RELATIVE JUDGMENTS

Adi Leibovitch*

ABSTRACT

The paper presents a theory of relative judgments, suggesting judges are evaluating individual cases based on how those cases are ranked in comparison to the other cases in their caseloads. As a result, a case can be ranked and viewed more severely by judges when judicial caseloads contain cases of lower gravity, and more leniently when judicial caseloads contain higher gravity cases. The paper develops a novel empirical identification strategy that exploits the properties of caseloads distribution under random assignment of cases as a source of exogenous variation in judicial exposure to gravity. Using data on criminal sentencing decisions, the paper constructs a matched sample of judges randomly located on different ends of the caseloads distribution, and demonstrates the existence of relative judgments bias in the sentencing decisions of judges exposed to different gravity of criminal behaviors in their caseloads. Judges exposed to lower levels of criminal gravity order higher absolute sentences, order sentences located higher in relation to the sentencing guidelines range, and are more likely to use the aggravated sentencing guidelines range or to depart above the sentencing guidelines recommendations, than judges exposed to higher levels of criminal gravity. The theory and results presented in the paper can have important implications for the structure of divisions in the criminal justice system. They may also have wider applicability to study judicial behavior in other questions and legal fields.

I. INTRODUCTION

Is the outcome of a particular case influenced by how that case fares in comparison to the other cases in a judge's caseload? Studies of context-dependence in decision making by juries found that jurors' decisions can be erratic, because jurors lack sight of

* Academic Fellow, Columbia Law School; JSD Candidate, The University of Chicago Law School. I am gratefully indebted to my JSD advisors, Omri Ben-Shahar and William Hubbard, for numerous conversations and invaluable advice. For helpful comments and discussions I thank Lisa Bernstein, Emiliano Catan, Adam Chilton, Lee Epstein, Valerie Hans, Josh Kleinfeld, Bill Landes, Saul Levmore, Jim Lindgren, Maria Macia, Anup Malani, Alan Miller, Ed Morrison, Jonathan Nash, Ariel Porat, Jeff Rachlinski, John Rappaport, Max Schanzenbach, Holger Spamann, Avishalom Tor, Stephane Wolton, and participants at the University of Chicago Law School Workshop on Judicial Behavior, Northwestern Law School Legal Scholarship Workshop, Haifa University Law and Economics Workshop, the American Law and Economics Association 25th Annual Meeting (Columbia Law School, May 15-16, 2015), the Tenth Annual Conference on Empirical Legal Studies (Washington University in St. Louis, October 30-31, 2015), and the First SELS Global Workshop for Junior Empirical-Legal Scholars (The Hebrew University of Jerusalem, December 16-17, 2015). I also thank the judges and court administrators at the courts of Cook County First and Sixth Municipal Districts for welcoming me into their courts. For support I thank the John M. Olin Foundation and the Coase-Sandor Institute for Law and Economics at the University of Chicago Law School. E-mail: aleibovitch@law.columbia.edu.

other cases against which they can form their evaluations of particular damage awards.[1] Very little is known, however, about decisions by judges. Some scholars speculate judges could outperform juries because unlike jurors they encounter multiple and varying sets of cases.[2] But that assumption has never been directly tested, and even its proponents caution that they propose judges *could* make more coherent decisions than juries, but they "are not aware of any direct evidence that judges *do* spontaneously consider particular cases in a broad context."[3]

Further, judicial exposure to the context of other types of cases can not only distinguish judges from jurors, but may also vary across judges. Judges in courts with different jurisdiction, in specialized courts, and even in divisions or dockets within a court can have quite different cases in their caseloads. If exposure to wider context through caseloads indeed affects judicial decisions, this invites the question of whether judges exposed to different cases in their caseloads will reach different decisions.

This question is material to the study of judicial behavior and to the operation of the courts. As far as different caseload compositions can be a source of bias leading to disparities of judgments, such a phenomenon has implications to any administrative decision changing the jurisdiction or case assignments across judges and courts, and to any study comparing case outcomes of judges sitting in different courts or handling different dockets. Yet despite its importance, the question how might judicial caseloads shape the decisions rendered by judges has received little scholarly attention.

Behavioral scholars found in experimental settings that the immediate comparison across cases can lead to local arbitrary judgments influenced by two possible opposite effects: contrast and anchoring. Several studies have found that contrast effect can cause the same behavior or harm to appear worse when compared against less serious cases, and vice versa. Parducci presented students with two lists containing eighteen wrongdoings each, and found the six misdeeds appearing in both lists were rated more leniently by students when judged in a context of other graver acts than when encountered in the context of relatively mild wrongdoings.[4] Kelman, Rottenstreich and Tversky showed how contrast and compromise effects can influence mock jurors' choice from a menu of punitive options for verdicts (choosing between possible offences for conviction) and sentencing (choosing between several options for

---

[1] Cass R. Sunstein, Daniel Kahneman and David Schkade, *Assessing Punitive Damages (With Notes on Cognition and Valuation in Law)*, 107 YALE L. J. 2071 (1998); Daniel Kahneman, David Schkade, and Cass Sunstein, *Shared outrage and erratic awards: The psychology of punitive damages*, 16 J. RISK AND UNCERTAINTY 49 (1998), 77-78; Cass R. Sunstein, Daniel Kahneman, David Schkade, and Ilana Ritov, *Predictably Incoherent Judgments*, 54 STAN. L. REV. 1153 (2002).

[2] *Id*.; Jeffrey J. Rachlinski and Forest Jourden, *The Cognitive Components of Punishment*, 88 CORNELL L. REV. 457 (2003), 477; Theodore Eisenberg, Jeffrey J. Rachlinski and Martin T. Wells, *Reconciling Experimental Incoherence with Real-World Coherence in Punitive Damages*, 54 STAN. L. REV. 1239 (2002) indeed provide empirical evidence suggesting the comparison between decisions by jurors and judges follows the predicted pattern if jurors' decisions are made in isolation and judicial decisions are made against a wider context of cases.

[3] Cass R. Sunstein, Daniel Kahneman, David Schkade, and Ilana Ritov, *Is Incoherence Outrageous?* 54 STAN. L. REV. 1293 (2002), 1295.

[4] Allen Parducci, *The Relativism of Absolute Judgments,* 219 SCIENTIFIC AMERICAN 84 (1968)

sentences).[5] Sunstein et al. similarly report that the immediate comparison between cases "makes a serious case appear more serious than it would on its own and makes a milder case appear milder" thereby affecting punitive awards decisions,[6] and Rachlinski and Jourden found a similar effect on what sentences students thought were appropriate.[7] However, for binary decisions, Rachlinski and Jourden found no contrast effect on the evaluation of justification for the death penalty by students when the same case was evaluated together with a weaker or a stronger case.[8]

An opposite effect that can arise out of context-dependent decision making is of anchoring – when an irrelevant numerical reference point serves as an anchor affecting following numerical estimations. Rachlinski, Wistrich and Guthrie recently proposed that in sentencing decisions, the sentence ordered in one case could serve as such anchor that affects the determination of a following sentence. In an experiment conducted with real judges, the authors found that "a lengthy sentence might serve as an anchor that lengthens a subsequent sentence assigned to a less serious crime,"[9] but only found partial support for the effect of a short sentence as an anchor that shortens a subsequent sentence of a more serious crime.[10]

Lab experiments, however, can only partially inform the predictions regarding judicial behavior. The experimental studies rely mostly on lay subjects, and only look at the immediate comparison across cases in settings controlled to manipulate the facts relevant for triggering the effect of interest. Such immediate short-term comparison can lead to arbitrary judgments. But it offers less predictive value for what would be the effect of context-dependence in the circumstances characterizing judicial decision making - when judges handle a wider spectrum of cases over a longer period of time.

Bringing real world evidence to bear on the question, qualitative case studies of the criminal justice system support the existence of contrast effect on the evaluation of behaviors, documenting how each agency always have some crimes defined as "the most severe" and some as "most trivial" in relation to its entire caseload.[11] The case

---

[5] Mark Kelman, Yuval Rottenstreich and Amos Tversky, *Context-dependence in legal decision making*, 25 JOURNAL OF LEGAL STUDIES, 287 (1996)

[6] Sunstein, Kahneman and Schkade, *supra* note 1, 2104. *See* also Sunstein, Kahneman, Schkade, and Ritov, *supra* note 1, 1176 – suggesting that when the graver harm is also from a more prominent category and the lesser harm is from a less prominent category: "the more prominent harm is assigned a lower rating and a lower dollar value when judged by itself than when directly compared to a harm of a less prominent kind."

[7] Rachlinski and Jourden, *supra* note 1 (finding that student subjects assigned higher sentences to a robbery case when compared against a fraud case than when judges alone, and lower sentences to the fraud case when compared against the robbery case than when judges by itself).

[8] *Id.*

[9] Jeffrey J. Rachlinski, Andrew J. Wistrich and Chris Guthrie, *Arbitrary Adjudication: How Anchoring and Scaling Distort Awards and Sentences* (presented at the University of Chicago conference on "Rational Choice Approach to judging", 2014)

[10] *Id.*, 33-34.

[11] For an extensive review of the literature, *see* Robert M. Emerson, *Holistic Effects in Social Control Decision-Making*, 17 LAW AND SOCIETY REVIEW 425, 428 (1983) (concluding that: "In a variety of social control settings, assessments of the "seriousness" of particular cases (on whatever organizationally relevant dimensions) tend to be made in relation to the kinds of cases regularly encountered in that particular setting. Thus, the decision to treat a case as an instance of something serious depends in part on the overall range and character of the case set processed by the agent or agency.")

studies, however, cannot completely distinguish to what extent might such classifications be attributed to different norms across different agencies or to allocating limited resources, rather than to purely psychological effects,[12] and they focus on non-judicial bodies such as police officers, prosecutors and parole boards.[13]

From the large literature on court communities comparing courts of different subject matter or geographic jurisdiction, a similar pattern emerges of a negative relationship between the scope of criminal gravity under the jurisdiction of the court and sentencing outcomes; courts that are exposed to offenses of lower gravity often have higher sentencing levels than courts hearing offenses of higher gravity.[14] While such studies cannot distinguish between the effect of exposure to gravity and other confounding variables characterizing the different organizational or local communities,[15] this paper is able to isolate the causal link between exposure to gravity and judicial behavior in real sentencing decisions, and can offer an additional, unifying, explanation to their findings.

