ESSAYS ON THE JUDICIARY

by

MAXWELL ROBERT MINDOCK

A DISSERTATION

Presented to the Department of Economics
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2020

DISSERTATION APPROVAL PAGE

Student: Maxwell Robert Mindock

Title: Essays on the Judiciary.

This thesis has been accepted and approved in partial fulfillment of the requirements of the Doctor of Philosophy degree in the Department of Economics by:

Glen Waddell	Chair
Michael Kuhn	Member
Ed Rubin	Member
Kristen Bell	Institutional Representative

and

Kate Mondloch	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2020

DISSERTATION ABSTRACT

Maxwell Robert Mindock

Doctor of Philosophy

Department of Economics

June 2020

Title: Essays on the Judiciary.

The complexities of the Judicial system provides unique research opportunities to model and learn about human behavior. Whether it be the opinions of United States Supreme Court Justices or the length of time a defendant is sentenced to incarceration, judicial outcomes are of extreme importance to all involved. In this thesis, I study vote determination on the Supreme Court, finding evidence of systematic variation in vote dependencies that align with Justice partisanship, sentencing cohort effects within criminal sentencing, finding evidence judges do not sentence defendants independently of other defendants, and multiplicity effects within criminal sentencing, finding evidence judges do not sentence offenses independently among defendants with multiple offenses and that the black-white racial gap in sentencing is larger than previously thought. This dissertation includes previously unpublished co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Maxwell Robert Mindock

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, Oregon

Grinnell College, Grinnell, Iowa

DEGREES AWARDED:

Doctor of Philosophy, Economics, 2020, University of Oregon

Master of Science, Economics, 2016, University of Oregon

Bachelor of Arts, 2015, Economics and Political Science, Grinnell College

AREAS OF SPECIAL INTEREST:

Applied Microeconomics

PROFESSIONAL EXPERIENCE:

Student Trainee, U.S. Census Bureau, Washington D.C. 2018

Instructor, Department of Economics, University of Oregon, Eugene, 2016-2020

GRANTS, AWARDS, AND HONORS:

Graduate Teaching Fellow, Department of Economics, University of Oregon, 2015-2020

Ph.D. Research Paper Award, Department of Economics, University of Oregon, 2018

ACKNOWLEDGMENTS

I thank my family and committee for continued guidance and support.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

CHAPTER I

VOTE INFLUENCE IN GROUP DECISION-MAKING: PARTY IMBALANCE AND INDIVIDUAL IDEOLOGY ON THE SUPREME COURT

The first chapter of this dissertation, titled, "Vote Influence in Group Decision-Making: Party Imbalance and Individual Ideology on the Supreme Court," and the third chapter, titled, "Multiple Offenses, Concurrent Sentencing, and Racial Gaps in Sentencing Outcomes," have a co-author, Glen Waddell. I have fully participated in every aspect of the process—initial research design, data acquisition, empirical methods, and both the writing and continued revision of research output.

1.1 Introduction

Collaborative decision-making is typically a process of aggregating conflicted positions into a single position. Within many organizations, however, while there are individual ideological positions at play, there are also implicit associations, or potential vote-sharing relationships that may change the nature of deliberation. Corporate Boards, for example, often have both inside directors (e.g., large stakeholders, the Chief Executive Officer, other executives of the organization) and outside directors, with no direct connections to the organization but with experience that may represent associated interests, who can bring balance to the interests of insiders as they are unlikely to tolerate "insider dealing." Academic units can even experience conflict when making hiring decisions, as fundamental positions are often in competition for fixed resources—the seeming imbalance of one over the other may well influence the nature of those decisions or the relationships implicated in coming to those decisions.

A large literature exists in which researchers theoretically examine collaborative decision-making in committees (e.g., see Buchanan and Tullock (1999); Li et al. (2001); Ottaviani and Sørensen (2001); Levy (2007)). Yet, empirical applications remain relatively unexplored. In this paper we consider one such decision-making environment—the Supreme Court of the United States (SCOTUS)—and the nature of decision making as the balance of the court changes. In doing so, we find empirical evidence that the greater is the imbalance in party affiliation (a six-three court is more imbalanced than a five-four court, for example) the less within-party dependency we find in Justice votes—a simple model that explains this would have Justices more-likely anticipate being the marginal voter on a five-four court, for example. To the contrary, as imbalance rises, we find a larger role for individual ideology in determining Justice voting behavior, as though the *a priori* closeness of votes based on political positions had shut down on information sharing across party, and moving priors away from five-four decisions opens up the potential to be influenced by those across the party divide but close in ideology space. The Supreme Court provides a unique opportunity to test such a hypothesis, as it offers exogenous changes in affiliation balance that corporate boards would not, for example.[1]

In the last 60 years, the country has shifted more broadly toward political polarization— differences in the ideological positions of the median Democrat and the median Republican in the U.S. Congress have increased by 53 percent (Poole, 2005), and Americans are themselves increasingly more

[1] See Manski (2000) for discussion of the challenges of empirically estimating social interactions.

politically polarized (Center, 2014).[2] At some level, this shift is also evident at the Supreme Court.

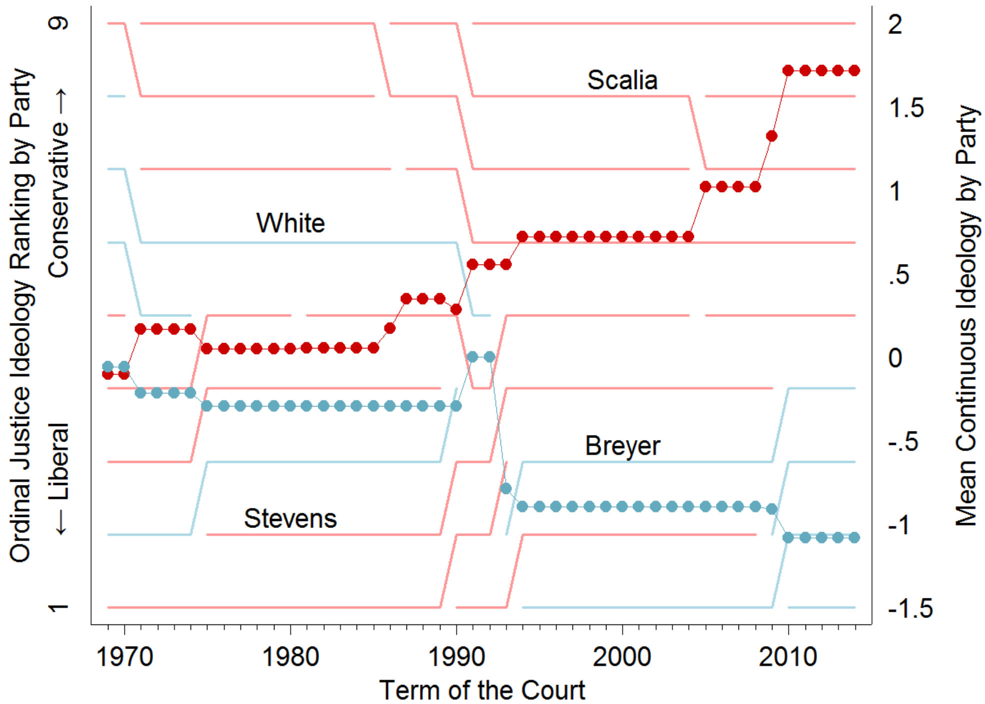
In Figure 1 we plot the appointing-President's party affiliation for each of the Justices on the Court between 1969 and 2014. In so doing, we also rank-order the Justices by a measure of their political ideology, illustrating the relative ideologies of the appointments to the Court over time.[3] At the same time we plot the mean ideology over the time-series, which highlights the overall rise in the Court's ideological polarization, with newly appointed Justices tending to increase the ideological distance between the average Republican and Democrat appointees. This figure also makes evident that the separate identification of the roles of party affiliation and Justice ideology is increasingly challenging over time. However, for as long a time series as is estimable (i.e., there is identifying variation in terms 1969 through 1990), our analysis will reveal a story that implicates party and ideology differently.

That said, it is uninteresting to demonstrate that Justice votes have an ideological component—in fact, throughout our analysis we will absorb all such justice-specific leanings into Justice fixed effects. We will also estimate separate models by term of the Court, which implicitly relaxes even the restriction that the ideologies of individual Justices map into votes similarly over time. (For example, though Souter was appointed by a Republican, he is thought to have become a reliable liberal vote on the Court over time.) Instead, then, we contribute to understanding *the interactions* of decision makers in groups with potentially conflicting interests, and to knowledge of Supreme Court decision-making in particular.

[2] Roughly 92 percent of those who identify as Republican now measurably more conservative (on issues) than the median Democrat, and 94 percent of those who identify as Democrat now more liberal than the median Republican. (Only twenty years ago, these same metrics were 64 and 70 percent.)

[3] This measure, introduced in Bonica et al. (2017b), is in no part determined by their actions as a SCOTUS Justices.

Figure 1: Justice ideology and party affiliation



Notes: In each term, the Justices are ranked from most liberal (1) to most conservative (9). Each line represents the ordinal ranking for a single Justice throughout her career. As Justices have a fixed ideology across all terms, changes in rank occur to the arrival and departure of Justices. The connected scatter plot, associated with the right vertical axis, displays the mean (cardinal) ideology of Justices, separately for Republican and Democrat appointees to the Court.

Although Supreme Court decisions are public, as are the final outcomes of most collaborative decision-making bodies, the actual deliberations and voting are quite secretive, occurring to the exclusion of all but the nine Justices themselves. While unobservable, there are well-developed techniques that allow one to retrieve measure the potential co-variation in the outcomes of those deliberations that can be informative.

We do so here, modeling the votes of Supreme Court Justices inclusive of a "spatial-lag" parameter—a parameter for each term of the Court that reflects the degree to which the direction of a Justice's vote on a given case is predictable by the votes of other Justices.[4] In so doing, we are estimating a dependency of sorts, within subsets of "spaces" that define potential relationships among the nine Justices. Our econometric procedure eliminates the concern the estimated dependencies are influenced by shared unobserved elements, which impact multiple votes simultaneously, allowing for a causal interpretation of the effect of one vote on another—a vote dependency. Our identification is not defeated by sources of unobserved heterogeneity across case- or justice-specific attributes—these sources of variation are absorbed into case-specific and justice-specific fixed effects. Neither do our estimates identify off of unobserved case-by-justice heterogeneity—any correlation in unobserved case-by-justice heterogeneity is removed from identifying variation through our empirical approach.

In the end, we estimate the extent to which shared party affiliations shape vote dependencies among Justices on the Court.[5] Doing so, we find strong evidence of such dependencies. More striking, even, is that the degree of measurable voting dependency is systematic with the *a priori* imbalance of potential voting blocks on the Court. That is, the revealed strength of vote dependency within groups of party affiliated Justices decreases abruptly when the party imbalance of the Court increases from five Republicans and four Democrats (from 1969 to 1970), to six and three (from 1971 to 1974), and again when the Court becomes seven and two (from 1975 to 1990). Moreover, we find no such evidence around changes in the Court's makeup that do not imply changes to party imbalance.

This is consistent with party affiliation playing a smaller role in voting

[4] A rather large literature in spatial econometrics follows the advances in Anselin (1988), LeSage and Pace (2004), and LeSage and Pace (2009) and elsewhere.

[5] We use the appointing President's party as a measure of the political party of the Justice, common in the literature (e.g., see Sunstein et al. (2006)).

when there is more imbalance in the party affiliation of Justices. For example, in 1969 and 1970, party affiliations would leave any individual Justice as the potential marginal vote, an expectation that could lead Justices to be more aware of their role in determining the aggregate outcome, and their party affiliated Justices' votes. While there need not be measurable dependencies of any kind, coincident with dependencies within party affiliations attenuating as the structure of the Court changes over time, we find an increase in *ideology-driven* vote dependencies. In particular, we find dependencies *across* party affiliations between Justices who are individually the most similar to each other in their ideology.[6] -squares estimates can exhibit bias in the presence of spatial dependency, spatial dependency is both measurable and interpretable.

As all voting dependencies can represent various forms of information sharing or learning, normative evaluations of dependency itself should be made with great care. As such, we do not ourselves take a position on whether partisan or ideological dependency is itself to be praised.[7] However, as we uncover systematic variation in how votes co-vary, and how this co-variation *changes* as the structural makeup of the court itself changes, we are tempted to interpret the data as suggestive of mechanisms other than simple notions of information sharing and learning, which should not vary with the Court's partisan structure. In particular, it cannot be ignored that the relationship between the votes of party affiliated Justices is strongest when the structure of the Court leaves the highest potential for one Justice to be the marginal vote on a case (i.e., a five-four margin).

In Section 1.2 we provide a brief summary of the Supreme Court as an

[6] Berdejó and Chen (2017) similarly finds that partisan voting among US Court of Appeals Judges varies with whether or not it is a presidential-election year.

[7] See Nivola (2009) and Galston (2009) for normative discussions regarding partisan politics.

institution, and the lifecycle of a SCOTUS cases. In this section, we also provide a review of the relevant literature to which we contribute. In Section 1.3, we formally motivate our estimated equation and describe our data, which we follow in Section 1.4 with a presentation of empirical results and discussion. In Section 1.5, we offer concluding remarks.

1.2 Background

Here, we offer context for the empirical application, and follow up with a review of the relevant literature, to which we contribute.

1.2.1 The institution of the Supreme Court

Each of the nine Justices of the Supreme Court of the United States (SCOTUS) is appointed by the President, confirmed by the Senate, and expected to hold office for life. Among the oaths taken upon confirmation, Justices commit to faithfully and impartially perform the duties of the Court. Yet, the rancor associated with judicial appointments to the Court suggests that some question this impartiality.[8] It is as though judicial appointments are political in part, and a lifetime of court rulings may be moved one way or the other by the political persuasions of the Justices, either individually or in the collective.[9]

Once a case is submitted to the Supreme Court, there are potentially four stages to the progression of the typical case. First, in one of the

[8] See, for example, Bonica and Sen (2017) for an empirical examination of partisan and ideological considerations in the Judicial selection process.

[9] After the Republican-controlled Senate refused to consider a 2016 Obama-nominated replacement for the deceased Justice Scalia, the nomination process appeared to reach a new level of polarization. Since the 2017 installation of President Trump, the nomination process for Federal judges at all levels has progressed with an alarming lack of bipartisan support, a significant departure from the preceding 100 years (Dash, 2017). In President Trump's first nine months, his nominees to federal courts are thought to be both increasingly partisan and younger, implying a lasting effect in the Judicial Branch (Klain, 2017).

twice-weekly Justices' conferences, the Justices collectively determine if the case is to be adjudicated. If four or more Justices vote to hear the case, the case is added to the Court's docket for the term.[10] Second, the Court hears oral arguments in the case, which are open to the public and consist of each party to the case making their argument before the Court. Oral arguments are completed in the beginning of the term, while the last few months of the term are dedicated solely to conference and opinion writing.

The third stage of any case is the convening of Justices in conference. As is the tradition, Justices vote on cases they've heard on that Monday and Tuesday at their Wednesday afternoon conference. Likewise, they vote on cases they've heard on the preceding Wednesday at their Friday afternoon conference. In order of seniority, each Justice is given the opportunity to express their view on each case, after which votes are verbally cast in the same order, with the most-senior Justice casting the first vote. A majority opinion writer and dissenting opinion writer, if applicable, is immediately assigned by the Chief Justice and highest-ranking Justice in the dissent (if applicable).[11]

When all opinions are written, the Justices meet in Conference for the fourth and final stage of a case to finalize their collective vote.[12] Case decisions and opinions are typically delivered to the Court during the last weeks of the term, in late June or early July. The fundamental privacy of all conference deliberations and voting that implies the need to model

[10] The votes of the Justices are taken privately, and the votes of the individual Justices are never published (Fisher, 2015).

[11] As majority opinions issued by the Supreme Court establish precedent, the reasons for the Court's decision are just as important as the decision itself. As such, *concurring opinions* can also be offered—a written opinion of a judge that agrees with the majority, but offers different or additional reasons as the basis for support.

[12] Over the course of writing the opinions, the Justices continue to deliberate with one another and see other cases. As the initial votes are never released to the public, it is unclear how often Justices switch their votes in this stage. While understood to be rare, the dissenting opinion has on occasion become the majority opinion as late as this stage (see `http://www.uscourts.gov/about-federal-courts/educational-resources/about-educational-outreach/activity-resources/supreme-1`).

case-specific voting as having a spatial component, which we justify below.[13]

1.2.2 Literature

A large literature exist across economics, political science, and law in which researchers have considered the determinants of SCOTUS voting. Within that broad arena, some will see our modeling approach most closely matching what is known as the "Attitudinal Model." Advocates of this approach suggest that not only do case characteristics influence how Justices vote, but the interaction between Justice-specific characteristics (e.g., their individual ideologies) and case characteristics also enter the Justices' decision-making process. For example, Justice characteristics are often found to be more influential than legal characteristics of the case in determining Justices' votes—Segal and Spaeth (2002) and Sunstein et al. (2006) interact "ideology" with legal co-variates that are thought to capture the important determinants of Justice votes (e.g., the extent of legal precedent). Moreover, Epstein and Posner (2016) finds SCOTUS Justices more likely to side with the government when the sitting President has appointed the Justice to the Court and Caldeira et al. (1999) finds Justices deliberately vote with an eye for how it will be seen given the anticipated outcome. Additionally, Hall (1992) concludes that State Supreme Court Justices vote strategically to increase their likelihood of reelection. Peppers and Zorn (2008) and Bonica et al. (2017a) (using what is our preferred approach to measuring Justice Ideology at

[13] The timing of the court's ruling, as well as the publication of the ruling, is little understood. Anecdotally, there is some evidence that Justices do not even begin deliberating some cases until all oral arguments are completed for the term (Levy, 2015). There is also evidence that the Court withholds from publishing certain rulings to ease the reporting process (Palmer, 2013). Not only is there little evidence that the within-term timing of Court decisions impacts Court rulings, the within-term timing is unobserved to the econometrician. Lastly, it should be noted that in an act of profound symbolism, SCOTUS rules strictly prohibit any photographs to be taken of the Court, with only two photographs of the Court's proceedings having ever been (illegally) published, the most recent of the two taken in 1937 (West, 2012).

the SCOTUS level) find that law clerks influence Justices' votes. Harmon et al. (2019) also suggests peer effects are influential in legislative voting in the European Parliament.[14]

The mapping of individual ideology to actual votes is complex, however, even before considering potential codependencies. Fischman (2013) suggests that the votes of Circuit Courts judges *directly* influence the votes of other judges—while their institutional setting differs from ours, they also estimate an instrumental-variables regression where colleagues' characteristics are used as instruments for colleagues' votes. While they do control for case attributes, our identification allows for the inclusion of case-specific fixed effects, which absorbs any unobserved case-specific heterogeneity that could influence the votes of multiple Justices—this will constitute our preferred specification. Though not considering how interactions may depend on the political balance of the court, Holden et al. (2019) considers the question of peer effects among SCOTUS Justices, using Justice turnover and absences as identifying variation. However, case-specific fixed effects again are not utilized in their specification. We consider SCOTUS Justices and, in particular, how the potential interactions and influences across Justices may depend on the party alignment of the Supreme Court. In our analysis, we absorb any leaning of a particular case (liberal or conservative, for example) into an estimated case-specific parameter, so to not confound the relationship we identify across Justices between the average vote on a particular case and the endogenous relationship between votes on that case. In so doing, we also approach the problem with a large degree of flexibility—we will run everything separately by term, for

[14] The literature also includes results like Danziger et al. (2011) (showing that judges are less likely to grant an individual parole the more time has passed since their last meal) and Eren and Mocan (2018) (that judges tend to grant longer sentences to certain defendants after the football team of their alma mater is defeated unexpectedly) and others (Cohen and Yang, 2019; Spamann and Klöhn, 2016) suggest that care be taken in identifying judicial behavior.

example—but restrict identifying variation to strictly within-case variation.[15]

1.3 Empirical methods

Here, we discuss our empirical methodology and discuss the data used in the analysis.

1.3.1 Model Specification

Even though Justice interactions occur within closed-door conference negotiations and are therefore unobserved by the econometrician, they do manifest in Justice votes. Given these votes, it is the implicit *interactions* that occur in that data-generating process that we are interested in measuring.[16] As it turns out, SCOTUS deliberations are precisely the sort of data-generating process that the inclusion of a spatial-lag coefficient would capture. This association is not always made explicit—Fischman (2013) does not mention the spatial econometric literature, though it is very much in keeping with the method—but there is an existing apparatus developed around the intent of extracting such information. In fact, some of the pioneering work is very much in keeping with the data-generating process of the Supreme Court—Anselin (2003) describes the methodology's value when investigating strategic interactions, social norms, neighborhood and peer-group effects, and how individual interactions can lead to emergent collective behavior and aggregate patterns. In this way, our methodology follows standard practice in the spatial-econometrics literature. Moreover, it highlights the endogeneity problem somewhat more formally.

[15] As will be made clear with the articulation of our specifications below, we include both case- and justice-specific fixed effects, which leaves within-case variation in justice ideology to identify the relationship between Justice votes. In so doing, we assume that Justice votes are not influenced by the ideologies of other Justices on the Court in the same term, other than through their votes.

[16] See Ladha (1995) for theoretical dissucssion of information sharing in group voting.

By way of example, consider a Justice who hears arguments in conference that are offered by other Justices. If the other Justices' arguments have no influence on her vote, on average, there should likewise be no explanatory power in the observed votes of those others when we come to predict her vote—that is, more generally, there should be no measurable dependency between realized votes. In the notation of a typical "spatial-lag" model, there is a parameter (typically notated ρ) that measures this dependency, as captured by a spatial weighting matrix (typically notated as W). If $\hat{\rho} = 0$, there is evidence the relationship allowed for in W is not a determinant of her votes, on average, as would be the case if Justice arguments were not systematically informative to her. Of course, even if the vote of the Justice is not influenced by the arguments of her peers, there may be correlation in the observed votes of Justices due to shared unobservable case attributes. As we discuss in detail below, we avoid this type of correlation from influencing our estimate of ρ by instrumenting for the votes of her peers. If, on the other hand, the arguments of other Justices did tend to influence her vote, then the extent to which $\hat{\rho}$ deviates from zero will be informative to the strength of those underlying mechanisms within the observed data-generating process. We do this, having absorbed any variation in realized votes that can be explained by case- and Justice-specific parameters, and an error process itself.[17] In particular, we anticipate that within-party vote-dependency is highest when

[17] Note that if all Justices are similarly persuaded by an argument, votes would collapse on unanimity and a case-specific parameter would sufficiently capture that realization without any need to appeal to a spatial process of vote dependency. Indeed, we might be inclined to infer that there was something unobservable about that case that best explained the unanimity, easily bypassing any appeal to "spatial dependencies" to explain such a tendency. This is the sort of mechanism that we will unfortunately not be able to speak to—spatial dependencies that are so strong that they collapse on unanimity can not identify the "spatial-lag" coefficient. In this way, we might identify a lower-bound of the true relationship.

the majority margin is thin.[18]

We now proceed to introduce additional formality in setting up the empirical model. As we estimate the model separately for each term of the Court (between 1969 and 1990), in no way do we restrict how these relationships appear across terms, which will allow for interpreting any changes in these interactions over time. Following standard notation, for each term we model Justice j's vote as

$$V_{jc} = \kappa_c + \lambda_j + \beta_c Ideology_j + \rho\, WV_{jc} + \varepsilon_{jc}, \tag{1}$$

where the elements of V_{jc} are $\{0, 1\}$ and represent the votes of justices j on cases c. We code $V_{jc} = 1$ if Justice j voted in the "liberal" direction on the case. In explaining the variation in V_{jc}, we allow votes to have their own level difference across cases, κ_c, capturing any case-specific characteristics that might influence votes across Justices for a given case. (For example, were the composition of cases influenced by the balance of the Court, level differences in Republican or Democrat Justice's inclinations would be captured in κ_c—the average liberal/conservative leaning, in a way.) Similarly, we capture any unobserved Justice-specific heterogeneity in λ_j, thereby absorbing any tendency for individual Justices to vote in the liberal or conservative directions, generally—to guard against any j-specific leanings inadvertently identifying dependency. While Justice ideology is considered fixed within a given term, and therefore captured in λ_j, we allow individual ideologies to map into voting differently across cases, through $\beta_c Ideology_j$. Given our identification strategy (below) this flexibility is what enables our estimation of

[18] The dynamic also applies for Justices within the minority-party. If four Justices belong to the minority-party and vote in the same direction, they are within one Justice of taking the majority opinion. However, the smaller the minority, the farther from establishing the majority position they will be, and the less significant will be the need to have strong within-party vote dependency.

