Judges on the Benchmark: Developing a Sentencing Feedback System

Viet Nguyen[1], Greg Ridgeway[1,2]

1 Department of Criminology, University of Pennsylvania

2 Department of Statistics, University of Pennsylvania

Abstract:

Research Summary: *We use a doubly robust, internal benchmarking method to measure the effect of each judge on sentencing outcomes relative to a set of cases that are handled by the judge's peers and that are statistically similar on all observable case features. Judges with the highest propensity in using custodial sentences were 22 percentage points more likely to impose an incarceration sentence and 5 percentage points more likely to use a prison sentence compared to their benchmark. Judges with lower propensity to incarcerate used alternative sanctions at a higher rate, whereas judges who deviate from their peers in assigning longer prison sentences were less likely to issue downward departures from sentencing guidelines. If the top 20 percentile of judges with respect to custodial sentences reduced their differences with their benchmark by 50 percent, the statewide incarceration rate and prison rate would decrease by 2 percentage points and 0.5 percentage points respectively.*

Policy Summary: *Judges receive limited information on how their sentencing practices contribute to inter-judge sentencing disparities which can undermine equity and the perceptions of legitimacy in the criminal justice system. Internal benchmarking methods like the one used in this study could inform judges that their sentencing practices are disparate from their peers and potentially mitigate sentencing disparities. Depending on the type of disparity that a judge contributes to, an internal benchmark system would highlight discretionary decisions and provide a feedback system from which courts and sentencing commissions could evaluate efforts to reduce inter-judge disparities. The approach used here could be adapted in states to provide regular feedback throughout a judge's tenure and to help move judges that contribute most to disparities to have sentencing practices more similar to their judicial peers.*

1 | INTRODUCTION

Judges wield a wide array of discretion. From the weight judges place on different case factors to the amount of perceived risk that judges see in a defendant, the variation and complexity in judicial discretion can produce inequities among similarly situated defendants. Because judges manage caseloads with varying compositions (e.g., percentage of violent crime cases, criminal history of defendants, etc.), it is difficult to assess whether differences in sentencing outcomes are attributable to a judge or the features of the cases for which they make legal decisions. Prior studies have measured overall variation among judges or inter-judge disparities in narrow settings, such as jurisdictions where cases are randomly assigned to judges (Yang, 2014; Scott & W., 2010). However, the current methods are not designed to assess how individual judges give rise to broader sentencing disparities across entire state or federal court systems. Moreover, while prior studies highlight sources of inter-judge disparity, most of these sources do not provide a concrete pathway for reducing disparities. This article addresses the gap by devising a sentencing feedback system that aims to move judges that substantially contribute to statewide sentencing disparities closer to the sentencing patterns of their peers. At the foundation of the feedback system, we use doubly robust, internal benchmarking to estimate the effect of individual judges on sentencing by comparing the sentencing outcomes of a judge to a benchmark that is composed of cases with closely, matched features that other judges adjudicated. The feedback system contains two components. The first component is an individualized, sentencing report that measures how a judge differs from his or her benchmark on a range of outcomes. The second component provides guidance on discretionary decisions that judges can consider for mitigating sentencing disparities. Pairing these two pieces of information can help judges diagnose their contribution to disparities and take steps to reduce inter-judge disparities.

Shifting judges whose sentencing practices considerably deviate from their peers closer to the norm is critical to maintaining the legitimacy of the criminal justice system and ensuring that defendants convicted of similar crimes will receive similar sentences. A judge's sentencing decision not only affect defendants but also the broader views of the legitimacy of the criminal justice system. Differential use of incarceration among judges can ultimately undermine trust and compliance with the criminal justice system if defendants feel like their sentences are determined by the lottery of judges. Moreover, the transactional nature of criminal courts may result in attorneys making different plea arrangements depending on what judge a defendant's case is being heard in front of. Attorney may ask for motions of continuances to schedule a hearing with a more

favorable judge. Since 1975, about one-third of states have adopted sentencing guidelines with many guidelines oriented towards reducing disparities and making sentencing fairer, more transparent, and more equitable (Mitchell, 2017; Tonry, 2013). Accordingly, judicial actors operating under guidelines have an incentive to reduce sentencing disparities, but it is often unclear how judges can assess their conduct when their case composition is different or the body of comparable colleagues is unknown. Because judges tend to receive evaluations or recommendations near the end of their term, the lack of a robust feedback system can lead judges to unknowingly contribute to sentencing disparities for extended periods of time.

We look at cases handled by 424 judges working across all courts in Pennsylvania's Court of Common Pleas to create a feedback system, focused on disparities on the use of jail or prison. We first identify judges who substantially contribute to sentencing disparities. Judges with the highest propensity to incarcerate imposed incarceration sentences at a rate that was more than 22 percentage points higher than their benchmark on average. Among judges with the highest propensity to use prison sentences, their prison sentence rate was 5 percentage points higher than their benchmark on average. We then underline discretionary decisions that can be given more weight in order to move judges closer to the norm. We find that judges with lower propensities to incarcerate use alternative sanctions, such as electronic monitoring and house arrest, as a substitute. Judges that impose long incarceration spells tend to use downward departures or apply mitigating factors at a lower rate. Concurrently, heavier fines and fees do not appear to be related to the rate of incarceration sentences.

The next section of the paper outlines the theoretical framework that connects equity, public safety, and judicial decision-making. Section 3 provides an overview of the literature on judges and their relationship with sentencing disparities. Section 4 describes the existing approaches for evaluating judges. Section 5 describes the context and data where benchmarking is used. Section 6 outlines the methodology behind internal benchmarking. Section 7 covers the results. Section 8 provides a brief discussion and conclusion.

2 | JUDICIAL DECISION-MAKING, EQUITY, AND PUBLIC SAFETY

Theories of procedural and distributive justice point out that the administration of justice influences people's perception of fairness, legitimacy and trust which can consequently moderate the level of compliance. When perceptions of legitimacy are low, criminal justice actors face

higher levels of on job risk and require more resources to maintain public safety (Tyler, 2001). Sherman's defiance theory (1993) suggests that procedural and distributive justice throughout the sentencing process can condition the effectiveness of sanctions. Individuals will respond with defiance to sanctions when the sanction is unfair, the individual is poorly bonded with society, the individual perceives the sanction as stigmatizing, and the individual refuses to acknowledge shame. Through this lens, judges that impose arbitrary, excessive, or undeserved sanctions will produce higher levels of defiance and anger (Sherman, 1993). Thus, in the process of using more criminal justice resources to punish an individual, it is possible that the individual becomes less responsive to rehabilitation and other criminal justice measures that aim to limit recidivism. Individuals who interact with judicial actors may become more resistant and combative, especially if they think that the sentence they received from a given judge was much harsher than they would have received if their case were tried before another judge.

A series of studies support the notion that quality of treatment and the degree of fairness in the criminal justice process affects perception of legitimacy and trust which can then influence compliance with criminal justice actors. Perceptions of disrespect generate feelings of anger and injustice; individuals respond to these feelings with retaliation and defiance (Bornstein, Marcus, Curtis, Rivera, & Swaner, 2001). As such, interactions between individuals and court actors directly influence perceptions of legitimacy and trust. Surveys show that high-risk youth and incarcerated inmates develop negative perceptions of legitimacy throughout different stages, from the preliminary hearing to the sentencing stage to the appeals process (Vuolo, Wright, & Lindsay, 2019; Sprott, 2010). Conversely, among community courts that directly seek to improve procedural justice, participants report higher levels of satisfaction driven by the demeanor of judges (Bornstein, Marcus, Curtis, Rivera, & Swaner, 2016; Dollar, Ray, Hudson, & Hood, 2018). These studies suggest that the criminal justice process and behavior of judges plays a role in shaping the perceptions of legitimacy. When judges vary widely in their sentencing patterns, being assigned to a punitive judge could give rise to feelings of inequity and unfairness.

Importantly, these perceptions of procedural justice, distributive justice, and legitimacy can impact public safety and the future conduct of individuals with criminal justice contact. At the front end, police officers who handled domestic violence incidents with higher levels of procedural justice as measured by representation, constituency, and impartiality led to significantly lower

subsequent assault incidents (Paternoster, Brame, Bachman, & Sherman, 1997). Similarly, DUI offenders sentenced to reintegrative shaming conference which aims to bolster procedural and distributive justice yielded long-term improvements in public safety. Participants in these conferences developed stronger perceptions of legitimacy, reported lower levels of drunk driving, and stronger efforts not to drive under the influence (Tyler, Lawrence, Strang, Barnes, & Woods, 2007). In the correctional setting, procedurally just interactions between officers and inmates in Dutch prison were associated with a 19.4 percent reduction in the likelihood of reconviction within 18 months (Beijersbergen, Dirkzwager, & Nieuwbeerta, 2016). To the authors' knowledge, there are no studies that estimate the effect of inter-judge disparities and the perceptions of procedural/distributive justice on public safety. But given the impact of procedural and distributive justice in other contexts, it is logical that differential treatment among judges can potentially affect perceptions of justice and future compliance with the law.

3 | LITERATURE ON JUDGES

Judges and their decision-making have implications for both equity and public safety. The literature on judges, sentencing, and decision-making can be organized under three broad groups: the influence of judicial attributes on sentencing, the effect of judicial biases on sentencing, and the impact of policy shifts on sentencing. While there is an extensive body of literature that assesses how case attributes affect sentencing and disparities, this paper will focus on the literature that examines judges (Blackwell, Holleran, & Finn, 2008; Koons-Witt, 2002; Starr, 2015).

Judicial attributes have varying effects on sentencing. A judge's unique, personal experience bears on sentencing decisions that can exacerbate or mitigate sentencing disparities. With respect to political affiliation, studies have found that Republican judges impose longer sentences or that the political affiliation of judges has no effect. Republican-appointed judges impose incarceration terms that are three months longer for black defendants and two months shorter for female defendants (Cohen & Yang, 2019). Republican-appointed, female federal judges are less likely to impose a downward departure (Tiede, Carp, & Manning, 2010). Schanzenbach's district level analysis finds that Republican appointees are more punitive than Democratic appointees, but racial disparities do not vary by political affiliation (Schanzenbach, 2018). Other studies in more limited jurisdictions find that the political affiliation of judges has no effect on sentencing (Ashenfelter, Eisenberg, & Schwab, 1995; Lim, Silveira, & Snyder, 2016).

At the same time, a judge's gender, race, and religion appears to influence sentencing and disparities in different ways. Using multi-level models that account for judge and county-level factors, Johnson finds that minority judges are less likely to impose incarceration sentences (Johnson, 2006). Moreover, minority judges are 93 percent more likely to use alternative sanctions in lieu of a prison sentence (Johnson & DiPietro, 2012). Among trial cases where judges hold more discretion, minority and female judges are less likely to incarcerate offenders (Johnson, 2014). Schanzenbach's study of judicial demographics at the federal district level suggest that judicial demographics have no impact on prison sentences, but they do influence racial and sex disparities (Schanzenbach, 2005). The most consistent finding is that judges exhibit in-group bias that leads to harsher sentences within race and more lenient sentence within religious group. (Depew, Eren, & Mocan, 2017; Gazal-Ayal & Sulitzeanu-Kenan, 2010; Shayo & Zussman, 2011; Emeriau, 2019; Welch, Combs, & Gruhl, 1988). Patterns of differential sentencing based on discernible attributes are problematic because they convey to the convicted individual that their penalties are based on extralegal factors. If judges with discernible attributes reside on the extreme ends of the sentencing continuum, it can further reinforce the notion that sentencing outcomes are a product of extralegal rather legal case factors.

Judicial discretion is also influenced by external factors beyond a judge's attributes. These biases contribute to disparities in sentencing severity and race (Abrams, Bertrand, & Mullainathan, 2012; Pina-Sanchez, Grech, Brunton-Smith, & Sferopoulos, 2019). At the micro level, the composition of cases can influence sentencing decisions. Judges exposed to an initial set of cases with lower levels of offense severity order longer sentences and are more likely depart upward above the sentencing guidelines; the inverse relationship holds for judges exposed to an initial set of cases with higher levels of offense severity (Leibovitch, 2016). In addition, being sentenced on the same day as defendants with longer sentences will drag a defendant's sentence length upward (Mindock, 2019).