The first contribution of the paper is to develop a theory of relative judgments, through its application to criminal sentencing decisions. Under the theory, judicial evaluation of criminal behavior – and the sentencing decision that follows – is made based on the relative ranking of the particular case in comparison to the other cases in the caseload of each judge. The same criminal behavior may be viewed as graver when compared to relatively mild offenses, and as milder when compared to more serious cases. Such relative comparison can lead judges who are exposed to lower levels of criminal behavior to evaluate the same cases more harshly and order more severe sentences than judges exposed to higher levels of crime. Further, under the theory, caseload effects on the evaluation of particular cases are manifested not only through the immediate comparison across cases, but also contribute to shaping judicial sentencing practices more generally. Consequently, context-dependence in judicial decision making can lead not just to arbitrarily different judgments, but to judgments that are biased in a predictable direction.

The paper then empirically tests the linkage between caseload exposure and sentencing outcomes. Testing for relative judgments in actual judicial decisions is of importance, since it is sometimes suggested that experimental findings do not

---

[12] For example, Emerson, *id.*, refers to the findings of William B. Sanders, DETECTIVE WORK: A STUDY OF CRIMINAL INVESTIGATIONS (1977) with regard to police officers – finding that in a major crime detail battery cases were viewed as "small cases" and rarely investigated, but at the juvenile detail battery cases were regarded as "big cases" in comparison to the usual crimes they handled (mainly petty theft and malicious mischief). Such different prioritizing can result from the psychological impact of frequency of exposure to different scope of crime, but also from efficient resource allocation of police officers within each unit to investigate crimes based on decreasing ranking of importance.

[13] While Emerson, *id.*, extensively reviews the effect of the exposure to different scope of criminal behavior by prosecutors, defense attorneys, police officers and probation authorities, it does not analyze possible behavioral influences on judges, and only regards courts briefly focusing on judges' strategic (deliberate) decisions about the sequence of hearing cases.

[14] See the discussion *infra* in Section V, in particular notes 57-60, 64, and accompanying text.

[15] The literature itself does not purport to explain the findings through contextual inference, and varying explanations include different perceptions by generalist judges of the uniqueness of cases from the specialized subject-matter, political capture, caseload pressure, and the influence of court communities ties and norms.

necessarily coincide with the patterns observed from real cases.[16] Moreover, experimental studies by design use settings of immediate parallel or sequential comparison and are limited in the number of scenarios each subject can effectively grade during a given session; they are therefore inapt to test for the effects of larger caseloads compositions over a longer period of time.

Empirically estimating the impact of differences in exposure to criminal behavior on judicial decision-making, however, is complicated for three important reasons. The main obstacle to such a study is the problem that the differences required for the identification of judges exposed to different gravity levels through their caseloads are the same differences that thwart meaningful comparison across judges. When judges systematically differ in their caseload composition, one cannot tell if observed differences in sentencing outcomes are the result of judges treating *different* cases differently or of judges treating *similar* cases differently. We may be seeing both, but a methodology is needed to separate the two effects. In addition, even when confining the analysis to a mutual subset of cases, comparison across judges sitting in different divisions or courts thus having caseloads differences raises concerns that judges might be selected (or self-selecting) to different dockets based on some prior underlying characteristics or different propensities to punish, potentially creating reverse causation. Lastly, such a comparison may also pose restrictions on the ability to make a causal inference in light of the concern of confounding factors characterizing the different court communities that may bias the estimates.

To circumvent those difficulties I develop an identification strategy that exogenously varies judicial exposure to criminal gravity. I do so by exploiting a natural experiment enabled by the properties of caseloads distribution under random assignment of cases. The second contribution of the paper is to highlight the fact that under random assignment, judicial caseloads should be balanced on average only when the caseloads contain a large enough number of cases. However, in the short term, with small numbers, even under random assignment we can observe a wider distribution of caseload compositions across judges. Tracking judicial caseloads composition over time, therefore, allows one to identify judges who initially randomly "drew" different types of cases from the overall distribution, but have balanced caseloads in the long term, and compare their sentencing outcomes only during the period for which their caseloads are balanced.

The novelty of this approach is that it enables estimating the causal link between exposure to gravity and judicial behavior in real cases. Existing scholarship faced a tradeoff between using experimental settings - that allow for a causal inference but are limited to the lab and to measuring short-term effects – and field studies looking at real cases that can only provide evidence of observed disparities but not of their causes. By overcoming this tradeoff, this paper is the first to document context-dependence in real judicial behavior. While the paper applies the method to studying the impact of initial exposure to gravity on later judicial sentencing practices, it can more generally

[16] *See, e.g.*, Eisenberg, Rachlinski and Wells, *supra* note 2; Shari Seidman Diamond, Mary R. Rose, Beth Murphy and John Meixner, *Damage Anchors on Real Juries*, 8 J. EMPIRICAL LEGAL STUD. 148 (2011); Steven D. Levitt and John A. List, *What do laboratory experiments measuring social preferences reveal about the real world?*, THE JOURNAL OF ECONOMIC PERSPECTIVES (2007): 153-174.

offer researchers a new way to study how caseload factors influence the formation of judicial attitudes.

Based on this methodology the paper finds that exposure to different levels of criminal behavior through caseloads can lead to a substantial and significant impact on sentencing outcomes. On average, judges who were initially exposed to one standard deviation lower average caseload gravity order sentences that are approximately two months longer than those ordered by judges exposed to higher levels of criminal behavior in their caseloads. Judges exposed to milder criminal gravity also have approximately 1.6 percentage points higher probability to depart above the sentencing guidelines range, and are 7.3 percentage points more likely to order sentences within the aggravated sentencing range or above the applicable sentencing range.

The effect is different from that of an immediate contrast effect, and lasts for approximately 40 court hearing days,[17] equivalent on average to six calendar months, before decaying, throughout a period for which judicial caseloads are balanced with similar exposure to gravity. This suggests exposure to different levels of criminal behavior can affect sentencing decisions not only through the immediate contrast effect on the comparison of cases within the same time period, but also through the formation of judicial sentencing preferences that judges continue to apply in future cases. When the composition of judicial caseloads is updated so that judges are exposed to similar gravity, previously formed judicial sentencing practices take a long time to update.

The effect is also different from that of anchoring or translation problems, and allows identifying the influence of context on the evaluation of behaviors, not just of punishments. If sentencing is based on a shared moral judgment of the underlying behavior, translation difficulties and anchoring should lead judges exposed to lower levels of gravity in their caseloads to evaluate the same sentences as relatively harsh when compared to the lower sentences imposed in milder offenses, and sentence the same cases more *leniently*. The finding of *increased* sentences by judges exposed to lower levels of criminal behavior can indicate that surrounding context affects the moral judgment of the underlying behavior itself, not just the translation of that sentiment into sentences.

The paper's findings may warrant further thought on the proliferation of specialization in the courts, and shed new light on the role of jurisdictional rules and case assignment mechanisms, not only as operational, but as having potential implications for the substance of legal outcomes as well. As far as different rules for case assignment can change the evaluation of particular cases against the backdrop of judicial caseloads, the results of the paper are relevant to practically any administrative decision that changes the type of cases under the jurisdiction of particular courts.

In sum, this paper makes three contributions. First, it advances a theory, previously unaddressed by legal scholars, on the relationship between the contours of cases in the caseloads of judges and the outcomes reached in judicial decisions. Second, it develops an identification strategy for measuring the impact of differences in caseloads

[17] As I explain *infra* in Section III.B.2, since the frequency of sentencing days varies across judges in different districts, I use net days on the bench rather than calendar time for the analysis.

composition that accounts for the otherwise possible concerns from an unbalanced comparison sample, selection effects and confounding variables bias. Third, with that methodology, the paper is the first to demonstrate contextual inference by judges outside an experimental setting, using actual sentencing data from real court cases. While the paper focuses on decisions of sentences, both the theoretical framework and the empirical methodology can be applied to other issues and legal fields. This paper, hopefully, opens the door for such future research.

The paper proceeds as follows: Section II develops the theory of the relative evaluation of legal cases and the possible relationship between the scope of cases in judges' caseloads and sentencing outcomes. Section III describes the data and the empirical methodology. Section IV presents the results. Section V briefly discusses some policy implications. Section VI concludes.

II. THE RELATIVE EVALUATION OF LEGAL CASES

A. Ranking Behaviors against Caseloads

Legal cases stem from disputes regarding a particular behavior and its outcomes – a tortious act, a violation of constitutional rights, or a criminal offense, to name prominent examples. When a behavior is deemed illegal or non-normative to justify an intervention, decisions on remedies or punitive outcomes require attention to the gravity of the behavior, the circumstances surrounding it, the harm suffered, and the characteristics of the offender. All of these aspects can be affected by the existence and magnitude of such features in the relevant comparison group.

Punitive decisions in particular – such as in the case of punitive damages or criminal sentences – are heavily based not only on a concrete quantifiable harm caused by a particular behavior that needs to be remedied, but also on an evaluation of the punishment suitable to answer the gravity of the defendant's action. Different justifications for punishment all lead in the direction of relative comparison across cases. Efficient deterrence requires marginal deterrence between acts of increasing gravity, as well as personal deterrence relative to the costs and benefits associated with the act. Retributive justifications dictate that punitive severity must accord with the relative gravity of the act, and rehabilitative goals need to be tailored in relation to the gravity of deviation from normative behavior and circumstances of the offender. When the sanction involves imprisonment, the length of incapacitation should similarly be correlated with the gravity of the behavior that society is to be protected from. All of these inject a relative aspect to the task of constructing punishments, requiring them to be computed in a way that takes the relative gravity of behaviors, circumstances and defendants into account.

The necessity and desirability of such comparisons mean that judges are required to assess not only the absolute gravity of a case but its relative gravity as well. The question is how would such comparison be made? Ideally we might want any comparison to take into account all relevant behaviors committed (or potentially committed) by offenders in the relevant jurisdiction – and the relevant jurisdiction may be at the court level, the city, the county, the state and even further based on theories

of jurisdictional competition and immigration of crime.[18] Clearly though, it is extremely and likely prohibitively difficult even for legislators to conduct such an inquiry.[19] Further, changes in law, in crime trends, or in societal views, render such an effort to be an ongoing updating process not a one-time task. At the same time, judges face limited physical and cognitive resources to engage in such inquiries.

Psychological theory prescribes such comparisons are heavily influenced by the proximate context in the makeup of cases before the decision-maker.[20] Judicial evaluations of a defendant's behavior in one case, in order to determine the existence of justification for punishment, and if justified - its magnitude, will likely be done not in relation to the entire possible spectrum of human behavior, but to the one they view on an ordinary basis. In other words, judges will likely be comparing a defendant in one case to defendants in other cases before them. As a result, judges who are exposed to different scope of behaviors in their caseloads may rank the gravity of the defendant's behavior differently, and this will lead to disparities in the judgments imposed by them.