$\hat{\rho}$. We report standard-error estimates having allowed for clustering on cases.[19]

In Equation (1) we also include the spatial-lag itself, $\rho\, WV_{jc}$, where ρ is to be estimated, and reflects the degree to which the voting of other Justices (where the "others" are captured in W) explains V_{jc}. We consider two such weighting matrices, in particular. First, we consider the potential relationship between votes of Justices who were appointed to the Court by a president with similar party affiliation. (We will notate this with the estimation of ρ^{Party}.) Second, we consider the potential for Justice-specific ideologies to form the basis for vote dependency—specifically, we will measure the extent to which the voting behavior of the ideologically closest Justice who does not share Justice j's political party affiliation explains Justices j's voting. (We will notate this with the estimation of ρ^{NO}, where "NO" is read as "Nearest Other.")

As Equation (1) allows V_{jc} to depend on a weighted transformation of V_{jc} itself, we follow Kelejian and Prucha (1998) in identifying $\hat{\rho}$—we instrument for WV_{jc} with the W-weighted exogenous variables in Equation (1).[20] By using the weighted exogenous variables to instrument for WV_{jc} we predict the average vote of j's peer-justices (defined through W) using the the average ideology of that group interacted with case-specific fixed effects. This is standard in the spatial econometrics literature, and simply amounts to modeling the endogenous peer effect through an instrumental variable technique, where the ideologies of other Justices (those serving in the same term) are used as instruments for the votes of other Justices. It is in this that we first see our eventual exclusion restriction—including both case- and

[19] Robust standard errors are similar, as are those allowing for clustering on justice (which accounts for correlation of votes across cases for given Justices, which would otherwise lead to misleadingly small standard errors).

[20] Following Kelejian and Prucha (1998), we also include W^2-weighted variables as instruments in the first stage.

justice-specific fixed effects, and still allowing for individual ideologies to influence own voting, we assume that Justice votes are not influenced by the ideologies of other Justices on the Court in the same term *other than through their votes*. Put differently, we retrieve an estimate of the influence of peer votes on own voting using information retrieved from estimating case-specific relationships between peer ideology and peer voting, while controlling for the general relationship between individual ideology and individual voting. We therefore assume that the impact of colleagues is purely an endogenous effect, since the instruments are invalid in the presence of contextual effects—justices can only be influenced by their colleagues through their votes, and their characteristics cannot have a direct impact. Following others (e.g., Harmon et al. (2019), Fischman (2013), Chupp (2014)), we estimate linear probability models in both first and second stages.[21]

Allowing for vote dependencies within party affiliated Justices

In considering the explanatory influence of party affiliated Justices, we define the political party of the appointing president as $Party_j \in \{Democrat, Republican\}$, with the elements w_{jk} of the 9×9 spatial-weight matrix W^{Party} defined $w_{jk} = 1$ if $Party_j = Party_k$ and $j \neq k$, and zero otherwise.

With λ_j absorbing across-Justice variation, κ_c absorbing across-case variation, and $\beta_c\ Ideology_j$ controlling for any relationship between individual ideology and voting in case c, the identifying variation contributing to $\hat{\rho}$ in our baseline specific of Equation (1) is that originating from variation in the votes (on the same case) of Justices who share the same $Party_j$. As our estimate of WV_{jc} is obtained from a regression that controls for both $\beta_c\ Ideology_j$ and $\beta_c\ W Ideology_j$ (our instrument), our exclusion restriction is *not* that a

[21] In Monte Carlo simulations, Beron and Vijverbeg (1999) finds that spatial linear-probability models are close approximations of the true data-generating processes.

justice's own ideology (interacted with case-specific fixed effects) has no influence on her vote. Rather, we must assume that the average ideology of the other Justices in her party (interacted with case-specific fixed effects) only influence her vote through their votes. Namely, we identify off of variation in the *predicted* votes in a first-stage regression that uses the relationship between voting and justice ideology on average... not actual peer-justice votes, but what we would anticipate one's peer-justice votes to be given the relationship between the ideology of those peers and their case-specific votes. Thus, while we would worry about identifying off of whatever causes actual peer votes to deviate from what one might anticipate given the mapping of ideology into voting, we have no such fear when instrumenting for actual votes with predicted votes.

As is customary (see Anselin (1988) and following), we row-normalize the weighting matrix (i.e., normalize weights to sum to one for each justice-case) to prevent the introduction of variation that confounds the actual behavioral response to other-Justice votes with the *number* of other Justices to which it is possible to best respond. This normalization also allows us to retrieve estimates of the response to the average affiliated Justice independently of structure that will subsequently allow us to compare parameter estimates across terms, as the number of party affiliated Justices changes.Though inference statements adjust accordingly (from "in response to the average Justice voting in the liberal direction" to "in response to one more Justice voting in the liberal direction," for example) results of the analysis are qualitatively similar when we do not row-normalize the spatial weighting matrix. [22]

As elements of W^{Party} that correspond to $\{j,k\}$—pairings across party lines are assigned a value of zero, $\hat{\rho}$ reflects the responsiveness to the average information exchanged between pairs of Justices of the same $Party_j$—indicative of the votes of j's party affiliated Justices having useful information in predicting j's vote.

With the inclusion of ρ in the estimation equation, we capture the vote-determination process more flexibly than would an $\rho = 0$ restriction. Specifically, $\hat{\rho}$ captures the *average* best response between Justices j and those Justices given weight in j's vote through W—those with similar appointees in this first case. For example, if $\hat{\rho}^{Party} = 0.20$ then the associated inference statement would be that were all of j's party affiliated Justices to move from a vote of "zero" (i.e., the conservative position) to a vote of "one" (i.e., the liberal position), the probability that Justice j would vote the liberal position would increase by 0.20, or 20 percentage points. Below, we will find estimates of ρ^{Party} as high as 0.75—this happens in the earliest years of our sample, when the party affiliations of the Justices are most in balance. When the Republican-affiliated majority is strongest, however, we retrieve estimates of ρ^{Party} as low as -0.50, suggesting that a "conservative" vote from an affiliated

[22] For example, if Justices 1, 2, and 3 (corresponding to the first three rows of W) were appointed by a Democratic president and Justices 4, 5, 6, 7, 8, and 9 were appointed by a Republican president, as was the case between 1971 and 1974, the 9×9 weight matrix, W^{Party}, can be defined as, for example

$$W^{Party}_{1971} = \begin{bmatrix} 0 & .5 & .5 & 0 & 0 & 0 & 0 & 0 & 0 \\ .5 & 0 & .5 & 0 & 0 & 0 & 0 & 0 & 0 \\ .5 & .5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & .2 & .2 & .2 & .2 & .2 \\ 0 & 0 & 0 & .2 & 0 & .2 & .2 & .2 & .2 \\ 0 & 0 & 0 & .2 & .2 & 0 & .2 & .2 & .2 \\ 0 & 0 & 0 & .2 & .2 & .2 & 0 & .2 & .2 \\ 0 & 0 & 0 & .2 & .2 & .2 & .2 & 0 & .2 \\ 0 & 0 & 0 & .2 & .2 & .2 & .2 & .2 & 0 \end{bmatrix}$$

Justice is associated with a lower probability that others sharing the same party affiliation would vote in the conservative direction. The average (across term) vote in our sample is in the conservative direction 51 percent of the time. As such, a -0.5 to 0.75 swing is equivalent to -100 percent to 150 percent of a standard deviation increase in the likelihood the Justice votes in the conservative direction.

Allowing for vote dependencies between unaffiliated ideological "neighbors"

Although a spatial dependence based on the partisanship of the appointing president seems natural, it is possible other spatial dependencies inform the votes of Supreme Court Justices. Specifically, in a second approach to modeling the dynamics of Justice votes, we consider looking to the ideologically like-minded Justices who do not share a party affiliation. This follows from our intuition, that votes of those closest in some ideological space but "across the aisle" may have explanatory power in predicting Justice j's vote.

In an alternative weight matrix, then, we allow j's vote to be influenced by k's vote (i.e., $w_{jk} = 1$), when k is the closest to j in ideology among all available k that satisfy $Party_j \neq Party_k$. We notate this "nearest-other" weighting matrix as W^{NO}. Under such a weighting rule, $\hat{\rho}^{NO} = 0.2$ would imply that, on average, when the oppositely-appointed Justice closest to j's ideology switches their vote from the conservative to the liberal direction, the likelihood Justice j votes in the liberal direction increases by 20 percentage points.

1.3.2 Data

Justice votes

The voting data we use in our analysis originates from the Supreme Court Database–Justice Centered Data, managed by Washington University Law School (Spaeth et al., 2017). We consider the terms of the Supreme Court from 1969 through 1990. The dataset contains information at the justice-by-case level.[23] The specific variable of interest is each justice's case-specific vote, which is classified as either being the "liberal" or "conservative" position on the matter. Spaeth et al. (2017) classifies each vote in this manner using an extensive hierarchical rubric, determined independently of how each specific Justice voted.[24] For example, if a Justice votes in a "pro-injured person" direction for a case on unions or economic activity, or the "pro-female" direction in cases regarding abortion, the vote is classified as liberal.[25]

Justice ideology

With respect to the measurement of Justice ideology, we adopt the index developed in Bonica et al. (2017b), which stands as one of the few measures of ideology that is independent of contemporaneous vote decisions, maintaining identification and mitigating the concern one would have over introducing

[23] While SCOTUS primarily functions as an appellate court, it has had original jurisdiction roughly 200 times since its founding in 1789. We exclude all cases in which SCOTUS has original jurisdiction as the votes in these cases are not classified as "liberal" or "conservative."

[24] That is, Justice votes cannot influence the classification of a vote direction as "liberal" or "conservative". Though, Harvey and Woodruff (2011) do suggest that there may be a small amount of confirmation bias in determining the case-type category for certain cases. We do not worry that this plays an important role in our analysis, and anticipate that any such bias would be absorbed into case and/or Justice fixed effects. We do redefine V_{jc} to capture Justice j voting to reverse the ruling on case c, and our results are both qualitatively (and even "quantitatively") unchanged.

[25] The measurement of political direction of each case has been extensively utilized within the field. For example, see Katz et al. (2017).

endogenous measures of ideology into our modeling of voting (Bailey, 2007; Martin and Quinn, 2002).[26] Fundamentally, this index is based on the political-campaign contributions of Justices' law clerks, yielding "Clerk-Based Ideology" (CBI) scores for each SCOTUS justice.[27] Each Justice in our sample has a fixed CBI score. Due to data limitations, Bonica et al. (2017b) does not calculate CBI scores for all Justices present in terms prior to 1969, thus limiting our analysis to the 1969, and later, terms.[28]

Sample restrictions

We estimate Equation (1) separately for all SCOTUS terms between 1969 and 1990. Recall that we consider dependencies both within party (W^{Party}) and between ideological neighbors (W^{NO}). Our considerations of potential vote dependencies are therefore limited to the number of terms for which there is informative variation. For example, we cannot estimate Equation (1) between 1991 and 1993, as the structure of the Court includes only a single Democrat-appointed Justice on the bench—there is no within-party variation.

[26] For additional detail, see Bailey (2017).

[27] CBI scores are calculated by averaging the CFScores, a measurement of ideology based off of political contributions, of all individuals who clerked for a given justice. See Bonica (2014) for the development of CFScores, and Bonica et al. (17su) for their application to law clerks. As a Justice's ideology correlates positively with that of their law clerks (Liptak, 2010; Nelson et al., 2009), the political donations of law clerks arguably predict but are exogenous to Justice ideology. While it is possible that Justice votes influence the political contributions of future clerks (e.g., potential clerks may donate to certain candidates to increase the likelihood they will be awarded a clerkship), Bonica (2014), Bonica et al. (17su), and Bonica et al. (2015) find no evidence that individuals contribute to candidates strategically.

[28] Segal and Cover (1989) develops an alternative measures of Justice ideology. However, these measures are intended to capture attitudes towards civil liberties, as opposed to overall ideology, and vary in informativeness, dependent on the appointing President. See Epstein and Mershon (1996) and Segal, Epstein, Cameron, and Spaeth (1995) for discussion of these and other limitations. Nonetheless, when the analysis is completed using these measurements of ideology, results (not reported) are qualitatively similar. Epstein et al. (2007) also develops a partially exogenous measure of Justice ideology, informed by Justice votes in their first year on the court. It is time-invariant and could be used for our analysis, but we should then also discard the six terms over our sample in which there is a first-term Justice. This measure also weighs the political ideology of the appointing president directly, introducing common factors across Justices, which would further confound inference. For these reasons, we do not adopt this measure.

However, once the court returns to having two democrats in 1994, Equation (1) can not be estimated with W^{NO}, as all Republican-appointed Justices share the nearest Democrat-appointed Justice and each Democrat-appointed Justice shares the nearest Republican-appointed justice. Moreover, after 2009 (and evident in Figure 1), the structure of the Court leaves no overlap in Justice ideology across party affiliation. That is, the least-liberal Democrat-appointed Justice is more liberal than the most-liberal Republican-appointed justice, leaving no variation among Republican (Democrat) Justices in their differently affiliated nearest ideological neighbor.

We also discard observations under three conditions. First, if fewer than nine justice's vote on a given case, it introduces a case-specific (c-specific) dimensionality to the weight matrix and identifying variation that is then also potentially endogenous. We therefore discard all such cases (861 out of 3,542) and the associated votes on the cases (7,534 of 31,663).[29] Second, we give special attention to the 1975 term, the term in which Justice Douglas retired after participating in only six cases. Instead of losing the information for the entire term, we model 1975 after discarding these first six cases and the 12 that cleared the Court's docket before the appointment of Justice Stevens. Last, in approximately two percent (602) of the remaining votes, the classification of votes as either "liberal" or "conservative" is indeterminable.[30] So to not impose this indeterminacy on related votes, we discard all (149) cases for which there is any Justice vote that is not classifiable as "liberal" or "conservative."

Other than the 1969 term, after the sample restrictions are applied, the

[29] In particular, note that only eight Justices were on the bench for the majority of the 1969 term. Despite the small sample size, and thus imprecise point estimates, the information in the term provides insight into possible spatial dependencies.

[30] For example, if in a justice's opinion for the case, whether the opinion is in the majority, dissent, or concurrence, she sides with both liberal and conservative points of view, as defined by the Spaeth et al. (2017) rubric, the direction of the Justices' vote is undefined.

fewest number of cases heard was 62 (in 1987) and the highest number of cases heard was 153 (in 1972). As the Supreme Court saw fewer cases over time during the terms in our sample (Moffett et al., 2016), there is a corresponding decline in the number of cases in our sample each term.

Descriptive Statistics

In Table 1, given the structure of the Court—the number of Republican and Democrat appointees over time—we report the fractions of times the various party specific majorities were realized. For example, of the five Republican-appointed Justices on the Court in 1970, in 54 percent of cases they voted together (i.e., a majority of five), in 27 percent of cases they voted four-to-one, and in 19 percent of cases they voted three-to-two.

Across all terms with seven Republican-appointed justices (1975-1990 and 1994-2008), on average, 21.2 percent of cases were ruled with the Republican-appointed justices divided by one vote (i.e., four-three) and across all terms with five Republican-appointed justices (1969 to 1970 and 2010 to 2014), the average percent of cases in which the Republican appointed justices were divided by one vote (three-two) is 13.1 percent. Similarly, in terms with two Democrat-appointed justices (1975 to 1990 and 1994 to 2008), the two Democrats voted in the same direction on average in 64.4 percent of cases, while in terms with four Democrat appointed justices (1969 to 1970 and 2009 to 2014), the four Democrats voted in the same direction on average in 75.0 percent of cases. We also show the percent of votes cast each term that were in the "conservative" direction. While there is variation across terms, there is no clear link between the average percent of votes in the "conservative" direction and the number of republicans on the court.

Table 1: Party Line Voting on the Supreme Court, 1969-2014

Number in majority =	Percent voting with the party majority: Republican 3	4	5	6	7	8	Democrat 1	2	3	4	"Conservative" votes (percent)
1969	0	17	83					0	17	83	70
1970	19	27	54					19	39	42	50
1971	8	26	18	48				42	58		46
1972	9	24	26	41				54	45		49
1973	6	18	34	42				45	55		48
1974	7	20	27	45				46	55		43
1975		18	22	20	40		33	67			49
1976		18	21	21	39		33	67			58
1977		21	33	14	31		27	73			48
1978		23	23	20	34		28	73			55
1979		24	27	20	29		38	62			47
1980		20	22	23	34		37	63			53
1981		26	20	16	38		33	67			53
1982		22	18	15	45		34	66			51
1983		20	18	17	45		34	66			53
1984		23	18	20	40		38	63			52
1985		29	22	14	35		44	56			54
1986		36	24	9	31		51	49			48
1987		27	15	11	47		32	68			48
1988		29	19	10	41		46	54			52
1989		33	17	13	37		46	54			50
1990		20	21	17	41		36	64			46
1991		8	22	24	8	38	100				49
1992		8	15	18	12	47	100				48
1993		5	19	24	9	43	100				53
1994		22	18	18	42		11	89			57
1995		15	17	23	44		17	83			52
1996		16	14	18	52		10	90			58
1997		16	16	17	52		13	88			56
1998		12	23	27	39		17	83			58
1999		16	27	17	41		18	82			45
2000		18	22	10	50		7	93			49
2001		25	25	9	42		9	91			60
2002		19	23	12	45		13	87			59
2003		21	21	10	49		8	92			56
2004		23	28	10	38		12	88			47
2005		13	28	8	51		23	77			50
2006		18	27	15	40		12	88			55
2008		19	33	10	39		21	79			54
2009	4	20	27	49				19	81		46
2010	6	24	71					6	16	78	56
2011	10	19	71					6	28	67	55
2012	16	16	67					3	12	85	50
2013	16	6	78					4	6	83	45
2014	25	24	51					1	12	87	46

Notes: For each term-party pairing, the value indicates the percent of cases in which were ruled by a within-party majority of the given number. For example, of the five Republican-appointed Justices on the Court in 1970, in 54 percent of cases they voted together (i.e., a majority of five), in 27 percent of cases they voted four-to-one, and in 19 percent of cases they voted three-to-two. For each term, we also report the percent of votes that were in the "conservative" direction.

1.4 Results

We estimate the models represented in Equation (1) separately for each year of data, for SCOTUS terms 1969 through 1990. In Section 1.4.1 we allow for vote dependencies between Justices of whom were appointed to the Court by Presidents of the same party and in Section 1.4.2 we allow for vote dependencies between ideologically similar Justices whom were appointed to the Court by Presidents of different parties. Recall, across the sample, there are three distinct structural compositions of the court. In 1969 and 1970, the Court was populated with five Republican-appointed Justices and four Democrat-appointed Justices—this is the strongest *a priori* voting block the Republican-appointed Justices will see in the time series. Between 1971 and 1974, the Court was populated with six Republican-appointed Justices and three Democrat-appointed Justices, and between 1975 and 1990, the Court was populated with seven Republican-appointed Justices and two Democrat-appointed Justices.

As an ideologically based $\hat{\rho}^{Party}$ (i.e., based on W^{NO}) is only estimable through 1990, we report on the analyses of party and ideology together in Figure 2 from 1969 through 1990. We discuss the post-1990 behavior separately in Section 1.4.3.

1.4.1 Are there evident relationships in the votes of party affiliated Justices?

In Panel A of Figure 2 we plot separate estimates of $\hat{\rho}^{Party}$ over time, with 95-percent confidence intervals. With no restrictions on the estimates across terms, the structural breaks in the composition of the Court are evident in the estimated $\hat{\rho}^{Party}$, one occurring between the 1970 and 1971 terms and one occurring between the 1974 and 1975 terms. As the Court becomes more imbalanced in terms of the party of the appointing presidents, the measurable vote dependency within party affiliated Justices declines. Point estimates of ρ^{Party} in 1970 and 1971, and in 1974 and 1975, are statically different from each other at the 1-percent level, while no other term-consecutive estimates of ρ^{Party} are statistically different from each other, even at the 10-percent level. This is consistent with more internal pressure (or bargaining) for Justices to vote in the same direction when the margins are thin—when the potential consequence of party affiliated Justices splitting their votes are higher. Note, in particular, that no discontinuities are evident around within-party changes in the Court's justices—same-party Justice turnover occurred in 1980-1981, 1985-1986, 1986-1987, and 1989-1990—consistent with our results being driven by the structural makeup of the Court relating to party affiliation.[31]

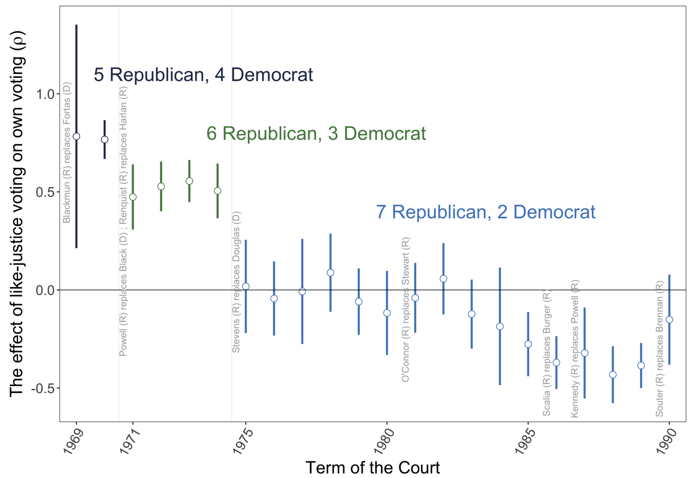
While the differences in $\hat{\rho}^{Party}$ between 1970-to-1971 and 1974-to-1975 are striking, it is possible that they are not a result of changes in the partisan-divide of the court and instead reflecting other changes. If SCOTUS cases vary systematically with changes in the partisan-divide of the court, for example, the parameter estimates reflect both the relationship between the partisan-divide and voting behavior of Justices and the selection of cases. In Figure 3 we explore this possibility.

In Panel A of Figure 3 we report the percent of cases in which the lower court's ruling was in the liberal direction. Importantly, we restrict the sample

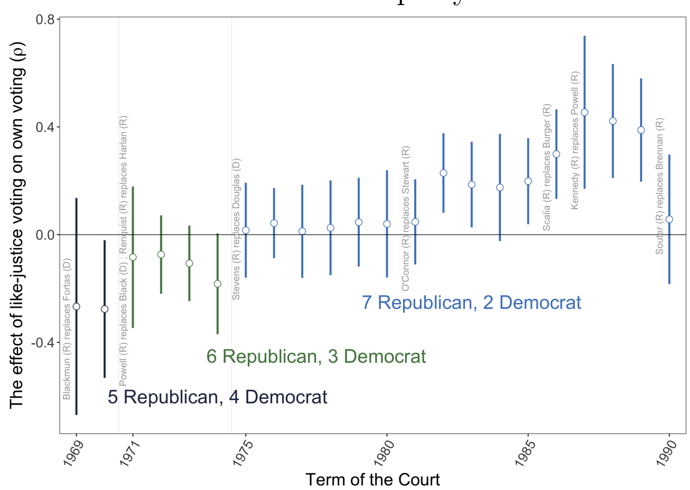
[31] The negative estimates of ρ^{Party} in the late 1980s deserve note. Throughout much of the 1980s, only two Democrat-appointed Justices (i.e., Justice White and Justice Marshall) sat on the bench. Mechanically, each vote cast by Justice White and Justice Marshall in this time span receives six times the weight of the average republican vote in the estimation of ρ^{Party}, as the total effect of six Republican-appointed Justices' votes on the remaining Republican-appointed Justice is normalized to one. While this does not inherently lead to a reduction in $\hat{\rho}^{Party}$, Justice White and Justice Marshall happened to have voted in opposite directions somewhat regularly—increasingly so in the late 1980s. For example, as suggested in Table 1, between 1975 and 1984 (when $\hat{\rho}^{Party}$ is not statistically different from zero), the two Democrat-appointed Justices voted in the same direction in 48 percent of non-unanimous cases. Yet, between 1985 and 1990, they voted in the same direction in only 33 percent of non-unanimous cases.