The broader environment surrounding a court can also alter sentencing decisions. Rotating judges have smaller sentencing disparities for violent offenses which may be a product of exposure to more local court actors (Pina-Sanchez, Grech, Brunton-Smith, & Sferopoulos, 2019). Research finds that judges incorporate local norms into their sentencing decisions, and over time, their sentencing decisions reflect the sentencing decisions of more senior judges (Abrams, Galbiati,

Henry, & Philippe, 2019). In settings where local norms or senior judges substantially contribute to disparities, a sentencing feedback system that draws from statewide cases could provide judges with another reference point to compare their sentencing practices to. External constraints and events also have an effect on sentencing decisions. For instance, more judicial vacancies in the federal court system cause case delays and lead prosecutors to dismiss more cases and move forward with different cases (Yang, 2016). Trial courts that experience an increase in private prison capacity in 14 states show a 1.3 percent estimated increase in sentence length when compared to bordering trial courts that do not see an increase in private prison capacity (Poyker & Dippel, 2019). Even emotional shocks, stemming from events as small as a football game loss to events as serious as terrorist attacks, give rise to anger and frustration that lead to harsher punishment toward minorities in several settings, such as Louisiana or Israel (Eren & Mocan, 2018; Shayo & Zussman, 2011).

Judges also show sensitivity to political and media factors that affect their electoral ability or reputation. When judges' election cycles are competitive, they exhibit more punitive sentencing behavior (Dippel & Poyker, 2019). Concurrently, when local media influence is strong, criminal justice events covered by media exert pressure on judicial actors, especially those in competitive districts. Among jury trials, recorded events of judicial error led to shorter sentences whereas more crime events lead to long sentences (Philippe & Ouss, 2018). Among civil cases in state courts, more robust media coverage reduces the disparities of damage awards between conservative and liberal districts (Lim, 2015). With criminal cases, newspaper coverage tends to increase the imposed sentence length by nonpartisan, elected judges for violent crimes (Lim, Snyder, & Strömberg, 2015). The challenge with these sources of inter-judge sentencing disparities is that the majority of them cannot be manipulated.

Consequently, sentencing guidelines have served as the main policy channel for curtailing inter-judge sentencing disparities. Numerous studies have examined how inter-judge sentencing disparities fluctuate under different federal sentencing guidelines that reduced judicial discretion. An early study comparing inter-judge sentencing disparities before and after the creation of the federal sentencing guidelines found that the average difference in prison sentences between two judges fell from 4.9 months to 3.0 months (Anderson, Kling, & Stith, 1999). Despite these improvements, racial disparities still persisted, especially among drug offenders (Mustard, 2001).

In the 2000s, a series of cases adjusted judicial discretion and reduced the importance of sentencing guidelines leading to wider inter-judge and inter-district disparities (Scott & W., 2010; Yang, 2014). Yang accounted for the increased charging of mandatory minimums following the weakened federal sentencing guidelines and continued to find inter-judge disparities (Yang, 2014). Moreover, inter-judge disparities and racial disparities were concentrated among judges who disagreed with the guidelines or judges who accumulated more of their experience under guidelines with wider discretion (Yang, 2015). The fluctuation of inter-judge sentencing disparities in relation to changes in the federal sentencing guidelines call attention to two points. While sentencing guidelines can reduce inter-judge disparities, their efficacy depends on whether the guidelines are more advisory or mandatory in nature. And even when the sentencing guidelines have a more mandatory design, it still leaves room for inter-judge disparities. Judges can fully comply with sentencing guidelines and contribute to sentencing disparities. For example, if the presumptive recommended sentence range is between 12 months and 24 months, a judge that imposes sentences only at 24 months would be compliant with the guidelines but likely exacerbate disparities if all other judges impose sentences at 18 months.

These observations repeat themselves at the non-federal level. Non-federal jurisdictions that adopted sentencing guidelines displayed decreases in sentencing disparities, but those with voluntary guidelines saw limited or no reduction in disparities (Grffin, Rushin, & Colquitt, 2019; Weisberg & Hunt, 2007; Bogan & Factor, 1997; Tonry, 2013). In Alabama which implemented both voluntary and presumptive guidelines, voluntary guidelines corresponded to smaller decreases in sentencing disparities compared to the presumptive guidelines (Grffin, Rushin, & Colquitt, 2019). In short, sentencing guidelines can reduce inter-judge disparities to varying degrees. A sentencing feedback system could help cover some of the limitations of sentencing guidelines. When guidelines are more advisory in nature or guideline compliance rates provide limited information about sentencing disparities, then a sentencing feedback system can potentially be a useful tool for nudging judges closer to the center.

4 | JUDICIAL EVALUATION AND RATINGS

Trial judges exercise tremendous authority in the criminal justice system and sentence defendants to millions of years in state jail and prisons (Huber & Gordon, 2004). On a year-to-year basis, recruiting judges with specific attributes, controlling external environmental factors, or

implementing policy reform that can address inter-judge sentencing disparities is unrealistic. In this respect, one possible venue for addressing sentencing disparities is to inform judges of their impact on the criminal justice system and sentencing outcomes.

Judges typically receive information on their conduct through judicial performance evaluations or ratings given by the state or local bar association. Bar association ratings provide recommendations on judicial candidates at judicial elections or retention elections. Judicial performance evaluations, which are funded through the legislature or judicial branch in 17 states, deliver feedback throughout a judge's tenure, and in a handful of states, the assessments are used to determine recommendations for retention elections. Both approaches share the same foundation where they aim to deliver feedback conducive to self-improvement as well as inform the electorate. To measure judicial quality, the evaluations survey a broad sample of attorneys, jurors, and other relevant parties and ask them to anonymously rate judges that they have interacted with on dimensions such as legal ability, integrity, impartiality, communication skills, professionalism, temperament, and administrative capacity. These ratings are then aggregated to inform the judge of his or her conduct which will ideally facilitate constructive criticism and feedback. Providing voters with this information can serve to improve judicial accountability without eliminating judicial independence (Andersen, 2001).

However, there are several limitations that highlight the need for additional, objective measures. First, judges are skeptical of performance evaluations because the sampling of respondents may not be representative, the response rate is low, non-attorney responses are not distinguished by their role, and non-attorneys may not have the expertise sufficiently evaluate judges (Brody, 2008). Second, conducting surveys to rate judges is resource intensive and time intensive which can make the assessment infrequent. Kourlis and Singer argue that judicial performance evaluation programs require more regular evaluations, a neutral criterion, and more transparency (Kourlis & Singer, 2007). Third, bar association ratings and judicial performance evaluations are susceptible to racial and gender bias. Gill et al. found that minority and female judges received significantly lower retention scores after accounting for other measures of judicial quality such as reversal rate, discipline records, scandal involvement, experience, and law school quality (Gill, Lazos, & Waters, 2011). Along similar lines, Sen (2014) found that minority and female nominees received lower ratings after controlling for education, experience, and

partisanship; the lower ABA ratings were not predictive of poorer judges as measured by reversal rates. Fourth, these evaluations provide a binary indicator of quality; the assessments tend to be universally positive where only a miniscule number of judges are not recommended. The lack of variation dampens the amount of information a judge can work with to review his or her conduct. Scholars have proposed other objective measures, such as citation rates, reversal rates, and independence from political affiliation, which all have their own limitations and challenges (Levi & Gulati, 2009). Robbennolot has suggested that judges be benchmarked by comparing their decisions with juries, but it is unclear which group is making the appropriate decision since both groups are susceptible to biases (Robbennolot, 2005).

Above all, the existing evaluations and proposed objective measures are not designed to address a judge's impact on the criminal justice system. If the goals of ratings and evaluation programs are to help judges with self-improvement, then judges should be given feedback that reflects a major component of their day-to-day work. Measures of inter-judge disparities can move the discussion beyond what makes a "good" or "bad" judge and help judges and court actors diagnose sentencing practices that lead to undue disparities. Judges manage varying caseload compositions and have different resources at their disposal. Judges who exclusively handle misdemeanor cases should not be compared to judges who handle mostly felony cases. A method that can be used to compare the sentences of a given judge to a reference group of judges that have similar caseloads could help flag judges that substantially contribute to sentencing disparities.

In the present paper we showcase an internal benchmarking method that can be used to inform judges of their impact on disparities and reduce inter-judge sentencing disparities. The internal benchmarking method uses existing measures of cases and a transparent criterion that can be conducted on a regular basis. This method is less susceptible to biases and measurement error found in surveys. Because of the heterogeneity in cases seen before judges, the assessment of a judge's sentencing record needs to disentangle between differences in underlying case compositions from their peers from their individual judicial discretion. While drawing a comparison group of judicial peers with similar cases in a large urban court setting is possible, when judges are working in smaller counties there needs to be a method to compare them to a similar pool of judges. We demonstrate the utility of an internal benchmarking method by comparing trial judges in Pennsylvania making sentencing decisions in different contexts.

5 | POLICY CONTEXT AND DATA

5.1 | Policy Context

In Pennsylvania, the majority of criminal cases are disposed of in the Court of Common Pleas, the trial courts of Pennsylvania. Pennsylvania uses a sentencing system where judges must consider the sentencing guidelines. If they depart from the guidelines, they must state a reason for the sentence imposed. A defendant can appeal a sentence based on the fact that the judge "departed from the guidelines and imposed an unreasonable sentence". Judges effectively reference a sentencing matrix with multiple cells outlining the possible sanctions as well as a range for the recommended sentence length. Each cell contains a minimum and max sentence length where the minimum sentence cannot be more than half the maximum sentence. Appendix Table A2 displays an example of Pennsylvania's sentencing matrix. The starting point is determined by a combination of the offender's current offense and the offender's criminal history. From here, judges have the option to 1) apply mitigating or aggravating factors, 2) depart above or below the recommended range, 3) apply enhancements, 4) impose additional fines, and 5) impose an intermediate sanction in lieu of a traditional sanction such as prison or probation. Depending on the combination of offense severity and criminal history, judges who find evidence of aggravating/mitigating circumstances may apply aggravating/mitigating factors which adjust the sentence between 3 to 12 months. A judge departs from the sentencing guideline if the sentence is above the aggravated range or below the mitigated range. In addition, if judges find evidence related to use of a deadly weapon or crimes related to youth or schools, they may choose to utilize a separate enhancement matrix with longer sentences. Judges also have discretion to impose fines. The fines may not exceed the statutory maximum, but there are broad recommended ranges based on the offense severity and criminal histories. Finally, for the majority of offenses, judges have the legal option to impose an alternative sanction, such as electronic monitoring or a boot camp program, in lieu of incarceration or probation. Alternative sanctions can still be used even if the offense carries a mandatory minimum. Altogether, judges have discretion over a range of decisions which can influence the final sentence. Focusing on one measure will fail to provide a comprehensive picture.

5.2 | Data and Measures

Our analysis uses data on 776,626 cases from the Pennsylvania Commission on Sentencing from 2009-2012 and 2014-2018, excluding 2013 since that year is missing information on alternative sanction eligibility. Each observation in the dataset is a case, described by a case identifier, an offense code, the corresponding sentence for the offense, the identity of the sentencing judge, and a broad range of variables related to case attributes. These cases are solely compromised of cases with convictions. For cases with multiple offenses, we use the most serious offense as the top charge and summarize the remaining offenses by counting them by offense grade. We drop 1.2% (n=9,965) of cases where defendants are sent to a state, mental hospital; these cases do not receive a sentence. Altogether, the benchmarking analysis is applied to 761,661 unique cases that contains information on sentencing outcomes, discretionary decisions, and case information considered during sentencing. We benchmark judges whose cases have sufficient overlap with their colleagues and judges who have sentenced as least 50 cases, a reasonably sized sample that can represent their sentencing practices. This restricts the benchmarking process to 424 out of the 690 judges (61%) observed in the time period who account for 98.3% of all cases; the remaining 266 judges only account for 13,232 cases.

Sentencing outcomes and discretionary decisions

We focus on the following primary sentencing outcomes: any prison sentence, any incarceration sentence (which includes jail and prison), prison sentence length, and incarceration length. We distinguish prison from any incarceration for three reasons. First, prisons cost more to operate and require more resources. Second, judges may operate under different constraints where prison capacity is larger than jail capacity. Third, the design and location of prisons versus jails suggest different levels of rehabilitative support. Prisons can offer long-term rehabilitative programming, but they also can produce adverse peer effects or reduce pro-social family contact due to remoteness of some facilities. If the individual did not receive a custody sentence, we code the minimum and maximum sentence length as 0 months. For individuals who receive a life sentence, we code their minimum sentence length as 360 months and their maximum sentence length as 720 months[1]. We transform prison and jail length as the log of the amount of sentenced time in days plus 1 so that relative differences can be expressed in percentage changes. We use 0/1

[1] Life sentences do not have a specified range. We use 360 months as it represents the midpoint between two upper boundary sentences for attempted murder (20 years for attempted murder without non-serious bodily injury and 40 years for attempted murder).

indicator variables to record the use of aggravating/mitigating factors, upward/downward deviations, enhancements, fines, and alternative sanctions. Upward/downward deviations encompass the application of an aggravating/mitigating factor and the use of an upward/downward departure. Appendix Table A1 shows that over ninety percent of alternative sanctions are some form of house arrest, electronic monitoring, work release, or a substance abuse related programming. We also operationalize an intensive measure of fines as the log of the fine amount plus one.