Importantly, there are reasons to expect such relative judgments will not only be different but also biased in a predictable direction. People tend to over estimate how much the samples they are exposed to actually represent the distribution of cases in the world. Consequently, the estimation of a particular case can be affected by the relative ranking of that case in comparison to the range and skewness of the distribution of other cases with which it is presented.[21] As illustrated by figure 1, if judge A's caseload

---

[18] *See, e.g.,* Doron Teichman *The Market for Criminal Justice: Federalism, Crime Control, and Jurisdictional Competition*, 103 MICHIGAN LAW REVIEW (2005).

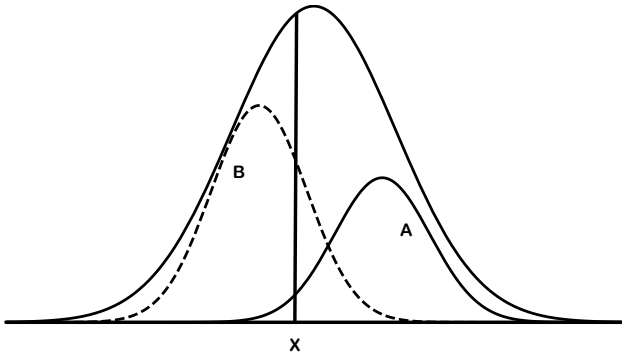
[19] Sunstein, Kahneman, Schkade, and Ritov, *supra* note 1, quote Justice Stephen Breyer on the inapplicability of such an evaluation: "Why didn't the Commission sit down and really go and rationalize this thing . . .? The short answer to that is: We couldn't. . . . Try listing all the crimes that there are in rank order of punishable merit. . . . Then collect results from your friends and see if they all match. I will tell you they don't."

[20] *See*, *e.g.*, Oren Shapira, Nira Liberman, Yaacov Trope and SoYon Rim, *Levels of mental construal*, in THE SAGE HANDBOOK OF SOCIAL COGNITION 229 (2012) ("effects of the context of the stimulus and the context of the perceiver – the constructs that happen to be accessible in his or her mind, his/her expectancies, and motivation – have been found in every subdiscipline in psychology"); Rachlinski and Jourden, supra note 9 ("Several decades of research on human judgment and choice indicate that human judgment is profoundly sensitive to context"); Amos Tversky, *Features of similarity*, 84 PSYCHOLOGICAL REVIEW 327 (1977) ("diagnostic factors are highly sensitive to the particular object set under study"); as well as the references and accompanying text, *supra* notes 1, 4-10, and *infra* note 21.

[21] Allen Parducci, *Category judgment: a range-frequency model*, 72 PSYCHOLOGICAL REVIEW 407 (1965). In psychological studies the effect of the relative ranking of a case in the distribution of cases has been demonstrated in a variety of domains, from the estimation of physical magnitudes, through feelings of happiness and satisfaction, to judgments of merit, non-normative behavior and even to medical diagnoses. See, e.g., Allen Parducci, *Contextual effects: A range-frequency analysis*, in E. Carterette & M. Friedman (eds.), HANDBOOK OF PERCEPTION 127 (1974) (providing an overview of experimental findings regarding loudness of sounds, size of squares, length in inches and weight in ounces); Parducci, *supra* note 4 (size of numerals, size of squares; satisfaction with lottery drawings; non-normative behaviors); Barbara A. Mellers and Michael H. Birnbaum, *Loci of Contextual Effects in Judgment*, 8 JOURNAL OF EXPERIMENTAL PSYCHOLOGY 582 (1982) (darkness of colored dots); Barbara A. Mellers, *Equity judgment: A revision of Aristotelian views*, 111 JOURNAL OF EXPERIMENTAL PSYCHOLOGY: GENERAL 242 (1982) (professional merit); Barbara A. Mellers, *Fair allocations of salaries and taxes*, 12 JOURNAL OF EXPERIMENTAL PSYCHOLOGY: HUMAN PERCEPTION AND PERFORMANCE 80 (1986) (professional

consists of graver cases with over-representation of cases from the right tail of the distribution, while judge B's caseload includes milder cases over-representing the left tail of the distribution, the ranking of the same case X will be as relatively worse by judge B than by judge A.[22]

Figure 1. The relative evaluation of cases against different judicial caseloads[23]



In other words, similarly to how contrast effect can change the relative evaluation of a case viewed against another more or less serious case, contrast effect can also arise from the comparison between a particular case and the general caseload before the judge. If judicial sentencing decisions are based on the relative evaluation of case gravity, contrast effect can lead judges exposed to more serious cases in their caseloads to sentence the same cases less harshly, and vice versa. The hypothesis generated by this theory is that *sentences by judges exposed to higher levels of criminal gravity in their caseloads will be (significantly) shorter than sentences by judges exposed to lower levels of criminal gravity.*

A potential caveat is that the suggested analysis focuses on the impact of context on the evaluation of graveness of the underlying *behavior* itself. Other scholarship has focused on the context-dependent evaluation of *outcome* severity, in which case the effect can move in the opposite direction. When the surrounding context is compounded of cases of lower gravity (and lower sentences attached to them)

---

merit); Douglas H. Wedell, Allen Parducci and Michael Lane, *Reducing the dependence of clinical judgment on the immediate context: Effects of number of categories and type of anchors*, 58 JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY 319 (1990) (diagnoses of psychopathology).

[22] Even if judges do not base their evaluations of case gravity solely on their caseloads, but also have some priors about the existence of cases from varying degrees in the world, judicial posteriors that are based on overweighing caseload representativeness will still move in the same direction as the distribution of cases that judges are exposed to. Different judges may hold different priors, but as long as such priors are not systematically and negatively correlated with the judge's likelihood to encounter a certain caseload composition, on average, judges exposed to graver types of behaviors through their caseloads will update their priors more toward the right tail of the distribution, while judges encountering relatively milder cases in their caseloads will update their priors towards the left tail of the distribution.

[23] Figure 1 is only illustrative. For convenience normal distributions are displayed, but the same logic applies to any shape or different shapes for the distributions of cases. In particular, psychologists have demonstrated that the rating of a particular object will be affected by its relative *location* in the distribution, and that different judgments are made when distributions have different range and skewness even when distributions means are the same.

translation problems may shift the sentencing scale of judges exposed to milder cases downwards – adding or subtracting months in comparison to the menu of lower sentences.[24] Anchoring based on the lower sentences ordered in less serious cases should similarly lead to a negative effect on sentencing.[25] If, however, as the paper suggests, it is the underlying behavior that is judged to be more serious in comparison to milder cases, judges exposed to lower levels of gravity in their caseloads would view the same behaviors as graver and sentence them more harshly. Translation and anchoring problems might be more acute for jurors' decisions than for judicial decisions, as judges do have better knowledge regarding the sentencing scales offered by other cases and by the sentencing guidelines. Especially in the setting examined by this paper, during the comparison period judicial caseloads are balanced; judges are exposed to a similar spectrum of overall cases, and are familiar with the punishments and sentencing ranges of other offenses.[26] To the extent both effects might be present, this will make the results of the analysis only a lower bound estimates for the effect of caseload composition on the relative evaluation of behaviors.

B. Judicial Behavior over Time

Different exposure to gravity can affect not only the immediate comparison across cases, but also contribute to the formation of more general judicial punitive practices. Initial exposure can contribute to forming a schema (or a theory) leading people to misinterpret new information and process it in a way more attuned toward indications consistent with the previously formed beliefs and expectations.[27] If judicial sentencing practices are formed and updated according to such models, then judges exposed to initial different gravity in their caseloads may develop different views on the relative gravity of different cases leading to sentencing disparities that can persist even in the face of new cases.

Since the empirical strategy utilized by this paper looks only at the period following the initial disparate exposure, during which judicial caseloads are balanced with similar exposure to gravity, it offers a test for the theory. If the only contextual mechanism affecting judicial behavior is of immediate comparison across cases, I should not be able to detect a significant difference in the sentencing decisions of judges handling similar caseloads, regardless of their past exposure to gravity. If, however, initial exposure to gravity affects judicial behavior for a longer time, I can test the hypothesis

---

[24] Sunstein et al. discuss both translation difficulties (Sunstein, Kahneman and Schkade, *supra note 1*; Kahneman, Schkade, and Sunstein, *supra* note 1) and the relative evaluation of harm (Sunstein, Kahneman, Schkade, and Ritov, *supra* note 1). In the setting studied by Sunstein et al. translation difficulties are relevant to jurors' decisions absent any context about other damage awards, but their findings are in line with the impact of contrast effect on the evaluation of case (or in their case, harm) gravity.

[25] *See* Rachlinski, Wistrich and Guthrie, *supra* note 19.

[26] The paper cannot reject the possibility that translation and anchoring problems might be present when judges are consistently facing completely different caseloads, for example, when a judge assigned to a serious felony docket is required once to sentence a misdemeanor case and vice versa.

[27] *See*, *e.g.*, Matthew Rabin and Joel L. Schrag , *First Impressions Matter: A Model of Confirmatory Bias*, 144 THE QUARTERLY JOURNAL OF ECONOMICS 37 (1999); Shapira, Liberman, Trope and Rim, *supra* note 20.

even during the period when judicial caseloads are balanced. A refinement of the hypothesis is therefore that after an initial disparate exposure to gravity, *sentences by judges initially exposed to higher levels of criminal gravity in their caseloads will be (significantly) shorter than sentences by judges initially exposed to lower levels of criminal gravity, even after such disparate exposure ends.*

This is not to say that such disparities will necessarily last indefinitely. Schemas, or previously held opinions, can persist for long periods of time, but can also be updated even if at a slower pace. Changing the contours of cases in a judge's caseload, by exposing a judge to new cases not previously in her caseload or to changed frequency of certain cases in the caseload, might as well change judicial behavior. If judges were previously overweighing the cases they saw – assuming their cases are more centered on the distribution than they really are – addition of cases lying to the right (left) of their previous sample can fix that error toward a more accurate evaluation of the right (left) hand tail, and shift their relative ranking and sentencing levels for any particular case. If over time both judge A and judge B are exposed to similar caseloads, judge A will rank a given case as relatively worse than she previously did, and judge B will rank the same case as relatively milder than before.