Figure 2: Measurable voting dependencies of Justice votes and the structure of the Supreme Court, 1969–1990

(a) Does within-party vote dependency change with how secure the party vote is?



(b) Does vote dependency among ideologically similar Justices change with how secure the party vote is?



Notes: Each estimate of $\hat{\rho}$ is derived from a separate model. The weighting matrices in Panel A, W^{Party} allows for the votes of a Justice to vary with the votes of all other Justices who were appointed to the Court by a president of the same party. For example, $\hat{\rho}^{Party} = 0.2$ would imply that if one Justice in Justice j's party changed their vote from the conservative to the liberal direction, the likelihood Justice j would vote in the liberal direction increases by 20 percentage points. The weighting matrices in Panel B, W^{NO} allows for the votes of a Justice to vary with the votes of the Justice who is closest in ideology, but affiliated with the other party. Confidence intervals (95%) are derived from errors which are allowed to cluster by case.

of cases to those SCOTUS cases that identify $\hat{\rho}^{Party}$—namely, cases with non-unanimous SCOTUS votes. While the percent of cases that were ruled in a liberal direction by the lower Court is a somewhat noisy process, no systematic variation appears and there is clearly no pattern coincident with structural changes to the party affiliations surrounding the 1970-to-1971 and 1974-to-1975 terms. In Panel B of Figure 3 we report the percent of ($\hat{\rho}^{Party}$ identifying) cases in which the lower court had disagreement in their ruling—that is, the percent of lower court rulings that were not unanimous. While a slight increase seems to occur over time, there is again an absence of shape that could be consistent with the structural changes in the court. In Panel C of Figure 3 we report the percent of ($\hat{\rho}$ identifying) cases originating from the Ninth Circuit Court, often believed to be the most liberal Circuit. Again, there is no indication of changes aligning with the structural changes in the court. Last, we re-estimate Equation (1) with the exclusion of the spatial lag (i.e., dropping the spatial component, or estimating Equation (1) with the restriction that $\rho^{Party} = 0$). In Panel D of Figure 3 we report the mean-squared error for each of these specifications, which steadily declines in a way that is consistent with the remaining covariates better capturing the decision-making process over time.[32] Importantly, there are no discontinuities in the MSEs across the structural divides of the court.

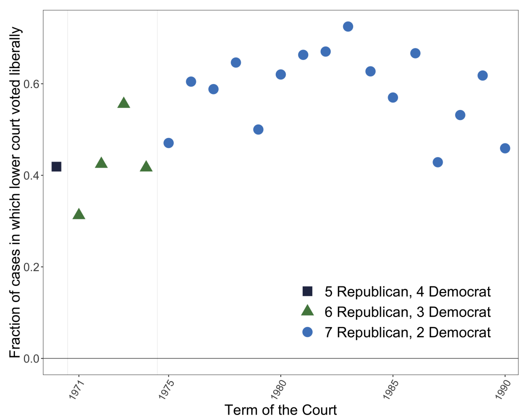
1.4.2 Has ideology replaced partisanship in voting?

In Panel B of Figure 2 we report the results of having re-estimated Equation (1), but with a weighting matrix that reflects ideology. Specifically, ideology that informs voting from "across the aisle," as we capture in

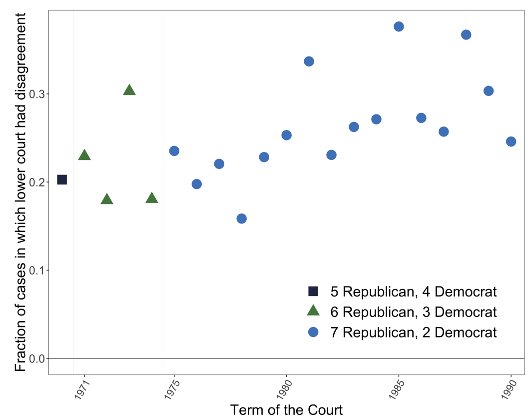
[32] This suggests that variation across Justices on ideology has itself played a more important role in explaining variation in voting over the twenty-year period we consider. This corroborates a rich literature that suggests that the Court has become increasing ideologically focused (Nelson et al., 2009; Landes and Posner, 2009).

Figure 3: Are there potential confounders that move similarly with the Court's structure?

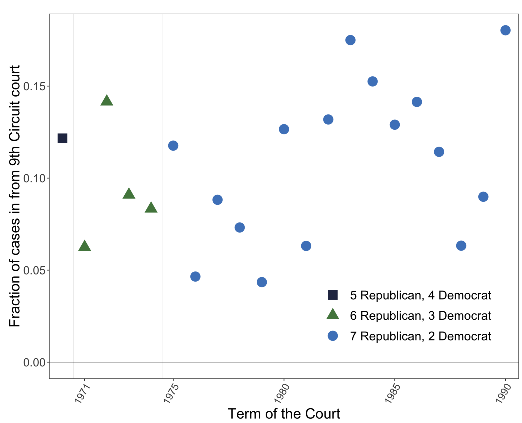
(a) Are there similar discontinuities in the ideology of lower-court decisions?



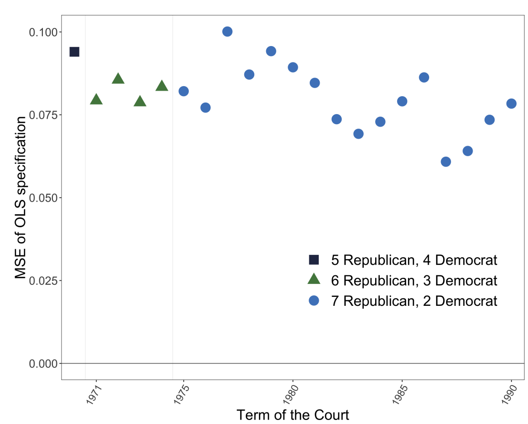
(b) Are there similar discontinuities in the unanimity of lower-court decisions?



(c) Are there similar discontinuities in the share of cases originating in the 9^{th} Circuit Court?



(d) Are other (non-spatial) covariates explaining votes differently?



Notes: In Panels (a), (b), and (c), each percentage point displays the percent of cases seen by SCOTUS in a term in which the lower court issued a liberal ruling, was unanimous, and was on appeal from the Ninth Circuit (often thought of as the most liberal Circuit), respectively, taken from the subset of cases in which SCOTUS's ruling was divided. In Panel (d), each point displays the mean squared error from a term-specific OLS regression of justice-level fixed effects, case-level fixed effects, and Justice ideology, which is allowed to interact with each case independently, on Justice vote. No spatial dependencies are allowed. In all Panels, values for the 1969 term are excluded due to the small number of observations.

$W = W^{NO}$.[33]

Interestingly, as the Court itself becomes more imbalanced with respect to party affiliation (and the measurable within-party vote dependencies decline), seemingly in accord with these structural changes, the influence of ideologically-similar Justices of *different* party affiliation... increases. While the step-function is somewhat less pronounced in $\hat{\rho}^{NO}$ (in Panel B) than it is in $\hat{\rho}^{Party}$ (in Panel A), the pattern is evident, suggesting that these two processes together tell an important story of how partisan and ideological dependencies move together.

There is an upward trend in $\hat{\rho}^{NO}$ in the period following the seven-two imbalance of the Court, which one could interpret as an eventual learning of sorts. One possible interpretation is that less dependency within party affiliated Justices (Figure 2, Panel A) allows Justices to weigh other sources of information in order to assist in their decision-making, and the views of likeminded "others" gradually takes on an importance of sorts. As Justice-specific ideology is controlled for, the results suggest that Justices' votes were positively influenced by the vote of their "other" in a way that is consistent with a discarding of the partisan-divided Courts of the early 1970s (when $Party_j$-type bargaining was highly influential), suggestive of the rise of an ideologically-guided court.

[33] In the first stage of the 2SLS approach we estimate the spatial lag of V_{jc} weighted by W^{NO}, using weighted versions of $\beta_c Ideology_j$. As W^{NO} is a nearest-neighbor weighting matrix that conditions on neighbors j and k being appointed by presidents of *different* parties, $Ideology_j$ and $Ideology_k$ are highly correlated. Thus, as $\beta_c Ideology_j$ is included in the first stage to predict WV_{jc}, there is little variation in WV_{jc} for the instruments $\beta_c Ideology_k$ to explain. The first-stage F-statistic is on the small side. However, we do not interpret this as indicative of $\beta_c Ideology_k$ being irrelevant instruments, but that $\beta_c Ideology_k$ are highly correlated with the included exogenous parameters—specifically, with $\beta_c Ideology_k$. Nonetheless, the results are suggestive.

1.4.3 The post-1990 Court

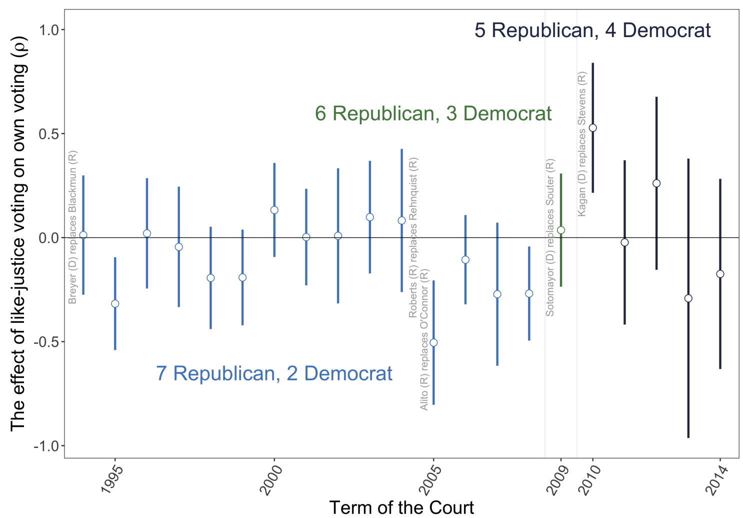
In 1991, Justice Marshall (a Democratic appointee) retired and was succeeded by Justice Thomas, who was appointed to the bench by a Republican, leaving just one Democrat appointed Justice on the Court. This defines the end of the set of terms for which a parameter on $W = W^{NO}$ is estimable, given the singular Democrat appointee. The eight-one divide continued until the 1994 term, when Justice Blackmun (a Republican appointee) replaced with the Democrat appointee Justice Breyer. This seven-two split continued through 2008, followed by a single term of six-three split, and five-four split thereafter.

This development provides an opportunity to examine a new set of transition periods and observe if the strength of dependency of within party affiliated Justice votes increased as the Court's party divide returned to a more-equal division. In Figure 4 we report the estimated spatial-lag coefficients associated with within-party vote dependency ($W = W^{Party}$). We do so separately for terms 1994 through 2014.[34]

While the estimate of ρ^{Party} does not change in a statistically significant manner with the Court's transitions from a seven-two (2008) or to six-three (2009), there is a large and statistically significant increase in $\hat{\rho}^{Party}$ as the Court transitions from six-three back to five-four (2010). The estimate of

[34] Recall, the single Democrat-appointed Justice in 1991 through 1993 precludes the estimation of both ρ^{Party} and ρ^{NO}, while the political political polarization after 1990 prohibits the estimation of ρ^{NO} for all subsequent years. Additionally, as Justice Scalia was only present for 18 cases in the 2015 term before passing away, we do not report analysis for the 2015 term, as only eight Justices were present on the Court. Moreover, regarding the 2015 term, Justice Elena Kagan remarked that "[The Court] didn't want to look as though [it] couldn't do [it's] job. ... And so we worked very, very hard to reach consensus and to find ways to agree that might not have been very obvious" (Biskupic, 2018). As such, the data-generating process in 2015 seems somewhat unique, *a priori*, as Justices not only cared about the case outcome, but specifically the makeup of the Justices' votes. (While tied votes are possible with eight Justices, it appears Justices actively avoided such outcomes in an attempt to uphold the integrity of the Court.)

Figure 4: Measurable party dependencies in Justice votes and the structure of the Supreme Court, 1994–2014



Notes: Each estimate of $\hat{\rho}$ is derived from a separate model. In each, the weighting matrix, W^{Party} allows for the votes of a Justice to vary with the votes of all other Justices who were appointed to the Court by a president of the same party. Confidence intervals (95%) are derived from errors which are allowed to cluster by case.

ρ^{Party} in 2010 (.53), which is the first year the Court returns to a five-four Republican Court, is moderately smaller than the estimates we retrieve for 1969 (.78) and 1970 (.77), the most-recent terms in which the Court was similarly structured. If anything, this suggests that the dependency between party affiliated Justices has weakened over time. The quick convergence of $\hat{\rho}^{Party}$ to zero further suggests that any dependency has lost importance in recent years, despite *a priori* balance returning, which is consistent with our finding that vote dependencies began to take on more of an ideological focus in the 1980s.

1.5 Conclusion

As the final arbiter of US law, the Supreme Court is charged with ensuring the American people the promise of equal Justice under law and, thereby, functions as guardian and interpreter of the Constitution. Given the importance, it is surprising to find that our understanding of how Justices of the Court function is relatively unexplored.

In this paper we examine the voting behavior of Supreme Court Justices in a way that also estimates potential co-dependencies in the votes of the Justices. Doing so, we identify causal relationships in the voting of Justices who share party affiliation (i.e., those having been appointed by Presidents of the same political party)—not merely level shifts in the conservativeness or liberalness of votes, which we absorb in Justice fixed effects each term and case-specific parameters, but measurable dependency that goes beyond any average inclinations of the Court.

Sharp discontinuities in these dependencies are evident, and are clearly coincident with changes in the party imbalance of the Court over time. Specifically, we find that in terms of larger imbalance—terms when the Republican-affiliated Justices outnumber Democrat-appointed Justices by a larger number—measurable co-dependencies between party affiliated Justices are attenuated. We also identify a larger role for information sharing among ideological neighbors in these terms. Interestingly, we find no similar pattern around the turnover of Justices that do not trigger a change in the balance of the Court, further suggesting that we have identified a mechanism through which the party imbalance of the Court operates.

Overall, voting patterns suggest a tradeoff—a tradeoff between political affiliation and ideology, with more-equal party representation on the Court encouraging greater party awareness in Justice voting, and less-equal party representation allowing those across party lines but with similar ideologies to inform each other's votes. As the polarization of today's Court inhibits researchers from separately identifying the roles of party affiliation and ideology, it seems party related dependencies, even as the Republican-Democrat imbalance has subsided, have not returned, which we see as a topic of future research.

Moreover, while the periodic changes in the structure of the Supreme Court allows for a unique opportunity to consider the roles of affiliation and individual ideologies, we see an intuition here that may inform other empirical applications—those more-complex and less able to allow for the identification of causal relationships. Though individual votes or actions are not always observable—the Supreme Court offers something of a unique opportunity in that dimension, since individual votes are recorded—such tensions may well be anticipated in decision-making environments more broadly, and inform the makeup of committees and decision-making authorities generally.

While my first chapter of my dissertation explores the voting behavior of Justices of the Supreme Court, I now turn my attention towards the decision-making behavior of judges within the criminal sector of the judicial branch. Both voting and criminal sentencing, the focus of my second and third chapter, are a like in crucial ways, as they both are activities which *should* be determined by relevant attributes of the setting at hand. However, in my second and third chapter, I again find unexpected characteristics often are influential in judges' decision making processes.

CHAPTER II

RACE, FAIRNESS, AND CO-DETERMINATION IN CRIMINAL SENTENCING: EVIDENCE FROM SENTENCING COHORTS IN PENNSYLVANIA

2.1 Introduction

Judges are believed to be legal experts and are, in turn, given great discretion in their sentencing decisions in criminal trials. While what information specifically should or should not influence a judge's discretion is a matter of opinion, it is generally assumed such information should not be arbitrary to the defendant in question. However, judges typically do not sentence one defendant a day, which poses a challenge to the implicit assumption that judges' decisions are independent across defendants. In fact, sentencing multiple defendants on the same day may inadvertently increase the possibility of other-defendant-specific factors (that are arbitrary to the defendant in question) impacting a judge's sentencing-determination processes, and ultimately leading to a defendant receiving an unjustified sentence.[35]

In this research, I utilize the fact that trial judges preside over multiple cases at once and sentence numerous defendants on the same day (creating what I define as a "sentencing cohort") to estimate dependencies between the sentences of different defendants. As judges receive information about defendants' cases concurrently, it is natural to believe judges may determine the sentencing outcomes of multiple defendants' cases simultaneously and even

[35] While this research is focused on judicial decision-making, as opposed to the prosecutorial decision-making, the findings of the paper can be abstracted away from the judge and simply be thought of as analysis of the actions of sentencing decision-makers, whomever they may be.

co-dependently.[36] A breadth of related literature exists within economics, political science, and law that examines what information judges use in sentencing, but it has largely focused on examining which types of characteristics, either of the judge (e.g., sex), the case (e.g., crime type), the defendant (e.g., race), or the state of the world (e.g., weather), influence sentencing outcomes for any given defendant. The literature has yet to fully examine an important aspect of the sentencing procedure: whether judges' sentencing decisions are independent across defendants.

I find that being sentenced alongside (same judge, same day) defendants with one-year longer average sentences leads to four-day increases in a defendant's sentence. As my estimate allows individual sentences to influence the sentences of multiple other defendants in the same sentencing cohort, cumulative effects stemming from an individual sentence can create large variation in the sentences of others. My estimate controls for endogeneity between sentences and potential sorting of defendants (either on observables or unobservables) into sentencing cohorts by using the average *predicted* sentence of other defendants in a sentencing cohort based on observable characteristics of the other defendants and their cases, rather than their *observed* average sentence. As such, my empirical specification does *not* rely on an assumption of random assignment of cases to judges, or random assignments of cases to sentencing days. In racial-heterogeneity analysis, I find that while the sentences of both black and white defendants are positively determined by the average sentence of defendants of the same-race in their sentencing cohort, neither are determined by the average sentence of defendants of the other-race in their sentencing cohort.

The results expand on the previous literature by shedding light on a new

[36] In fact, in an experimental setting with hypothetical cases, Rachlinski et al. (2015) finds awarded sentence lengths to be influenced by previous cases "sentenced" that day.

channel that affects judges sentencing decisions: sentence-to-sentence effects. The existence of inter-defendant effects in judicial sentencing suggests defendants receive shorter or longer sentences simply by chance with respect to the other defendants in their sentencing cohort. In addition to the moral implications of unjustified sentence lengths, as sentence length may have profound economic ramifications for defendants' futures (Aizer and Doyle, 2015; Mueller-Smith, 2015), any elements of judges' sentencing decisions which are not based on the defendants' alleged activity should be eliminated.[37] Moreover, the same-race versus other-race heterogeneity suggests judges are behaving in differing manners based on the racial-composition of the defendants seen on a particular day, potentially adding to overall racial discrepancies in sentencing.

In Section 2.2, I summarize relevant literature and review the institutional processes of sentencing within Pennsylvania, the setting of my empirical research. In Section 2.3, I describe the estimated equation and provide summary statistics of the data. In Section 2.4, I present baseline results and a variety of heterogeneity analyses. In Section 2.5, I present results of robustness exercises and in Section 2.6, I conclude.

2.2 Background

Here, I briefly describe the relevant literature and provide institutional knowledge as context for the empirical application.

[37] Note, the literature is evolving and mixed on the long-run effects of incarceration. See e.g., Bhuller et al. (2019), Rose and Shem-Tov (2019), Landerø (2015), Green and Winik (2010), and Kling (2006) for varying results.

2.2.1 Literature

It is well established that defendants receive differentiated sentences based on their demographic profile; being black or male and having low income or educational status have all been shown to increase the severity of defendants' sentences (see, e.g. Mustard (2001); Steffensmeier et al. (2006); Steen et al. (2005)).[38] In addition, many judge-specific attributes have also been found to be an important part of the sentencing procedure. For example, judge race (Steffensmeier and Britt, 2002), political partisanship (Cohen and Yang, 2019; Schanzenbach, 2015), and family structure (Glynn and Sen, 2015) all have been shown to influence sentencing behavior.[39]

Extra-legal influences have also been shown to dramatically affect judges' sentencing decisions. For example, Eren and Mocan (2018) find judges assign harsher sentences to black defendants after unexpected football loses the week before and Chen et al. (2016) and Heyes and Saberian (2019) find the results of baseball games and the temperature on the day of sentencing affect sentencing outcomes. In an experimental setting with legal experts, Englich et al. (2006) finds (literally) random information presented to legal experts influences recommended sentences length, even when the legal expert knows the information is generated at random. As it is clear judges use non-defendant-specific information in their decision making processes, it raises the question as to whether they also use different-defendant-specific

[38] Alesina and Ferrara (2014) additionally shows patterns of racism when looking at sentencing appeals. Moreover, Blair et al. (2004) and Eberhardt et al. (2006) show that when comparing defendants of the *same* race, defendants that have more Afrocentric facial features receive harsher sentences. However, the presence of differentiating sentencing outcomes among defendants of varying demographics does not itself prove judges act in an unfounded "racist" manner; Park (2017) fails to reject the null hypothesis that judges solely use statistical discrimination, as opposed to taste-based discrimination.

[39] Note, other research has found judge race and political partisanship to not influence judicial decision making (see, e.g. Lim et al. (2016)) or do so in more nuanced ways (see, e.g. **?**), and judge gender has often been shown to not impact sentencing outcomes (see, e.g. Gruhl et al. (1981)).

information in their processes.

2.2.2 Pennsylvania Criminal Sentencing

In most settings, including Pennsylvania, there are official limits to which judges must adhere during sentencing. The guidelines are extensive and have two crucial benchmarks: maximum and minimum limits to the punishment.[40] These limits provide a range of potential outcomes, and judges are required to use their discretion on a case-by-case basis to determine the specific sentence for each defendant. In addition to legally binding limits, judges are also provided with non-binding sentencing guidelines which are intended to serve as benchmarks in sentencing across defendants.

Trial judges are required to sentence defendants shortly after a defendant has been found guilty of their accused crime.[41] Sentences can include a wide variety of penalties, with fines, probation, and incarceration time being among the most common. If incarceration time is assigned, both a maximum and minimum sentence must be given. Judges have wide discretion in their sentencing behavior, but are encouraged to use set guidelines in their decision-making. Sentencing guidelines are produced by the Pennsylvania Commission on Sentencing, which was established in 1978 with the intent to standardize sentencing practices within the state.[42] Sentencing guidelines are largely based off of the seriousness of the offense (quantified as an "Offense Gravity Score") and the defendant's criminal history (quantified as a "Prior Record Score").

[40] For example, in Pennsylvania, a defendant found guilty of identity theft can not be sentenced to more than 10 years in prison.

[41] The average defendant is sentenced 335 days after their date of offense, with fewer than 1% of defendants sentenced within a month of their date of offense.

[42] There are mixed results as to whether the guidelines have succeeded in their goal (Blackwell et al., 2008; Gorton and Boies, 1999), and some believe they may actively increase racial disparities in sentencing (Wykstra, 2018)

In the majority of instances, judges reveal the sentences of multiple defendants on a single day-of-sentencing (DOS). This occurs largely for practical reasons in an attempt to maximize the efficiency of judges' time, as judges may be sentencing cases while at the same time presiding over other cases. Often, the specific cases of defendants sentenced on the same day share no connections; it is simply a matter of chance the defendants are sentenced on the same day. However, as discussed in Section 2.3.1, my empirical analysis does not rely on the random assignment of cases to specific days-of-sentencing (nor the random assignment of cases to judges).

2.3 Empirical Analysis

Judges may determine defendants' sentences co-dependently, both before the days-of-sentencing while judges weigh their options, and on days-of-sentencing themselves when judges announce the sentences to the defendants. While controlling for standard case-, judge-, and defendant-specific attributes, I allow the sentences of other defendants to influence judges' sentencing decisions. Specifically, my model specification allows the mean sentence length of other defendants in defendants' sentencing cohorts to have a direct impact on their sentences. I define this variable as "Mean-Other-Sentence."

The predicted sign of the coefficient on Mean-Other-Sentence in a regression of Sentence on Mean-Other-Sentence is theoretically ambiguous. For instance, if defendants share a sentencing cohort with someone found guilty of murder (which would lead to a large value of Mean-Other-Sentence), judges may reduce the other defendants' sentences because they do not want to award high sentences to everyone. However, it is equally plausible that judges may extend the other defendants' sentences as assigning a high sentence to the defendant found guilty of murder has temporarily made judges "tough-on-crime" for all defendants. Alternatively, if judges sentence defendants independently, the coefficient on Mean-Other-Sentence should be zero.

