences specification replacing residual sentence length for a defendant characteristic. Since Black, Female, and High-class are all binary variables, I compute residuals by estimating logit models rather than a poisson model. Each coefficient gives the mean difference in the predicted probability of a defendant characteristic between treated and control judges. If the 95% confidence interval includes 0, then I cannot reject that the samples are balanced along that demographic characteristic after controlling for observables.

I do not find evidence of sorting on demographic characteristics. I cannot reject 0 difference in the predicted % of defendants who are Black, female, or older around event time after controlling for observables. I do find some evidence of sorting on charge severity: high-class cases make up 2% more of the cases judge by treated relative to control judges in the 10 days after a first-degree murder sentencing. Because I directly control for class, this sorting does not drive my results. However, the motivation for a test of sorting on observables is that documented sorting on observables could be indicative of sorting on unobservables which cannot be controlled for and would bias estimates. I also observe differences in the share of high-class felonies of a similar magnitude during event time $[-80, -70]$ and event time $[-40, -30]$, but do not observe elevated sentencing during those time periods in Figure 1. This is reassuring as it suggests that unobservable sorting that is correlated with sorting on class is unlikely to drive my results, because if sorting on class was indicative of unobservable sorting, then we would expect that this effect should also influence these earlier time periods. Still, the evidence of sorting on class is concerning, and I cannot rule out the possibility that my estimates may be biased due to unobservable sorting around event time.

3.3 Interpretation of Treatment Variation and Limitations

A limitation of the Cook County data is that I do not observe detailed information on the timing of events within each court case. This complicates the interpretation of treatment because the process of judging a court case consists of a series of tasks (e.g., hearing arguments, viewing evidence, holding a trial, sentencing), which could each be influential. My treatment variation, the timing of the sentencing hearing, is a noisy measurement for when many tasks related to a defendant will occur in a short span of time, and so I cannot attribute the treatment effect to sentencing alone.

The murder cases I study often sit on a judge's docket for a year or more during which time a judge holds occasional hearings and both sides gather evidence. I observe the time when each Class M cases enters the prosecutor's office, but the extended lag between intake

and sentencing makes producing credible estimates using this variation difficult because a judge may not meaningfully interact with a case on their docket while some legal processes (e.g., discovery) occur. Given how long cases stay on the docket, the treatment effect should be thought of as capturing something over and above any effects of having a Class M felony on the docket, which is likely to be the case for all of the pre-treatment period.

A related limitation of my strategy is that it can only be used to identify the effects of crimes that occur rarely. Lower class felonies all occur frequently so I cannot construct clean pre-periods. However, the question of whether or not non-M felony sentences have scarring effects on nearby sentences is obviously of interest to the broader question of behavioral biases in judicial sentencing.

A natural specification would have been to estimate effects for a judge's first murder case in order to avoid the issue of defining an arbitrary clean pre-period. This is not possible in my setting due to a data limitation: most judges are already judging before the start of my sample. This is because it is typical for judges to serve multiple six-year terms, and I only observe eleven years of data. Because of this limitation, I cannot be sure whether the first murder I observe is the first murder that a judge presides over for the majority of judges.

3.4 Anchoring and Other Potential Mechanisms

I refer to scarring as an anchoring-like effect because a severe case with a long sentence results in over-sentencing. However, sentencing a murder case is a bundled treatment that involves a variety of information, so it is likely that the mechanism behind the scarring effect encompasses more than just the large numeric value of the sentence.

One alternative mechanism is that a murder case might temporarily influence beliefs. For example, sentencing a case might change a judge's perceptions regarding the prevalence of violent crime or the benefits of taking potential offenders off the streets.

Similarly, an intuitive alternative mechanism that I cannot rule out is an emotionally driven story. Judging a violent murder could leave a judge angry or scared resulting in over-sentencing. While this is plausible, it is worth noting that felony sentencing judges regularly sentence other violent crimes and the average judge is experienced, so these emotional effects would have to exist over and above a baseline emotional state. Additionally, the main treatment effect appears to fade out over 40 days. This effect length is inconsistent with typical descriptions of emotional states, which are usually thought to last for shorter periods of time.