Such updating will likely not happen instantly, but over time the flow of new information could eventually lead to updating exactly because of the sensitivity of the initial beliefs to the contextual background. To test for the dynamic effect of caseload exposure on judicial behavior, I also account for the lapse of time and for the effect of time on judges with initial higher gravity exposure in the regressions. If exposure to new cases leads to updating judicial behavior, then *over time the gap in the sentencing decisions of judges initially exposed to higher or lower gravity should be mitigated.*

III. RESEARCH DESIGN

A. Data

To estimate the impact of exposure to gravity on judicial decisions, I use sentencing data from the Pennsylvania Commission on Sentencing (PCS). Sentencing data is particularly suited to test for the impact of relative judgments on the evaluation of behaviors due to the existence of both objective (offense gravity score assigned by the guidelines) and subjective (the sentence imposed) measures for case gravity. The PCS data covers a twelve-year period between 2001-2012 and includes all misdemeanor and felony offences in which the offender was convicted and sentenced by Pennsylvania Courts of Common Pleas. The PCS data is widely used in studies of sentencing outcomes due to its inclusiveness and the rich information provided about offender and case characteristics, sentencing guidelines range and sentence imposed by the court, as well as identifiers for the judge who imposed the sentence – thus allowing an analysis at the individual judge level. I complement the data with information on judges' election dates from the Pennsylvania Manual.[28] To accurately measure judicial

[28] The biennial Pennsylvania Manuals can be found at: http://www.portal.state.pa.us/portal/server.pt/community/pa_manual/1294

caseload exposure to criminal behavior from the beginning of their judicial career, I only include in the analysis judges elected after 2001, the first year covered by the data.

Under Pennsylvania's unified court system the Courts of Common Pleas have general jurisdiction over felony and misdemeanor cases. Generally, cases in Pennsylvania judicial districts are randomly assigned across judges in the Courts of Common Pleas at the district level through a computerized system, or in some districts based on a daily or weekly rotation. I also empirically test for random assignment, and only include in the analysis districts in which the caseloads of judges are balanced in the long term with regard to exposure to gravity, and I confirm that their caseloads are also balanced for different case and defendant characteristics.[29]

In Pennsylvania's sentencing system, when imposing a sentence of partial or total confinement, Pennsylvania statute requires the court to impose both a minimum and a maximum sentence.[30] The sentencing guidelines and most mandatory sentencing provisions address only the minimum sentence.[31] In following with the convention in prior research, the analysis looks at the decisions on minimum sentences, as this part of the sentence reflects the court's discretion over sentencing, while the range between the minimum and maximum sentence is considered a part of the sentence to be later decided under the Parole Board discretion.

When deciding on the minimum sentence in a misdemeanor or felony offense, Pennsylvania prescribes sentencing guidelines that judges must consider. The sentencing guidelines set a range for the minimum sentence between a lower and an upper limit, both stated in months of incarceration, that is the result of two scores assigned by the guidelines for each offense: (1) an "Offense Gravity Score" (OGS), which takes into account the gravity of the current conviction offense, and ranges between 1 and 14[32]; and (2) a "Prior Record Score" (PRS), which weighs the seriousness and extent of the offender's prior criminal record, and is divided into 8 categories.[33] Both scores are predetermined by the sentencing guidelines for lists of particular offenses. Based on a matrix of OGS and PRS combinations the sentencing guidelines categorize all offenses into five levels.[34] Within each level, the guidelines

---

[29] See *infra* Part III.B.3.

[30] 42 Pa.C.S. §§9755, 9756.

[31] The trial judge is free to impose any maximum sentence that is at least double the minimum sentence, and no longer than the statutory maximum established by 18 Pa.C.S. §§1103-1105 for the grade of the conviction offense.

[32] All misdemeanor and felony offenses have an OGS between 1-14. Each offence has a baseline OGS; when applicable, statutory enhancement prescribe number of points to be added to the baseline OGS for each offence, and are included in the OGS reported for each offence in the data. Murder 1 and Murder 2 felonies do not have OGS, and are prescribes mandatory life or death sentences. As is further explained below, I omit murder cases from the regression analysis (as those are cases where judges lack discretion over sentencing), and there are no murder cases heard by the judges in the sample during the initial period used for calculating judges' exposure to gravity in their caseloads.

[33] The categories, by order of increasing severity, are: No prior record (0), 5 categories based on the number and severity of prior record (1-5), Repeat Felony Offender (RFEL), and Repeat Violent Offender (REVOC). I recode the PRS for repeat felony offender as 6 and for repeat violent offender as 7.

[34] Except for murder of the first and second degree, which are prescribed mandatory life/death and life sentences, respectively.

prescribe three ranges: (1) a standard range, for use under normal circumstances; (2) an aggravated range, for use when the judge determines that there are aggravating circumstances; and (3) a mitigated range for use when the judge determines that there are mitigating circumstances.

The sentencing guidelines suggest recommended sentences, but judges retain discretion whether to impose a sentence within the guidelines range or not in particular cases, subject to mandatory minimum and maximum statutory sentences when such apply.[35] Where a case includes several counts, the judge must sentence each count separately, and decide whether to impose the sentence concurrently or consecutively with sentences on other counts in the same case. For cases with multiple counts I use the most serious offense in the judicial proceeding and calculate the total overall sentence.

As will be further explained in the next sections, I use two datasets for different parts of the analysis. For the first part, to calculate judges' exposure to gravity through their caseloads, I use all cases heard by each judge (hereinafter: "*identification sample*"). For the regression analysis, in following with the convention in the literature, I exclude charges that are subject to mandatory life or death sentence, since the judge has no discretion in sentencing such offences, as well as charges for which incarceration sanction is inapplicable, since the dependent variable looks only at sentencing decisions of incarceration (hereinafter: "*comparison sample*").

B. Empirical Methodology

1. Identification Strategy

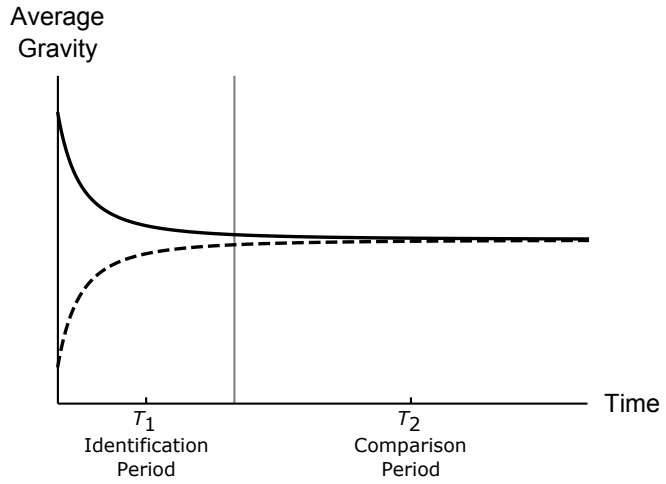
The common problem in analyzing real world data is the difficulty in isolating a particular causal effect from other confounding factors. In order to estimate causal relationships, researchers seek to identify natural experiments that create an exogenous variation in the variable of interest. But what should one do when reality does not provide a natural experiment? The methodology developed in this paper shows that where random assignment exists one could create a natural experiment by exploiting the fact that under random assignment there is a distribution of values for the variable of interest.

The key is to recognize that under random assignment, with a large number of draws judicial caseloads should be balanced on average. However, in the short term, with small numbers, we can expect a wider distribution of caseload compositions across judges. Even judges sitting in the same district courthouse on the same year may "draw" different cases from the overall distribution leading to different average gravity in their caseloads. Nevertheless, if cases are randomly assigned across judges those differences will decrease over time toward convergence, as judges handle more and more cases.

[35] If the judge deviates from the guidelines range, she must specifically state the reasons for such departure in a written statement. Sentences inside the guidelines also pose a heightened burden on the defendant for appeal than sentences outside the guidelines range (42 Pa.C.S. § 9781).

To illustrate, Figure 2 presents the cumulative averages of caseload gravity in a hypothetical district courthouse with two judges. Imagine one judge happened to initially draw harsher cases from the overall distribution, and the other happened to initially draw milder cases. For an initial short period, marked in Figure 2 as T_1, the two judges experience different exposure to gravity. Over time, as judicial caseloads contain more and more cases, the cumulative exposure to gravity of the two judges moves closer together, until, by period T_2, average exposure to gravity is similar for both judges.

Figure 2. Variance in Caseload Gravity Under Random Assignment of Cases



This tendency is equivalent to simulating an experiment randomly assigning judges to different conditions of gravity exposure. The initial period (T_1) – when judges' caseload compositions are more widely distributed – serves as the "treatment" period, when judicial exposure to higher or lower criminal gravity is identified (hereinafter: "*identification period*"); The later period (T_2) – after judges' caseload compositions are balanced on average – serves as the "comparison" period, during which the impact of the initial different exposure on sentencing outcomes can be measured (hereinafter: "*comparison period*").

To ensure the differences in exposure identified during the identification period are not the result of systematic differences due to geographic or time trends, I use a matched sample. I identify matches of judges from the same district who start hearing criminal cases on the same year.[36] I then identify within each match which of the

[36] I use the year of the first criminal case heard by the judge as the start year. Since elections in Pennsylvania take place on November, for judges who start hearing cases on November or December of a given year, I use the following calendar year as their start year (but I keep the cases heard prior to it for the exposure calculation). In addition, judges may assist on occasion with criminal cases even when this is not their assigned docket, which leads to very few cases heard by those judges in a given year. To keep the sample annually balanced, I only consider as a judge's start year the year in which that judge heard at least 25 criminal cases, but I include in the exposure calculation cases heard in the prior six months as well if such exist.

judges happen to initially fall on opposite ends of the distribution of exposure to criminal gravity in their caseloads, but over the long run have balanced caseloads. In such way, using a matched sample that is balanced for judges from the same district courthouses and in the same time period, allows one to capture for the comparison period judge differences in sentencing rather than different caseloads or local customs.

2. The Identification Period

From the PCS data I generate the identification sample for the identification of judicial exposure to gravity. For each judge I calculate the daily cumulative average of gravity in the judge's caseload, based on the OGS for each offense.[37] The gravity cumulative average is based on all offenses sentenced by the judge, including offenses not eligible for incarceration. The reported OGS for each offense offers an exogenous measure for case gravity, independent of the judge's perception of the case. It therefore allows one to compare how exposure to different levels of gravity in caseloads (through the exogenous measure of OGS) might affect judges' subjective perceptions of gravity in subsequent cases – through the way judges' evaluations are reflected in their sentencing decisions. Since the time unit reported by the data is of date of sentence, but with no indication for the sequence of cases within each hearing day, I treat all cases sentenced on the same day as exposed to simultaneously. Also, since frequency of sentencing days varies across judges in different districts I use net days on the bench rather than calendar time (hereinafter: "*net hearing days*").