2.3.1 Model Specification

To account for the endogenous nature of co-dependencies, my empirical specification follows a straight-forward two-stage approach and takes caution from the warnings put forth in Angrist (2014) on how to empirically estimate peer effects. First, I use a variety of observable characteristics to obtain a predicted value of each defendant's sentence using 10-fold cross validation.[43] Equation 2 describes the estimated equation,

$$\text{Sentence}_{ijt} = \delta_j + \delta_y + \beta_1 \text{Crime}_i + \beta_2 \text{DOS}_{jt} + \beta_3 \text{Defendant}_i + \varepsilon_{ijt} \qquad (2)$$

where $Sentence_{ijt}$ represents the awarded maximum sentence for defendant i, sentenced by judge j on day t (in year y).[44] I control for the average sentence of each judge and year with δ_j and δ_y, judge fixed effects and year fixed effects. $Crime_i$ includes indicator variables for crime type and the grade of the crime, and a continuous measure of the maximum and minimum legally allowed sentence.[45] DOS_{jt} includes a day-of-week fixed effect (non-judge specific) and the number of black and white defendants sentenced by judge_j on DOS_t.

[43] With the cross validation, I estimate the model using 90% of the observations (randomly sampled at the sentencing cohort level) and use the estimated coefficients to predict the sentence lengths for the 10% out-of-sample observations. I then repeat this ten times to obtain predicted sentences for the entirety of my sample. This approach helps avoid over-fitting by disallowing defendants' sentences to impact their predicted counterpart through their influence on the estimated coefficients. Note, results are qualitatively similar when I do not use cross validation.

[44] As reported in Figure A1 in the Appendix, results of the analysis are qualitatively similar when the awarded minimum sentence is used as the dependent variable.

[45] There are 62 different crime types in the analysis and are narrowly defined. For example, there are six crime types for assault and four for theft. Grade of crime represents severity and can vary within crime type.

$Defendant_i$ includes the defendant's drug-dependency status, Offense Gravity Score, Prior Record Score, age, gender, recommended maximum and minimum sentence, and type of disposition (type of plea, trial, etc). To more robustly control for differences in sentencing outcomes based on defendant race, all control variables, excluding the judge fixed effect, are allowed to interact separately by defendant race.[46] The results of the empirical application are strongly robust to the specific set of controls utilized to generate the predicted values.

For each defendant, I then find the average predicted sentence of *other* defendants in the defendant's sentencing cohort and define it as Mean-Other-Sentence$_{ijt}$. To be clear, Mean-Other-Sentence$_{ijt}$ is not obtained by using defendant$_i$'s attirbutes, but instead from the attributes of the other defendants in defendant$_i$'s sentencing cohort. In the second stage of my empirical procedure, I use defendant$_i$'s estimated value of Mean-Other-Sentence$_{ijt}$ as an explanatory variable in my primary estimated equation, Equation 3:

$$\begin{aligned} \text{Sentence}_{ijt} = \delta_j + \delta_y + \beta_1 \text{Crime}_i + \beta_2 \text{DOS}_{jt} + \beta_3 \text{Defendant}_i + \\ \theta \text{Mean-Other-Sentence}_{ijt} + \varepsilon_{ijt}. \end{aligned} \tag{3}$$

By including Mean-Other-Sentence$_{ijt}$ as my measure of co-dependency across sentences, I am allowing the average predicted sentence of Defendant$_i$'s peers (who were sentenced by the same judge on the same day) to impact Defendant$_i$'s sentence. Crucially, *all* controls which are included in the first stage (Equation 2) to obtain predicted sentences are included in the second

[46] In order to gain statistical efficiency, I do not allow judge fixed effects to vary by defendant race. However, the results are qualitatively similar when I allow the fixed effects to vary in such way.

stage (Equation 3), with Mean-Other-Sentence$_{ijt}$ as the sole addition. The two-stage approach helps account for endogeneity concerns in the estimation of Mean-Other-Sentence$_{ijt}$ on Sentence$_{ijt}$, as defendants' predicted sentences are obtained using only observable attributes of the judge, case, and defendant, with any co-dependencies or shared unobservables between sentences not influencing predicted sentence, and thus not being captured in the prediction of Mean-Other-Sentence$_{ijt}$. Recall, a defendant's own sentencing outcome is not included in the calculation of Mean-Other-Sentence$_{ijt}$ (as the variable specifically captures the average sentence of *other* defendants in a sentencing cohort), so each defendant has a unique value of Mean-Other-Sentence$_{ijt}$. In heterogeneity analyses, I allow cohort-race-specific versions of Mean-Other-Sentence$_{ijt}$ to impact sentencing outcomes. Due to the cross validation and two-stage approach, I estimate bootstrapped standard errors with 1,000 repetitions, allowing for clustering by judge.[47]

In order to attribute a causal relationship between Mean-Other-Sentence$_{ijt}$ and Sentence$_{ijt}$, one must make an assumption regarding the exogeneity of other defendants' characteristics on defendants' sentences. As Mean-Other-Sentence$_{ijt}$ is calculated from first-stage predicted values of other defendants, which are estimated using the observables of other defendants, for the estimate of Mean-Other-Sentence$_{ijt}$ to act as a valid instrument for the true average of other defendants' sentences, the observables of other defendants in a sentencing cohort must be exogenous to a defendant's sentence. This assumption is common throughout the spatial econometric literature and implies any possible impact from other defendants' observables onto a defendant's sentence manifest themselves only through the other defendants' sentences (see, e.g. Kelejian and Prucha (1998)). For example, if

[47] Results are qualitatively similar when I run the analysis using a random forest to obtain the predicted value of Mean-Other-Sentence$_{ijt}$.

the ages of *other* defendants in defendant$_i$'s sentencing cohort impact defendant$_i$'s sentence, they do so only indirectly through the average sentence of those other defendants, and the other ages themselves do not have a direct impact on defendant$_i$'s sentence.

To better understand what the estimate of θ in Equation 3 captures, and what it does not, consider the following examples. Consider if judges group defendants of a specific observable type onto the same DOS. Not only would this lead to positive correlation between Sentence$_{ijt}$ and the average other sentence length among defendants in a sentencing cohort, it would also lead to positive correlation between Sentence$_{ijt}$ and the predicted value of Mean-Other-Sentence$_{ijt}$. However, as the source of correlation is observable characteristics which are controlled for in the second stage regression, there would be no correlation between Sentence$_{ijt}$ and the predicted Mean-Other-Sentence$_{ijt}$ once control variables are used. Likewise, consider if judges sentence defendants with similar unobservables on the same DOS. Again, there would be positive correlation between Sentence$_{ijt}$ and the average other sentence among defendants in a sentencing cohort, but now there would not be correlation between Sentence$_{ijt}$ and the estimated Mean-Other-Sentence$_{ijt}$, as the unobservables are not captured in the first stage prediction of Sentence$_{ijt}$, and thus not captured in the calculation of Mean-Other-Sentence$_{ijt}$.[48] To allow for causal inference, my empirical specification accounts for these possible on-observables grouping effects with the inclusion of the control variables, and the on-unobservable grouping effects by allowing errors to cluster by judge. In either case, the estimate of θ would not be influenced by the grouping behavior.

[48] See the results of the Monte-Carlo described in Section 2.5.2 for further evidence shared unobservables among defendants in the same sentencing cohort do not influence my estimate of θ.

2.3.2 Data

The data used in the analysis covers adult criminal sentencing decisions in Pennsylvania from 2007 to 2016, covering all crime types.[49] The data is structured at the judicial proceeding offense level, which describes the sentencing for a single offense. As my data begins at the sentencing level, I do not observe information regarding alleged criminal incidents which did not result in a sentence being assigned. As such, any inference statements made throughout the paper are conditional on the defendant already reaching the sentencing stage of the judicial process.[50] There are occasionally multiple judicial proceeding offenses for the same incident, and therefore there can be multiple observations for every defendant-judge-DOS combination. As my analysis examines effects across defendants, as opposed to effects within-defendants, across-offenses, I only include one observation per judicial proceeding. Specifically, I include the observation associated with the judicial proceeding offense categorized as the most serious offense per defendant.[51]

I simplify the analysis by only including black and white defendants (94% of observations) and drop observations with missing values or with rare characteristics which make the estimation of bootstrapped standard errors implausible. Next, I drop all observations in which two or more defendants within the same sentencing cohort shared a common date of offense (6% of remaining observations). This reduces the chance any two defendants seen by a judge on the same day were involved in the same incident, greatly increasing the likelihood defendants seen by the same judge on the same DOS are

[49] The data was provided by the Pennsylvania Commission on Sentencing.

[50] It is easy to imagine other forms of co-dependneices between possible defendants (for example, a police officer may decide whether to arrest someone based on how many other individuals they have arrested that day). Thus, my estimates should be thought of a small part of a potentially much larger story of inter-defendant effects within the judicial process.

[51] The most serious offense per person is determined by the Pennsylvania Commission on Sentencing, prior to sentencing.

unrelated.[52]. Last, I drop observations in which a judge sentences a single defendant on a given DOS, as the observations can not identify any inter-defendant relations (10% of remaining observations). After data trimming, I am left with 621,190 observations. I display summary statistics in Table 3.

Table 2: Summary Statistics

		Mean	Std. Dev	Min	Max
Defendant	Defendants	621,190			
	Black (%)	25			
	Max Sentence	12	28	0	2,333
	Min Sentence	4	13	0	300
Judge	Judges	404			
	Cases	1,538	1,470	101	10,750
	Cases-Year	302	206	101	1,987
	Cases-DOS	6	4	2	63
DOS	Days-of-Sentencing	110,991			
	Only Black (%)	7			
	Only White (%)	35			
	Identical Crime-Type (%)	5			
	Mean Max Sentence	14	20	0	480
	Min Max Sentence	2	10	0	480
	Max Max Sentence	37	51	0	2,333

Notes: The observational unit is the defendant, judge, and day-of-sentencing in the above three groupings of summary statistics. Time is in months.

As my analysis focuses on within-DOS interactions, I particularly highlight summary statistics at the DOS level. Notably, of the 100,000 plus days-of-sentencing, approximately 60% include the sentencing of defendants of various races, only 5% include defendants which all have the same crime-type,

[52] Note, my econometric procedure makes this sample restriction unnecessarily, but it leads to greater standardization of the relationships between defendants on a DOS, easing interpretability of possible inter-defendant effects. Results are unchanged when I do not drop these observations

and the average spread between the largest and smallest sentence awarded on a day is approximately three years (which is over three times the length of the average sentence in my sample). The summary statistics provide support that the defendants sentenced on a particular DOS are different on both observables and outcomes, suggesting enough variation exists in the data for my econometric analysis to identify within-DOS relationships with adequate precision.

2.4 Results

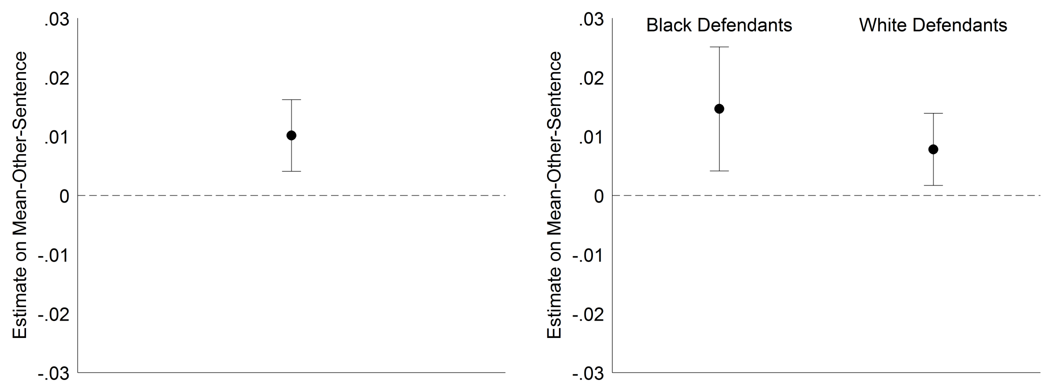
I present results of the analysis in four subsections. First, I estimate baseline model as described in Equation 3 and plot the coefficient on Mean-Other-Sentence. I also examine racial heterogeneity by allowing the coefficient on Mean-Other-Sentence to vary by defendant race. Second, I further examine racial heterogeneity by allowing for racial heterogeneity among the defendant's sentencing cohort. That is, I allow a defendant's sentence to be impacted separately by black and white defendants in the defendant's sentencing cohort. Third, I examine non-racial sources of defendant-type heterogeneity, and fourth, judge-type heterogeneity. Last, I explore possible pathways in which judges may use the sentences of other defendants in their sentencing-decisions.

2.4.1 Baseline Results

I display the baseline results of Equation 3 in Panel A of Figure 5. One-month (one-year) higher average predicted sentences among other defendants in defendants' sentencing cohorts lead to defendants receiving 0.01 month (3.70 day) longer sentences, on average. While a one-year increase in Mean-Other-Sentence is a realistically sized increase given the data, a four-day increase in sentence is relatively small, compared to the average sentence length (the average sentence is approximately 12 months). Despite the coefficient on Mean-Other-Sentence being small in magnitude, the fact that it is statistically significantly different from zero is both economically and morally important; defendants receive variation in their sentence lengths due to the sentences of *other* defendants.

Figure 5: The Role of Sentencing Cohorts

(a) Baseline Results (b) Defendant-Race Heterogeneity



Notes: Panel A displays the coefficient on Mean-Other-Sentence in a regression on a defendant's maximum sentence. The coefficients capture the average change (in months) in a defendant's sentence due to being sentenced alongside defendants with one-month longer average sentences. In Panel B, the coefficient is allowed to vary by defendant race. 95% confidence intervals derived from bootstrapped standard errors, allowed to cluster by judge, are shown.

The positive coefficient on Mean-Other-Sentence does not imply on average, defendants are receiving longer sentences than they would otherwise. The positive coefficient implies the sentences of different defendants are positively co-determined, meaning a longer (shorter) average sentence among defendants in a defendant's sentencing leads to a longer (shorter) sentence for the defendant, on average. While there may be correlation among the unobservable aspects of sentences for defendants in the same sentencing cohort, the coefficient of interest does *not* capture this form of correlation, and instead captures the average effect of the causal chain linking other defendants' sentences to particular defendants' sentences.[53] This relationship expands upon the types of elements that have previously been found to influence sentencing outcomes, from defendant-, crime-, judge-, and time-level, to now include peer-level effects.

It is useful to compare the size of the effect to factors previously found to influence sentence outcomes. Eren and Mocan (2018) find a collegiate football upset loss leads to a 35-day increase in sentence length (on average, across races). Given the estimate of the coefficient on Mean-Other-Sentence, it would take an increase of almost nine years in Mean-Other-Sentence to lead to a 35-day increase in defendants' sentences, which exceeds the 99th percentile of values of Mean-Other-Sentence. However, while upset losses occurred 14 times over the 16 year sample (potentially affecting fewer than 1,000 defendants) in Eren and Mocan (2018), the vast majority of all defendants in Pennsylvania are sentenced on the same DOS as at least one other defendant, suggesting the impact of Mean-Other-Sentence on sentences may be small, but is essentially universally present across all defendants. Moreover, numerous sentences are assigned on the average DOS, a single sentence has the potential to influence multiple other sentences. For a specific example, these estimates suggest that if a defendant found guilty of rape is added to a sentencing cohort with the average number of defendants with an average value of Mean-Other-Sentence, the other defendants sentence would increase approximately two months due to the addition of the defendant found guilty of rape.[54]

Motivated by the various forms of heterogeneity across race found in the literature, I rerun the second-stage equation, now allowing

[53] For example, the role of temperature on sentencing behavior, as found in Heyes and Saberian (2019), would manifest itself as positively correlated errors among defendants with the same DOS, given my specification.

[54] The average sentence of a defendant found guilty of rape is 207 months, the average (across defendants) value of Mean-Other-Sentence is 12 months, and the average (across defendants) number of defendants sentenced on a DOS is 9.

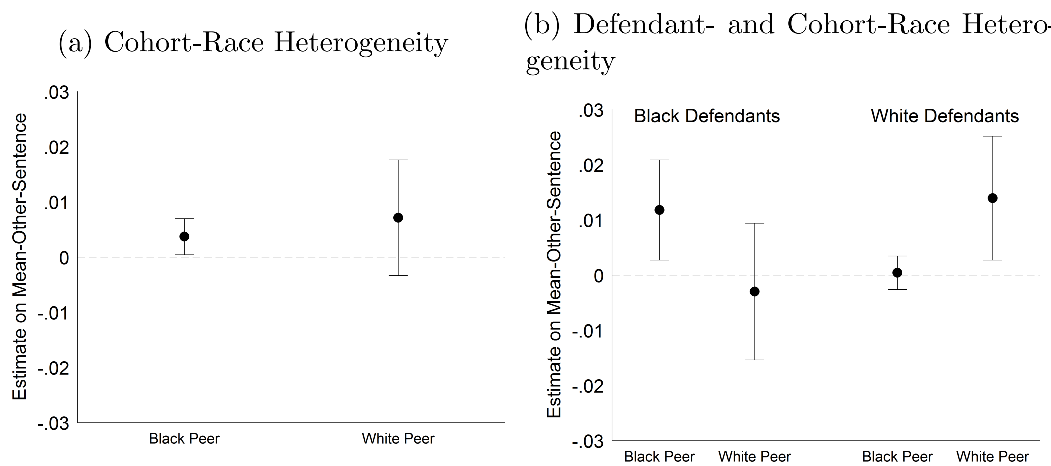
Mean-Other-Sentence to interact with the defendant's race. I display the results in Panel B of Figure 5. Black and white defendants' sentences increase on average 5.35 and 2.84 additional days (not statistically different from each other) due to one year longer average predicted other sentences, respectively. Again, even if the estimated coefficients were statistically different from each other, the fact that the black-specific coefficients on Mean-Other-Sentence is larger than the white-specific coefficients would not necessarily imply sentencing cohort effects penalize black defendants more than white defendants. Instead, it would imply that on average, black defendants are subject to more variation in their sentence as a result of their sentencing cohort.

2.4.2 Peer-Characteristic Heterogeneity

Next, I allow the races of a defendant's peers to alter the effect Mean-Other-Sentence has on the defendant's sentence. I display the coefficients on Mean-Other-Sentence when the peer groups are separated by race in Panel A of Figure 6. Defendants' sentences increase 1.35 and 2.60 days due to one year higher values of Mean-Other-Sentence of black and white peers, respectively. To calculate the effect of an overall (across peers of both races) increase in Mean-Other-Sentence, one must sum both race-specific effects. The coefficients are not statistically different from each other at conventional levels. Admittedly, the interpretation of Figure 6 is limited without allowing the coefficients on Mean-Other-Sentence of black (white) peers to vary by the defendant's race.

Thus, I display the coefficients on Mean-Other-Sentence when both the peer groups and defendants are separated by race in Panel B of Figure 6. Black defendants' sentences increase 4.30 days due to one-year longer average predicted sentence of black peers, and white defendants maximum sentences increase 5.08 days due to one-year longer average predicted sentence of white peers, on average. The results suggest the presence of within-race effects; the coefficient on Mean-Other-Sentence of black peers for black defendants and the coefficient on Mean-Other-Sentence of white peers for white defendants are both statistically significantly different from zero, but not from each other. Moreover, the coefficient on Mean-Other-Sentence of black peers for white defendants and the coefficient on Mean-Other-Sentence of white peers for black defendants are not statistically different from each other, nor zero.[55]

Figure 6: The Role of Sentencing Cohorts by Cohort Race



Notes: Panel A displays the coefficients on cohort-race-specific Mean-Other-Sentence in a regression on a defendant's maximum sentence. The coefficients capture the average change (in months) in a defendant's sentence due to being sentenced alongside defendants of a given race with one-month longer average sentences. In Panel B, the coefficients are allowed to vary by defendant race. 95% confidence intervals derived from bootstrapped standard errors, allowed to cluster by judge, are shown.

Whereas the results of the baseline analysis (Panel A of Figure 5) illustrate that the average sentence of defendants' peers directly affects their sentences, the current results may more precisely highlight the pathways at work. The sentences of defendants' same-race peers directly affect their sentences, but the sentences of their other-race peers do not. The results sit

[55] The across-race coefficients on Mean-Other-Sentence are statistically different from the within-race coefficients at at least the 5% level in all cases.

within a wide breadth of literature of behavioral economics, law, and psychology (see, Jolls (2007) for an overview) that shows judges use a variety of mental heuristics in their sentencing decisions (see, e.g., Guthrie et al. (2001); Choi and Pritchard (2003)) and said heuristics are often connected with race (see, e.g., Taylor et al. (1978); Fiske et al. (1991); Levinson (2007)).

However, it is important to note the cohort-race heterogeneity in the estimated coefficients on Mean-Other-Sentence may be driven by a different, unobserved form of heterogeneity, which happens to correlate with defendants' races. Recall though, my empirical specification controls for many obserables, all but the judge fixed effects which are allowed to vary by race. For example, if there is correlation between defendants' races and the quality of their attorneys, and judges mentally group defendants with similar quality attorneys, I would not capture attorney quality heterogeneity in my estimated coefficients for the defendant, cohort rate-specific estimates of Mean-Other-Sentence, as I allow my plea-deal indicator variable to interact with defendant race.[56] Nonetheless, I interpret the racial heterogeneity as descriptive and do not claim a race-specific causal mechanism relates the sentence lengths of defendants of the same race, but instead simply that there is something which relates the sentences of defendants of the same race, but not those of other races.

2.4.3 Defendant-Type Heterogeneity

While defendant and cohort race is a salient form of heterogeneity, countless other forms exist. In each of this set of three heterogeneity analyses, labeled A through C, I divide all defendants into one of three mutually

[56] Moreover, Berdejò (2018) finds evidence that said correlation exists; there are significant racial disparities in the quality of plea deals reached through bargaining.

exclusive groups based on various characteristics.[57] I then allow the coefficient on Mean-Other-Sentence to vary by group.

Defining Defendant-Type

The first two heterogeneity analyses cut the data in ways that roughly capture the expected length of incarceration time, with defendants of type 1 expected (based on observables) to receive the shortest sentence and defendants of type 3 expected to receive the longest sentence. In Analysis A, I divide defendants by their maximum legally allowed sentences; and in Analysis B, by their offense gravity score. In Analysis C, I divide the defendants into three groups based on whether they were sentenced with a plea deal: defendants of type 1 did not have a standard plea deal, defendants of type 2 had a plea deal that was not negotiated, and defendants of type 3 had a plea deal that was negotiated. Column (A) of Figure 7 displays histograms of the heterogeneity examined in each analysis, with defendants of type 1 having the lightest background, defendants of type 2 having the medium background, and defendants of type 3 having the darkest background.

Defendant-Type Results

I present the results of the defendant-type heterogeneity analyses in Figure 7. The results are consistent across heterogeneity analyses A and B, with the estimated coefficient on Mean-Other-Sentence increasing by defendant type. In these two divisions of defendant type, the estimated coefficients on Mean-Other-Sentence for defendants who are the most likely to receive low incarceration times are negative, suggesting said defendants' sentences are inversely impacted by those of their peers; being sentenced alongside

[57] For the sake of statistical power, I limit myself to defining three defendant types for each heterogeneity analysis.

defendants with longer sentences significantly decreases sentence length for defendants who are expected (based on statutory maximum sentence or offense gravity score) to receive a short sentence. However, the estimated coefficients on Mean-Other-Sentence for defendants who are the most likely to receive high incarceration times are large and statistically different from zero. Lastly, when defendant type is measured by the defendant's plea deal status, the estimated coefficients are similar in magnitude and do not statistically vary.