Case attributes

The data contain a wide range of case information that judges can legally consider: the specific offense code, the specific mandatory minimum offense code, the offense gravity score, the prior record score, the overall sentencing level, the offense grade, the offense character (i.e. conspiracy, attempted), the relevant drug for a drug offense, the offender's eligibility for alternative sanctions, indicators for substance abuse issues, age at the time of the committed offense, and age at sentencing. Rather than aggregate the offense codes into broader categories, we use the specific offense title and statute to maintain the most granular level of information. For example, if we were to aggregate two robbery charge codes with different recommended penalties and contexts, we would conceal key information that judges potentially consider in sentencing. In total, the data contains 1,003 unique offense codes and 83 mandatory minimum codes.

Next, we turn to the set of variables that explicitly influence a person's position on the sentencing matrix. The offense gravity score constitutes the severity of the offense and is treated as a categorical variable that ranges from 1 to 14. The prior record score is categorized as repeat felony offenders, repeat violent offenders, or a value that ranges from 0 to 5. A higher prior record score gives way to a longer recommended sentence. Being categorized as a repeat felony or violent offenders shifts the recommended penalty even higher. Given that the prior record score can be driven by varying compositions of misdemeanor and felony offenses, we include 180 prior record variables that count the number of priors for specific offenses or misdemeanors. For example, these prior record variables will count the number of prior robberies with serious bodily injury convictions or the number of prior sexual assault conviction. 90 of these variables are for adjudications at the juvenile level, and the remaining 90 variables are convictions as an adult. In this way, we are able to account for two offenders with the same prior record score but with

different compositions in prior offenses or different distributions of prior offenses between the juvenile and adult system. The overall sentencing level, which is determined by the combination of the offense gravity score and prior record score, is also treated as a categorical variable that ranges from 1 to 5. The overall sentencing level provides guidelines on the recommended sanction type. The data also include a categorical measure for inchoate offenses that describes whether the offense was attempted, solicited, or conspired; inchoate offenses will adjust the offense gravity score by 1 point. The offense grade outlines whether the offense is misdemeanor, felony, or homicide charge, and these grades are broken down by their grade level (i.e. F, F-1, F-2, F-3).

In addition, we incorporate several measures that can influence sentencing decisions beyond the sentencing matrix. We operationalize eligibility for boot camp, state intermediate punishment, and county intermediate punishment, as 0/1 indicator variables, set to 1 if they are eligible and 0 if they are not. County intermediate punishments include alternatives such house arrest, intensive supervision, work release, substance abuse treatment, and electronic monitoring. We use a categorical variable to describe the type of drug involved in a drug offense: cocaine, methamphetamine, phencyclidine, heroin, marijuana, narcotics, opioids, or other drugs. We created 0/1 indicator variables for whether or not a pre-sentencing report was completed, a drug and alcohol assessment was completed, the individual is drug or alcohol dependent, and the individual is a sexually violent predator. To account for the other offenses that are not designated as the most serious offense within a case, we include counts of the remaining offenses broken down by offense grade. Last of all, we create a categorical variable for the year when the sentence occurred which will account for yearly differences in guidelines and policy contexts. The data also contain demographics of the offender and the method of disposition (i.e., jury trial, bench trial, plea), but these measures are excluded as they should not have a legal bearing on the sentence. To put it another way, if we were to include the defendant's race, our analysis would legitimize a judge handing out harsher sentences because of the race distribution of defendants in court. Similarly, if we included covariates on whether a trial occurred, our analysis would legitimize a judge imposing a "trial penalty" just because defendants exercised their right to trial. As a result, we include as covariates only case features that offer legitimate explanations for sentencing decisions. In addition, we exclude the county where the case is sentenced which effectively allows all judges across Pennsylvania to contribute to each other's benchmarks.

Table 1 displays the descriptive statistics aggregated at the judge level on sentencing outcomes, discretionary decisions, and select case attributes. For each of the 424 judges, we first calculate averages and percentages on the sentencing decisions and attributes of their cases. Using these 424 averages and percentages, we then produce descriptive statistics at the judge level. These descriptive statistics indicate a high level of variation across judges. The incarceration sentence rate among judges ranges from 5 to 85 percent, and the prison sentence rate ranges from 0 to 60 percent. Judges also exhibit variation in discretionary decisions. The mean alternative sentence rate is 11 percent and the standard deviation is 11 percent. The median judge imposes a fine in 36 percent of cases, and the standard deviation on the fine amount is 201 dollars. The most common discretionary decision is the application of mitigating factors with a mean of 14 percent and standard deviation of 13 percent. For downward departures which lead to sentences below the mitigating range, the mean among judges is 6 percent and the standard deviation is 7. Conversely, for discretionary decisions leading to harsher punishment, judges exhibit lower levels of variation. Of course, judges handle cases with different offense severities and criminal histories which drive differences in sentencing decision. The range for persons offense is between 0 and 50 percent and for weapon offenses is between 0 and 43 percent. With respect to eligibility for alternative sanctions, judges dispose cases where roughly 92 percent of defendants are eligible which suggest that alternatives should be a viable option to most judges.

Table 1: *Descriptive statistics of the average rate of sentencing decisions, discretionary decisions, and select case attributes for each of the 424 benchmarked judges.*

Measures	Mean	SD	Min	Median	Max
Primary Sentencing Outcomes					
Incarceration Rate	0.50	0.14	0.05	0.50	0.85
Prison Rate	0.13	0.09	0.00	0.11	0.60
Inc. Length - Minimum Months	5.85	6.52	0.04	4.45	65.44
Pris. Length - Minimum Months	4.47	6.22	0.00	3.20	63.49
Discretionary Decisions					
Fine Rate	0.47	0.36	0.00	0.36	1.00
Fine Amount	292.42	201.11	0.00	285.21	1168.33
Alternative Rate	0.11	0.11	0.00	0.09	0.69
Enhancement Rate	0.02	0.03	0.00	0.01	0.26
Aggravated Factor	0.10	0.06	0.00	0.09	0.38
Upward Departure	0.03	0.03	0.00	0.02	0.26
Mitigating Factor	0.14	0.13	0.00	0.10	0.64
Downward Departure	0.06	0.07	0.00	0.04	0.39

Measures	Mean	SD	Min	Median	Max
Criminal History					
Prior Record History 0	0.51	0.09	0.22	0.50	0.95
Prior Record History 1	0.13	0.03	0.01	0.13	0.21
Prior Record History 2	0.10	0.02	0.00	0.11	0.20
Prior Record History 3	0.07	0.02	0.00	0.07	0.12
Prior Record History 4	0.06	0.02	0.00	0.06	0.12
Prior Record History 5	0.12	0.05	0.00	0.11	0.38
Current Offense					
Offense Gravity Score	4.55	1.08	3.01	4.23	9.27
Misdemeanor	0.69	0.17	0.13	0.74	1.00
Felony	0.31	0.17	0.00	0.26	0.87
Mandatory Minimum	0.23	0.14	0.00	0.23	0.98
Persons	0.15	0.07	0.00	0.14	0.50
Property	0.23	0.07	0.00	0.23	0.55
Drug	0.22	0.09	0.00	0.21	0.67
Weapons	0.03	0.04	0.00	0.02	0.43
Sex Offense/Rape	0.02	0.02	0.00	0.02	0.25
Eligibility and Drug/Alcohol Indicators					
Offense Eligible for State Intermediate Punishment (Alternatives)	0.84	0.08	0.41	0.85	1.00
Offense Eligible for County Intermediate Punishment (Alternatives)	0.92	0.07	0.53	0.94	1.00
Defendant Eligible for Bootcamp	0.05	0.03	0.00	0.04	0.16
Drug Dependent	0.08	0.12	0.00	0.01	0.50
Multiple Drug Priors	0.01	0.02	0.00	0.00	0.12
Comprehensive Drug and Alcohol Evaluation	0.07	0.10	0.00	0.01	0.48

Notes: The variables shown in this table are a selection of the numerous case features included in the analysis. Some variables aggregated in this table are retained at a more granular level in the benchmarking process. For example, the analyses use the specific offense title and code rather than the broad offense categories displayed in this table. State intermediate punishment involves a combination of substance-abuse related treatment and interventions. County intermediate punishment encompass broader alternatives such as house arrest and electronic monitoring.

6 | Methodology

The goal of internal benchmarking in this context is to assess how an individual judge sentences relative to colleagues that have cases with similar features. Once the benchmark is created, we can estimate how much the individual judge differs in sentencing practices from their benchmark. Compiling these estimates can create a sentencing report for each judge with measures on relative difference, percentiles, and indicators on whether the judge differs from his or her colleagues on similar cases. In addition, we complement the sentencing report with a guidance

component that describes the general association between discretionary decisions and the final sentencing outcomes. Ideally, these two components form a feedback system that can help judges evaluate their sentencing practices.

To create benchmarks and to estimate the effect of each individual judge, we apply a doubly robust, internal benchmarking process, which has been used to assess racial profiling among police officers and racial disparities in sentencing among counties (Ridgeway & MacDonald, 2014; Ridgeway & MacDonald, 2009; Ridgeway, Moyer, & Bushway, 2020). Doubly robust, internal benchmarking is composed of two stages: propensity score weighting and doubly robust estimation. Propensity score weighting reweights a comparison set of cases seen by other judges so that the joint distribution of their features matches the case features of an individual judge. We estimate the propensity score with the method described in McCaffrey, Ridgeway, and Morral (2004), which does not require assumptions about variable selection, the functional form and distribution of variables, and the specification of interactions. The method performs better than other propensity score estimation methods with respect to covariate balance, standard error, per cent absolute bias, and 95 percent confidence interval coverage (Lee, Lessler, & Stuart, 2009). In addition, compared to approaches like propensity score stratification and matching, weighting does not require specifications for different matching types (i.e. one-to-one versus one-to-many) and retains more of the sample size. Once the propensity score is estimated, cases for the individual target judge are given a weight equal to 1, and cases from the comparison set of cases seen by other judges are given weight equal to their propensity score divided by one minus the propensity score, the standard weighting for estimating the average treatment effect on the treated. Effectively, we iterate through each judge and weigh the cases handled by other judges so that the distribution of their case features matches the individual target judge.

Table 2 provides an example of the propensity score weighting for one specific target judge. Cases that are similar to the target judge's cases will be given a larger weight, and cases that are not similar will be downweighed. The offense in the last row of Table 2 is sales of a weapon to minor. Because the target judge does not handle this specific offense, the last row is given a weight of zero. On the other hand, the target judge handles one case charged with robbery in the third degree with a prior record history score of 5, and the remaining judges handle 2 of the cases with same case attributes. Each of these comparison cases are given a weight of 0.5. In this

simplified example, the total weight for every crime type and criminal history combination for the comparison cases equals the number of cases for the target judge.

Table 2: *Illustrative example demonstrating the use of weights to align case features*

Cases for target judge	**History**	**Incarcerate**	**Case for remaining 423 comparison judges**	**History**	**Incarcerate**	**Weight**
Robbery 1st Degree	3	Yes				
Robbery 3rd Degree	5	Yes	Robbery 3rd Degree	5	Yes	0.5
			Robbery 3rd Degree	5	No	0.5
			Robbery 3rd Degree	4	Yes	0
Robbery 3rd Degree	0	No	Robbery 3rd Degree	0	Yes	1
Theft, taking movable property	2	No	Theft, taking movable property	2	No	2
Theft, taking movable property	2	Yes				
			Sale of starter pistols to minor	1	Yes	0

Note that in Table 2 the target judge handles a 1st degree robbery case, a case that no other judge handled. Cases such as these for which we can find no other judge handling a similar case must be dropped from the analysis. Without making tenuous assumptions, like grouping all robbery cases together regardless of degree, we do not have any information on how other judges would handle such a case. The analysis has to focus on the types of cases that are not in the exclusive domain of a single judge.[2] Accordingly, for judges whose features do not match their colleagues, we remove cases from the target judge where the propensity score is greater than the 99th percentile of the comparison group's propensity scores and then re-estimate the weights. These are cases that the comparison group are unlikely to handle. Even after propensity score weighting and dropping difficult to match cases, we may still have a judge with a case mix for which we cannot find a suitable set of comparison cases. Specifically, if we have a judge with cases that have features on which they differ from the comparison set of cases by more than 5 percentage points, then we do not create a benchmark for the judge[3].