I employ several sample restrictions for the matched sample. First, to ensure a sufficient length for the comparison period and a sufficient number of cases per judge for statistical power I drop judges who sat through less than 30 net hearing days for the comparison period. Second, I drop judges who heard less than 25 cases in the district during the comparison period, to weed out cases heard by judges who might receive few cases from another district that is not the district they regularly sit in.[38] Finally, I drop district-start year matches containing only one judge starting to hear criminal cases at the given district on a given year.

From the PCS data 17 judicial districts had more than one judge newly assigned to the criminal docket on the same year, creating overall 33 district-start year combinations. For each match I calculate the time until judicial caseloads' gravity cumulative averages converge, which I define as the time at which the gap between the caseload gravity average of the judges with the lowest and highest scores is less than half a standard deviation of all judicial gravity averages in the sample. From the convergence calculation and the cumulative averages plots, 10 district-year combinations from 8 districts display a pattern in line with the paper's experiment setting. For those 10 matches initially judges' gravity averages diverge considerably, but differences across judges diminish over time until judges' average gravity scores

[37] The reported OGS is for felonies and misdemeanors on a scale of 1-14.

[38] On occasion a case may be transferred to a judge of another district, if the judge of that district cannot sit in judgment or if a judge serves as a backup judge for the district in small districts. Those cases were included in the first stage calculation of judicial exposure to gravity, but dropped from the matched sample. This led to dropping between 1-7 cases for such judges, but all judges remained in the sample for their main district.

converge within 8-25 net hearing days. Out of the other 23 matches, 13 matches did not display variance across judges during the initial period and judges' averages move closely together[39] – thus not demonstrating the natural experiment sought by the paper, and for 10 matches the differences in the initial period persisted over time[40] – thus raising concerns of differences in judges' caseloads for the comparison period. For these reasons, I focus the analysis on the 10 matches that hold to the natural experiment conditions.

Figure 3 displays the gravity cumulative averages plots for judges within each match, where the time of convergence is set to zero. The period until convergence ($T \leq 0$) serves as the identification period; it lasts for 8-25 net hearing days in different matches, which are equivalent to between one to seven months of hearing criminal cases, and three months on average. The period after convergence ($T > 0$) is the comparison period for the analysis.

For each day throughout the identification period, I code the judges within each match as exposed to "high" or "low" gravity in their caseload in relation to the other judge or judges in the match. To maintain at least a minimal magnitude of treatment, judges are only coded for exposure under the condition that the gravity average of the judge with the lowest score in the "high" group is at least half a standard deviation higher than the gravity average of the judge with the highest score in the "low" group. I construct the independent variable *Exposure* – to equal 1 if the judge had relatively high exposure to gravity in comparison to the other judge or judges in the match, and zero if that judge exposure to gravity was relatively low in comparison to the other judge or judges in the match.

The final sample includes 20 judges from 10 matches: 10 judges in the "low" exposure condition (*Exposure* = 0*)* and 10 judges in the "high" exposure condition (*Exposure* = 1*)*. The identification strategy by definition imposes a constraint on the number of judges fitting all criteria: elected after 2001, started hearing cases at the same district on the same year as at least one other judge, and having different caseload composition during the identification period but similar caseload composition during the comparison period. This is the often-existing tradeoff between the desire to identify conditions simulating a natural experiment and sample size, and it is not an uncommon sample size for studies that take a first step at a previously unexplored question. [41]

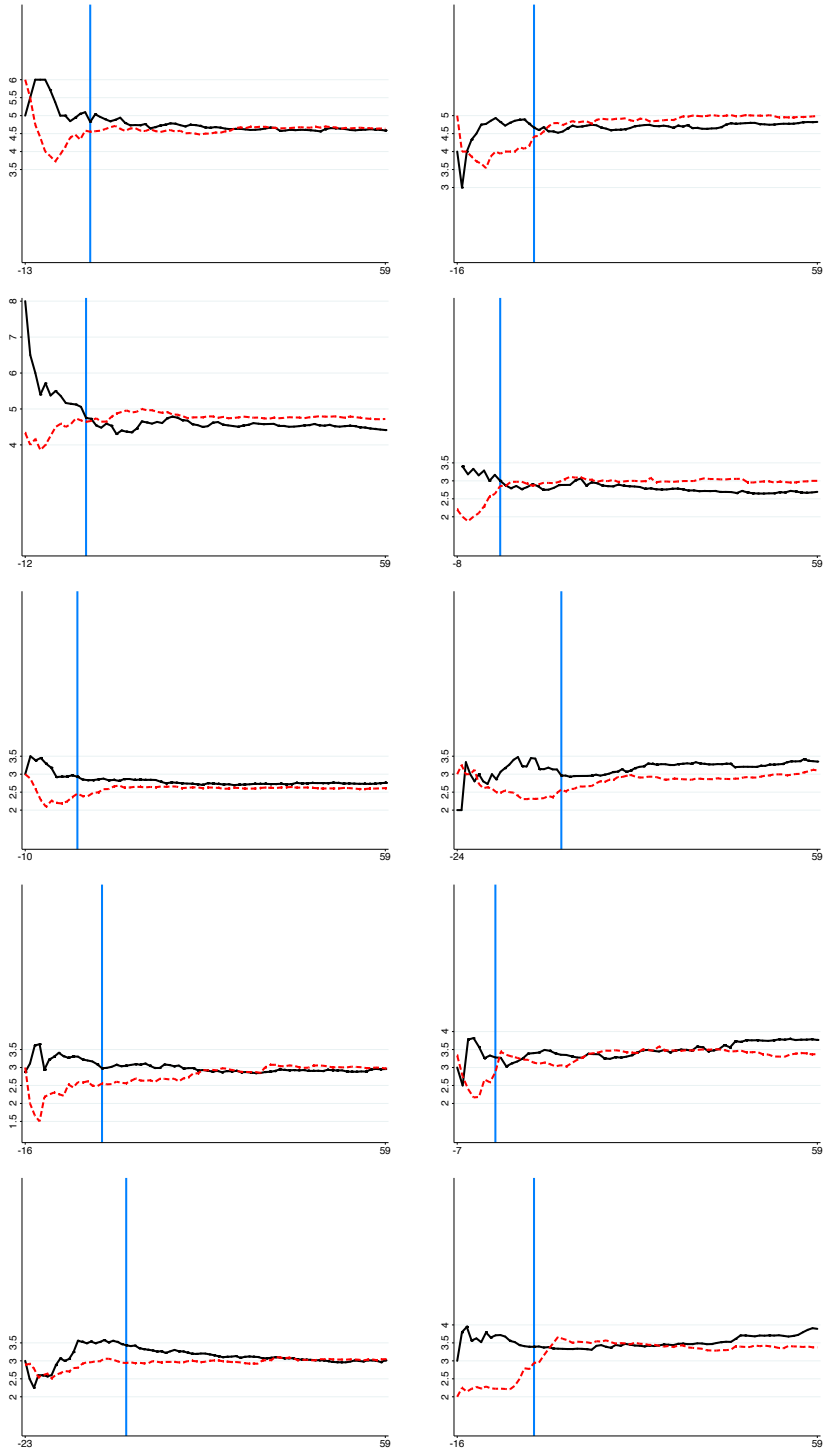
---

[39] For 9 of the 13 matches, judicial gravity cumulative averages converge already within the first 1-4 days. For 4 of the 13 matches, convergence is reached after 5-12 net days, but judicial cumulative gravity averages intersect throughout the identification period and do not display a consistent difference.

[40] In 7 of the 10 matches the distance between the gravity cumulative averages of judges did not consistently decreased to less than ½ standard deviation of judicial gravity averages in the sample throughout the entire period reported in the data. In 3 matches, the gravity cumulative averages of judges decreased to less than ½ standard deviation only after a period of 97-147 net days.

[41] For example, Ryan W. Scott, *Inter-judge Sentencing Disparity After Booker: A First Look*, 63 STAN. L. REV. 1 (2010), uses a sample of ten judges in the Boston courthouse to study the impact of striking down the sentencing guidelines by *United States v. Booker,* 543 U.S. 220 (2005) on sentencing disparities. Danziger, Shai, Jonathan Levav and Liora Avnaim-Pesso, *Extraneous Factors in Judicial Decisions*, 108 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES 6889 (2011), use a sample of eight judges to study the impact of taking food breaks on Parole Board decisions.

Figure 3. Judicial Gravity Cumulative Averages Over Time



Y-axis is caseload gravity cumulative average.
X-axis is net hearings days. The day of convergence is set to zero and marked by a blue line.
"High" Exposure Condition "Low" Exposure Condition

3. The Comparison Period

The crucial assumption underlying the validity of the natural experiment is the random assignment of cases to judges. I therefore test whether the caseloads of judges in the "high" and "low" exposure groups are balanced during the comparison period on different case and defendant characteristics. Since the maximal length for the comparison period for which all judges in the matches hear cases is 59 net hearing days, the main period analyzed is for the time between 1 and 59 net days (I later test that results are robust when using different time periods).

If case assignment across judges in the district courthouse is random, one expects the caseloads of judges to be balanced in the long term, and for there to be no significant difference in average offense gravity in the caseloads of judge from the "low" and "high" exposure groups during the comparison period. Table 1 summarizes the comparison of means for the average gravity score for judges in the "high" and "low" groups across the two periods, based on the full exposure to cases in the identification sample. Statistical significances of the differences were determined by Student's t-test. Since the length of the identification period varies across matches, I display judicial gravity averages for the period quartiles.

Over the first half of the identification period, judges' average gravity exposure between the "low" and "high" groups is significantly different with a gap of approximately 1 point, equivalent to about 1 standard deviation of judicial gravity averages in the sample. At the middle of the identification period, judicial cumulative gravity averages are 2.87 and 3.93 respectively (p=.016). The gap decreases over time and by the end of the identification period the difference is small and insignificant (3.28 and 3.63 respectively, p=.371). For the comparison period, on the other hand, judges' gravity averages are practically identical. For the period between 1 and 59 net days, the average gravity exposure is 3.67 in the "low" group and 3.65 in the "high" group (p=.948). These trends are also evident from Figure 3.