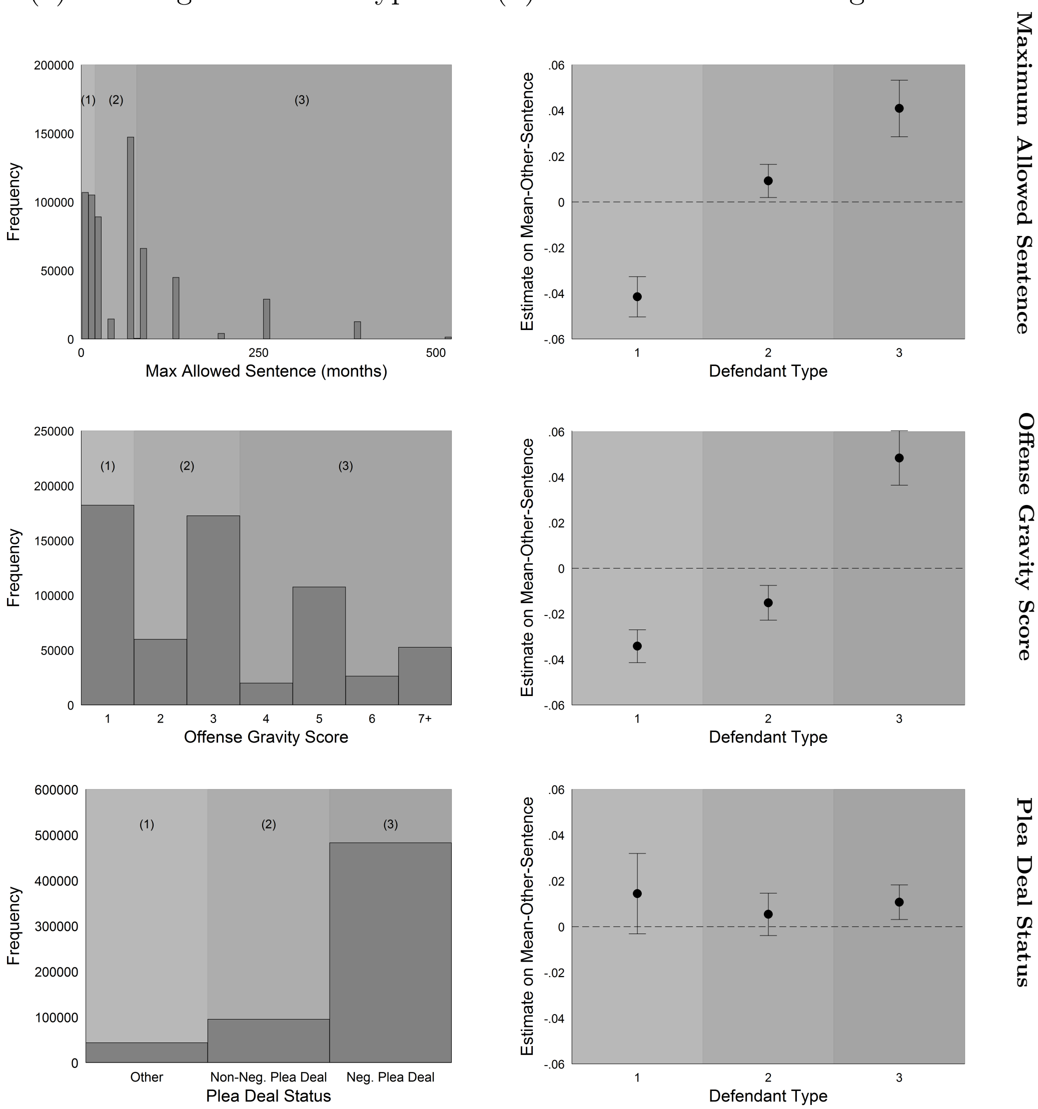
The results suggest the positive and statistically significant coefficient on Mean-Other-Sentence in the baseline regression is primarily being driven by defendants who are expected, based on observables, to receive long incarceration time. The result have suggestive policy insights. If judges were to sentence defendants who are expected to receive similar-length sentences on the same DOS, the average sentence length for defendants of 2 would likely be similar. However for defendants of type 1, as removing defendants of type 2 and 3 from sentencing cohorts would decrease the value of Mean-Other-Sentence and the coefficient on Mean-Other-Sentence is negative, sentence lengths would increase, on average. Similarly for defendants of type 3, as the value of Mean-Other-Sentence would likely increase with the omission of defendants of type 1 and 2 from sentencing cohorts and the coefficient on Mean-Other-Sentence is positive, sentence lengths would increase, on average. These patterns suggest there is not a clear policy change, which if implemented, would predictably either increase of decrease average sentence length.

2.4.4 Judge-Type Heterogeneity

Due to the set of controls used in the estimation of Equation 3, the variation that allows the coefficient on Mean-Other-Sentence to be estimated originates *within* judge. However, the impact of sentencing cohort effects need

Figure 7: Differential Effects by Defendant Type

(a) Defining Defendant-Type (b) The Role of Sentencing Cohorts



Notes: Each row in Column (A) shows the division of defendants into defendant types. The background color represents the defendant type, with defendant type increasing in background color darkness. Each row in Column (B) shows the estimated coefficient on Mean-Other-Sentence when allowed to vary by defendant type. 95% confidence intervals derived from bootstrapped standard errors, allowed to cluster by judge, are shown. Defendant type is defined by the legally allowed maximum sentence (top), Offense Gravity Score (middle), and plea deal status (bottom).

not be equal across judges. I do not have data on the demographic characteristics of judges, but I can estimate an effect for various judge types. Following the previous defendant-type heterogeneity analysis, I separate judges into one of three types along two dimensions.

Defining Judge-Type

To separate judges into groups in the first judge-type heterogeneity analysis, I first estimate the first stage regression as previously discussed (without including Mean-Other-$\widehat{\text{Sentence}}$ as an explanatory variable) and store the residuals, ε_{ijt}. I then calculate the within judge average value of the absolute value of ε_{ijt} to obtain a measurement of how well the model fits the data for different judges. Based on the judge average values, I divide judges into three types indicating relatively how much of the judge's variation in sentencing is controlled for using the explanatory variables.

In my second judge-type heterogeneity analysis, I store the residuals, ε_{ijt} obtained after the first stage regression, renaming them $Error_{it}$. For each judge independently, I then regress $Error_{it}$ on the set of judge-DOS fixed effects, δ_t. Equation 4 describes the estimation:

$$Error_{it} = \delta_t + \varepsilon_{it} \tag{4}$$

where all subscripts are as previously noted.[58] As $Error_{it}$ is the summation of numerous unobserved effects, some defendant-specific and some DOS-specific, solely using δ_t, to capture $Error_{it}$ sheds light on what proportion of $Error_{it}$ can be explained by day-of-sentencing effects. If judges exhibit no DOS effects in their sentencing behavior, δ_t would be mean zero and the adjusted R^2 for

[58] Recall, my sample selection procedure omitted all observations in which a judge saw a single defendant on a DOS, so δ_t does not mechanically perfectly predict any $Error_{it}$.

the regressions would be approximately zero. On the other hand, if all unobserved effects stem from DOS-specific, and not defendant-specific, variation, the set of δ_t would perfectly absorb all variation in $Error_{it}$ and the adjusted R^2 would be one. In reality, $Error_{it}$ is made up of both defendant-specific and DOS-specific variation, but of varying proportions across judges.

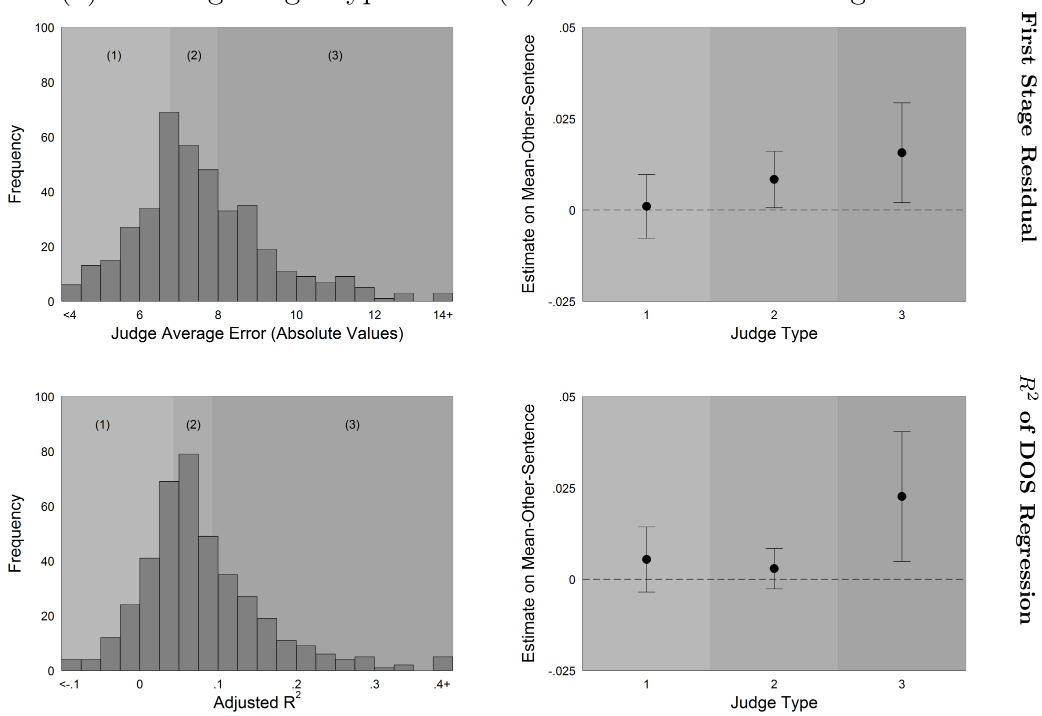
Column (A) of Figure 8 displays two histograms, which I use to define judge type. As shown in the second row, day-of-sentencing fixed effects do contribute to $Error_{it}$, but as the mean (median) adjusted R^2 is only 0.08 (0.07), there clearly is defendant-specific variation as well. The dispersion of the adjusted R^2's highlights an important source of heterogeneity across judges; the proportion of the average $Error_{it}$ for a judge that is due to DOS-level variation varies by judge. Given the histograms, I define three judge types of approximately equal size based on: judges with little to no DOS components of $Error_{it}$ (displayed with the lightest background), judges with moderate DOS components of $Error_{it}$ (displayed with the medium background), and judges with the most DOS components of $Error_{it}$ (displayed with the darkest background). I define judge type in an analogous manner for the other type of heterogeneity. The model as described in Section 2.3.1 is again ran for each measure of heterogeneity, now allowing the effect of Mean-Other-Sentence to vary by judge type. I present results in Column (B) of Figure 8.

Judge-Type Results

When judge heterogeneity is measured by the average magnitude of the first stage errors (first row) and how accurately judge-DOS fixed effects predict first stage errors (second row), the estimated coefficients on

Figure 8: Differential Effects by Judge Type

(a) Defining Judge-Type (b) The Role of Sentencing Cohorts



Notes: Each row in Column (A) shows the division of judges into judge types. The background color represents the judge type, with judge type increasing in background color darkness. Each row in Column (B) shows the estimated coefficient on Mean-Other-Sentence when allowed to vary by the corresponding judge type. 95% confidence intervals derived from bootstrapped standard errors, allowed to cluster by judge, are shown. Judge type is defined by the magnitude of the within judge average first stage residual (top) and the explanatory power of judge-DOS fixed effects in an estimation of first stage residuals (bottom).

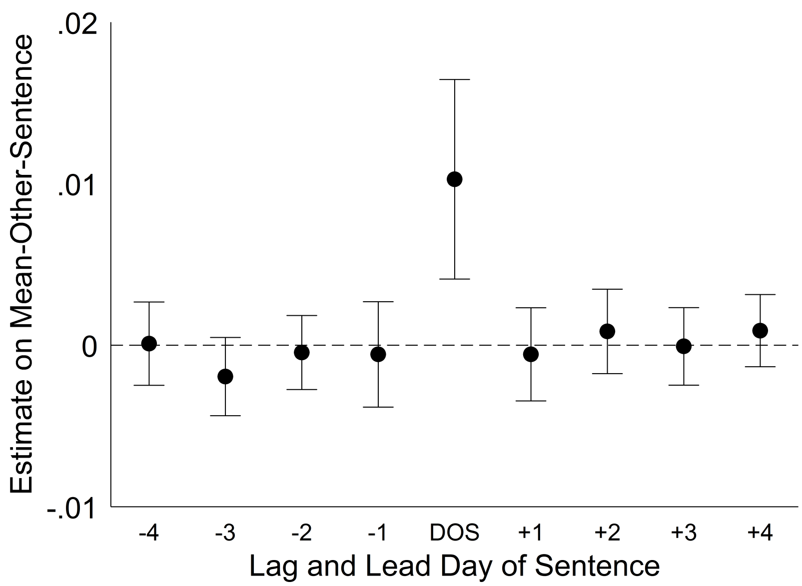
Mean-Other-Sentence vary by judge type. When divided by the magnitude of the average error term, it is not surprising the coefficient on Mean-Other-Sentence increase by judge type, as the amount of variation left to be captured by Mean-Other-Sentence mechanistically increases by judge type. However, said relationship is not present when judges are divided by the predictive power of judge-DOS fixed effects in explaining first stage errors, as they are divided by values of adjusted R^2, which measure explanatory power in percent terms. Nonetheless, judges whose unexplained elements of sentencing decisions are best explained by DOS fixed effects also tend to exhibit the strongest inter-defendant tendencies in their rulings, while judges whose first stage errors are less predictable given DOS fixed effects exhibit a considerably weaker, if any at all, level of inter-defendant effects.

The results are consistent with the notion that the judges who are most heavily influenced by aggregate day of sentence effects are also the judges who are most heavily influenced by inter-defendant effects. However, as aggregate day of sentencing effects include the average co-dependency given my specification, the two measures are not easily untangled. Nonetheless, the finding suggests the same types of defendants may be more prone to receive variation in their awarded sentence length along a variety of dimensions, meaning the total amount of non-defendant-specific variation in defendants' sentenced are unlikely to even out in sum, and in fact may be greater that first appears for some defendants when solely considering a single source of non-defendant-specific variation.

2.4.5 Further Exploration of Co-Dependencies

In addition to judges using information about other defendants sentenced on a particular DOS when assigning sentences, they may also use information about defendants sentenced on a *different* DOS. To examine this possibility, I rerun the empirical specification, but now include four leads and lags of Mean-Other-Sentence as explanatory variables. Each captures the average sentence length of defendants sentenced on a DOS l days away from DOS_t, where l is the degree of lead or lag. For example, the first lag of Mean-Other-Sentence allows for the average sentence of defendants sentenced on the DOS immediately prior to defendants' days-of-sentencing to influence their sentences.[59] I display the estimated coefficients in Figure 9.

Figure 9: The Role of Time Varying Sentencing Cohorts



Notes: The figure displays the coefficients on Mean-Other-Sentence and four leads and lags in a regression on a defendant's sentence. For example, the second lag captures the average sentence of defendants sentenced by a judge two sentencing days prior to the sentencing day of the defendant. 95% confidence intervals derived from bootstrapped standard errors, allowed to cluster by judge, are shown.

As illustrated, the average sentences awarded on different days-of-sentencing (same judge) do not appear to impact defendants'

[59] On average, there is approximately one week between days-of-sentencing for a given judge.

sentences, on average. The result is consistent with judges being able to mentally separate cases across days-of-sentencing, potentially eliminating co-dependencies between sentences. Moreover, this strongly implies the coefficient of interest in the baseline specification captures an effect which specifically connects multiple defendants sentenced on the same DOS, and not simply defendants sentenced by the same judge, or even by the same judge around the same time. Moving forward, my co-dependency exploration focuses estimating patterns of judicial behavior within days-of-sentencing.

As I do not observe the within-day ordering of defendants, I am only able to capture the *average* inter-defendant effect between sentences of all defendants sentenced on the same DOS. This is opposed to, for example, measuring the impact of the specific sentence which was awarded immediately prior to a defendant's sentencing. Moreover, my baseline specification does not flexibly allow for the number of "others" who contribute to Mean-Other-Sentence to differently influence the impact Mean-Other-Sentence has on sentencing outcomes. However, this restriction can be modified, and exploring heterogeneity in the estimates on Mean-Other-Sentence based on the number of defendants sentenced on the DOS can give support for, or against, various possible pathways.

Single Dependency

One potential type of dependency which would result in a positive coefficient on Mean-Other-Sentence is that of judges being impacted by a single other sentence they award on a DOS while awarding other sentences throughout that day. This specific influential sentence could be "chosen" in a variety of ways; for example, it may be first sentence awarded on a DOS which influences all subsequent sentences. If this type of dependency is driving my baseline result, there would be no co-dependencies between the majority of defendant-to-defendant pairs. Thus, as the number of defendants sentenced on a DOS increases, the coefficient on Mean-Other-Sentence should mechanistically decrease and converge to zero, as on average, the number of pairs of defendants with co-dependencies would increase and the single source of co-dependency would play a smaller role in the entirely of the defendant-to-defendant relationships.

Prior Cumulative Dependency

Another plausible type of dependency which would result in a positive coefficient on Mean-Other-Sentence is that of judges being impacted by every sentence they previously awarded on a DOS. In this scenario, the within-DOS ordering of cases is influential in the determination of a particular defendant's sentence, as sentences awarded later in the day are impacted by more other sentences. This type of dependency would lead to variation in the coefficients on Mean-Other-Sentence across the number of sentences awarded on a DOS, as now indirect defendant-to-defendant effects could manifest.[60] As the number of sentences assigned on a DOS increases, the number of indirect effects also increases, and thus the average defendant-to-defendant effect would grow in magnitude, leading to a larger coefficient on Mean-Other-Sentence. However, as long as the direct defendant-to-defendant effect is smaller than one, the series of coefficients on Mean-Other-Sentence would eventually converge to a positive constant, as the average additive importance of new indirect effects formed by an additional sentence being awarded on a DOS would converge to zero as the number of links between the defendants increased.

[60] For example, the first sentence would directly influence the second and third sentence, but also indirectly effect the third through its effect on the second sentence (which then also has a direct influence on the third sentence).

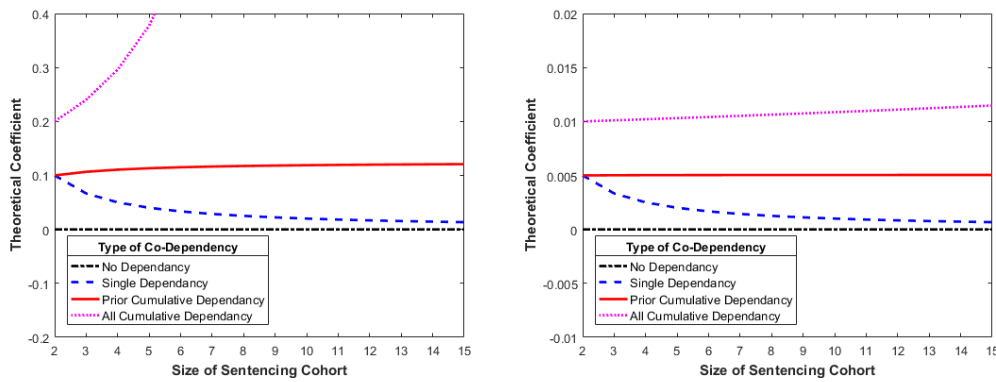
All Cumulative Dependency

Last, it is possible judges are not only influenced by every sentence they previously awarded on a DOS (Prior Cumulative Dependency), but also all future sentences they award that day. This could occur if judges determine all sentences simultaneously, prior to the actual DOS, and simply use the DOS to report the sentences to the defendants. Similar to Prior Cumulative Dependency, as the number of sentences awarded increases, so does the number of indirect effects between defendants. While each additional sentence is impacted by previously sentenced cases, it also directly and indirectly impacts all previously assigned sentences. This type of dependency would result in a strikingly different pattern in the coefficients on Mean-Other-Sentence, as the estimated average effect would grow exponentially as sentences are added onto a DOS, as opposed to eventually converging.

Each of the three proposed pathways would lead to a different pattern across the estimated coefficients on Mean-Other-Sentence when allowed to vary by the number of defendants sentenced on a DOS. I show the theorized estimated coefficients on Mean-Other-Sentence across varying cohort sizes for each of the three pathways in Figure 10. Additionally, I display the theorized coefficients on Mean-Other-Sentence when sentences are awarded independently. In Panel A, the direct sentence- to -sentence effect is set to equal 0.2 (intuitively meaning, if the average other sentence increases by 1 month, a defendant's sentence would increase by 0.2 months, ignoring any spill-over effects), while in Panel B it is set to equal 0.01. As illustrated in Panel B, which is generated using a direct effect size which more closely matches the estimated coefficient on Mean-Other-Sentence in the baseline specification, the patterns derived from Prior Cumulative Dependency and All Cumulative Dependency are similar for the range of values in question, but distinctly different from that derived from Single Dependency. As such, it may be statistically difficult to distinguish between Prior and All Cumulative Dependency when examining the empirical results.

Figure 10: Theoretical Trends of Mean-Other-Sentence Across Size of Sentencing Cohort

(a) When Direct Effect Equals 0.2 (b) When Direct Effect Equals 0.01

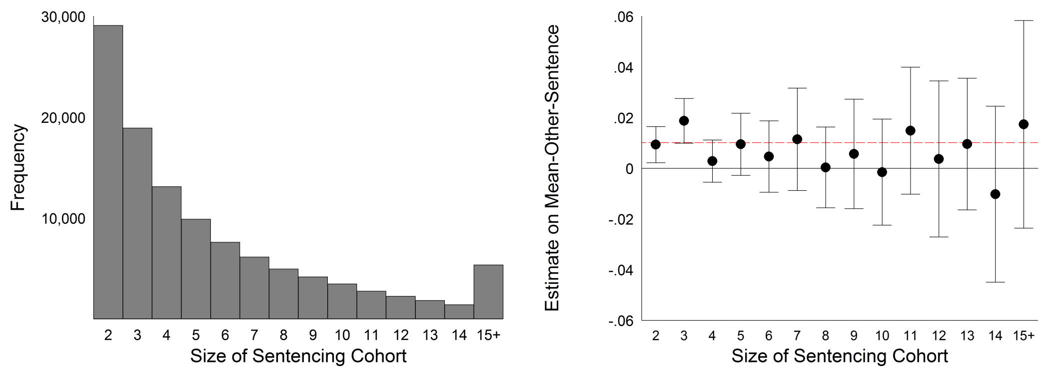


Notes: Panels A and B display three patterns of theorized coefficients on Mean-Other-Sentence when allowed to vary by sentencing cohort size. The three dependencies examined are when effects occur between each defendant in a sentencing cohort (All Cumulative Dependency: Dotted Black Line), effects occur between defendants in a chronological order (Prior Cumulative Dependency: Solid Red Line), and an effect occurs between a single sentence and another sentence (Single Dependency: Dashed Blue Line). A reference line displaying the pattern of coefficients when no co-dependencies are present is also shown (dashed green line). The direct sentence-to-sentence effect is assumed to be 0.2 in Panel A and 0.01 in Panel B.

Next, I rerun the baseline regression, allowing the coefficient on Mean-Other-Sentence to vary by the number of sentences awarded on the DOS. I display a histogram of the number of cases sentenced per DOS in Panel A and the estimated coefficients in Panel B of Figure 11. The red dashed reference line in Panel B illustrates the point estimate on Mean-Other-Sentence from the baseline specification, when the coefficient is not allowed to vary by the number of sentences awarded on the DOS. While some heterogeneity exists among the coefficients, none of the coefficients are statistically different from each other, nor the coefficient estimated in the baseline model. Even disregarding lack of statistical precision, there does not appear to be a meaningful trend in the estimates.

Figure 11: Empirical Examination of the Role of Sentencing Cohort Size on Estimates of Mean-Other-Sentence

(a) Number of Cases per Judge-Day (b) Estimates of Mean-Other-Sentence



Notes: Panel A displays a histogram of the number of sentences awarded by each judge on a DOS. Panel B shows the coefficients of Mean-Other-Sentence when allowed to vary by the number of sentences awarded on the DOS. 95% confidence intervals derived from bootstrapped standard errors, allowed to cluster by judge, are shown. The point estimates show the effect (in months) of a one-month increase in the average sentence of other defendants in the defendant's sentencing cohort on a defendant's sentence. A reference line displaying the estimate of Mean-Other-Estimate when not allowed to vary is also displayed (dashed red line).

The pattern of the estimates strongly suggest judges do not operate under a routine of Single Dependency, as the estimates do not decline as the number of defendants in a sentencing cohort increases. The results are consistent with the notions of judges operating with a Prior Cumulative Dependency or an All Cumulative Dependency, and as expected, I do not have enough statistical power to meaningfully distinguish between the two. Overall, the results provide further evidence that co-determination of sentences across defendants are present in a wide range of sentencing scenarios.