One mechanism that I can rule out is learning or changes to preferences because the scarring effects I find appear to fade out. This means the effect cannot be driven by updates to priors unless these updates also fade which is not typically what is meant by learning.

4 Conclusion

I document a sizable, ongoing behavioral bias in felony sentencing in Cook County. Identifying and correcting biases in this setting is important given the stakes: unjustly harsh sentences rob defendants of freedom, and equal justice under the law is a foundational legal principle.

Judges sentence defendants to an average of 97 additional days in prison after a first-degree murder sentencing. Since judges repeatedly sentence first-degree murders, scarring affects 6% of defendants on an ongoing basis. Effects are about twice as large for defendants who are the same race as the murderer and for defendants who face a high class felony charge. Coupled with the spike-and-fade pattern of the main effect, these results suggest that defendants who call to mind a recent murderer face harsher sentences. These differential treatment effects exacerbate existing racial disparities, temporarily increasing the Black sentencing penalty by 91%.

As a piece of law and economics, these results are important because they shed light on a novel and sizable violation of the equal justice principle: defendants are sentenced differently due to the luck of timing. From a behavioral perspective, this is the first paper to demonstrate an anchoring-like effect in the judicial context in the field.

The scarring effect emphasizes how psychological biases interact with the institutional design of the court system to produce injustice. This is important because the way that sentencing decisions are scheduled is under the purview of judges and policymakers. Institutional reforms could plausibly mitigate scarring: it would be feasible to institute tighter sentencing guidelines, to have judges specialize in sentencing high class felonies, or to mandate cooling off periods after a judge handles a particularly challenging case, though a cost-benefit analysis of such policies is beyond the scope of this paper.

References

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey Wooldridge. When should you adjust standard errors for clustering? Technical report, National Bureau of Economic Research, 2017.

David S Abrams, Marianne Bertrand, and Sendhil Mullainathan. Do judges vary in their treatment of race? *The Journal of Legal Studies*, 41(2):347–383, 2012.

Masoud Asadi. Essays on intersection of economics and judicial decision-making. *Working Paper*, 2022.

John G Beebe-Center. The law of affective equilibrium. *The American Journal of Psychology*, pages 54–69, 1929.

Alan Beggs and Kathryn Graddy. Anchoring effects: Evidence from art auctions. *American Economic Review*, 99(3):1027–39, 2009.

Saurabh Bhargava and Ray Fisman. Contrast effects in sequential decisions: Evidence from speed dating. *Review of Economics and Statistics*, 96(3):444–457, 2014.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Salience theory of judicial decisions. *The Journal of Legal Studies*, 44(S1):S7–S33, 2015.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Salience. *Annual Review of Economics*, 14:521–544, 2022.

Abel Brodeur and Taylor Wright. Terrorism, immigration and asylum approval. *Journal of Economic Behavior & Organization*, 168:119–131, 2019.

Grace W Bucchianeri and Julia A Minson. A homeowner's dilemma: Anchoring in residential real estate transactions. *Journal of Economic Behavior & Organization*, 89:76–92, 2013.

Brantly Callaway and Pedro HC Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230, 2021.

Daniel L Chen, Tobias J Moskowitz, and Kelly Shue. Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *The Quarterly Journal of Economics*, 131(3):1181–1242, 2016.

Kyoungmin Cho, Christopher M Barnes, and Cristiano L Guanara. Sleepy punishers are harsh punishers: Daylight saving time and legal sentences. *Psychological science*, 2016.

Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892, 2011.

Briggs Depew, Ozkan Eren, and Naci Mocan. Judges, juveniles, and in-group bias. *The Journal of Law and Economics*, 60(2):209–239, 2017.

Christoph Engel and Keren Weinshall. Manna from heaven for judges: Judges' reaction to a quasi-random reduction in caseload. *Journal of Empirical Legal Studies*, 17(4):722–751, 2020.

Birte Enough and Thomas Mussweiler. Sentencing under uncertainty: Anchoring effects in the courtroom 1. *Journal of applied social psychology*, 31(7):1535–1551, 2001.

Ozkan Eren and Naci Mocan. Emotional judges and unlucky juveniles. *American Economic Journal: Applied Economics*, 10(3):171–205, 2018.