[2] Under finite samples with sufficient overlap between the target and comparison group, normalized reweighting will exhibit smaller bias and variance reweighting than matching, but poor overlap will lead to less effective estimates (Busso, DiNardo, & McCrary, 2014).

[3] For categorical case features we compare the percentage of cases with that feature for the target judge and the comparison cases and compute the difference. For continuous case features we use the analogous Kolmogorov-Smirnov test statistic, which computes the largest percentage point difference in the cumulative distribution

Following the propensity score weighting stage, we utilize doubly robust estimation to produce a standardized score for the effect of the target judge on each sentencing outcome and discretionary decision. Doubly robust estimates use the propensity score weights as sampling weights in a regression model that includes the potential confounders in order to estimate the judge effect. This approach protects against model misspecification and provides a consistent estimate of the treatment effect with a correctly specified propensity score model or correctly specified outcomes regression model (Bang & Robins, 2005; Ho, Imai, King, & Stuart, 2007; Bang & Robins, 2008). These regression models produce doubly-robust z-scores measuring the how much the target judge deviates from their benchmark on the outcome. For each of the 424 judges we produce ten doubly robust z-scores, one for each of the sentencing outcomes and discretionary decisions. For dichotomous sentencing outcomes and discretionary decisions, we use a propensity score weighted linear probability model. For logged sentence length and fines, we use a propensity score weighted OLS model. After iterating through each judge, we produce a judge-level dataset with standardized z-scores for each sentencing outcome and discretionary decision. Table 3 provides an example of the judge-level dataset with z-scores. Each row represents a single judge that was benchmarked. Each column contains a z-score related to one of the sentencing decisions, which measures the magnitude to which a judge's decision rate differs from his or her benchmark. For example, Judge 10 has an incarceration z-score of -2.5, indicating that they are substantially less likely than their peers to use incarceration in sentencing. Judge 10 also shows that they have a z-score of 3.5 for downward departures. In short, this judge appears to be much more lenient in their use of incarceration and more likely to use downward departures or mitigating factors than their statistical benchmark of judicial peers.

Table 3: *z-scores for an example set of judges representing the magnitude of difference between the target judge and his or her benchmark on sentencing decisions*

Judge	Incarceration rate.	Incarceration. Length	Prison rate	Prison Length	Fine rate	Enhancement rate	Use of Alternative rate	Upward Deviation	Downward Deviation
1	0.9	0.8	0.5	0.5	1.1	-1.3	0.3	0.2	-0.5

functions for the target judge's cases and the comparison cases. We do not benchmark 54 judges who had percentage point differences or KS statistics exceeding 0.05. These judges account for only 2.5% (n=11,259) of cases (2.5%) in the analysis.

3	0.1	0.2	-0.4	-0.4	-1.1	-2.4	-0.9	0.4	0.9
6	1.0	0.7	0.0	0.0	0.1	0.4	-1.8	-0.6	-0.1
7	1.6	1.6	-0.2	-0.1	-0.5	0.1	-4.9	-0.7	0.1
8	0.0	-0.1	-0.2	-0.3	-0.2	0.6	0.2	0.2	0.4
10	-2.5	-3.4	-3.4	-3.2	-0.8	-0.6	0.4	-1.3	3.5

Notes: This table shows judges who have been benchmarked. The z-scores are scaled to have a mean of 0 and standard deviation of 1. The z-score for fine amount is not shown.

To assess the association between discretionary decisions and deviations from a judge's benchmark, we use the judge-level dataset and scale the z-scores to have a mean of 0 and standard deviation of 1. We scale the z-scores so each component is standardized on the same metric. We then regress the scaled scores for the outcome variables on the scaled scores of the discretionary variable. For each primary sentencing outcome, we use the regression model shown in (1) to examine the relationship between a judge's sentencing z-score (a measure of how much of an outlier they are on a sentencing outcome) and their z-scores on discretionary case decisions.

$$z_{\text{sentence},i} = \beta_0 + \beta_1 z_{upward,i} + \beta_2 z_{\text{downward},i} + \beta_3 z_{\text{alternative},i} + \beta_4 z_{\text{enhancement},i} + \beta_5 z_{\text{fine},i} + \varepsilon_i \quad (1)$$

$z_{\text{sentence},i}$ is the standardized score for the use of incarceration, the use of prison, incarceration length, or prison length for judge *i*. It measures whether and by how much Judge *i* is an outlier on the use of that sentencing option compared to Judge *i*'s benchmark. $z_{upward,i}$ is the z-statistic that measures whether Judge *i* is an outlier on upward deviations, including an aggravated sentence or a sentence above the recommended range. The remaining covariates measure whether Judge *i* is an outlier on the other discretionary decisions.

Given the heterogeneity in cases, estimating the effect of judges on sentencing decisions requires an approach that can disentangle the impact of observable, case attributes. Prior studies have used multi-level models, leveraged the random assignment of cases, or used exogenous changes in policies to estimate inter-judge disparities (Johnson, 2006; Yang, 2014; Anderson, Kling, & Stith, 1999; Yang, 2015). Studies using multi-level models can decompose the amount of sentencing variation attributable to judges; however, they do not provide precise estimates for each individual judge's effect on sentencing. Similarly, studies leveraging random case assignment or exogenous changes in policies can provide estimate of inter-judge disparities, but these estimates are restricted to certain geographic locations or time frames. Simply fitting a model to

estimate judge-fixed effects in jurisdictions without random assignment will lead to biased estimates due to the lack of common support (i.e. varying case compositions). Moreover, the judge-fixed effects will boil down to an arbitrary effect based on the omitted, reference judge. Telling Judge "A" that he sentences more punitively than Judge "B" and equally with Judge "C" provides limited value. Rather, it is more beneficial for a judge to understand how his or her sentencing practices differ from judges who handle the same type of cases. Benchmarking provides another approach for evaluating inter-judge disparities by peeling back the layers and measuring how judges differ on discretionary decisions and the eventual sentencing outcomes.

7 | RESULTS

7.1 | Balanced Comparison Group

After weighting the cases for the comparison group of each target judge, the case features of the comparison groups closely match the features for each target judge. The average, maximum difference on case features for each judge drops from 0.348 to 0.023. This means that the case feature for which the target judge matches the benchmark cases the least differs by 2.3 percentage points. Table 4 provides an example of balance between a target judge and comparison cases after weighting. The target judge almost exclusively handles DUI cases under Title 75 Pa.C.S.A. Vehicles § 3802. Weighting brings down the maximum difference on case features from 0.52 to 0.01. The target judge's benchmark is now effectively 97.2 percent DUI cases. The KS statistics show that after weighting the comparison cases have features whose distributions are closely matched to the target judge.

Table 4: *Example Balance Table for an Example Target Judge on a Selection of Case Features*

Variable	Target Judge (n = 1,091)	Comparison Cases (n = 175,430)	Weighted Comparison Cases (ESS = 11,974)	Weighted KS
Age at Date of Sentence (average years)	36.98	34.71	36.94	0.01
Age at Date of Offense (average years)	36.23	33.87	36.00	0.01
Year when case was sentenced (%)				
2009	0.33	0.26	0.33	0.00
2010	0.39	0.25	0.38	0.01
2011	0.28	0.24	0.28	0.00

2012	0.00	0.25	0.01	0.01
Offense Code				
75 3802a1 (DUI general impairment)	0.14	0.08	0.14	0.00
75 3802a2 (DUI w/ BAC 0.08% to 0.10%)	0.02	0.02	0.02	0.00
75 3802b (DUI w/ BAC 0.10% to 0.16%)	0.20	0.10	0.20	0.00
75 3802c (DUI w/ BAC over 0.16%)	0.52	0.19	0.52	0.00
75 3802d (DUI controlled substance)	0.08	0.07	0.07	0.00
18 2701b (Simple Assault)	0.00	0.10	0.00	0.00
18 3929biii (Retail Theft – Misd 1st)	0.00	0.05	0.00	0.00
18 3929biv (Retail Theft – Fel 3rd)	0.00	0.04	0.00	0.00
35 780-113 16 (Possess controlled substance not registered)	0.00	0.11	0.00	0.00
35 780-113 32 (Use or possession of drug paraphernalia for purpose of planting)	0.00	0.08	0.01	0.00
Mandatory Minimum				
30 5502c11i	0.00	0.00	0.00	0.00
75 3802	0.98	0.46	0.97	0.01
Attempted Offense	0.00	0.00	0.00	0.00
Offense Grade				
F	0.00	0.00	0.00	0.00
F-3	0.00	0.07	0.00	0.00
M	0.65	0.50	0.66	0.01
M-1	0.34	0.24	0.33	0.01
M-2	0.01	0.16	0.01	0.01
M-3	0.00	0.03	0.00	0.00
Offense Gravity Score				
1	0.65	0.42	0.65	0.00
2	0.00	0.08	0.01	0.00
3	0.01	0.33	0.02	0.01
5	0.33	0.17	0.32	0.01
6	0.00	0.00	0.00	0.00
Prior Record Score				
0	0.75	0.54	0.75	0.00
1	0.09	0.13	0.09	0.00

2	0.05	0.10	0.05	0.01
3	0.03	0.06	0.03	0.00
4	0.02	0.05	0.02	0.00
5	0.05	0.09	0.04	0.01
Repeat Felony	0.01	0.02	0.01	0.00
Sentencing Level				
1	0.54	0.30	0.55	0.00
2	0.31	0.52	0.32	0.01
3	0.12	0.15	0.11	0.01
4	0.03	0.03	0.02	0.00
Other Case Attributes				
Offender Eligible for State Intermediate Punishment	0.99	0.86	0.99	0.01
Offender Eligible for County Intermediate Punishment	1.00	1.00	1.00	0.00
Offender Eligible for Boot Camp	0.01	0.01	0.01	0.00
Presentence Investigation	0.65	0.21	0.64	0.01
Sexual and Violent Predator	0.00	0.00	0.00	0.00
Full Drug and Alcohol Evaluation	0.02	0.10	0.03	0.01
Preliminary Drug and Alcohol Evaluation	0.06	0.23	0.07	0.01
Drug Dependent	0.01	0.09	0.02	0.01
Drug Type - None	0.99	0.81	0.99	0.00
Drug Type – Other	0.01	0.19	0.01	0.00
Prior # of Misdemeanors	0.65	1.18	0.65	0.01
Other Convictions Within Case (average count)				
Felony 3	0.01	0.03	0.01	0.00
Felony Ungraded	0.00	0.00	0.00	0.00
Misdemeanor 1	0.02	0.06	0.03	0.00
Misdemeanor 2	0.03	0.09	0.04	0.01
Misdemeanor 3	0.02	0.05	0.02	0.00
Misdemeanor	0.08	0.12	0.09	0.01

Notes: We do not show in this table 13 offenses that account for less than 2 percent of the target judge's offenses. We also do not show in this table the 180 variables prior history variables.