2.5 Robustness Exercises

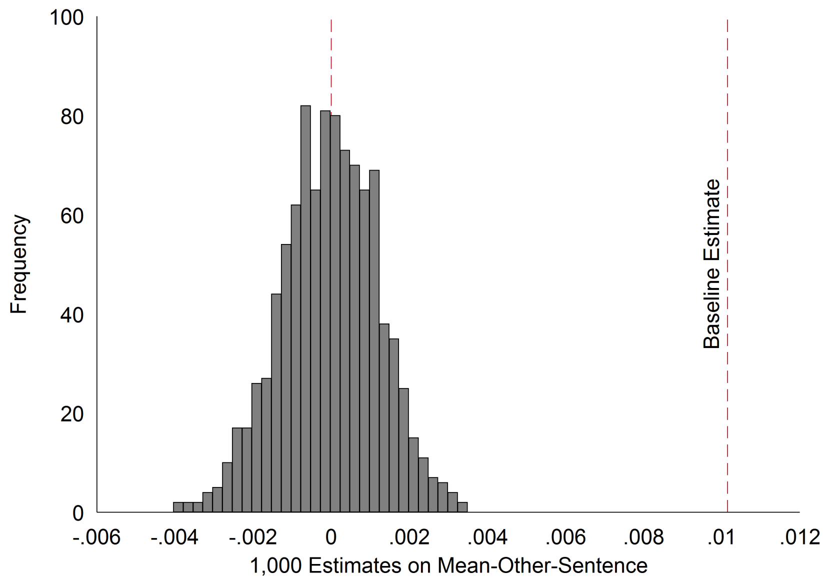
I undertake two robustness exercises to further support the validity of the results and provide evidence the non-zero coefficient on Mean-Other-Sentence is not derived from my econometric procedure. As Mean-Other-Sentence is calculated using the same data used in the second stage regression, it is possible values of Sentence and Mean-Other-Sentence are correlated within judge, and not specifically judge-DOS. Furthermore, if said correlation does exist, it is possible including Mean-Other-Sentence as an explanatory variable in the second stage regression may lead to a non-zero coefficient, even if no causal relationship exists.

2.5.1 Using Other Defendant Mean-Other-Sentence

In the first robustness check, I begin by estimating the first stage regression as described in Section 2.3.1, and obtain the estimated values of Mean-Other-Sentence. Before continuing to the second stage regression, I assign each observation a new value of Mean-Other-Sentence, randomly chosen from the set of all estimated values of Mean-Other-Sentence associated with $judge_j$ and DOS_{-t}. That is, each observation is randomly given the predicted value of Mean-Other-Sentence of a *different* defendant sentenced by the same judge (on a different DOS). If the previously estimated coefficient on Mean-Other-Sentence truly reflects sentencing cohort interactions, the estimated coefficients when the incorrect values of Mean-Other-Sentence are used in the second stage should be mean zero.

I display a histogram of said coefficients, generated from 1,000 iterations, in Figure 12. The coefficients have a mean value of zero. Additionally, the range of estimated coefficients is far below that of the estimate on Mean-Other-Sentence in my baseline specification of 0.01 (0.01 is over 8 standard deviations away from the mean of the distribution). The results provide support for the interpretation that the positive coefficient on Mean-Other-Sentence is driven by judge-DOS specific interactions. My next robustness exercise furthers this support, and additionally provides evidence that unobservable effects shared by defendants in a sentencing cohort do not influence the estimation of the coefficient on Mean-Other-Sentence.

Figure 12: Estimates of θ Using Random Mean-Other-Sentence



Notes: The above histograms display 1,000 point estimates on Mean-Other-Sentence. Each simulated point estimate is generated from the estimation of Equation 3, where the values of Mean-Other-Sentence are randomly chosen from the set of all Mean-Other-Sentence's associated with judge$_j$. Thus, the coefficients capture the effect the average other maximum sentence among defendants sentenced by judge$_j$ on DOS$_{-t}$ have on a defendant's maximum sentence length while sentenced by judge$_j$ on DOS_t. For reference, the estimated coefficient on Mean-Other-Sentence in the baseline specification is 0.011.

2.5.2 Monte Carlo Simulation

My second robustness exercise exists in a purely simulated environment; no real-world data is utilized. In short, I randomly create data that exhibits known co-dependencies in sentences, and then I test whether my empirical specification correctly identifies the parameters.

My simulated data set matches the structure of the Pennsylvania data; the observational unit is a defendant, who was sentenced, along with numerous other defendants, by a judge in a year on a day.[61] I randomly generate race (bi-variate), generic explanatory variables (combination of uniformly and normally distributed), a random unobservable term (normally distributed) for each defendant, and a both a judge and judge-DOS fixed effect for each judge (both normally distributed). With the created variables, I then begin to calculate the "Y" variable (simulated sentencing outcome) for each observation. I also force the average other "Y" for a given sentencing cohort to directly impact "Y" itself. Equation 5 displays the true data generating process.

$$Y_{ijt} = \beta X_i + \delta_j + \delta_{jyt} + \theta \overline{Y_{-ijt}} + Random_{ijt} \tag{5}$$

where all subscripts are equivalent as previously noted.[62]

I estimate the model using the two stage approach discussed in Section 2.3.1. Recall, my empirical specification does *not* include judge-DOS fixed effects, while they are part of the true data generating process in the simulated data. The judge-DOS fixed effects can be thought of as capturing, for example, the average mood of the judge, which is a shared component of

[61] In the presented results, each simulation is run on 200,000 observations (80 judges, 5 years per judge, 50 sentencing days per judge-year, 10 individuals per judge-DOS). Results are robust to varying sample parameters.

[62] In order to create Y that in itself depends on a weighted version of Y, I undertake an iterative process. I create an initial version of Y that does not depend on the weighted version of Y (call the vector Y^i), calculate $\overline{Y_{-ijt}}$ for each observation, and use them to create a new version of Y, as described in Equation 5 (call the vector Y^{i+1}). I then compute the difference between associated elements of Y^i and Y^{i+1}. If either the 5th or 95th percentile difference is greater than a specific threshold in absolute value, I repeat the process, now using Y^i as the initial version of Y. Note, this is more selective than solely ensuring the average difference is smaller than a threshold. In the presented simulated results, I use a threshold of 0.0001 (the average element of Y is approximately 20) and the process typically converges after seven iterations.

all sentences awarded by that judge on that day. If the empirical specification is correctly specified and $\hat{\theta}$ captures the causal effect of the average other sentence on a defendant's sentence, the omission of the judge-DOS fixed effects from the estimated equation should not impact the estimation of θ, as the judge-DOS effects should not be captured in the first-stage prediction of $\overline{Y_{-ijt}}$. I run the simulation twice, once with θ set to equal 0, and once with θ set to equal 0.02.

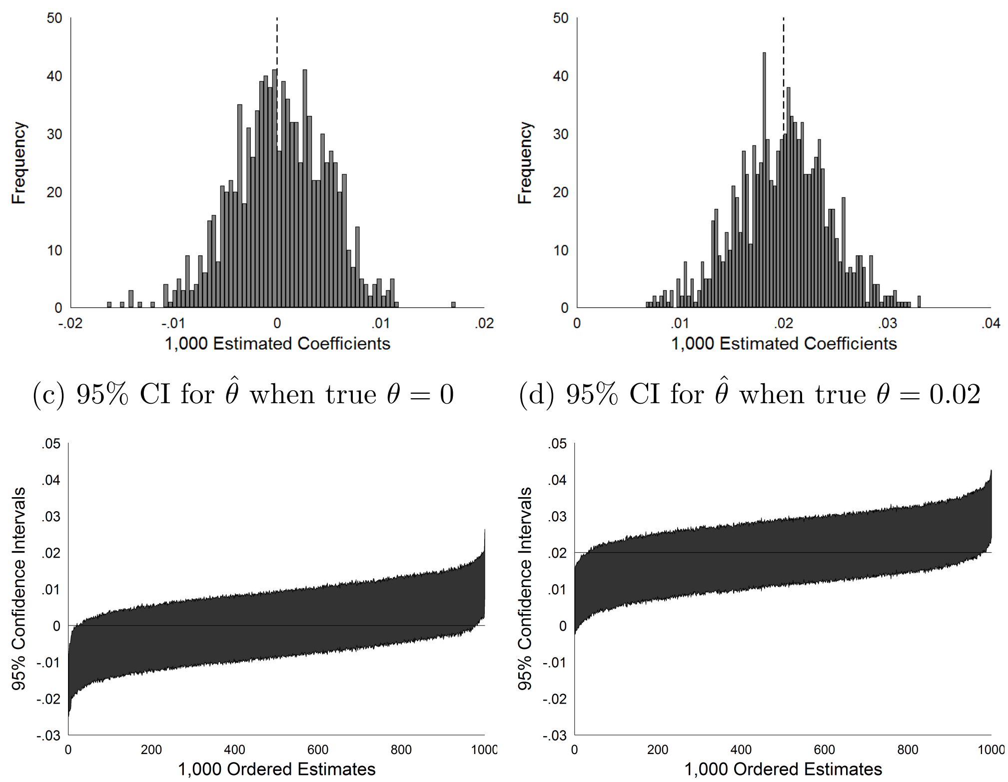
I display the histograms of the estimates of θ in Panels A and B of Figure 13, each compiled from 1,000 simulations. The estimated coefficients are narrowly dispersed, with mean values equal to the true parameters. This provides evidence the prediction procedure is working as expected, and shared unobservables among a sentencing cohort (the judge-DOS fixed effects) do not impact the estimate of the coefficient of interest. Furthermore, I display the 95% confidence intervals for errors allowed to cluster by judge for estimates of θ, which are ordered in size, in Panels C and D. In both cases, the true value of θ is captured within the confidence intervals in approximately 95% of the simulations.

Next, I run the robustness analysis described in Section 12 on the simulated data. That is, instead of using the correct value of $\overline{Y_{-ijt}}$ in the estimation, I use a (within judge, across DOS) random value of $\overline{Y_{-ijt}}$. I again run the simulation twice, once with θ set to 0, and once with θ set to 0.02. I report results in Figure 14, which provides evidence the first robustness analysis behaves as expected.

2.6 Conclusion

Judges hold the incredible power to incarcerate citizens for years on end, and for that power, their decisions should be scrutinized from every angle.

Figure 13: Simulated Estimates of θ



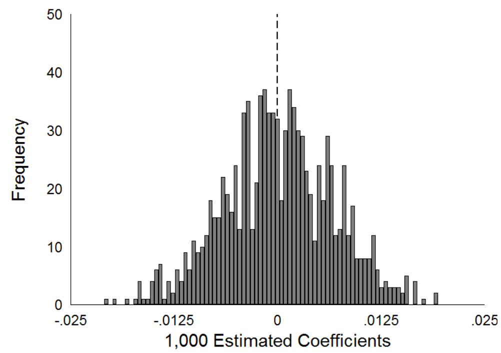
(a) Distribution of $\hat{\theta}$ when true $\theta = 0$ (b) Distribution of $\hat{\theta}$ when true $\theta = 0.02$

(c) 95% CI for $\hat{\theta}$ when true $\theta = 0$ (d) 95% CI for $\hat{\theta}$ when true $\theta = 0.02$

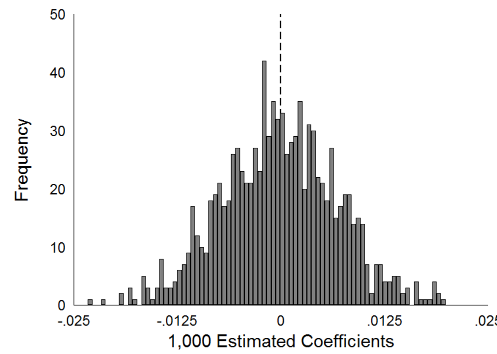
Notes: Panel A displays the estimates of θ from 1,000 simulations, in which the true value of θ is set to 0. Panel B again displays the estimates of θ, now compiled from 1,000 simulations in which the true value of θ is set to 0.02. Panel (C) and (D) display 95% confidence intervals for the ordered estimates in Panel A and (B), respectively.

Figure 14: Robustness within Simulated Environment

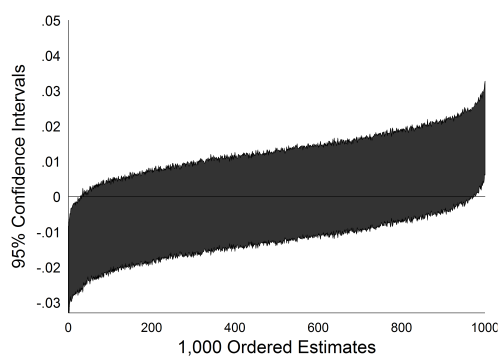
(a) Coefficient on $\hat{\theta}$ when true $\theta = 0$, while using Within-Judge Random $\overline{Y_{-ijt}}$



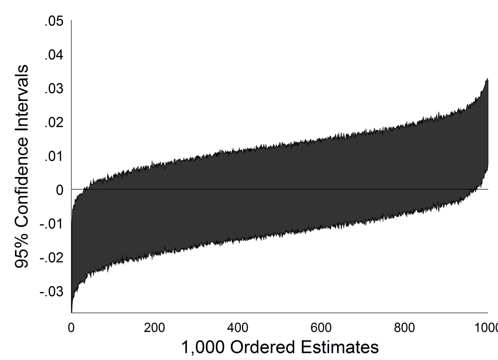
(b) Coefficient on $\hat{\theta}$ when true $\theta = 0.02$, while using Within-Judge Random $\overline{Y_{-ijt}}$



(c) 95% CI for $\hat{\theta}$ when $\theta = 0$, while using Within-Judge Random $\overline{Y_{-ijt}}$



(d) 95% CI for $\hat{\theta}$ when $\theta = 0.02$, while using Within-Judge Random $\overline{Y_{-ijt}}$



Notes: Panel A displays the estimates of θ from 1,000 simulations, in which the true value of θ is set to 0. Panel B again displays the estimates of θ, now compiled from 1,000 simulations in which the true value of θ is set to 0.02. In both Panel A and (b) a within-judge random value of $\widehat{\overline{Y_{-ijt}}}$ is used in replace of the true estimate. Panel (C) and (D) display 95% confidence intervals for the ordered estimates in Panel A and (B), respectively.

However, researchers have so far made a seemingly innocuous, but potentially misleading assumption that judges' decisions for one case do not impact their decisions for others. This assumption deserves to be explored.

In this paper, I examine the sentencing decisions of Pennsylvania judges, allowing for the possibility sentencing decisions for various defendants do not occur independently. Doing so, I find defendants' sentence lengths are influenced by variation in the average sentence length of other defendants sentenced by the same judge on the same day; a one-year higher average of the predicted sentences of other defendants leads to a four-day increase in defendants' sentences, on average. My econometric approach accounts for both potential endogeneity between sentences and judges grouping defendants onto specific days of sentencing by using predicted, as opposed to observed, values of other defendants sentences. Racial heterogeneity analysis further explores the types of dependencies at play and reveals that while the sentences of black (white) defendants are influenced by the average sentence of other black (white) defendants, judges do not appear to take the average sentence of other-race defendants into account when sentencing. Additional heterogeneity analysis reveals differing effects based on both judge, and defendant type (defined along a variety of measures). Lastly, a robustness exercise provides evidence that judges' sentencing decisions are not influenced by the average sentence of defendants they sentenced on other days, and a Monte-Carlo simulation finds results that are consistent with my empirical procedure producing unbiased estimates of the parameter of interest.

The results have far-reaching moral implications and should act as a springboard for further research into the co-determination of sentences across defendants. Future work should investigate additional pathways in which the procedures judges utilize in their sentencing decisions for specific defendants overlap with one another and cause defendants' sentencing outcomes to be co-determined. Research in behavioral economics offer may provide insight into the specific mental-heuristics at play, which will assist in implementing changes to courts' sentencing procedures to reduce the amount defendants' sentences are influenced by random variation, ultimately leading to a more just judicial system.

While the existence of sentencing cohorts provides opportunity for co-dependencies between the sentences of different defendants who happen to be sentenced by the same judge on the same, the fact some defendants are sentenced for multiple offenses at once provides opportunities for *within*-defendant dependencies. In the third chapter of my dissertation, I examine this possibility and consider its implications for the understanding of racial biases in sentencing.

CHAPTER III

MULTIPLE OFFENSES, CONCURRENT SENTENCING, AND RACIAL GAPS IN SENTENCING OUTCOMES

The research described in this chapter was developed by myself and a co-author, Glen Waddell. I have fully participated in every aspect of the process—initial research design, data acquisition, empirical methods, and both the writing and continued revision of research output.

3.1 Introduction

Across the United States, criminal-sentencing guidelines typically inform judges about the offense and offender attributes that should be considered in assigning sentences to those for which guilt has been determined. While states differ in procedure—some use grids or worksheet scoring systems, some employ sophisticated algorithms—the goals of guidelines typically collapse around offenders with similar offenses and criminal histories being treated alike.[63]

That said, judges often find little guidance on the sentencing of defendants who face coincident sentencing decisions. There is also a surprising lack of empirical work on the role of multiplicity in sentencing outcomes. As such, the literature leaves us largely uninformed about the treatment of those facing multiple sentences, either with respect to their sentencing of specific crimes or with respect to their ultimate sentence, arrived at through both

[63] Guidelines vary considerably across states. In Pennsylvania, which is the state from which we draw our data, the 1978 "Pennsylvania Commission on Sentencing" created a statewide sentencing policy that would increase sentencing severity for serious crimes and promote fairer and more-uniform sentencing practices. Judges are required to complete a sentencing judgment form—a grid with 14 offense and 8 prior record levels that maps into suggested sentencing outcomes. The language does not indicate that the guidelines are mandatory, and defense counsel can appeal based on the fact that a judge "departed from the guidelines and imposed an unreasonable sentence."

sentence length and the judge's determination that some or all should be served concurrently. It is not an insignificant number of cases for which this is relevant. For example, across all sentencing hearings from the Common Pleas Courts of Pennsylvania between 2007 and 2016, over 30 percent of defendants (over 50 percent of offenses) enter sentencing hearings facing multiple sentence determinations.

Most empirical analysis of sentencing is performed at the offense level, with analysis typically being performed on the most-severe crime for each defendant (Berdejò, 2018; Aizer and Doyle, 2015; Steffensmeier and Britt, 2002; Mustard, 2001; Gruhl et al., 1981).[64] Thus, there is typically no regard for the reality that defendants often face multiple offenses and therefore multiple and simultaneous sentences. Even by the inclusion of multiple sentences—documenting patterns in the sentencing of criminals with multiple offenses, and the systematic application of concurrent sentencing—our analysis contributes to a large literature on sentencing.

Before accounting for concurrent sentencing, we find mixed evidence of independence in sentencing across multiplicity—while judges are relatively less likely to incarcerate defendants with multiple offenses given their multiple offenses, they impose slightly longer sentences to what one would expect if each of the defendant's offenses were treated as if it were the only offense. That said, it proves important to account for the use of concurrent sentencing, where we find a strikingly different pattern. Compared to what one would expect under an assumption of independent and consecutive sentences, the

[64] Chen (2008) is an exception—in California, an additional charge is associated with a three-percent increase in sentence length. Lovegrove (2004) summarizes interviews ($n = 8$) that suggest that Australian judges consider the overall sentence length in their determination of sentences for individual offenses for those who face multiple sentences—if the null hypothesis is that judges make independent decisions across multiple sentences, we interpret this as a proof of concept. See Ryberg et al. (2017) for an overview of a small theoretical literature that considers issues related to defendants with multiple offenses.

average sentence length is 56 percent as long for those with even a second guilty verdict, reaching only 35 percent as long for those with seven guilty verdicts.

Directly considering the role of multiplicity will also inform our understanding of racial discrepancies in criminal justice. It is well documented that black individuals face more-severe punishments, all else equal—in traffic stops (Horrace and Rohlin, 2016), bail (Arnold et al., 2018), plea bargaining (Berdejò, 2018), jury decisions (Flanagan, 2018), and death sentences (Eberhardt et al., 2006). While the unconditioned gap in black–white sentencing is partially contributed to by differences in the underlying offenses, conditioning on offense and criminal history, the literature supports black–white sentencing gaps on the order of 10-to-15 percent (Rehavi and Starr, 2014; Mustard, 2001).[65]

We also find important and troubling patterns in the effect of multiple sentences on the gap in sentencing outcomes for white and black defendants, with evidence that both multiplicity itself, and the use of concurrent sentencing, further gaps in sentencing outcomes. After accounting for both multiplicity in sentencing and controlling for the role of concurrent sentencing, we find overall racial discrepancies in defendants' total sentences to be 28-percent larger for black defendants with two offenses compared to those experienced by black defendants with a single offense, with the. Conditional on those incarcerated, discrepancies are 346-percent larger for black defendants with two offenses compared to black defendants with a single offense.

In Section 3.2 we describe our data and document the prevalence of multiple-sentence determination. In Section 3.3 we present our empirical

[65] As racial disparities in sentencing are often driven by differences in a defendant's likelihood of being jailed at all (Abrams et al., 2012; Demuth and Steffensmeier, 2004), throughout our analysis we will examine both the intensive and extensive margins of sentencing.

analysis, which we follow with a discussion of the policy implications in Section 3.4.

3.2 Data

To consider the role of multiple-sentencing in judicial outcomes, we use administrative data files from the Pennsylvania Commission on Sentencing, spanning all criminal sentencing in the Common Pleas Courts of Pennsylvania between 2007 and 2016. Selection into the sample is conditioned on guilt having been determined—we have every guilty verdict within the Court system. These guilty verdicts are then sentenced by an individual judge on a given day, regardless of the dates the offenses were committed. Our data do not include arrest information, so we will be careful to limit our inference to the sentence behavior conditional on guilt.[66]

We drop all observations for defendants who receive a minimum or maximum offense sentence that is longer than the 99^{th} percentile of offense sentence length (roughly two percent of the sample), and all observations for defendants with any missing or infeasible values (fewer than one percent of remaining observations).[67] We also discard defendants with more than seven guilty offenses—the 99^{th} percentile in the distribution of defendants in the Pennsylvania data. We provide summary statistics at the offense and defendant levels in Table 3. As shown in Table 3, there are considerable

[66] As judge's sentencing behavior can vary with defendant gender (Butcher et al., 2017), we drop all female defendants from the analysis—they account for twenty percent of guilty-verdict observations in the dataset. We also drop defendants who are not white or black—this amounts to six percent of all defendants, and seven percent of all offenses. Patterns evident in other races are similar to those of black defendants.

[67] Dropping the top-one percent of offense sentence lengths amounts to roughly two percent of observations because we drop all observations for defendants, not just their offense in the 99^{th} percentile, but also because offenses with *minimum*-sentence lengths greater than the 99^{th} percentile are not the very same set of offenses with *maximum*-sentence lengths greater than the 99^{th} percentile. In rare cases, we drop observations with impossible values, such as when *minimum* sentences are longer than *maximum* sentences, again dropping all observations for defendants.

differences in racial discrepancies when using offenses versus defendants as the unit of analysis— for example, while black defendants are four percentage points more likely than white defendants to be incarcerated for any offense (47- vs. 43-percent), they are eight percentage points more likely to be incarcerated for at least one offense (57- vs. 49-percent).

Table 3: Descriptive statistics

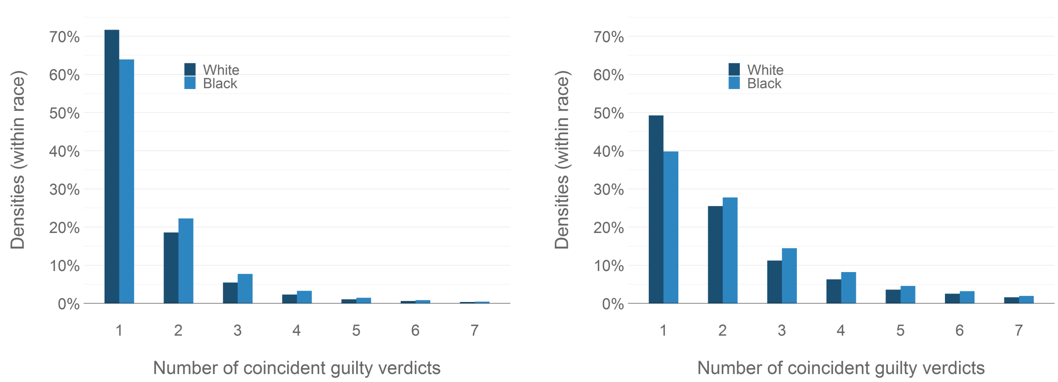
	Offense (n=946,393)		Defendant (n=631,487)	
	Mean	Std Dev	Mean	Std Dev
White	69%		71%	
Number of Offenses	1	0	1.45	0.92
Receive incarceration	43%		49%	
Sentence length				
Unconditional	98	248	143	447
Conditional on jail time	228	336	292	603
Black	31%		29%	
Number of Offenses	1	0	1.61	1.04
Receive incarceration	47%		57%	
Sentence Length				
Unconditional	198	391	318	744
Conditional on jail time	422	482	558	915

Notes: Time is in days.