Paul J Ferraro, Kent D Messer, Pallavi Shukla, and Collin Weigel. Behavioral biases among producers: Experimental evidence of anchoring in procurement auctions. *The Review of Economics and Statistics*, pages 1–40, 2022.

Adrian Furnham and Hua Chu Boo. A literature review of the anchoring effect. *The journal of socio-economics*, 40(1):35–42, 2011.

Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277, 2021.

Chris Guthrie, Jeffrey J Rachlinski, and Andrew J Wistrich. Inside the judicial mind. *Cornell L. Rev.*, 86:777, 2000.

Allison P Harris and Maya Sen. Bias and judging. *Annual Review of Political Science*, 22: 241–259, 2019.

Samuel M Hartzmark and Kelly Shue. A tough act to follow: Contrast effects in financial markets. *The Journal of Finance*, 73(4):1567–1613, 2018.

Anthony Heyes and Soodeh Saberian. Temperature and decisions: evidence from 207,000 court cases. *American Economic Journal: Applied Economics*, 11(2):238–65, 2019.

Michael Jetter and Jay K Walker. Anchoring in financial decision-making: Evidence from jeopardy! *Journal of Economic Behavior & Organization*, 141:164–176, 2017.

Andrew Jordan, Ezra Karger, and Derek A Neal. Heterogeneous impacts of sentencing decisions. 2021.

Adi Leibovitch. Relative judgments. *The Journal of Legal Studies*, 45(2):281–330, 2016.

Claire SH Lim, James M Snyder Jr, and David Strömberg. The judge, the politician, and the press: newspaper coverage and criminal sentencing across electoral systems. *American Economic Journal: Applied Economics*, 7(4):103–35, 2015.

Charles E Loeffler. Does imprisonment alter the life course? evidence on crime and employment from a natural experiment. *Criminology*, 51(1):137–166, 2013.

Patrick McAlvanah and Charles C Moul. The house doesn't always win: Evidence of anchoring among australian bookies. *Journal of Economic Behavior & Organization*, 90:87–99, 2013.

Brendon McConnell and Imran Rasul. Contagious animosity in the field: Evidence from the federal criminal justice system. *Journal of Labor Economics*, 39(3):739–785, 2021.

David B Mustard. Racial, ethnic, and gender disparities in sentencing: Evidence from the us federal courts. *The Journal of Law and Economics*, 44(1):285–314, 2001.

Albert Pepitone and Mark DiNubile. Contrast effects in judgments of crime severity and the punishment of criminal violators. *Journal of Personality and Social Psychology*, 33(4): 448, 1976.

Arnaud Philippe and Aurélie Ouss. "no hatred or malice, fear or affection": Media and sentencing. *Journal of Political Economy*, 126(5):2134–2178, 2018.

M Marit Rehavi and Sonja B Starr. Racial disparity in federal criminal sentences. *Journal of Political Economy*, 122(6):1320–1354, 2014.

Moses Shayo and Asaf Zussman. Judicial ingroup bias in the shadow of terrorism. *The Quarterly journal of economics*, 126(3):1447–1484, 2011.

Muzafer Sherif, Daniel Taub, and Carl I Hovland. Assimilation and contrast effects of anchoring stimuli on judgments. *Journal of experimental psychology*, 55(2):150, 1958.

Clayson Shumway and Riley Wilson. Workplace disruptions, judge caseloads, and judge decisions: Evidence from ssa judicial corps retirements. *Journal of Public Economics*, 205: 104573, 2022.

Uri Simonsohn. New yorkers commute more everywhere: contrast effects in the field. *Review of Economics and Statistics*, 88(1):1–9, 2006.

Holger Spamann. Are sleepy punishers really harsh punishers? comment on cho, barnes, and guanara (2017). *Psychological science*, 29(6):1006–1009, 2018.

Holger Spamann. Comment on" temperature and decisions: Evidence from 207,000 court cases". *American Economic Journal: Applied Economics*, 14(4):519–28, 2022.

Liyang Sun and Sarah Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199, 2021.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185 (4157):1124–1131, 1974.