Table 1. Tests for Means: Exposure to Gravity

		Low	High	t
Identification period	T≤Q1	2.81*	3.84*	-2.295
		(.271)	(.358)	
	T≤Q2	2.87*	3.93*	-2.658
		(.288)	(.275)	
	T≤Q3	3.07+	3.79+	-1.819
		(.291)	(.267)	
Convergence Point	T≤0	3.28	3.63	-.917
		(.280)	(.251)	
Comparison period	1≤T≤59	3.67	3.65	.067
		(.268)	(.261)	

Results are based on the identification sample, including all offenses sentenced by judges. T indicates net hearing days intervals for each period. Q indicates the 1st, 2nd and 3rd quartiles during the identification period.
Standard errors in parentheses.
P-values from Student's t-test marked by + p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Testing for other caseload characteristics, presented in Table 2, judicial caseloads throughout the comparison period are balanced in terms of defendant and case characteristics as well. The average offense gravity score in the comparison sample is 4.14 for judges in the "low" exposure group, and 4.13 for judges in the "high" exposure group. The average prior record score for judges in the "low" and "high" exposure groups are 2.09 and 2.07 respectively. Judges of both groups also have similar racial and gender composition in their caseloads, a similar proportion of cases going to trial and similar average statutory mandatory minimum sentences in cases. Student's t-test results of the comparison of means for all variables are insignificant.

Table 2. Independent Variables Means by Exposure Group

	Comparison Period 1≤T≤59	
	Low	High
Offense Gravity Score (OGS)	4.14	4.13
	(.072)	(.079)
Prior Record Score (PRS)	2.09	2.07
	(.067)	(.071)
White Defendant	.497	.505
	(.016)	(.017)
African-American Defendant	.402	.388
	(.016)	(.016)
Hispanic Defendant	.056	.070
	(.008)	(.009)
Female	.142	.128
	(.011)	(.011)
Trial	.097	.104
	(.011)	(.011)
Mandatory Minimum (months)	1.03	.853
	(.164)	(.140)

Results are based on the comparison sample used for the regression, including only offenses eligible for discretionary incarceration for the 1≤T≤59 net hearing days period.
Standard errors in parentheses.
P-values from Student's t-test marked by + p<0.10, * p<0.05, ** p<0.01, *** p<0.001

I also complement the data with information about the individual characteristics of the judges in the sample, presented in Table 3. Both "low" and "high" exposure

groups contain an identical proportion of judges with prior experience as prosecutors (0.3) or as public defenders (0.2), and similar female representation (0.3 and 0.2 respectively). The vast majority of judges in both groups were endorsed during elections by both the Republican and the Democratic Parties, and the remainder are similarly distributed across the two parties.

Table 3. Judicial Characteristics by Exposure Group

	Low	High
Demographic Characteristics:		
Prior Prosecutor	0.3	0.3
	(.153)	(.153)
Prior Public Defender	0.2	0.2
	(.133)	(.133)
Female	0.3	0.2
	(.153)	(.133)
Political Affiliation:(1)		
Cross-Endorsed	0.7	0.8
	(.153)	(.133)
Republican	0.1	0.1
	(.316)	(.316)
Democrat	0.2	0.1
	(.133)	(.316)

(1) Political Affiliation is based on party support at election.
Standard errors in parentheses.
P-values from Student's t-test marked by + p<0.10, * p<0.05, ** p<0.01, *** p<0.001

C. The Model

Based on the matched sample identification strategy, I test how judicial exposure to gravity impact sentencing decisions in the following comparison period. The main specification is of the form:

$$Y_{ijdy} = \alpha + \beta_1 * Exposure_j + \beta_2 * Time + \beta_3 * Exposure_j * Time + X_{ijdy} + \delta_d + \gamma_y + \delta_d * \gamma_y + \varepsilon_{ijdy},$$

where Y is the outcome for defendant i, whose case is sentenced by judge j in district courthouse d on year y.

I test case outcomes using three variables:

(1) ***Sentence*** – incarceration sentence (in months).
(2) ***Deviation*** – the deviation of the incarceration sentence from the recommended sentence of the standard guidelines range (defined as the midpoint of the standard sentencing range). I use this measure to test whether

the gap in actual sentences imposed coincides with different sentencing in relation to the guidelines. An additional advantage of such a comparison is that unlike actual sentences, whose distribution is skewed to the left, the *Deviation* of sentences from the guidelines has a bell shaped distribution, which makes it more suited for OLS regression analysis.

I calculate the sentence deviation as: $Deviation = Sentence - \frac{Upper + Lower}{2}$, where *upper* is the upper limit of the sentencing guidelines standard range, and *lower* is the lower limit of the sentencing guidelines standard range (all in months). When a judge renders a sentence lower than the middle recommended sentence *Deviation*<0, when the sentence imposed equals the middle recommended sentence *Deviation*=0, and when the sentence is higher than the middle recommended sentence *Deviation*>0.

(3) ***Non-Conformity*** – the conformity of sentence with the sentencing guidelines. I use this measure to test for the tendency of judges to sentence outside the standard sentencing range, and for the direction of departures from the sentencing range. I test the decision on location within each sentencing category in relation to all other categories below or above it, using four separate dummy variables: *Non-Conformity Below* (below departures = 1); *Non-Conformity Low* (below departure or mitigated range = 1); *Non-Conformity High* (above departure or aggravated range = 1); and *Non-Conformity Above* (above departure = 1).

To avoid the impact of extreme values I trim the dependent variable of *Sentence* at the 99th percentile. The reported results for all variables are for the same cases analyzed for trimmed sentences. Results for all variables are consistent (and a little stronger) when including outlier cases as well.

The main coefficient of interest is β_1, which captures the impact of exposure to different levels of gravity on sentencing outcomes. $Exposure_j$ is an indicator for the scope of criminal gravity in judge *j* caseload during the identification period, equals 1 if judge *j* experienced high average gravity relative to the other judge in the same district-start year match, and 0 if judge *j* caseload had low average gravity relative to the other judge in the same district-start year match.

Time is net sentencing days (in days) from the beginning of the identification period. β_2 captures the general impact of time on all judges, and β_3 captures the impact of the interaction between exposure to gravity and time. Since the exposure to different gravity levels by judges is limited in time to the identification period, and during the comparison period judges are exposed to similar average gravity in their caseloads, β_3 allows me to test whether the impact of initial exposure changes over time.

X is a vector of case and defendant characteristics for defendant *i* sentenced by judge *j* in district court *d* on year *y*. Case characteristics include the offense gravity

score as prescribed by the sentencing guidelines, the type of offense[42], the length of mandatory minimum sentence (in months) when such applies, and whether the disposition of the case was by trial or by guilty plea. Defendant characteristics include gender, race (accounting for African-American and Hispanic defendants) and the defendant prior record score.

The model also includes district court fixed effects (δ_d), sentencing year fixed effects (γ_y), and district by year fixed effects.[43] Standard errors are robust and clustered at the judge level. Results are robust to clustering at the district level.

IV. RESULTS

A. Gravity Exposure and Sentencing Disparities

The maximal length period for which all judges in the matches hear cases is 59 net hearing days. This period is equivalent to between three to thirty calendar months in different matches, and to about 9.5 calendar months on average. Therefore, the main analysis focuses on how differences in *Exposure* during the identification period affect sentencing outcomes over the next 1-59 net hearing days (Table A1 in the Appendix presents robustness tests for different time periods). Summary statistics for the incarceration sample used for the regression are presented in Table 4.

Over the period between 1 and 59 net days there is a substantial and significant difference in sentencing outcomes across judges from the "high" and "low" exposure groups. Table 5 reports the regression results for *Sentence* (in months). First, comparing column (1) with the other columns in Table 5 indicates the identification strategy succeeds in simulating random assignment to the treatment of exposure to gravity, and results are stable across different specifications whether accounting for additional covariates or not. The initial sentencing gap between judges exposed to "high" and "low" gravity in their caseloads is -2.173 months (p=.030) under the basic model with no covariates (column 1), in comparison to -2.049 (p=.008) under the model accounting for all covariates (columns 5).

The result that judges in the "high" exposure condition impose sentences that are approximately -2.049 months shorter than judges in the "low" exposure condition, is very significant both statistically at the 1% level (p=.008) and economically. With an average sentence of 7.95 months in the sample, a difference of two months in incarceration time is a very substantial gap between judges in the two groups, equivalent to a difference of 25% of the average sentence imposed. For comparison,

[42] Type of offenses include: DUI, drug, offenses against property, offenses against person, and a general category of other.

[43] Because of the short time period tested and the empirical methodology of matching within district-years, there is practically no difference between accounting for year and district fixed effects alone or adding an interaction variable. The difference between models with or without year fixed effects is very minimal both in economic and statistical significance.

this gap is equivalent in magnitude to almost twice the effect of having one more point in the offender's prior record score (β=1.076, p=.000), or 73% of the effect of committing an offense with one-point higher gravity score (β=2.809, p=.000).

When I test for judicial sentencing decisions relative to the standard sentencing range prescribed by the sentencing guidelines for each offense, the results are similar. From Table 6, the gap in deviation from the standard guidelines range between judges initially exposed to high or low gravity in their caseloads is -1.731 months on average (p=.024), constituting a difference of 22% in relation to the average sentence in the sample of 7.95 months. Again, the coefficients are consistent throughout the different specifications, whether accounting for additional covariates or not.

Interestingly, since the range of the sentencing guidelines is derived from the combination of the offense gravity score and defendant prior record score, judges also seem to compensate for high OGS and PRS values that lead to higher sentencing recommendations, and both variables have a negative impact on the sentence relative to the sentencing guidelines. Also, since the guidelines recommended standard range already takes into account applicable mandatory minimum sentences,[44] the separate effect of mandatory minimums that was significant for months of *Sentence* in the main model (β=0.486, p=.000), is eliminated for *Deviation* from the guidelines (β=0.068, p=.327).

Judges in the "high" exposure condition are also more lenient in exercising their discretion to depart from the standard sentencing guidelines range. There are five possible conformity categories in relation to the sentencing guidelines recommendation: judges may sentence a defendant within the standard guidelines range, they may find the circumstances justify applying the mitigated or the aggravated sentencing range, or they can depart below or above whatever is the applicable guidelines range. Overall, in the comparison sample, 65.8% of the cases were sentenced within the standard guidelines range, 14.6% of the cases were sentenced within the mitigated range, 9.7% in the aggravated range, and the remainder involved departures below (5.5%) or above (4.4%) the applicable guidelines range. Unfortunately, to analyze such a categorical variable, sample size is too small to use a generalized ordered logistic regression model (GO-Logit), and the proportional odds assumption is violated making the dependent variable unfitting for a non-generalized ordered logistic regression model (O-Logit).[45]

---

[44] If the mandatory minimum sentence exceeds both lower and upper sentences recommended by the combination of OGS and PRS, the sentencing range will reflect the mandatory minimum sentence on both ends. Only if the upper sentence exceeds the mandatory minimum there will be a range of discretion for the judge within the guidelines, but still any mandatory minimum will replace the otherwise lower limit.