While the modal defendant faces sentencing on only one guilty verdict, a large fraction of defendants do face multiple, coincident sentencing decisions. In Panel A of Figure 15 we plot the incidence of coincident guilty verdicts, separately for white and black defendants—more than 30 percent of defendants face multiple sentencing decisions, with a non-negligible fraction (two percent) being sentenced for more-than-four offenses. In Panel B of Figure 15 we plot the same data at the offense level—over 50 percent of all guilty verdicts are attributable to defendants who face a coincident sentencing decision, and 8 percent of offenses are attributable to defendants who face more than four coincident sentencing decisions. These frequencies strike us as sizable—especially as the literature has largely ignored the potential implications of multiple sentencing. If over half of all offenses have a defendant in common, the scope for a richer understanding of sentencing outcomes is likewise sizable. Given a higher degree of multiple sentencing among black defendants, ignoring multiplicity in sentencing may also have important implications for our understand of racial disparities in criminal justice.

Figure 15: The frequency of multiple offenses, by defendant race

(a) Defendants (n = 631,487) (b) Offenses (n = 946,393)



Notes: The above histograms show the distribution of the number of coincident guilty verdicts, by defendant race, both in terms of both defendants and offenses.

3.3 Empirics

Below, we consider the relationship between the number of coincident guilty verdicts and the intensive (incarceration) and extensive (sentence length) margins of sentencing. We find evidence that judges use discretion while assigning sentences to be served consecutively or concurrently. We then isolate the racial effects separately from any direct impact the number of coincident guilty verdicts and we find systematic patterns of behavior that reveal a race-based gap in sentencing that is between 25 and 75 percent larger than what is commonly found in the literature.

3.3.1 The likelihood of incarceration

In Panel A of Figure 16 we plot the unadjusted mean probabilities that defendants receive jail time for at-least-one offense, across those who face different numbers of guilty verdicts. Part of this relationship is mechanistic, of course—with more opportunity to receive jail time, we anticipate at-least some jail time with higher probability. However, as we illustrate, the increase in the unadjusted probabilities fall short of what one would expect given the treatment of single-sentence defendants. For example, defendants with only one guilty verdict are incarcerated with a probability of 0.46, and to assume this probability across all offenses significantly overstates the likelihood that defendants with multiple offenses will receive any jail time. That jail time is everywhere less likely than what would result from independence across multiple sentences suggests either a direct role for multiple sentences in determining outcomes, or a selection into multiple crimes that drives outcomes. Though, selection here would be of a kind such that attributes associated with *less*-severe sentencing of defendants with multiple sentences, and selection into crimes that are *less* likely to be jailed. This is not the selection one anticipate, we believe, where "more-criminally inclined" defendants select into more guilty verdicts. The pattern evident in Panel A is consistent, however, with the multiplicity actively attenuating the severity of sentencing decisions.

In Panel B of Figure 16 we plot estimates of the effect of multiplicity on the probability a defendant receives any jail time. To fit the data-generating process flexibly, we consider the *offense*-level model

$$\mathbb{1}\left(\sum_{k=1}^{K_i} \text{Sentence}_{icj} > 0\right) = \alpha + \Omega_{K_i} + \mathbf{X}_{ic}'\boldsymbol{\beta} + \delta_c + \gamma_j + \epsilon_{icj} \ , \tag{6}$$

where $\mathbb{1}\left(\sum_{k=1}^{K_i} \text{Sentence}_{icj} > 0\right)$ captures whether defendant d seen by judge j receives jail time for *any* of the K_i offenses for which d has been found guilty. (By construction, this allows defendants with multiple offenses to have more weight in the estimator, which we adjust for by estimating standard errors allowing for clustering at the defendant-by-judge level.) Across defendants with varying number of guilty verdicts, we measure the differences in the probability of being incarcerated in Ω_K. In $\mathbf{X}_{ic}$ we include race fixed effects, age (in five-year bins), and i's "prior-record score" to absorb any effect of criminal history on sentencing.[68] We also include the statutory minimum-sentence length for crime c, and the offense-gravity score, which varies across i within c.[69] We also control for crime δ_c and judge κ_j fixed effects throughout the analysis.[70]

Reporting the estimated differences in the probability of receiving *any* jail time across defendants who face $K \in \{2, 3, 4, 5, 6, 7\}$ guilty verdicts—derived from the Ω_K fixed effects in Equation (6)—again identifies a role for the number of coincident guilty verdicts in sentencing decisions.

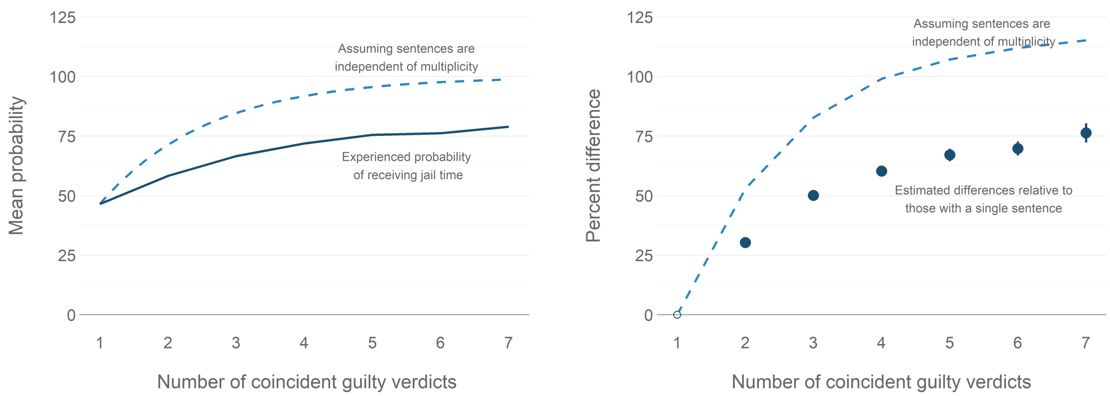
[68] The "prior-record score" is a measure of criminal history for instances that occurred *prior* to the judicial proceeding—contemporaneous offenses do not factor into this score. Moreover, of prior sentences which were assigned concurrently, just that of which was defined as "the most serious offense" in the judicial proceeding is used in the calculation of the prior-record score. 204 Pa. Code § 303.5(b)(1).

[69] The "offense-gravity score" measures the intensity of the offense and varies *within* crime-type so can be estimated even with the inclusion of crime fixed effects.

[70] Note that "crime fixed effects" are quite specific, encompassing 82 distinct crime types, and our specification therefore quite flexible. For example, we include thirteen categories of homicide and eight categories of assault.

Figure 16: How does a defendant's likelihood of being incarcerated change across the number of coincident guilty verdicts?

(a) Unadjusted means across multiplicity (b) Estimated effects of multiplicity



Notes: In Panel A we plot the mean probabilities that a defendant is incarcerated across the number of coincident guilty verdicts faced by defendants. In Panel B we plot the estimated impacts of multiplicity in offenses on the probability of incarceration (i.e., the percent difference relative to defendants with a single offense)—we control for race fixed effects, age (in five-year bins), each defendant's "prior-record score" to absorb any effect of criminal history on sentencing, the statutory minimum-sentence lengths for the crime, the offense-gravity score, and crime- and judge-fixed effects. In both Panel A and B, the dotted line represents the prediction under the assumption of independence across number of guilty verdicts, with differences in observables across the number of coincident guilty verdicts being accounted for in Panel B, but not Panel A.

Despite the many restrictions on the identifying information, the estimated differences associated with multiple guilty verdicts (in Panel B) follow a similar pattern to that of the unadjusted probabilities (in Panel A). In the end, the fitted model reveals a similar departure from the patterns of incarceration implied under independence.

In order to derive that prediction, which we plot as dashed line in panel B, we first fit a model from a sample we restrict to defendants who face only $K_i = 1$, as sentence multiplicity cannot explain variation in incarceration in this restricted sample or the weights give to the set of covariates. We then make the out-of-sample predictions of incarceration for each offense for those defendants who face multiple sentences. Given the predicted likelihood a given offense would result in jail time, we then calculate the probability that a defendant would be jailed for *at-least-one* offense.

Individual sentences are clearly not independent across the number of guilty verdicts. Having absorbed much of the variation in sentencing through our econometric specification, we are inclined to interpret this pattern as evidence of differential treatment rather than to selection into multiple offenses. Moreover, that estimates fall off of the application of single-verdict sentencing patterns to multiple-verdict sentencing we interpret as evidence of sentence severity *decreasing* in the number of coincident guilty verdicts.

3.3.2 Sentence Length

In most jurisdictions, when defendants face multiple sentences it may be determined that they are to be served at the same time (known as concurrent sentences), with the longest period implying the length of sentence to be serve. Around the world, judges are given wide discretion in determining whether a sentence is to be served consecutively or concurrently. This is true of Pennsylvania sentencing, with the relevant guidelines simply noting that “In determining the sentence to be imposed the court ... may impose them consecutively or concurrently,” (Cirillo, 1986).[71] However, it is generally believed that when sentencing defendants with multiple offenses, judges should not consider offenses independently—instead they are to assure that the total punishment matches the offending behavior. That said, the scope for discretion is relatively large, and the implications of the assignment of concurrent sentencing likewise.

While discretion over how to sentence multiple offenses is large, it is bounded by two types of sentencing behavior: that when all offenses served

[71] 42 Pa. Stat. and Cons. Stat. Ann. § 9721 (West).

consecutively or when all offenses served concurrently. Many jurisdictions have coded a preference for one of these boundaries into law as a "presumption" that operates as a default rule. In these jurisdictions, a judge must explicitly override such presumption in order to assign a different type of sentence. While the Federal courts and may state jurisdictions maintain a presumption in favor of concurrent sentences (LaFave et al., 2000), there is no codified presumption under Pennsylvania law, with the relevant rule of criminal procedure stating, "When more than one sentence is imposed at the same time on a defendant ... the judge shall state whether the sentences shall run concurrently or consecutively," (Pa. R. Crim. P. 705(B)).[72]

In Figure **??** we describe the assignment of concurrent sentences across defendants with different numbers of coincident guilty verdicts. Across all defendants convicted of at least two offenses receiving any jail time, roughly 30-to-70 percent receive a concurrent sentence for at least one offense (Panel A) and of those who receive at least one concurrent sentence, 40-to-70 percent of their total sentence length is to be served concurrently (Panel B). As we intend to pursue the modeling of race-specific effects of multiplicity, we've plotted these for black and white defendants separately—this suggests that there are differences in the allocation of concurrency across race, favoring white defendants in some dimensions, while favoring black defendants in others.

In Panel A of Figure 17 we plot unadjusted "consecutive" and "real" sentence lengths by the number of coincident guilty verdicts—real-sentence lengths are 80 percent of consecutive-sentence lengths. Of course, for those facing single sentences, there is no like notion of concurrent sentencing. However, we retain these defendants in fitting models of sentence lengths.[73]

[72] Prior to 1996, Pennsylvania did maintain a presumption that certain sentences should run concurrently unless the judge said otherwise (Pa. R. Crim. P. 705(B)).

[73] When we condition on defendants with more than one offense, real-sentence lengths are 64 percent of consecutive-sentence lengths, on average.

For context, we also plot (with dashed lines) the percent difference in predicted sentence lengths under the assumptions that all offenses are to be sentenced 1) independently and 2) consecutively (or concurrently)— that is, under the assumptions all offenses are sentenced under a mandated presumption of consecutive, or concurrent, sentencing. Here, that amounts to modeling sentence lengths for a sample of $K_i = 1$ defendants (consecutive and real sentence lengths are identical for the K_i=1 population), and using those coefficient estimates to project sentence lengths for all $K \geq 2$ defendants to determine the predicted values under the assumption of consecutive sentencing. To generate the predictions under the assumption of concurrent sentencing, we use the estimates generated from the $K_i = 1$ model to make out-of-sample predictions of offense sentence lengths for all offenses for defendants who face multiple guilty verdicts. We then set the predicted actual sentence length for the defendant as the maximum offense sentence length which was estimated for the defendant. We then plot the percent differences between the counterfactual consecutive and concurrent sentences and those found when $K_i = 1$.

Next, to find the direct effect of offense multiplicity on sentence length, we model the length of defendant i's total sentence across all K_i guilty verdicts as

$$\sum_{k=1}^{K_i} \text{Consecutive Sentence}_{icj} = \alpha + \Omega_{K_i} + \mathbf{X}_i'\boldsymbol{\beta} + \gamma_j + \epsilon_{ij} \ , \tag{7}$$

and

$$\sum_{k=1}^{K_i} \text{Real Sentence}_{icj} = \alpha + \Omega_{K_i} + \mathbf{X}_i'\boldsymbol{\beta} + \gamma_j + \epsilon_{ij} \ , \tag{8}$$

where $\sum_{k=1}^{K_i}$ Consecutive Sentence$_{icj}$ and $\sum_{k=1}^{K_i}$ Real Sentence$_{icj}$ captures the consecutive sentence assigned to defendant i seen by Judge j, and the associated sentence length that accounts for the judge's decision to allow some or all of the consecutive sentence to be served concurrently. Of interest to us, again, are the fixed effects that absorb how sentencing decisions change across multiple sentences (i.e., the Ω_{K_i}).[74]

In Panel B of Figure 17 we report the differences in sentence length due to the number of coincident guilty verdicts defendants face. Specifically, we plot the effect of multiplicity relative to the rates experienced by defendants who face single sentencing decisions—for both consecutive and concurrent sentences lengths.

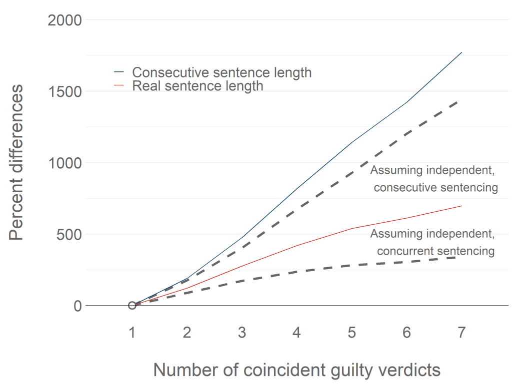
For context, we also again plot (with dashed lines) counterfactuals of the estimated effects under the assumptions that all offenses are to be sentenced 1) independently and 2) consecutively (or concurrently). To generate the counterfactuals effects, we net out the differences in the counterfactual consecutive and concurrent sentence lengths (those used to generate the counterfactuals in Panel A) that can be explained by differences in observables, leaving the remaining differences as the counterfactual effects due to multiplicity.

The results in Panel B closely match those found in Panel A. With respect to the consecutive sentences defendants receive, we see the monotonic increase we anticipate with additional sentences. With that, while we see general similarity in the model's predicted impact across multiple sentences and the derived impact under independence with an assumption that all

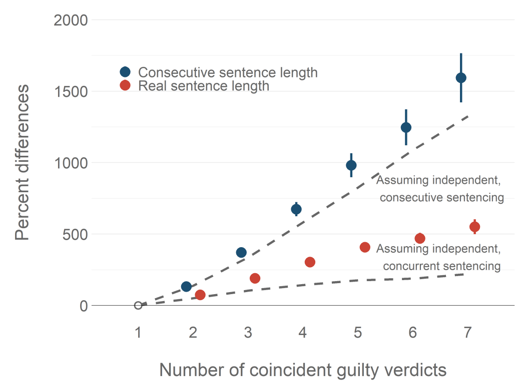
[74] In $\mathbf{X}_{ic}$ we again include race fixed effects, age (in five-year bins), and i's "prior-record score" to absorb any effect of criminal history on sentencing. We also include the average statutory minimum-sentence lengths for defendant i's crimes, his average offense-gravity score, and controls for the fixed contributions specific to crime types. We also control for judge γ_j fixed effects.

Figure 17: How do consecutive and concurrent-adjusted sentence lengths change across the number of coincident guilty verdicts?

(a) Unadjusted means across multiplicity



(b) Estimated effects of multiplicity



Notes: In Panel A we plot the average total consecutive and total real (concurrent-adjusted) sentence length that a defendants receive across the number of coincident guilty verdicts faced by defendants. In Panel B we plot the estimated impact of the number of coincident guilty verdicts on the total consecutive and total real (i.e., the percent difference relative to those facing only one sentencing decision)—we control for race fixed effects, age (in five-year bins), and i's "prior-record score" to absorb any effect of criminal history on sentencing. We also include the average statutory minimum-sentence lengths for defendant i's crimes, his average offense-gravity score, and controls for the fixed contributions specific to crime types. We also control for judge γ_j fixed effects. In Panel A, the dotted lines represents the counterfactual percent differences in sentence length, while in Panel B the counterfactuals show the percent differences in sentence length that is specifically due to multiplicity in sentencing, that one would find under independent consecutive or concurrent sentencing.

sentences are to be served consecutively, the estimated effects routinely run higher, which is consistent with complementarities in sentencing offenses increasing total length.[75]

With respect to *real*-sentence length—accounting for part of all of the multiple sentences to be served concurrently—we see strong evidence of the significance of these determinations. For those with two sentences, the allocation of concurrency amounts to multiplicity leading to a 44-percent shorter (52 days, on average) real sentence length compared consecutive

[75] We fail to reject at the 95% level that a quadratic trend better fits the data than a linear trend.

sentence length—this reduction increases to 65-percent for those defendants with seven offenses (930 days). However, it is also evident that concurrent sentence lengths significantly depart from what one would expect under independent concurrent sentencing, suggesting a lack of any (even unofficial) presumption for concurrent sentencing within the Courts. Across $K_i \in \{2, 3, \ldots 7\}$ the average effects of offense multiplicity on concurrent sentence lengths is between 43- and 58-percent smaller than one find under an assumption of independent consecutive sentencing and between 47- and 151-percent larger than one would find under an assumption of independent concurrent sentencing.

As it is evident judges use discretion while determining whether the sentences for a defendant are to be served concurrently or consecutively, and as the literature has well established judicial discretion can lead to racial gaps and bias (Rehavi and Starr, 2014; Mustard, 2001), it is natural to consider how discretion over how to sentence multiple offenses may influence racial discrepancies.

3.3.3 Racial bias in sentencing

While racial gaps in sentencing are routinely found, it is not always clear what racial gap is being measured, and how it relates to individual offenses or defendants as a whole. For example, in their sample when estimating sentencing outcomes, Mustard (2001) includes all offenses for defendants and implicitly assumes independence between them, and Demuth and Steffensmeier (2004) ignores multiplicity by discarding all but the most-serious of a defendant's offenses. Noting that "sentencing across charges for a given case will be highly correlated," Abrams et al. (2012) also discards all but the most severe charge per defendant.[76] Our analysis of Pennsylvania's Common Pleas Courts rulings between 2007 and 2016 is consistent with this overall expectation, revealing systematic gaps in overall defendant sentencing outcomes by race of roughly thirteen percent. However, we find significant variation in the magnitude of racial bias, both across defendants with varying number of coincident guilty verdicts, and across measurements of consecutive- and real-sentences.

In Figure 18 we plot estimated differences in sentencing outcomes by defendant race—here, we can again relax earlier restrictions on β across $K_i \in \{1, 2, \ldots 7\}$ sentencing decisions and estimate models separately for all K_i. While we implicitly control for the number of sentences the defendant faces, in producing differences by race we absorb any systematic differences in sentencing outcomes that are explained by the types of crime, and the number of each.[77]

First, as has been evidenced in the literature, the treatment of black defendants is consistently to their detriment. We demonstrate this empirical regularity generally holding across all levels of multiplicity. For example, across all K_i, black defendants face roughly 8- to 12-percent higher likelihood of receiving jail time than their white comparators (Panel A).

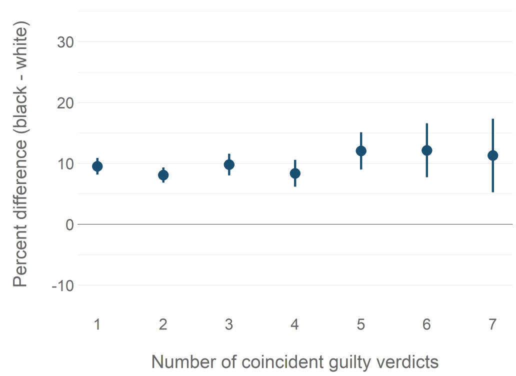
In panels B and C of Figure 18, we consider the extensive margin, plotting the estimated racial gaps in consecutive-sentence lengths and real-sentence lengths. This reveals an important distinction in the experienced differential in the sentencing outcomes of black defendants with multiple guilty verdicts.

[76] Lim et al. (2016) reveals that some defendants in their sample have multiple offenses—it is not clear how they are treated.

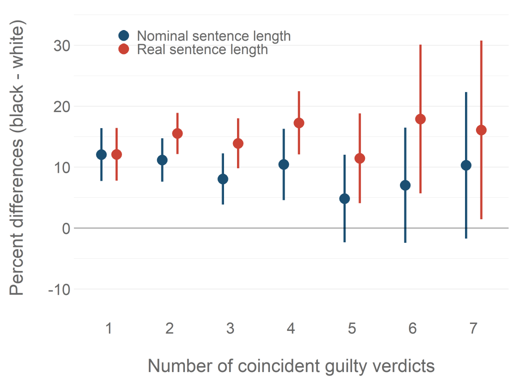
[77] In doing so we are implicitly assuming that outcomes are additive in the number of within-crime-type offenses a defendant has been found guilty of. We find similar results when we allow outcomes to by quadratic in counts by crime type.

Figure 18: How does the black–white gap in sentencing defendants change across the number of coincident guilty verdicts?

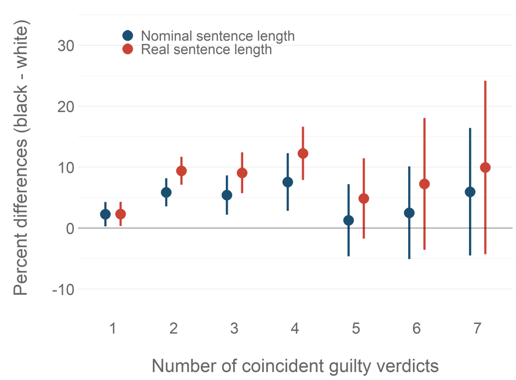
(a) Racial gap in incarceration (yes/no)



(b) Racial gap in total sentence lengths



(c) Racial gap in total sentence lengths, conditional on being incarcerated



Notes: In each panel, we plot impact estimates associated with race from seven different models of defense-level observations. In all panels, we control for race fixed effects, age (in five-year bins), and i's "prior-record score" to absorb any effect of criminal history on sentencing. We also include the average statutory minimum-sentence lengths for defendant i's crimes, his average offense-gravity score, and controls for the fixed contributions specific to crime types. We also control for judge γ_j fixed effects.

Namely, accounting for the assignment of concurrent sentencing reveals large *increases* in the differential treatment of black and white defendants. In fact, in consecutive sentence length one is forced to conclude that the gap in sentence length is insignificant among those facing five or more sentences.[78] At the same time, as the systematic assignment of concurrent sentencing is more generous to white defendants than to black, the "black–white" gap in real-sentence lengths is increasing through this same range—while black defendants face 12-percent longer sentences when $K = 1$, the gap in concurrently adjusted sentence length (Real) is 16 percent at $K = 2$, and increases to as much as 18 percent (at $K = 6$). Across $K_i \in \{2, 3, \dots 7\}$, we find 5-percent smaller to 48-percent larger racial discrepancies than that found when $K = 1$.