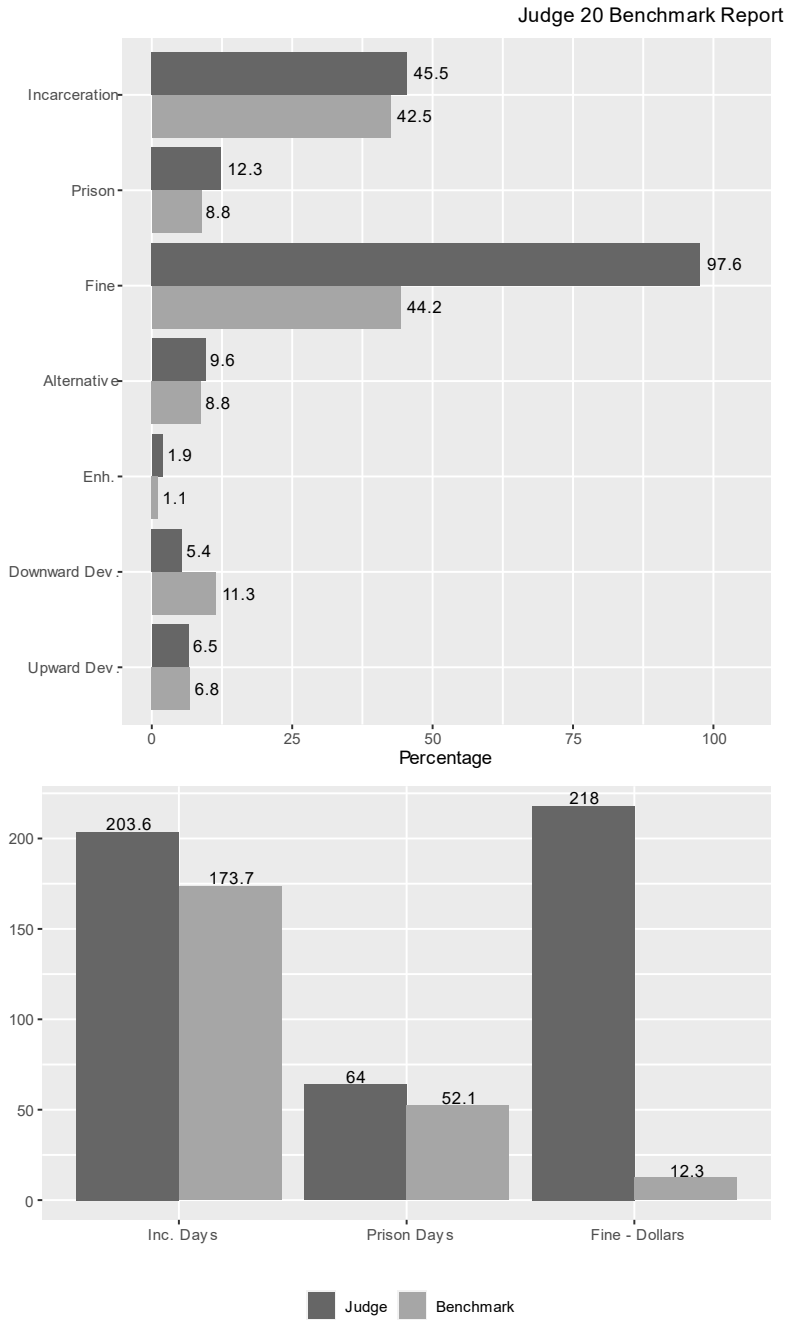
7.2 | Individualized Feedback Report

Once a benchmark is created for each judge, we can create individualized feedback reports that describe how judges differ from their benchmark and highlight sentencing decisions that contribute to disparities. These detailed reports allow judges to diagnose which margins can be adjusted and which margins are structural. Figure 1 provides an example of a benchmark report. The left-hand side of Figure 1 displays the difference between the judge and his or her benchmark. In this example, Judge 20 uses incarceration at a rate relatively close to the benchmark, but Judge 20's rate of prison sentences is about 3.5 percentage points above the benchmark. The judge also imposes fine in almost all cases in which the average fine amount is roughly 200 dollars more than the benchmark. The top, right-hand side corner reveals the percentile of Judge 20 among the 424 judges based on the z-score of each sentencing decision.

The percentile legend in Figure 1 provides another measure that quantifies the magnitude of difference across sentencing decisions. Judge 20 is in the 93rd percentile in both fines and prison usage. Judge 20 uses alternative sanctions at about the same rate as the benchmark and applies upward deviations at a rate similar to the benchmark. If Judge 20 wants to adjust the prison usage rate, Judge 20 can potentially substitute prison with jail sentences or consider cases that warrant a mitigating factor at a higher rate.

The bottom, right-hand side of Figure 1 provides context on the case composition that Judge 20 handles and the case composition of the benchmark. The case characteristics are closely matched with the percentage point difference no greater than 1. Putting the case compositions side-to-side helps communicate to the judge that he/she is being compared against a similar set of cases. This information can also be used by judges to identify cases where they feel their sanctioning options are limited or ineffective. Subsequently, the judge can work with the wider criminal justice system to inform how to address these specific case types. For instance, Judge 20 may be dealing with property cases that have more serious criminal histories and recidivism rates. The utility of the benchmark report stems from its capacity to report on a wide array of information that can assist judges in their assessment of sentencing decisions.

Figure 1: *Example benchmark report that demonstrates the variation in sentencing decisions.*



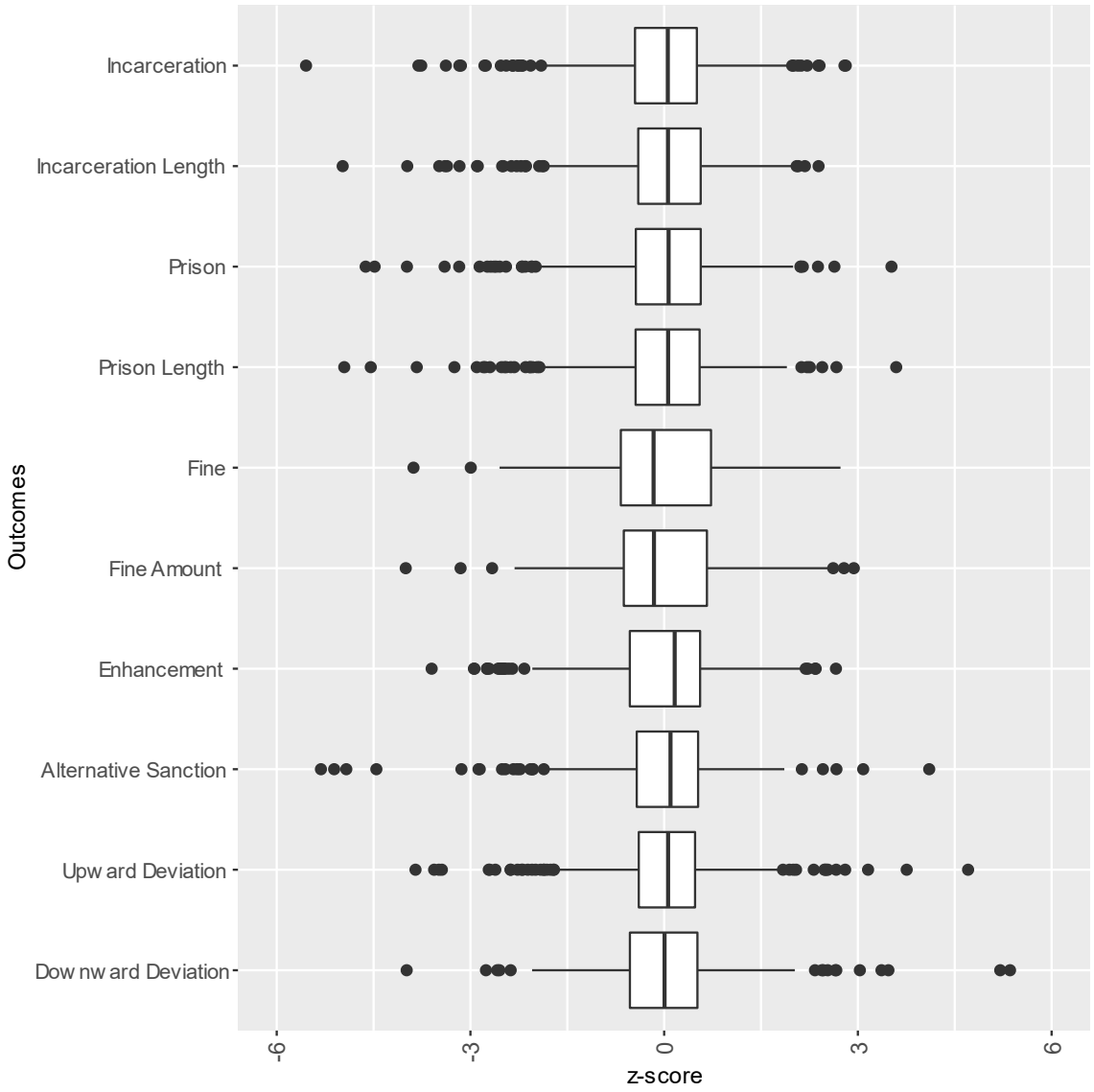
Outcomes	Percentile
Incarceration	57
Prison	93
Inc. Length	60
Prison Length	93
Fine	97
Fine Amount	97
Alternative	67
Enhancement	88
Upward Deviation	54
Downward Deviation	8

Case Attributes	Judge 20	Benchmark
Offense Gravity Score	3.84	3.86
PRS 0	0.46	0.45
PRS 1	0.13	0.14
PRS 2	0.12	0.12
PRS 3	0.07	0.08
PRS 4	0.06	0.06
PRS 5	0.13	0.13
PRS Repeat Violent	0.00	0.00
PRS Repeat Felony	0.02	0.03
Misdemeanor	0.79	0.79
Felony	0.21	0.21
Mandatory Minimum	0.22	0.23
Persons	0.15	0.15
Property	0.27	0.27
Drug	0.16	0.16
Weapons	0.01	0.01
Sex Offense/Rape	0.03	0.03
Eligible for SIP	0.83	0.83
Eligible for CIP	0.95	0.96
Eligible for Bootcamp	0.04	0.03
Drug Dependent	0.23	0.22
Mult. Drug Priors	0.00	0.00
Drug and Alcohol Eval.	0.18	0.18

Notes: The two, left-hand side panel compares the target judge's sentencing rate to the sentencing rate of his/her benchmark. The top right panel provide the judge's percentile for z-score on each sentencing decision. The bottom right panel compares the distribution of case features of the target judge to his/her benchmark.

We create individualized sentencing reports for all 424 judges. Figure 2 summarizes the extent to which judges deviate from the benchmark and displays boxplots of the 424 z-scores for the ten sentencing decisions. The black points highlight outlier judges with z-scores that are 1.5 times below or above the interquartile range of the 25th and 75th percentile. Outlier judges who deviate above and below the benchmark appear for all sentencing decisions except the use of fines. In the next section, we examine how deviations from the benchmark for different sentencing decisions are correlated with each other. By pairing the individualized sentencing reports with a guidance component, the sentencing feedback system aims to move outlier judges or judges that substantially impact sentencing disparities closer to the norm.

Figure 2: *Boxplots of z-scores for each of the sentencing decisions. Points beyond the whiskers indicate an outlier judge who substantially deviates from his/her benchmark.*



7.3 | Guidance Component:

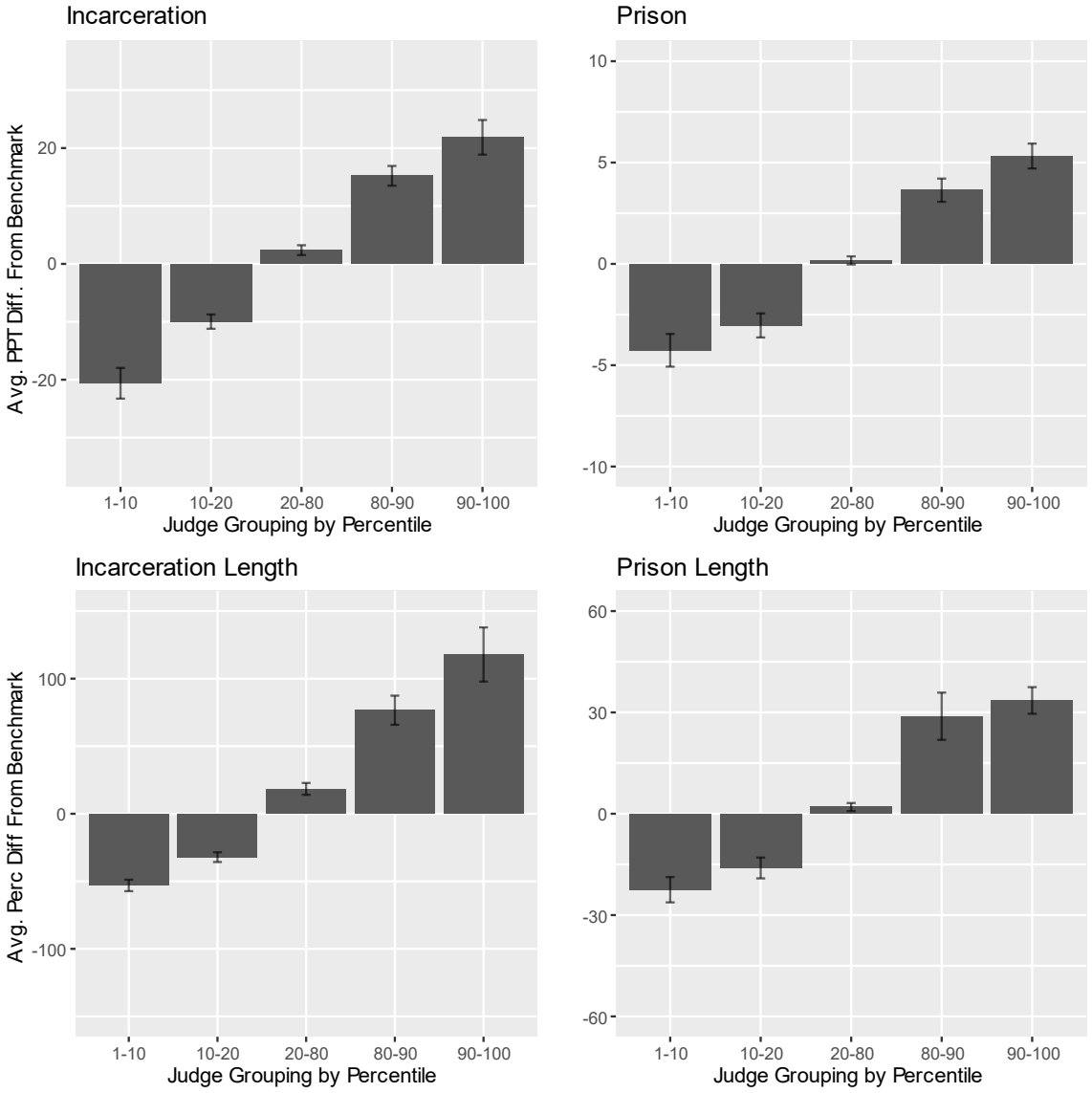
The guidance component of the sentencing report provides an overview of judges and the association between discretionary decisions and sentencing outcomes. The first part of the

guidance component descriptively compares judges with different propensities on the primary sentencing outcomes, and the second part uses regression models to assess how discretionary decisions influence sentencing outcomes. We start with the descriptive portion. For the four primary sentencing outcomes, we assign judges to five groupings based on the percentile of the corresponding z-score: 1st - 10th percentile, 10th - 20th percentile, 20th - 80th percentile, 80th-90th percentile, and 90th-100th percentile. Once judges are assigned to their grouping, we calculate the average percentage point difference from the benchmark.