Keren Weinshall-Margel and John Shapard. Overlooked factors in the analysis of parole decisions. *Proceedings of the National Academy of Sciences*, 108(42):E833–E833, 2011.

5 Figures

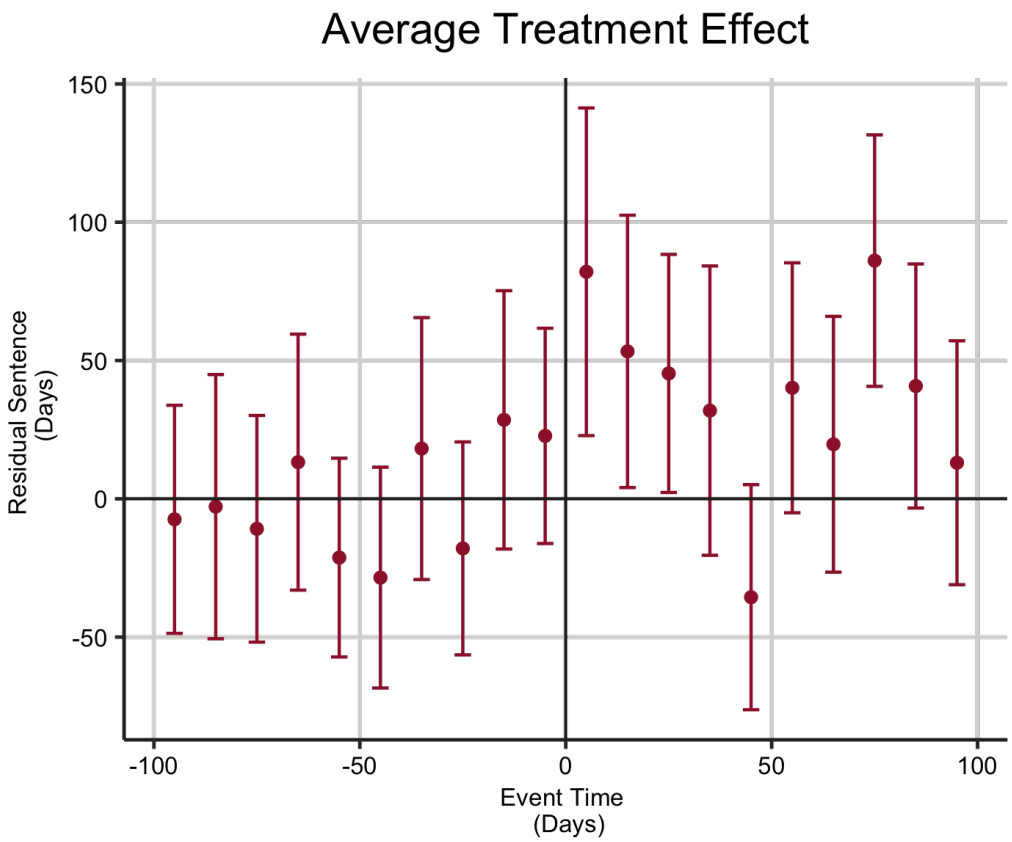

Figure 1: This figure visualizes a difference-in-differences design showing that mean prison sentence length spikes in the 10 days following a first degree murder sentencing. I define an event as the sentencing date of a first degree murder by a judge who has not sentenced a first degree murder in the prior 100 days. A treated judge spell is the set of sentences in a 100 day window around the event handed down by the judge who sentenced the first degree murder. A control judge spell is the set of sentences in a 100 day window around the event handed down by a judge in the same courtroom who did not sentence a first degree murder in the pre or post-period. Each point represents the difference in mean residual sentence length between treated and control judge spells in a 10 day event time bin, and bars represent 95% confidence intervals. Residual sentence length is the difference between actual sentence length and predicted sentence length after controlling for race, gender, age, felony class, judge and event fixed effects (see Equation 1).