[45] In other words, the relationship between each pair of outcome groups is not the same, therefore the coefficients that describe the relationship between one pair of outcomes (for example, below departure versus all higher categories of the response variable) are not necessarily the same as those describing the

To overcome these restrictions, I test conformity using four separate dummy variables, each splitting sentencing categories into two groups, so that: *Non-Conformity Below* – compares departures below the guidelines (equals 1) to all four higher categories (equals zero); *Non-Conformity Low* compares below departures and sentencing within the mitigated range (equals 1) against the three higher categories (equals zero); *Non-Conformity High* compares departures above the sentencing range and sentencing within the aggravated range (equals 1) with the three lower categories (equals zero); and *Non-Conformity Above* compares above departures (equals 1) against all four lower categories (equals zero). In such a way, each of the four dummy variables compares only two outcomes and can be analyzed using a simple logistic regression model (Logit).

Tables 7 and 8 report the logit marginal effects, evaluated at the sample mean. Judges exposed to higher initial gravity have significantly lower probability than judges with lower exposure to sentence a defendant within the aggravated sentencing range or to depart above the guidelines range, and results are overall stable across different specifications whether accounting for different covariates or not. At the margin, initial "high" exposure to gravity is associated with approximately 0.016 decrease in the probability to depart above the guidelines (p=.017),[46] and 0.073 decrease in the probability to sentence a defendant either above the guidelines or within the aggravated range (p=.006).

Results are less conclusive regarding judicial tendency to sentence in more lenient guidelines categories, but the signs of the coefficients are in the expected direction if judges from the "high" exposure group might be more likely to sentence in lenient categories than their colleagues exposed to lower gravity. At the margin, judges with "high" gravity exposure have approximately 0.021 higher likelihood to depart below the guidelines than judges with "low" exposure, a difference which is not significant under the basic model (p=.280), but is significant when accounting for additional covariates (p=.021). Looking at judicial tendency to sentence either below the guidelines or within the mitigated range, the marginal effect is of approximately 0.014 but is insignificant through all specifications.

The difference in the strength of the effect for sentencing in the mitigated guidelines range or below the guidelines might be due to the sensitivity of statistical significance to sample size, but it also seems reasonable under the circumstances. First, since offense gravity has a distribution with a long right tail (there are more drug cases than murder cases, for example) judges in both groups have a considerable portion of their caseloads containing milder offenses, but only judges in the high exposure condition were also initially exposed to more frequent serious cases. The difference in

---

relationship between another pair of outcomes (for example, sentencing within the mitigated guidelines range and all higher categories), etc.

[46] For above departures from the guidelines range the model include 18 judges from 9 out of the 10 matches, since in one match neither judge sentenced any cases above the guidelines recommendations.

exposure to graver offenses can lead judges in the "high" group to view as more common offenses or circumstances that judges in the "low" group might view as egregious at first. In addition, while the worst cases (except murder charges) are included in the analysis sample – allowing me to evaluate differences in the evaluations of the gravest cases in the caseloads – the mildest cases are not the cases analyzed but rather cases not eligible to an incarceration sanction at all that are excluded from the regression. Therefore, the comparison of the mildest incarceration cases is not in fact a comparison of cases in the lowest end of the caseloads.

As additional tests, I also run all models excluding cases with mandatory minimum sentences, and results remain similar for all variables (Table 9). I test both for excluding all cases in which mandatory minimum sentences exceed the lower guidelines limit, and for excluding only cases in which mandatory minimum sentences exceed the upper guidelines limit.

In light of the strong results for the impact of initial exposure to gravity on sentencing outcomes, I also run a "placebo" test. I use the matches that were previously dropped because judges did not experience initial significant differences in exposure to gravity through their caseloads, and test whether their sentencing outcomes are similar absent significant treatment. As is evident from Table A2, sentencing outcomes by judges in the sample who only experienced negligible differences in initial exposure to gravity are practically identical. The sentencing "gap" equals 0.108 (p=.852) for sentences imposed, and 0.153 (p=.854) for months of deviation from the guidelines.[47] The coefficients for sentencing across the different guidelines conformity categories are similarly insignificant. The "placebo" test indicates that when disparate exposure to gravity was negligible, there is no difference attributable to gravity exposure in the sentencing decisions of judges.

B. The Impact of Time on Updating Sentencing Practices

The exposure of judges to different levels of case gravity was only during the identification period. By the start of the comparison period and throughout it, the analysis in section III.B.3 ensures that the average exposure to gravity by judges in both groups is similar. To test whether the sentencing gap following the initial exposure to gravity is persistent or decays over time, the regressions account for the impact of time on all judges (*Time*) and on judges exposed to initial higher gravity levels (*Exposure*Time*), and results suggest the large effect of initial exposure to gravity is decaying over time.

[47] Table A2 displays the results for the 20 judges from the 10 matches who had at least 59 net hearing days after convergence, in order to compare an equivalent time period, as well as a period for which judicial caseloads can be reasonably expected to be balanced (otherwise the maximal comparison time for which all judges in the "placebo" matches hear cases is 46 net hearing days). Results are similarly insignificant for shorter and longer periods including the applicable matches for each period.

First, it is worth noting that although the *Time* variable by itself is significant only in some specifications, for all judges the lapse of time generally has a negative effect on sentencing, with every passing net hearing day associated with a decrease of between -0.011 to -0.013 months in sentence duration. Similarly, for the non-conformity dummy variables, the coefficients for the *Time* variable indicate decreasing likelihood to sentence above the guidelines range or in the guidelines aggravated range, and increasing likelihood to sentence below the guidelines range or in the guidelines mitigated range. The direction of the effect toward leniency with time is in line with psychological theory, suggesting that as judges gain more experience and are exposed to more criminal offenses, the frequently viewed cases might be routinized and lose some of their initial aura of severity. To the best of my knowledge, these are first empirical evidence of desensitization in judicial behavior.

The lapse of time, however, also significantly contributes to mitigating the sentencing gap between judges exposed to different initial gravity (*Exposure*Time*). For months of incarceration, every passing net hearing day decreases the sentencing gap between judges on average by approximately 0.032 (Table 6, p=.170) to 0.052 (Table 5, p=.007) months, or about 1-1.5 days. This effect is equivalent to a daily decrease of 2.5% in the sentencing gap of 2.049 months for *Sentence*, and of 1.8% in the sentencing gap of 1.731 moths for *Deviation*, suggesting that approximately by the passage of 40 net hearing days, the effect of initial exposure to gravity decays completely. A period of 40 net hearing days is equivalent to between two to sixteen calendar months across different matches, or about six calendar months on average. This indicates that throughout the period for which judicial caseloads are balanced with similar exposure to gravity, the effect of initial exposure persists for a substantial period of several months before decaying.

The significant differences in judicial likelihood to conform with different guidelines categories as well diminish over time. The difference in the probability to sentence above the guidelines decreases at a daily rate of 0.0005 (p=.023), equivalent to 3% of the -0.0165 initial difference. The difference in the probability to sentence above the guidelines or in the aggravated range decreases at a daily rate of 0.002 (p=.009), equivalent to 2.7% of the -0.0734 initial difference. Judicial likelihood to depart below the guidelines range decreases at a daily rate of -0.0005 (p=.010), which is about 2.4% of the initial difference of 0.0207.

The temporality of the sentencing gap is expected from the setting of the natural experiment utilized by the paper. Since after the end of the identification period the caseloads of judges from both "high" and "low" groups are balanced, over time judges have similar exposure to gravity and can be reasonably expected to update their sentencing behavior accordingly. Judicial reputation, court communities, learning from peers and references to precedents, can all further foster such learning. Nevertheless, the finding that initial exposure to gravity, over an average period of three months, can lead to a substantial and significant effect on sentencing outcomes over an average period of six months afterwards, even at the face of new information and balanced

caseloads, demonstrates that context-dependence can shape judicial behavior not only through immediate comparison across cases, but also have a longer lasting effect on judicial punitive attitudes and sentencing practices.

It is also important to note, that although the sentencing gap *under the terms of the natural experiment* is only temporary, it does not follow that the general concern from sentencing disparities driven by differences in caseload composition is temporal or of limited implications. Unlike the conditions set by the quasi experiment, in many courthouses differences across judicial docket compositions are systematic and persistent. While Pennsylvania Unified Courts of Common Pleas operate under a unified court structure, many states have two-tiered court systems in which different judges are presiding over misdemeanor and felony dockets. The increasing movement toward specialization in the courts and the proliferation of specialized courts lead to more and more judges with different scope of criminal gravity in their caseloads. Further, under the terms of the natural experiment, because of Pennsylvania's unified court model individual judicial dockets were drawn from a common pool of cases. But when judges are assigned to divisions with different subject matter jurisdiction the absolute magnitude of difference in their exposure to gravity might be greater than that experienced by judges in the setting set forth by the paper. When differences in judicial exposure to gravity are the result of a systematic diversion in the assignment of cases to judges, and are not corrected over time, the resulting differences in the evaluation of criminal cases may persist and pose a greater concern for sentencing disparities in the justice system. The next Section offers examples for such settings.

V. IMPLICATIONS

Relative judgments affected by judicial caseload composition can have normative and practical implications for the fairness of the justice system and the institutional design of the courts. Treating similar cases differently may not only undermine notions of equality and justice, but when unaccounted for can also hinder the fulfillment of social policy in particular areas or of wider policy goals such as deterrence. At the same time, the multiplicity of court models across states and fields suggest the potential implications of such bias can be broad.

A. Assigning Misdemeanor and Felony Cases

In the US states vary widely with regard to the structure of divisions in the criminal justice system. Some states use a two-tiered model of criminal jurisdiction in which a lower level court hears misdemeanor cases and a higher court sits in felony cases,[48]

[48] States in which there are two-tiered courts are, for example, Florida, Georgia, Kentucky, Michigan, New York, Ohio, and Virginia.

while other states have a unified court system - in which one state court has general criminal jurisdiction over all criminal offences.[49]

States may also change, at times, the court models they implement, usually for efficiency reasons. In 1998 California changed the entire structure of the state court system from a two-tiered to a unified model, in an effort to save on operating costs of the courts.[50] In 2004 the Criminal Court (previously hearing only misdemeanors) and the Supreme Court (previously hearing only felonies) in the Bronx borough in New York were merged to a single Supreme Court hearing misdemeanor and felony cases together, with the hope to expedite the disposition of the large misdemeanor backlogs that accumulated in the Criminal Court.[51] In 2012, in light of the creation of large felony backlogs[52] and due to constitutional restrictions on the assignments of judges across courts,[53] the Bronx courts reverted to the two-tiered model. While division of subject-matter jurisdiction across courts is implemented and changed by politicians, judges and administrators based on practical constraints and efficiency goals, the resulting change in the contours of judges' caseloads may impact not only the processing of cases but the substantive evaluation of cases as well.