Figure 3 shows the average percentage point difference for each judge grouping compared to his or her benchmark on the primary sentencing outcomes. The error bars encompass two standard errors from the groups' mean. We find substantial variation across judges. The top left panel of Figure 3 shows the percentage point difference in the use of incarceration compared to each judge's benchmark. The bottom 10th percentile of judges (n=43 judges) based on the z-score of the use of incarceration were 21 percentage points less likely than their benchmark to impose a jail or prison spell. Conversely, the top 10th percentile were 22 percentage points more likely to sentence an individual to incarceration compared to their benchmark.

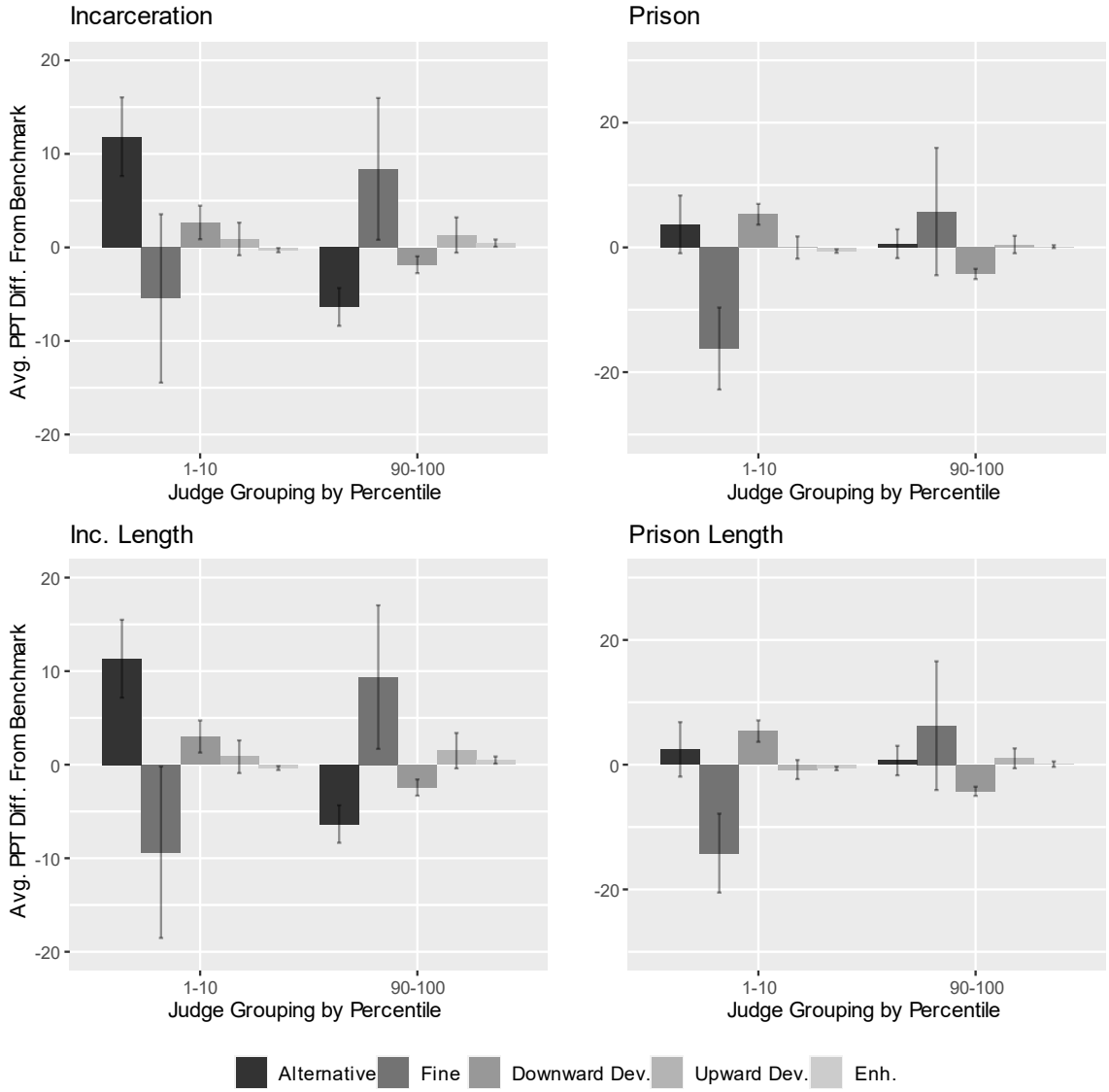
The bottom row of Figure 3 corresponds to sentence length as measured by percent differences. Judges ranked in the 90th percentile or above for incarceration length impose sentences that are 118 percent larger than their benchmark. This poses a question of how the bottom 10 and upper 90 percentile arrive at sentencing outcomes that are so different from their peers.

Figure 3: *Differences in sentencing outcomes compared to benchmark based on judges grouped by the percentile of the corresponding z-score.*



Note: Judges are grouped by the percentile of their z-scores for each of the primary sentencing outcomes. Error bars encompass a two standard error difference from the group means.

Figure 4: *Differences in discretionary decisions compared to benchmark for the 1st – 10th percentile and 90th – 100th percentile group of the primary sentencing outcomes.*



Note: Judges are grouped by the percentile of their z-scores for each of the primary sentencing outcomes. Error bars encompass a two standard error difference from the group means.

Figure 4 compares the average difference in discretionary decisions from a judge's benchmark for the top 90th percentile of judges versus the bottom 10th percentile of judges based on the earlier groupings of the four primary outcomes. The top left panel of Figure 4 shows that on average, judges with low propensities to incarcerate impose alternative sanctions 12 percentage points above the benchmark, give fines at a rate 5 percentage points below the benchmark, apply

mitigating factors or downward departures 3 percentage points above the benchmark, apply aggravating factors or upward departures 1 percentage points above the benchmark, and use the enhancement matrix 0.3 percentage points below the benchmark. The upper 90 percentile of judges show the inverse pattern where alternative sanctions and downward deviations are used at a lower rate than their benchmark. In the top right panel, judges varying in their propensity to impose prison sentences showed sizeable differences on their use of downward deviation. The two group of judges do not appear to differ on the use of alternative sanctions. For sentence length, the general patterns hold but the differences on downward deviations compared to the benchmark are slightly larger.

It is important to note that the use of jail, prison, or alternative sanctions is a function of both county resources and judicial discretion. Alternative sanctions can lead to lower incarceration rates while maintaining some degree of deterrence and accountability, but it needs to be an available sanction option. In counties with higher capacities, judges can adjust sentencing practices to influence both the number of people who serve custodial sentences and the length of custodial sentence. In counties where correctional resources are constrained, judges may have limited influence on either factor.

Table 4 displays the results for Equation 1 which uses a robust regression[4] and robust standard errors. Each column contains the estimates for the association between the discretionary decisions and one of the four primary sentencing outcomes. The coefficients can be interpreted as the standard deviation change in z-scores of the dependent variable given a 1 unit change in standard deviation of the z-score of the discretionary decision. Column 1 shows that a 1 standard deviation increase in the z-score that measures a judge's usage of alternative sanctions compared to a judge's benchmark is associated with a 0.692 standard deviation decrease in the z-score that measures the judge's use of incarceration compared to a judge's benchmark. If the coefficient on alternative sanctions is small or positive, this association would indicate that judges are using alternative sanctions in a way that aligns with net-widening. On the other hand, if the coefficient on alternative sanctions is negative, this association suggests that judges are using alternative sanctions as a substitute for incarceration. The results strongly indicate the latter. Fines do not

[4] The robust regression uses a M-estimator with the default Huber psi function which downweighs outliers so that the coefficients are not driven by outlier judges.

appear to affect the use of incarceration. Column 2 examines incarceration length and finds a similar relationship except that the coefficients for upward and downward deviations are larger.

In Column 3 of Table 4 which analyzes the use of prison, the effect of alternative sanctions becomes muted suggesting that alternative sanctions are primarily used as a substitute for jail time rather than prison time. Downward deviations are stronger drivers of differential prison rates compared to upward deviations. Column 4 focuses on prison length and reiterates the relationship found in Column 2; upward and downward deviations are more influential with sentence length. The small coefficients on fines across all four columns suggest that heavier fine usage is not being substituted for incarceration, and in the case of prison sentence, it could be used to increase the sanctioning penalty. Judges who are marked as substantial contributors to disparities can use this portion of the guidance component to see which discretionary decisions will be helpful in moving closer to the norm.

Table 4: *Regression tables decomposing the association between discretionary decisions and incarceration and incarceration length.*

	Dependent Variables			
	Incarceration (1)	Incarceration Length (2)	Prison (3)	Prison Length (4)
Enhancement	0.170**	0.174**	0.086*	0.084*
	(0.032)	(0.028)	(0.039)	(0.037)
Alternative	-0.692**	-0.675**	-0.109*	-0.114*
	(0.054)	(0.047)	(0.048)	(0.048)
Fine	0.050	0.04	0.101*	0.099*
	(0.029)	(0.027)	(0.047)	(0.047)
Upward Dev.	0.338**	0.407**	0.299**	0.309**
	(0.033)	(0.034)	(0.047)	(0.043)
Downward Dev.	-0.398**	-0.483**	-0.621**	-0.647**
	(0.033)	(0.039)	(0.043)	(0.042)
Observations	424	424	424	424

Notes: Standard errors are in parentheses. p<0.05*, p<0.01**

Table 5 repeats the prior analysis, but it substitutes fines amount for the use of any fines. The coefficients in Table 5 closely match the coefficients in Table 4. The coefficients for fine amount in Column 3 and Column 4 are slightly larger. This reiterates that judges do not consider

higher imposed fees as a balance against the use of jail and prison. If policy-makers are concerned about the adverse effects of incarceration, alternative sanctions may provide an option for reducing incarceration contact and levels, but these reductions will be concentrated in jail facilities. If policy-makers are concerned with the adverse consequences of prison, the impact of increased capacity or usage of alternative sanctions will be more limited.

Table 5: *Regression tables decomposing the association between discretionary decisions and prison and prison length*

	Dependent Variables			
	Incarceration	Incarceration Length	Prison	Prison Length
Enhancement	0.168**	0.172**	0.082*	0.080*
	(0.032)	(0.027)	(0.038)	(0.037)
Alternative	-0.692**	-0.675**	-0.108*	-0.113*
	(0.054)	(0.047)	(0.047)	(0.047)
Fine Amount	0.050	0.04	0.105*	0.101*
	(0.029)	(0.027)	(0.049)	(0.047)
Upward Dev.	0.337**	0.407**	0.299**	0.308**
	(0.033)	(0.033)	(0.046)	(0.043)
Downward Dev.	-0.397**	-0.482**	-0.619**	-0.645**
	(0.033)	(0.04)	(0.044)	(0.042)
Observations	424	424	424	424

Notes: Standard errors are in parentheses. p<0.05*, p<0.01**

8 | DISCUSSION & CONCLUSION

Providing judges with a benchmark to assess how their sentencing decisions differs may be one mechanism for curtailing sentencing disparities and higher levels of incarceration. In Pennsylvania, Court of Common Plea judges operate under a system where they only receive feedback when they run for a judgeship or when they are up for a retention vote every 10 years. Given that the feedback does not incorporate sentencing information and is synthesized into coarse ratings of highly recommended, recommended, or not recommend, judges receive limited information that would allow them to mitigate sentencing disparities[5]. If we consider that judges

[5] The Pennsylvania Bar Association provides an evaluation for any individual who is being considered for election, retention, or appointment on ten factors: sufficient legal ability, the amount of trial or comparable experience to ensure knowledge of the rules, a record and reputation for excellent character and integrity, financial responsibility judicial temperament, mental and physical capacity, record of community involvement, administrative ability,

with a high propensity to incarcerate imposed a custodial sentence at a rate that is 22-percentage points higher than their benchmark with an average sentence length that is 118 percent longer, it is possible that thousands of individuals are arbitrarily receiving longer and harsher sentences. Without any feedback mechanisms, judges will find it difficult not only to change course but to diagnosis the magnitude of change required for more equitable sentencing.