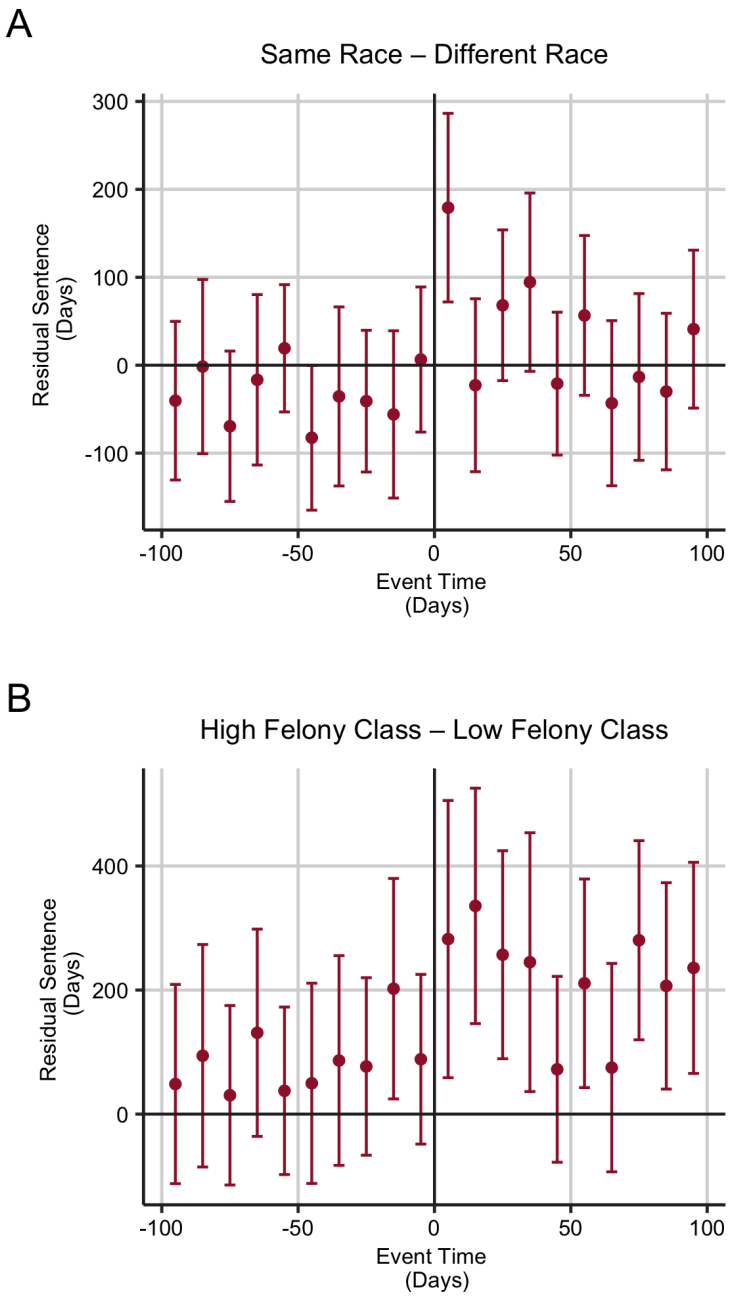

Figure 2: This figure visualizes a triple-difference design comparing treatment effects by similarity between the defendant and the murderer along the dimensions of race and felony class. The three differences are: (1) sentences that are handed down (before/after) a first degree murder sentence to (2) defendants who are (similar to/different from) the murderer by (3) judges who (did/did not) hand down the first degree murder sentence. Panel A differences treatment effects for defendants who are the same race as the murderer from treatment effects for defendants who are a different race than the murderer. Panel B differences treatment effects for defendants who face a high class felony charge (Class X or 1) from treatment effects for defendants who face a low class felony charge (Class 2, 3, or 4). Residual sentence length is the difference between actual sentence length and predicted sentence length after controlling for race, gender, age, felony class, judge and event fixed effects (see Equation 1). Each point represents 10 days of event time, and bars represent 95% confidence intervals.