In Panel C we condition again on the defendant receiving at-least some jail time—we find larger discrepancies in the sentence lengths received by black defendants when we account for the use of concurrent sentencing, though the confidence intervals largely overlap. Compared to the racial bias in sentence length of 2 percent found when $K = 1$, racial biases in real sentence lengths are 110 to 427 percent larger across $K_i \in \{2, 3, \dots 7\}$.

3.4 Conclusion

Between 2007 and 2016, thirty-one percent of defendants in the Common Pleas Courts of Pennsylvania were brought before a presiding judge with multiple sentences. Yet, existing analysis of criminal sentencing largely ignores any potential co-dependencies across multiple sentences—either assuming that all such sentences are determined independently and jeopardizing the internal validity of estimates independently across offenses for a given defendant. This

[78] The range in estimates across K_i is roughly 5- to 12-percent longer sentences for black defendants.

assumption, however, deserved to be explored.

We examine how multiplicity of offenses influence sentencing outcomes. We find strong evidence of judicial discretion in the decision to assign sentences to be served consecutively or concurrently, with multiplicity in offenses leading defendants to receive concurrent sentences significantly longer than what would occur under a scheme of concurrent sentencing, but significantly less than what would be found under consecutive sentencing.

We find disturbing evidence that racial gaps in sentencing also vary across defendants by their number of offenses. We know from prior literature that sentences are level different for black defendants, on average. We demonstrate that the use of concurrent sentencing—typically left to the discretion of individual judges—drives "black–white" gaps in outcomes further apart for defendants who face multiple sentences. Among those facing single sentences, black defendants face 12-percent longer sentences. Yet, accounting for the non-random allocation of concurrences, the gap in sentence length increases, with black defendants with two offenses experiencing 28 percent longer gaps.

The results have far-reaching implications—in the sentencing literature, but also in policy. We provide cautionary evidence that analyses can fall short of the fuller understanding of sentencing outcomes without addressing multiplicity i multiplicity. In our analysis, this reveals *larger* discrepancies between black and white defendants than would be found when both only looking at defendants with one offense, and when not accounting for concurrences in sentencing.

APPENDIX

Chapter II

Including Mean-Other-Sentence in the baseline estimation equation allows the average sentence of other defendants in defendants' sentencing cohorts to impact defendants' sentences. However, alternative elements of the sentences of other defendants may impact defendants' sentences. I re-run the baseline specification, additionally allowing the maximum and minimum sentence length of other defendants in defendants' sentencing cohorts to impact defendants' sentences. That is, using the previously discussed predicted sentence lengths, I calculate Max-Other-Sentence and Min-Other-Sentence as the maximum (minimum) predicted sentence length of other defendants in a sentencing cohort. Note, the values of Max-Other-Sentence, Min-Other-Sentence, and Mean-Other-Sentence are identical for defendants in sentencing cohorts of two, and linear combinations of each other for defendants in sentencing cohorts of three, leaving little variation for separate identification. To account for this and allow for a wider range of effect types, I allow the coefficients to vary by the number of defendants in a sentencing cohort. I present the results in Table A1.

As shown in Specification 1 in Table A1, the coefficient on Mean-Other-Sentence is slightly larger in magnitude, but does not statistically change with the inclusion of Max-Other-Sentence and Min-Other-Sentence as control variables. The coefficient on Min-Other-Sentence is negative and statistically significant at the 5% level, suggesting defendants receive *shorter* sentences as the "least bad" other defendant in their sentencing cohort receives a longer sentence. While statistical precision is lost when the variables

Table A1: Joint Effects of Mean-, Max-, and Min-Other-Sentence

$(\beta * 10^2)$	Sentencing Cohort Measure					
	Mean	*# Other	Max	*# Other	Min	*# Other
Specification 1	1.75*	–	0.01	–	-1.26*	–
	(0.69)	–	(0.15)	–	(0.55)	–
Specification 2	1.84	-0.06	-0.03	0.01	-0.30	-0.32**
	(1.00)	(0.22)	(0.28)	(0.02)	(0.74)	(0.11)

Notes: The coefficients capture the average change in months in a defendant's sentence due to a one-month increase in the mean (max, min) sentence of other defendants in the defendant's sentencing cohort. In Specification 1 (2), the parameters are not (are) allowed to vary by the number of defendants within a sentencing cohort. Estimated bootstrapped standard errors which are allowed to cluster by judge are shown in parentheses.
* $p<0.05$, ** $p<0.01$

of interest are allowed to vary by the number of "others" in the sentencing cohort (Specification 2 in Table A1), the overall pattern of results remains the same. However, the coefficient on Min-Other-Sentence decreases in magnitude and looses statistical significance, while its interaction with the number of others is statistically significant and negative, suggesting the negative relationship between defendants' sentences and the minimum sentence of other defendants in their sentencing cohort increases in strength as the size of the cohort increases. Nonetheless, Mean-Other-Sentence continues to be an impactfull measure.

Additionally, I display the results of the entire primary analysis with Max-Other-Sentence or Min-Other-Sentence as the primary variable of interest in Table A2. Results are qualitatively similar to the baseline results.

Table A2: The Role of Max-Other-Sentence and Min-Other-Sentence on Sentencing Outcomes

Estimate on Max-Other-Sentence

Cohort Race	**Defendant Race:** All	Black	White
All	.004** **A**	.005**	.003** **B**
	(.001)	(.002)	(.001)
Black	.002* **C**	.004**	.001 **D**
	(.001)	(.002)	(.001)
White	.004**	.000	.004**
	(.001)	(.002)	(.001)

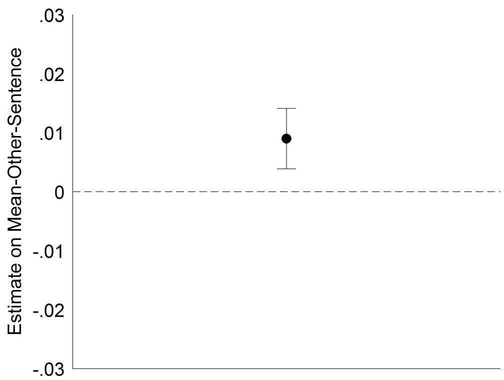
Estimate on Min-Other-Sentence

Cohort Race	**Defendant Race:** All	Black	White
All	.003 **A**	.008	.000 **B**
	(.004)	(.007)	(.004)
Black	.005* **C**	.013*	.001 **D**
	(.002)	(.007)	(.002)
White	.001	-.008	.005
	(.007)	(.008)	(.008)

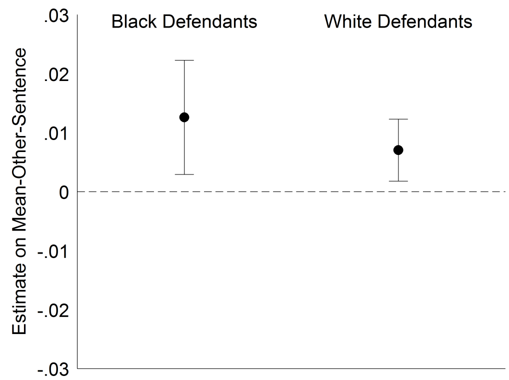
Notes: The left (right) table displays the estimated coefficients on Max (Min)-Other-Sentence in Analyses A, B, C, and D. In Analysis A, the coefficient captures the average change (in months) in a defendant's sentence due to the maximum (minimum) other sentence of defendants in their sentencing cohort being one-month longer. In Analysis B (Analysis C), the effects are allowed to vary by defendant (cohort) race, and in Analysis D, by both defendant and cohort race. Each of the four models are estimated independently for each dependent variable. Estimated bootstrapped standard errors which are allowed to cluster by judge are shown in parentheses.
* p<0.05, ** p<0.01

Figure A1: The Role of Sentencing Cohorts (Minimum Sentence)

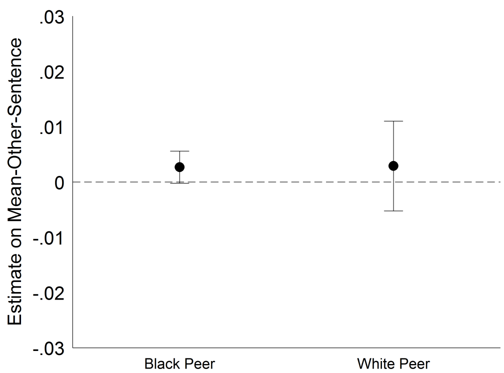
(a) Baseline Results



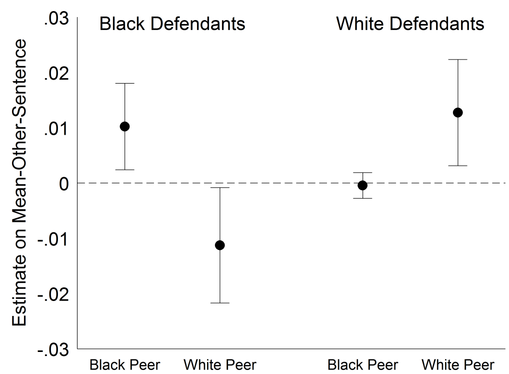
(b) Defendant-Race Heterogeneity



(c) Cohort-Race Heterogeneity



(d) Defendant- and Cohort-Race Heterogeneity



Notes: Notes: Panel A displays the coefficient on Mean-Other-Sentence in a regression on a defendant's maximum sentence. The coefficient captures the average change (in months) in a defendant's sentence due to a one-month increase in the mean sentence of other defendants in the defendant's sentencing cohort. In Panel B, the coefficient is allowed to vary by defendant race. Panel C displays the coefficients on cohort-race-specific Mean-Other-Sentence in a regression on a defendant's maximum sentence. The coefficient captures the average change (in months) in a defendant's sentence due to a one-month increase in the mean sentence of other defendants (of a given race) in the defendant's sentencing cohort. In Panel D, the coefficients are allowed to vary by defendant race. 95% confidence intervals derived from bootstrapped standard errors, allowed to cluster by judge, are shown.

REFERENCES CITED

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

Aizer, A. and J. Doyle (2015). Juvenile incarceration, human capital and future crime: Evidence from randomly-assigned judges. *The Quarterly Journal of Economics 130*(2), 759–803.

Alesina, A. and E. L. Ferrara (2014). A test of racial bias in capital sentencing. *American Economic Review 104*(11), 3397–3433.

Angrist, J. D. (2014). The perils of peer effects. *Labour Economics*.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers.

Anselin, L. (2003). *A Companion to Theoretical Econometrics*. Hoboken: Blackwell Publishing Ltd.

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

Bailey, M. A. (2007). Comparable preference estimates across time and institutions for the Court, Congress, and Presidency. *American Journal of Political Science 51*(3), 433–448.

Bailey, M. A. (2017). Measuring ideology on the courts. In R. M. Howard and K. A. Randazzo (Eds.), *Routledge Handbook of Judicial Behavior*, Chapter 4. Abingdon: Taylor & Francis.

Berdejò, C. (2018). Criminalizing race: Racial disparities in plea-bargaining. *Boston College Law Review 59*(4).

Berdejó, C. and D. L. Chen (2017). Electoral cycles among US Courts of Appeals Judges. *The Journal of Law and Economics 60*(3), 479–496.

Beron, K. J. and W. P. Vijverbeg (1999). *Probit in a Spatial Context: a Monte Carlo Approach In L. Anselin and R. Florax (Eds.), Advance in Spatial Econometrics*. Heidelberg: Springer.

Bhuller, M., G. B. Dahl, K. V. L. ken, and M. Mogstad (2019). Incarceration, recidivism and employment. *NBER Working Paper 22648*.

Biskupic, J. (2018). Supreme Court still feeling the impact of Antonin Scalia's death. *CNN*.

Blackwell, B. S., D. Holleran, and M. A. Finn (2008). The impact of the Pennsylvania Sentencing Guidelines on sex differences in sentencing. *Journal of Contemporary Criminal Justice 24*(4), 399–418.

Blair, I., C. Judd, and K. Chapleau (2004). The influence of afrocentric facial features in criminal sentencing. *Psychological Science 15*, 674–679.

Bonica, A. (2014). Mapping the ideological marketplace. *American Journal of Political Science 58*(2), 367–386.

Bonica, A., A. Chilton, J. Goldin, K. Rozema, and M. Sen (2017a). Influence and ideology in the American Judiciary: Evidence from Supreme Court law clerks. *Coase—Sandor Working Paper*.

Bonica, A., A. Chilton, J. Goldin, K. Rozema, and M. Sen (2017su). The political ideologies of law clerks. *American Law and Economics Review 19*(1), 96–128.

Bonica, A., A. Chilton, and M. Sen (2015). The political ideologies of American lawyers. *Journal of Legal Analysis 8*(2), 277–335.

Bonica, A., A. S. Chilton, J. Goldin, K. Rozema, and M. Sen (2017b). Measuring judicial ideology using law clerk hiring. *American Law and Economics Review 19*(1), 129–161.

Bonica, A. and M. Sen (2017). The politics of selecting the bench from the bar: The legal profession and partisian incentives to introduce ideology into Judicial selection. *The Journal of Law and Economics 60*(4), 559–595.

Buchanan, J. M. and G. Tullock (1999). *The Calculus of Consent: Logical Foundations of Constitutional Democracy.* Indianapolis: Liberty Fund.

Butcher, K. F., K. H. Park, and A. M. Piehl (2017). Comparing apples to oranges: Differences in women's and men's incarceration and sentencing outcomes. *Journal of Labor Economics 35*(S1), 201–234.

Caldeira, G. A., J. R. Wright, and C. J. W. Zorn (1999). Sophisticated voting and gate-keeping in the Supreme Court. *The Journal of Law, Economics, and Organization 15*(3), 549–572.

Center, P. R. (2014). Political polarization in the american public.

Chen, D. L., M. Locher, N. Barry, L. Buchanan, and E. Bakhturina (2016). Events unrelated to crime predict criminal sentence length. *NBER Working Paper*.

Chen, E. Y.-F. (2008). Cumulative disadvantage and racial and ethnic disparities in california felony sentencing. In B. Cain, J. Regalado, and S. Bass (Eds.), *Racial and Ethnic Politcs in California*, pp. 251–274. Berkeley, CA: Institute of Governmental Studies Press.

Choi, S. J. and A. C. Pritchard (2003). Behavioral economics and the SEC. *Standord Law Review 56*(1), 1–73.

Chupp, A. (2014). Political interactions in the Senate: Estimating a political "spatial" weights matrix and and application to lobbying behavior. *Public Choice 160*(3-4), 521–538.

Cirillo, V. A. (1986). Windows for discretion in the Pennsylvania sentencing guidelines. *Villanova Law Review 31*(3), 1309–1349.

Cohen, A. and C. S. Yang (2019). Judicial politics and sentencing decisions. *American Economic Journal: Economic Policy 11*(1), 160–191.

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

Dash, J. (2017). Trump is politicizing the courts – and our Judiciary is under threat. *The Guardian*.

Demuth, S. and D. Steffensmeier (2004). Ethnicity effects on sentence outcomes in large urban courts: Comparissons among white, black, and hispanic defendants. *Social Science Quarterly 85*, 994–1011.

Eberhardt, J. L., P. G. Davies, V. J. Purdie-Vaughns, and S. L. Johnson (2006). Looking deathworthy. *Psychological Science 17*(5), 383–386.

Englich, B., T. Mussweiler, and F. Strack (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin 32*(2), 188–200.

Epstein, L., A. D. Martin, J. A. Segal, and C. Westerland (2007). The judicial common space. *Journal of Law, Economics, & Organization 23*(2), 303–325.

Epstein, L. and C. Mershon (1996). Measuring political preferences. *American Journal of Political Science 40*(1), 261–294.

Epstein, L. and E. A. Posner (2016). Supreme Court Justices' loyalty to the President. *Journal of Legal Studies 45*(2), 401–436.

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

Fischman, J. B. (2013). Interpreting circuit court voting patterns: A social interactions framework. *The Journal of Law, Economics, and Organization 31*(4), 800–842.

Fisher, J. L. (2015). The Supreme Court's secret power. *The New York Times*.

Fiske, A. P., N. Haslam, and S. T. Fiske (1991). Confusing one person with another: What errors reveal about the elementary forms of social relations. *Journal of Personality and Social Psychology 60*(5), 656–674.

Flanagan, F. X. (2018). Race, gender, and juries: Evidence from north carolina. *The Journal of Law and Economics 61*(2), 189–214.

Galston, W. A. (2009). One and a half cheers for bipartisanship. *Brookings Institute*.

Glynn, A. and M. Sen (2015). Identifying judicial empathy: does having daughters cause judges to rule for women's issues? *American Journal of Political Science 59*(1), 37–54.

Gorton, J. and J. L. Boies (1999). Sentencing guidelines and the racial disparity across time: Pennsylvania prison sentences in 1977, 1983, 1992, and 1993. *Social Science Quarterly 80*(1), 37–54.

Green, D. P. and D. Winik (2010). Using random judge assignments to estimate the effects of incarceration and probation on recidivism among drug offenders. *Criminology 48*(2), 357–387.

Gruhl, J., C. Spohn, and S. Welch (1981). Women as policymakers: The case of trial judges. *American Journal of Political Science 25*(2), 308–322.

Guthrie, C., J. J. Rachlinski, and A. J. Wistrich (2001). Inside the judicial mind. *Cornell Law Faculty Publications 86*, 777–829.

Hall, M. G. (1992). Electoral politics and strategic voting in state supreme courts. *The Journal of Politics 54*(2), 427–446.

Harmon, N., R. Fisman, and E. Kamenica (2019). Peer effects in legislative voting. *American Economic Journal: Applied Economics, 11*(4), 156–180.

Harvey, A. and M. J. Woodruff (2011). Confirmation bias in the United States Supreme Court Judicial Database. *The Journal of Law, Economics, & Organization 29*(2), 414–460.

Heyes, A. and S. Saberian (2019). Temperature and decisions: Evidence from 207,000 court cases. *American Economic Journal: Applied Economics 11*(2), 238–265.

Holden, R., M. Keane, and M. Lilley (2019). Peer-effects on the United States Supreme Court. *UNSW Economics Working Paper*.

Horrace, W. C. and S. M. Rohlin (2016). How dark is dark? bright lights, big city, racial profiling. *The Review of Economics and Statistics 98*(2), 226–232.

Jolls, C. (2007). Behavioral law and economics. *NBER Working Paper Series 12879*.

Katz, D. M., M. J. Bommarito II, and J. Blackman (2017). A general approach for predicting the behavior of the Supreme Court of the United States. *PLoS One 12*(4).

Kelejian, H. H. and I. R. Prucha (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics 17*(1), 99–121.

Klain, R. A. (2017). The one area where Trump has been wildly successful. *The Washington Post*.

Kling, J. R. (2006). Incarceration length, employment, and earnings. *American Economic Review 96*(3), 863–876.

Ladha, K. K. (1995). Information pooling through majority-rule voting: Condorcet's jury theorem with correlated votes. *Journal of Economic Behavior and Organization 26*, 353–372.

LaFave, W. R., J. H. Israel, and N. King (2000). *Criminal Procedure*. Hornbook series. West Group.

Landerø, R. (2015). Does incarceration length affect labor market outcomes? *The Journal of Law and Economics 58*(1), 205–234.

Landes, W. M. and R. A. Posner (2009). Rational judicial behavior: A statistical study. *Journal of Legal Analysis 1*(2), 775–831.

LeSage, J. P. and R. K. Pace (Eds.) (2004). *Spatial and Spatiotemporal Econometrics*. Elsevier.

LeSage, J. P. and R. K. Pace (2009). *Introduction to Spatial Econometrics*. Chapman & Hall / CRC Taylor & Francis Group.

Levinson, J. D. (2007). Forgotten racial equality: Implicit bias, decisionmaking, and misremembering. *Duke Law Journal 57*, 345–424.

Levy, G. (2007). Decision making in committees: Transparency, reputation, and voting rules. *American Economic Review 97*(1), 150–168.

Levy, P. (2015). This is how the Supreme Court decides which cases to announce when. *Mother Jones*.

Li, H., S. Rosen, and W. Suen (2001). Conflicts and common interests in committees. *American Economic Review 91*(5), 1478–1497.

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

Liptak, A. (2010). A sign of court's polarization: Choice of clerks. *The New York Times*.

Lovegrove, A. (2004). Sentencing the multiple offender: Judicial practice and legal principle. *Australian Institute of Criminology*.

Manski, C. F. (2000). Economic analysis of social interactions. *Journal of Economic Perspectives 14*(3), 115–136.

Martin, A. D. and K. M. Quinn (2002). Dyamic ideal point estimation via markov chain monte carlo for the U.S. Supreme Court, 1953-1999. *Political Analysis 10*(2), 134–153.

Moffett, K. W., C. Shipan, and F. Maltzman (2016). The Supreme Court is taking far fewer cases than usual. Here's why. *The Washington Post*.

Mueller-Smith, M. (2015). The criminal and labor market impacts of incarceration. *Unpublished Working Paper*.

Mustard, D. B. (2001). Racial, ethnic, and gender disparities in sentencing: Evidence from the U.S. Federal Courts. *The Journal of Law and Economics 44*(1), 285–314.

Nelson, W. E., H. Rishikof, I. S. Messinger, and M. Jo (2009). The liberal tradition of the Supreme Court clerkship: Its rise, fall, and reincarnation? *Vanderbilt Law Review 62*, 1749–1804.

Nivola, P. S. (2009). In defense of partisan politics. *Brookings Institute*.

Ottaviani, M. and P. Sørensen (2001). Information aggregation in debate: who should speak first? *Journal of Public Economics 81*, 393–421.

Palmer, B. (2013). How does SCOTUS schedule its decisions? *Slate*.

Park, K. H. (2017). Do judges have tastes for discrimination? Evidence from criminal courts. *The Review of Economics and Statistics 99*(5), 810–823.

Peppers, T. C. and C. Zorn (2008). Law clerk influence on Supreme Court decision making: An empirical assessment. *DePaul Law Review 58*(51).

Poole, K. T. (2005). *Spatial Models of Parliamentary Voting*. Cambridge University Press.

Rachlinski, J. J., A. J. Wistrich, and C. Guthrie (2015). Can judges make reliable numeric judgments? distored damages and skewed sentences. *Indiana Law Journal 90*.

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

Rose, E. K. and Y. Shem-Tov (2019). Does incarceration increase crime? *Working Paper*.

Ryberg, J., J. V. Roberts, and J. W. de Keijser (Eds.) (2017). *Sentencing Multiple Crimes*. Oxford University Press.

Schanzenbach, M. M. (2015). Racial disparities, judge characteristics, and standards of review in sentencing. *Journal of Institutional and Theoretical Economics 171*(1), 27–47.

Segal, J. A. and A. D. Cover (1989). Ideological values and the votes of U.S. Supreme Court Justices. *The American Political Science Review 83*(2), 557–565.

Segal, J. A., L. Epstein, C. M. Cameron, and H. J. Spaeth (1995). Ideological values and the votes of U.S. Supreme Court Justices revisited. *The Journal of Politics 57*(3), 812–823.

Segal, J. A. and H. J. Spaeth (2002). *The Supreme Court and the Attitudinal Model Revisted*. Cambridge University Press.

Spaeth, H. J., L. Epstein, A. D. Martin, J. A. Segal, T. J. Ruger, and S. C. Benesh (2017). 2017 Supreme Court database. *Version 2017*(Release 01).

Spamann, H. and L. Klöhn (2016). Justice is less blind, and less legalistic, than we thought: Evidence from an experiment with real judges. *The Journal of Legal Studies 45*(2), 255–280.

Steen, S., R. L. Engen, and R. R. Gainey (2005). Images of danger and culpability: Racial stereotyping, case processing, and criminal sentencing. *Criminology 43*(2), 435–468.

Steffensmeier, D. and C. L. Britt (2002). Judges' race and judicial decision making: Do black judges sentence differently? *Social Science Quarterly 84*(4), 749–764.

Steffensmeier, D., J. Ulmer, and J. Kramer (2006). The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, black, and male. *Criminology 36*(4), 763–798.

Sunstein, C. R., D. Schkade, L. M. Ellman, and A. Sawicki (2006). *Are Judges Political? An Empirical Analysis of the Federal Judiciary.* Brookings Institute.

Taylor, S. E., S. T. Fiske, N. L. Etcoff, and A. J. Ruderman (1978). Categorial and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology 36*(7), 778–793.

West, S. (2012). Smile for the camera. *Slate*.

Wykstra, S. (2018). Just how transparent can a criminal justice algorithm be? *Slate*.