Simultaneously, judicial feedback systems need to account for the heterogeneity in case composition as well as the numerous sentencing decisions that judges consider. Individualized reports giving detailed information to judges can help move judges closer to the norm. It circumvents the need for a top-down approach or significant changes to the sentencing guidelines to mitigate sentencing disparities. Other domains have used similar feedback systems to adjust social norms or the use of resources. For example, in the utilities sector, a company named OPOWER designs and delivers reports to households comparing their energy usage to neighbors and providing guidance on how to conserve energy. These reports led to a 2% reduction in energy consumption where the effect was equivalent to increasing the price of electricity between 11 to 20% (Allcott, 2011). Our proposed feedback system operates in a similar manner. We provide general guidance on the potential impact of discretionary decisions coupled with information tailored for the target judge. Without this latter component, a judge is more likely to dismiss the guidance as out of context. As we highlight for the example judge in Figure 1, a judge who is more punitive on one dimension is not necessarily punitive across all sentencing decisions.

We conduct back of the envelope calculations estimating the change in sentencing patterns if the top 20 percentile of judges reduced their differences with their benchmark by 50 percent. Statewide incarceration rates would drop 2 percentage points (47 to 45 percent) and the variance in incarceration rates would decrease by 21 percent. The average sentenced incarceration time would drop from 4.9 to 4.6 months. Statewide prison rates would decrease by 0.5 percentage points and the variance would decrease by 8 percent. The average sentenced prison time would drop by roughly a week. When this number is scaled up by the number of cases observed, the reduction is equivalent to roughly 12,600 fewer one-year prison sentences over the past decade. These are

devotion to the improvement of the quality of justice, and the demonstration of sound judgement in one's professional life

sizeable disparities that can be reduced by moving judges at the most punitive end of the distribution closer to the norm.

Overall, judges with lower propensities to incarcerate used alternative sentences as a substitute for jail. Interestingly, judges with the highest rates of alternative sentences worked in small to mid-size counties. Judges may perceive alternative sanctions as options that balance the need to maintain public safety while preserving scarce, correctional resources. The analyses also showed that mitigating factors and downward departures as opposed to deviations above the guideline exerted more influence in the use of incarceration and incarceration lengths. If these decisions drive sentencing disparities, sentencing commissions can use this information to adjust the recommended sentence. Benchmarking can equip judges and judicial entities with information to more effectively manage disparities within the criminal justice system.

While benchmarking provides a method for assessing individual judges, we note a few limitations to this approach. First, we provide benchmarks for sentencing dimensions that do not contain an ideal norm that judges should strive for. With other dimensions such as racial disparities or trial penalties, the ideal norm is to have the smallest disparity or trial penalty possible. With the use of incarceration and prison, it is unclear what the objective, ideal norm is. This approach can be problematic if the benchmark is composed of cases where judges made "poor" decisions. But again, it is difficult to formulate a criterion for good or bad decisions when it comes to the sentencing outcomes at hand. At the very least, moving judges at the tail end closer to the norm can improve equity and reduce sentencing disparities that undermine distributive justice. Second, the final sentencing outcomes are a function of both judicial decision-making and prosecutorial decision-making. The differences among judges may actually reflect the behavior of prosecutors. We are able to account for the use of mandatory minimums. However, we cannot disentangle the full effect of the two actors without information that captures all prosecutorial behavior. Third, benchmarking only accounts for observable, case features. Case attributes, such as the defendant's income level or the victim's race, could influence sentencing decisions which we cannot incorporate when building each judge's benchmark. These unobserved case features will bias estimates; however, the bias will need to be sizable to explain away differences on judges at the extreme ends of the distribution. Fourth, benchmarking does not account for the resources, norms, and restraints in a judge's jurisdiction. For example, smaller counties may be heavily reliant on

fines and fees and have limited local, jail capacity. A judge's sentencing options would then obviously be limited and give rise to higher rates of prison sentences and fines. For this reason, we include measures corresponding to both sentence usage and sentence severity. Judges have better knowledge of their constraints and can identify where sentencing decisions are adjustable. Future studies can apply benchmarking to judges within the same county to account for this issue. We choose to exclude the county covariate in order to include judges who work in smaller counties. If these judges are benchmarked within the same county, their counterparts could potentially consist of 1 or 2 judges which is a poor benchmark. Finally, for judges that handle cases that significantly differ from their colleagues, benchmarking will only be able to provide an assessment on cases that overlap.

Despite these limitations, we believe that benchmarking can provide judges with critical information necessary to manage correctional resources and disparities. Future work can apply internal benchmarking to other dimensions. For example, this method can be extended to race, method of disposition, or offense categories to see if judges differ in their treatment of these case attributes. Sentencing reports can also be improved by assessing measures of public safety and community well-being. If judges find that an increase in the use of alternative sanctions does not correspond to higher rates of recidivism, they may be more willing to apply an alternative sentence. Altogether, internal benchmarking provides a feedback mechanism that could nudge judges that substantially contribute to sentencing disparities closer to their peers.

Appendix

Table A1: *Distribution of Alternative Sanctions*

Alternative Sanction	N	Percent
Electronic Monitoring	39800	42.7%
House Arrest	25812	27.7%
Work Release	11585	12.4%
Individualized Services	3850	4.1%
State Intermediate Punishment	3184	3.4%
Boot Camp	2566	2.8%
Residence Rehab/Halfway House	2470	2.7%
Intensive Supervision	1955	2.1%
D & A Inpatient	925	1.0%
Day Reporting	645	0.7%
Outmate Program	213	0.2%
Work Camp	121	0.1%
DUI Court	63	0.1%
Drug Court	53	0.1%

Table A2: Example of Sentencing Matrix

Level	OGS	Example Offenses	Prior Record Score								
			0	1	2	3	4	5	RFEL	REVOC	AGG/MIT
LEVEL 5 State Incar	14	Murder 3 Inchoate Murder (SBI) Rape (victim <13 yrs)	72-SL	84-SL	96-SL	120-SL	168-SL	192-SL	204-SL	SL	~/-12
	13	Inchoate Murder (No SBI) Weapons Mass Destr-Use PWID Cocaine (>1,000 g)	60-78	66-84	72-90	78-96	84-102	96-114	108-126	240	+/- 12
	12	Rape-Forcible Compulsion IDSI-Forcible Compulsion Robbery-Inflicts SBI	48-66	54-72	60-78	66-84	72-90	84-102	96-114	120	+/- 12
	11	Agg Assault-Cause SBI Voluntary Manslaughter Sexual Assault PWID Cocaine (100-1,000 g)	36-54 BC	42-60	48-66	54-72	60-78	72-90	84-102	120	+/- 12
	10	Kidnapping Agg Indecent Assault F2 Arson-Person in Building Hom by Vehicle-DUI & Work Zone PWID Cocaine(50-<100 g)	22-36 BC	30-42 BC	36-48 BC	42-54	48-60	60-72	72-84	120	+/- 12
	9	Sexual Exploitation of Children Robbery-Commit/Threat F1/F2 Burglary-Home/Person Present Arson-No Person in Building	12-24 BC	18-30 BC	24-36 BC	30-42 BC	36-48 BC	48-60	60-72	120	+/- 12
LEVEL 4 State Incar/ RIP trade	8 (F1)	Agg Assaul -Attempt/Cause BI w/DW Theft (Firearm) Identity theft (3rd/+ & Vic>=60 yrs) Hom by Veh-DUI or Work Zone Theft (>=$500,00) PWID Cocaine (10-<50 g)	9-16 BC	12-18 BC	15-21 BC	18-24 BC	21-27 BC	27-33 BC	40-52	NA	+/- 9
LEVEL 3 State/ Cnty Incar RIP trade	7 (F2)	Robbery-Inflicts/Threatens BI Burglary-Home/No Person Present Assault by Prisoner Theft ($100,000-<$500,000) Identity Theft (3rd/subq) PWID Cocaine (5-<10 g)	6-14 BC	9-16 BC	12-18 BC	15-21 BC	18-24 BC	24-30 BC	35-45 BC	NA	+/- 6
	6	Agg Assault-Cause Fear of SBI Homicide by Vehicle Burglary-Not a Home/Person Prsnt Theft (>$25,000-<$100,000) Arson-Endanger Property PWID Cocaine (2<5 g)	3-12 BC	6-14 BC	9-16 BC	12-18 BC	15-21 BC	21-27 BC	27-40 BC	NA	+/- 6
LEVEL 2 Cnty Incar RIP RS	5 (F3)	Burglary F2 Theft (>$2000-$25,000) DUI (M1) PWID Marij (1-<10 lbs)	RS-9	1-12 BC	3-14 BC	6-16 BC	9-16 BC	12-18 BC	24-36 BC	NA	+/- 3
	4	Indecent Assault M2 Forgery (Money, Stocks) Weapon on School Property Crim Trespass F2	RS-3	RS-9	RS-<12	3-14 BC	6-16 BC	9-16 BC	21-30 BC	NA	+/- 3
	3 (M1)	Simple Assault-Attempt/Cause BI Theft ($200-$2000) Carrying Explosives Simple Possession	RS-1	RS-6	RS-9	RS-<12	3-14 BC	6-16 BC	12-18 BC	NA	+/- 3

LEVEL 1 RS	**2 (M2)**	Theft ($50-<$200) Retail Theft (1st/2nd Offense) Bad Checks ($500-<$1,000)	**RS**	**RS-2**	**RS-3**	**RS-4**	**RS-6**	**1-9**	**6- <12**	**NA**	**+/- 3**
	1 (M3)	Most Misd. 3's;Theft (<$50) Disorderly Conduct Poss Small Amount Marij	**RS**	**RS-1**	**RS-2**	**RS-3**	**RS-4**	**RS-6**	**3-6**	**NA**	**+/- 3**

1. Designated areas of the matrix indicate restrictive intermediate punishments may be imposed as a substitute for incarceration.
2. When restrictive intermediate punishments are appropriate, the duration of the restrictive intermediate punishment programs are recommended not to exceed the guideline ranges.
3. When the range is RS through a number of months (e.g. RS-6), RIP may be appropriate.
4. All numbers in sentence recommendations suggest months of minimum confinement pursuant to 42 Pa.C.S. 9755(b) and 9756(b).
5. Statutory classification (e.g., F1, F2, etc.) in brackets reflect the omnibus OGS assignment for the given grade.

Key:

BC	=	boot camp
CNTY	=	County
INCAR	=	Incarceration
PWID	=	possession with intent to deliver
REVOC	=	repeat violent offender category
RFEL	=	repeat felony 1 and felony 2 offender category
RIP	=	restrictive intermediate punishments
RS	=	restorative sanctions
SBI	=	serious bodily injury
SL	=	statutory limit (longest minimum sentence)
~	=	no recommendation (aggravated sentence would exceed statutory limit)
<;>	=	less than; greater than

References

Abrams, D. S., Bertrand, M., & Mullainathan, S. (2012). Do Judges Vary in Their Treatment of Race? *The Journal of Legal Studies, 41*(2), 347-383.

Abrams, D. S., Galbiati, R., Henry, E., & Philippe, A. (2019). Decisions, When in Rome... on Local Norms and Sentencing. *SSRN Electronic Journal*, 1-52.

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics, 95*, 1082 - 1095.

American Bar Association. (2005). *Black Letter Guidelines for the Evaluation of Judicial Performance.* Chicago: American Bar Association.

Andersen, S. S. (2001). Judicial Retention Evaluation Programs. *Loyola of Los Angeles Law Review*, 1375-1389.

Anderson, J. M., Kling, J. R., & Stith, K. (1999). Measuring Interjudge Sentencing Disparity: Before and After the Federal Sentencing Guidelines. *The Journal of Law and Economics, 42*(S1), 271-308.

Ashenfelter, O., Eisenberg, T., & Schwab, S. J. (1995). Politics and the Judiciary: The Influence of Judicial Background on Case Outcomes. *The Journal of Legal Studies, 24*(2), 257-281.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal. *Biometrics, 61*(4), 962-972.