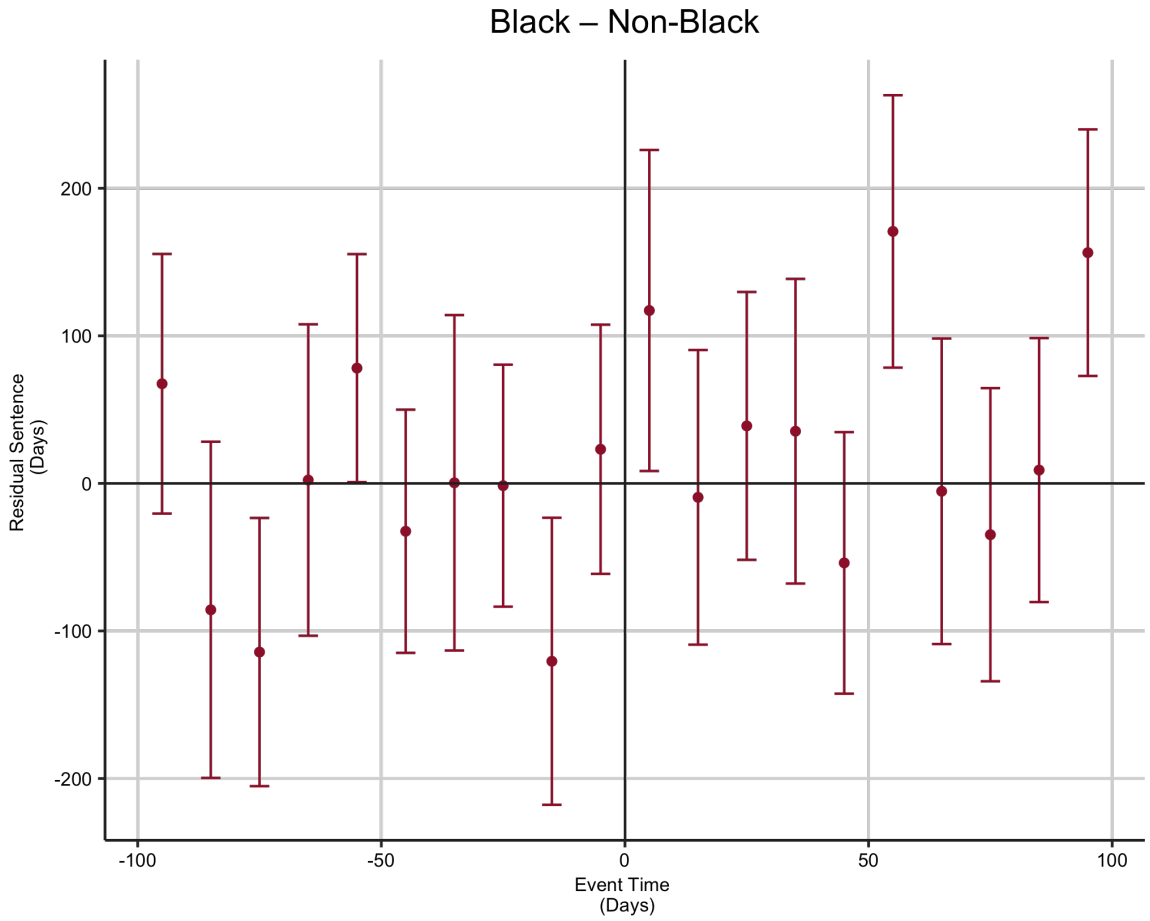

Figure 3: This figure visualizes a triple-difference design comparing treatment effects between Black and non-Black defendants. The three differences are: (1) sentences that are handed down (before/after) a first degree murder sentence to (2) defendants who are (Black/non-Black) by (3) judges who (did/did not) hand down the first degree murder sentence. Panel A visualizes treatment effects for each of the difference-in-difference designs: it plots the mean difference in residual sentence length between treated and control judge spells for Black defendants and non-Black defendants. Panel B plots the difference between these treatment effects. Residual sentence length is the difference between actual sentence length and predicted sentence length after controlling for race, gender, age, felony class, judge and event fixed effects (see Equation 1). Each point represents 10 days in event time, and bars represent 95% confidence intervals.

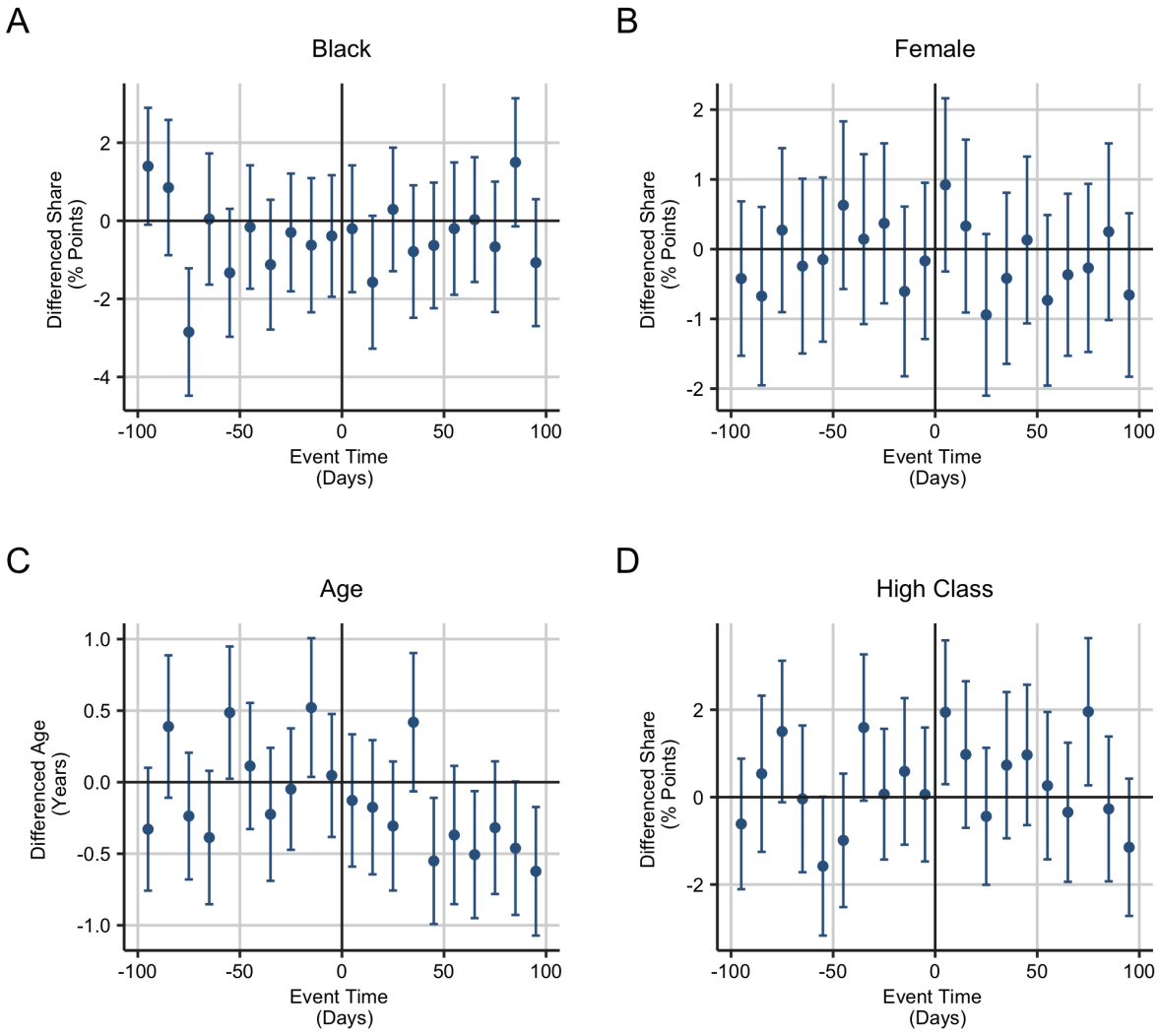

Figure 4: This figure plots tests for violations of random assignment of cases across event time. Each panel visualizes a difference-in-differences design equivalent to Figure 1 replacing the primary outcome variable (residual sentence length) for another defendant characteristic. Residual defendant characteristics are computed by estimating Equation 1 replacing sentence length for the characteristic as the left hand side variable, and taking the difference between actual and predicted defendant characteristic. I estimate a logit regression rather than a poisson regression when the characteristic is binary. Each point represents 10 days in event time, and bars represent 95% confidence intervals.