Bang, H., & Robins, J. M. (2008). Correction to "Doubly Robust Estimation in Missing Data and Causal Inference Models". *Biometrics, 61*, 962–972.

Beijersbergen, K. A., Dirkzwager, A. J., & Nieuwbeerta, P. (2016). Reoffending After Release: Does Procedural Justice During Imprisonment Matter? *Criminal Justice and Behavior, 43*(1), 63-82.

Bierschbach, R. A., & Bibas, S. (2017). Rationing Criminal Justice. *Michigan Law Review, 116*(2), 187-245.

Blackwell, B. S., Holleran, D., & Finn, M. A. (2008). The Impact of the Pennsylvania Sentencing Guidelines on Sex Differences in Sentencing. *Journal of Contemporary Criminal Justice, 24*(4), 399-418.

Bogan, K., & Factor, D. (1997). Oregon Guideline, 1989-1994. In M. Tonry, & K. Hatlestad, *Sentencing Reform in Crowded Times* (pp. 49-57). New York: Oxford University Press.

Bornstein, A., Marcus, A., Curtis, R., Rivera, S., & Swaner, R. (2001). Disrespect and the Experience of Injustice. *Annual Review of Psychology, 52*(1), 527-553.

Bornstein, A., Marcus, A., Curtis, R., Rivera, S., & Swaner, R. (2016). Tell It to the Judge: Procedural Justice and a Community Court in Brooklyn. *Political and Legal Anthropology Review, 39*(2), 206-225.

Brody, D. (2008). The use of judicial performance evaluations to enhance judicial accountability, judicial independence, and public trust. *Denver Law Review*, 1-42.

Bushway, S. D., & Owens, E. G. (2013). Framing Punishment: Incarceration, Recommended Sentences, and Recidivism. *Journal of Law and Economics, 56*, 301-331.

Busso, M., DiNardo, J., & McCrary, J. (2014). New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators. *The Review of Economics and Statistics, 96*(5), 885–897.

Cohen, A., & Yang, C. S. (2019, 11 1). Judicial Politics and Sentencing Decisions. *American Economic Journal: Economic Policy*, 160-191.

Depew, B., Eren, O., & Mocan, N. (2017). Judges, Juveniles, and In-Group Bias. *Journal of Law and Economics, 60*, 209-239.

Dippel, C., & Poyker, M. (2019). How Common are Electoral Cycles in Criminal Sentencing? *National Bureau of Economic Research*, 1-19.

Dollar, C. B., Ray, B., Hudson, M. K., & Hood, B. J. (2018). Examining changes in procedural justice and their influence on problem-solving court outcomes. *Behavioral Sciences & the Law, 36*(1), 32-45.

Emeriau, M. (2019). Learning to be Unbiased: Evidence from the French Asylum Office. *Working Paper*, 1-63.

Engen, R. L. (2009). Assessing determinate and presumptive sentencing-Making research relevant. *Criminology & Public Policy, 8*(2), 323-336.

Eren, O., & Mocan, N. (2018). Emotional Judges and Unlucky Juveniles. *American Economic Journal: Applied Economics, 10*(3), 171-205.

Fischman, J. B., & Schanzenbach, M. M. (2012). Racial Disparities Under the Federal Sentencing Guideliness: The Role of Judicial Discretion and Mandatory Minimums. *Journal of Empirical Legal Studies, 9*(4), 729-764.

Gazal-Ayal, O., & Sulitzeanu-Kenan, R. (2010). Let My People Go: Ethnic In-Group Bias in Judicial Decisions-Evidence from a Randomized Natural Experiment. *Journal of Empirical Legal Studies, 7*(3), 403-428.

Gill, R. D., Lazos, S. R., & Waters, M. M. (2011). Are Judicial Performance Evaluations Fair to Women and Minorities? A Cautionary Tale from Clark County, Nevada. *Law and Society Review, 45*(3), 731-759.

Grffin, E., Rushin, S., & Colquitt, J. (2019). The Effects of Voluntary and Presumptive Sentencing Guidelines. *Texas Law Review*(1), 1-66.

Haynes, S. H., Ruback, B., & Cusick, G. R. (2010). Courtroom Workgroups and Sentencing: The Effects of Similarity, Proximity, and Stability. *Crime & Delinquency, 56*(1), 126-161.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Modle Dependence in Parametric Causal Inference. *Political Analysis, 15*, 199-236.

Huber, G. A., & Gordon, S. C. (2004). Accountability and Coercion: Is Justice Blind when It Runs for Office? *American Journal of Political Science, 48*(2), 247-263.

IAALS. (2020, 10 06). *Judicial Performance Evaluation in the States*. Retrieved from https://iaals.du.edu/judicial-performance-evaluation-states

Institute for the Advancement of the American Legal System. (2016). *Transparent Courthouse Revisited: An Updated Blueprint for Judicial Performance Evaluation.*

Johnson, B. D. (2006). The Multi-level Context of Criminal Sentencing: Integrating Judge- and County-Level Influences. *Criminology, 44*(2), 259-298.

Johnson, B. D. (2014). Judges on Trial: A Reexamination of Judicial Race and Gender Effects Across Modes of Conviction. *Criminal Justice Policy Review, 25*(2), 159-184.

Johnson, B. D., & DiPietro, S. M. (2012). The Power of Diversion: Intermediate Sanctions and Sentencing Disparity Under Presumptive Guidelines. *Criminology, 50*(3), 811-850.

Koons-Witt, B. A. (2002). The Effect of Gender on the Decision to Incarcerate Before and After the Introduction of Sentencing Guidelines*. *Criminology, 40*(2), 297-328.

Kourlis, R. L., & Singer, J. M. (2007). Using judicial performance evaluations to promote judicial accountability. *Judicature, 90*(5), 200-207.

Lacasse, C., & Payne, A. A. (1999). Federal Sentencing Guidelines and Mandatory Minimum Sentences: Do Defendants Bargain in the Shadow of the Judge? *Journal of Law and Economics, 42*(S1), 245-270.

Lee, B. K., Lessler, J., & Stuart, E. A. (2009). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*, 337-346.

Leibovitch, A. (2016). Relative Judgments. *The Journal of Legal Studies, 45*(2), 281-330.

Levi, D. F., & Gulati, M. (2009). Judging Measures. *UMKC Law Review*, 381-413.

Lim, C. S. (2015). Media Influence on Courts: Evidence from Civil Case Adjudication. *American Law and Economics Review, 17*(1), 87-126.

Lim, C. S., Silveira, B. S., & Snyder, J. M. (2016). Do Judges' Characteristics Matter? Ethnicity, Gender, and Partisanship in Texas State Trial Courts. *American Law and Economics Review, 18*(2), 302-357.

Lim, C. S., Snyder, J. M., & Strömberg, D. (2015). The Judge, the Politician, and the Press: Newspaper Coverage and Criminal Sentencing across Electoral Systems. *American Economic Journal: Applied Economics, 7*(4), 103-135.

Mindock, M. R. (2019). Race, Fairness, and Co-Determination in Criminal Sentencing: Evidence From Sentencing Cohorts in Pennsylvania. *Working Paper*, 1-42.

Mitchell, K. L. (2017). State Sentencing Guidelines: A Garden Full of Variety. *Federal Probation*, 28-36.

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*, 285-314.

Paternoster, R., Brame, R., Bachman, R., & Sherman, L. W. (1997). Do Fair Procedures Matter? The Effect of Procedural Justice on Spouse Assault. *Law & Society Review, 31*(1), 163-204.

Pelander, J. A. (1998). Judicial Performance Review in Arizona: Goals, Practical Effects and Concerns. *Arizona State Law Journal, 30*(3), 643-760.

Philippe, A., & Ouss, A. (2018). “No Hatred or Malice, Fear or Affection”: Media and Sentencing. *Journal of Political Economy, 126*(5), 2134-2178.

Pina-Sanchez, J., Grech, D., Brunton-Smith, I., & Sferopoulos, D. (2019). Exploring the origin of sentencing disparities in the Crown Court: Using text mining techniques to differentiate between court and judge disparities. *Social Science Research, 84*, 1-13.

Poyker, M., & Dippel, C. (2019). Do Private Prisons Affect Criminal Sentencing? *National Bureau of Economic Research*, 1-29.

Ridgeway, G., & MacDonald, J. M. (2009). Doubly Robust Internal Benchmarking and False Discovery Rates for Detecting Racial Bias in Police Stops. *American Statistical Association, 104*(486), 661-668.

Ridgeway, G., & MacDonald, J. M. (2014). A Method for Internal Benchmarking of Criminal Justice System Performance. *Crime & Delinquency, 60*(1), 145–162.

Ridgeway, G., Moyer, R. A., & Bushway, S. D. (2020). Sentencing scorecards: Reducing racial disparities in prison sentences at their source. *Criminology & Public Policy*, 1-26.

Robbennolot, J. K. (2005). Evaluating Juries by Comparison to Judges: A Benchmark for Judging? *Florida State University Law Review, 32*(2), 1-44.

Schanzenbach, M. (2005). Racial and Sex Disparities in Prison Sentences: The Effect of District-Level Judicial Demographics. *The Journal of Legal Studies, 34*(1), 57-92.

Schanzenbach, M. (2018). Racial Disparities, Judge Characteristics, and Standards of Review in Sentencing. *Journal of Institutional and Theoretical Economics, 171*, 27-46.

Scott, & W., R. (2010). Inter-Judge Sentencing Disparity After Booker: A First Look. *Stanford Law Review, 63*(1), 1-66.

Sen, M. (2014). How Judicial Qualification Ratings May Disadvantage Minority and Female Candidates. *Journal of Law and Courts*, 33-65.

Sen, M. (2015). Is Justice Really Blind? Race and Reversal in US Courts. *Journal of Legal Studies, 44*(1), S187-S229.

Shayo, M., & Zussman, A. (2011). Judicial Ingroup Bias in the Shadow of Terrorism. *The Quarterly Journal of Economics, 126*(3), 1447-1484.

Sherman, L. W. (1993). Defiance, Deterrence, and Irrelevance: A Theory of the Criminal Sanction. *Journal of Research in Crime and Delinquency, 30*(4), 445-473.

Sprott, J. B. (2010). Trust and Confidence in the Courts: Does the Quality of Treatment Young Offenders Receive Affect Their Views of the Courts? *Crime & Delinquency, 56*(2), 269-289.

Starr, S. B. (2015). Estimating Gender Disparities in Federal Criminal Cases. *American Law and Economics Review, 17*(1), 127-159.

Steffensmeier, D., & Britt, C. L. (2001). Judges' Race and Judicial Decision Making: Do Black Judges Sentence Differently? *Social Science Quarterly, 82*(4), 749-764.

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

Tiede, L., Carp, R., & Manning, K. L. (2010). Judicial Attributes and Sentencing-Deviation Case: Do Sex, Race, and Politics Matter? *The Justice System Journal, 31*(3), 249-272.

Tita, G., & Ridgeway, G. (2007). The Impact of Gang Formation on Local Patterns of Crime. *Journal of Research in Crime and Delinquency, 44*(2), 208-237.

Tonry, M. (2013). Sentencing in America, 1975-2025. *Crime and Justice, 42*(1), 141-198.

Tyler, T. R. (2001). Trust and Law Abidingness: A Proactive Model of Social Regulation. *Boston University Law Review, 81*(2), 361-406.

Tyler, T. R., Lawrence, S., Strang, H., Barnes, G. C., & Woods, D. (2007). Reintegrative Shaming, Procedural Justice, and Recidivism: The Engagement of Offenders' Psychological Mechanisms in the Canberra RISE Drinking-and-Driving Experiment. *Law & Society Review, 41*(3), 553-586.

Vuolo, M., Wright, B. R., & Lindsay, S. L. (2019). Inmate Responses to Experiences With Court System Procedural and Distributive Justice. *The Prison Journal, 99*(6), 725-747.

Weisberg, F. H., & Hunt, K. S. (2007). Voluntary Sentencing Guidelines in the District of Columbia: Results of the Pilot Program. *Federal Sentencing Reporter, 19*(3), 208-218.

Welch, S., Combs, M., & Gruhl, J. (1988). Do Black Judges Make a Difference? *American Journal of Political Science, 32*(1), 126-136.

Yang, C. S. (2014). Have Inter-judge Sentencing Disparities Increased in an Advisory Guideline Regime? Evidence From Booker. *New York University Law Review, 89*(4), 1268-1342.

Yang, C. S. (2015). Free at Last? Judicial Discretion and Racial Disparities in Federal Sentencing. *Journal of Legal Studies, 44*(1), 75-110.

Yang, C. S. (2016). Resource Constraints and the Criminal Justice System: Evidence from Judicial Vacancies. *American Economic Journal: Economic Policy, 8*(4), 289-332.