Furthermore, in practice, even in states with a unified court system the allocation of cases across particular judges in the court may nevertheless result in different judges presiding over misdemeanors, felonies or both, and such decisions are often reached at the county or even the particular court level.[54] As Baum observed, for decisions about docket assignments: "the most meaningful level is the county or other unit of trial-court jurisdiction rather than the state."[55] Yet, to the extent that the decisions to assign judges to different dockets may lead to different sentencing outcomes, one might question the vast discretion granted to local administrators and judges in implementing

---

[49] States characterized by unified criminal courts systems are, for example, Alabama, California, Connecticut, Illinois, Indiana, Maryland, Massachusetts, Missouri, New Jersey, Tennessee, and Wisconsin.

[50] Proposition 220 Courts. Superior and Municipal Court Consolidation, June 1998 (http://www.lao.ca.gov/ballot/1998/220_06_1998.htm)

[51] 22 NYCRR §42.1.

[52] New York City Bar Report on the Merger of the Bronx Supreme and Criminal Courts, June 2009 (http://www.nycbar.org/pdf/report/20071735Merger_BronxSupreme_CriminalCourts.pdf).

[53] State of New York Unified Court System Report, "The Bronx Criminal Division: Merger After Five Years", October 2009 (https://www.nycourts.gov/publications/pdfs/BronxReport11-09.pdf)

[54] In states with two-tiered models some constitutional limitation on judges' authorities restrict the ability of misdemeanor judges to sentence cases outside the jurisdiction of their court, however felony judges might occasionally decide misdemeanor cases. In states with unified court models the situation is much more complex, as judges are authorized to hear all type of cases and in practice the allocation of cases across judges may vary considerably based on the discretion of local authorities and the presiding judges of the courts. For examples, *see*, Lawrence Baum, *SPECIALIZING THE COURTS* (2011), 20-21, depicting the great differences between state courts organizational charts and their operation in practice. According to Baum, in a survey conducted on 1977, 16% of judges in general jurisdiction trial courts were actually serving in specialized courtrooms and "that level almost surely is higher today."

[55] *Id*, *Id.*

such rules. Within a given jurisdiction concerned with equitable treatment across cases, greater emphasis should be granted to ensure that different courthouses within the jurisdiction implement similar case assignment rules, and to promote centralist decision-making at the county or state level, rather than at discretion of the presiding judges of particular courts.

More broadly, while the paper focuses on criminal sentencing decisions, relative judgments might be relevant to other subject matters as well. For example, when states divide civil jurisdiction based on the amount of claim, similar effects could be relevant to the evaluation of tort claims. The scholarship on behavioral biases in the evaluation of damage awards highlighted the problems faced by jurors in determining such awards absent the context of other cases.[56] Yet, the concern of relative judgments might suggest at least one dimension in which it is the decisions by judges, rather than juries, that might be biased, exactly because of the contextual framework provided by other cases they adjudicate.

B. Specialized Courts

Courts are also specialized around crime categories (such as drug courts or domestic violence courts), around the population of offenders (as in the Juvenile Courts) and of victims (as in the Elderly Courts). All such specializations are usually associated with a bounded scope of criminal gravity under the jurisdiction of the specialized court, mainly misdemeanors or lower-class felonies.[57] As a result judges in specialized courts will end up getting exposed only to cases of a limited scope of gravity in comparison to the variety of cases in general criminal courts, and the decision to assign certain cases to a specialized division or to a general criminal court can influence the relative assessment of severity of such cases and the sentences they receive. While specialization of the courts may carry many benefits, not accounting for the additional impact of the changed caseload exposure may lead to unintended consequences – opposite of those intended by policymakers.

For example, juvenile courts were established with an eye toward rehabilitation of adolescents, viewed as lesser offenders usually with no prior criminal record. Already in the juvenile courts of the progressive-era Baum identifies how, although created with rehabilitative and treatment-oriented goals, in practice many such courts took a punitive approach and "turned out to be a bad bargain for juvenile defendants."[58] In

---

[56] *See*, *supra*, note 1.

[57] Drug cases, even across the entire scope of criminal charges, are usually still relatively milder than the equivalent spectrum of general criminal activity. Domestic violence courts are limited to hearing only misdemeanors or class D felonies, and in many cases the criminal nature of such cases is derived more from the familial and mental aspect associated with them than would otherwise be outside a domestic setting. Juvenile courts were established with an eye toward rehabilitation of adolescents that are viewed as lesser offenders usually with no prior criminal record.

[58] Baum, *supra* note 50, 108-109. Baum suggests such shift might have resulted from the decrease in the status of juvenile courts and their staffing by judges less committed to the primary rehabilitative goal.

modern courts, a common practice is to transfer juvenile delinquents from the juvenile court to an adult criminal court in graver cases to allow sentencing them more harshly as adults. Yet, researchers are divided over the effectiveness of such transfers and some studies found that most juveniles transferred to adult court are not given longer punishments than if they had remained in the juvenile justice system,[59] that chronic young property offenders typically receive more lenient sentences when they appear in criminal court as first-time adult offenders than they would have received in juvenile court,[60] that juvenile defendants charged with violent crimes were more likely than those retained in juvenile court to be released pending adjudication.[61]

Evidence from Chicago Domestic Violence Division tells a similar story. In January 2010 Illinois amended its domestic violence laws in order to increase punishment for recidivist domestic violence offenders – enabling to charge recidivist offenders of domestic battery with class 4 felonies instead of misdemeanors, because of the aggravating prior record.[62] In the city of Chicago, the Domestic Violence Division had jurisdiction only over misdemeanor domestic violence cases, while felony charges were under the jurisdiction of the central Criminal Court. Following the amendment, when recidivist domestic violence offenders in Chicago were charged with felonies (at the Criminal Court) instead of misdemeanors, the felony charges received lower sentences than those that would have been ordered for equivalent misdemeanors by the Domestic Violence Division. Prosecutors' requests for harsher punishments did not succeed in increasing sentencing levels at the Criminal Court, despite the intended purpose of the amendment, until after a year it was decided to transfer class 4 felonies to the jurisdiction of the Domestic Violence Division.[63]

These examples cannot isolate the impact of caseload exposure from other characteristics of the specialized courts, and therefore should be interpreted with caution, but the direction of the effects, and the clear contrast between the intended

---

[59] For an overview of the literature *see*, *e.g.*: Fagan, J.A. *Separating the Men from the Boys: The Comparative Impacts of Juvenile and Criminal Court Sanctions on Recidivism of Adolescent Felony Offenders,* in SOURCEBOOK ON SERIOUS, CHRONIC AND VIOLENT JUVENILE OFFENDERS (Howell et al. eds.) (1995) 238-260, at p. 244. Also *see*, *e.g.*, Kristine Kinder, Carol Veneziano, Michael Fichter and Henry Azuma, *A Comparison of the Dispositions of Juvenile Offenders Certified as Adults with Juvenile Offenders Not Certified*, 46 JUV. & FAM. CT. J. 37 (1995); James C. Howell, *Juvenile Transfers to the Criminal Justice System: State of the Art*, 18 LAW & POLICY 17 (1996).

[60] S*ee, e.g.,* Carole Wolff Barnes and Randal S. Franz, *Questionably Adult: Determinants and Effects of the Juvenile Waiver Decision*, 6 JUST. Q. 117 (1989); Marcy Rasmussen Podkopacz and Barry C. Feld, *Judicial waiver policy and practice: persistence, seriousness and race*, 14 LAW & INEQ. 73 (1995): 73.

[61] S*ee, e.g.,* David L. Myers and Kraig Kiehl, *The Predisposition Status of Violent Youthful Offenders: Is There A "Custody Gap" in Adult Criminal Court?*, 3 JUSTICE RESEARCH AND POLICY 115 (2001).

[62] 720 ILCS 5/12-3.2

[63] Based on conversations of the author with judges and court administrators at the Chicago Domestic Violence Court, the Chicago Domestic Relations Court, and the Chicago Criminal Court. See also: The State of Illinois Cook County Press Release, 10/28/2010 (http://www.cookcountycourt.org/MEDIA/ViewPressRelease/tabid/338/ArticleId/459/Courts-new-Domestic-Violence-Division-expands-to-two-suburban-districts.aspx).

purposes and their outcomes, are in line with the bias predicted by this paper. They therefore might illustrate the importance of accounting for the impact of caseload composition on relative judgments in addressing policy goals in the criminal justice system. When policymakers wish to institute separate divisions for certain subsets of cases, careful attention should be given to the unique purposes such specialization seeks to fulfill, and to the possible effects of the bounded scope of cases under the jurisdiction of such divisions on the formation of judicial punitive approaches and the substantive resolutions of cases.

In certain cases, specialization might be more successful in achieving its intended purposes. For example, when the goal is to obtain different sentencing levels in particular fields, sometimes creating specialized courts can do so more easily and more effectively than changing legislation, guidelines or judicial instructions. This could be to fight a crime epidemic – as was at times the case with drugs or weapons crimes – requiring uniform harshening of sentencing levels; when policy makers are concerned that certain crimes have unique features whose gravity cannot be properly identified when they are immersed in the general criminal caseload – as is argued the case for domestic violence offenses; or even for cases that require a separate sentencing scale, such as traffic cases or crimes who carry mostly monetary fines which might be harder to scale in comparison to crimes that are mainly punished with imprisonment. However, such benefits can come at the price of assessing the appropriate relation between the court's cases and other types of crimes.

When the reason for separating the treatment of certain cases is due to special professional knowledge, training or resources that are required for the appropriate treatment of identified populations or subtypes of cases, specialized courts might have unintended consequences of harshening sentencing levels. In such instances, a better balance to enable the desired separate treatment while avoiding the uncalled for harsher punishments, could be to implement a model of "judge concentration" – having cases of a particular field concentrated around a limited number of judges – rather than of "case concentration" – when judges concentrate on a limited range of cases.[64] Some judges can be appointed to specialize in the subset of cases (based on their prior experience, or to receive the relevant training), instead of dividing them across all judges; but that subset of cases will constitute only part of their caseloads while they continue receiving cases of all other types in the remaining part of their caseload. In such a way, some of the benefits of specialization can be achieved, while the exposure to other types of cases as well can mitigate the contextual influences that exist when handling only narrower categories of cases of lower-gravity.

[64] Borrowing the terminology used by Baum, *supra* note 50.

C. Geographic Jurisdiction

Potential disparate exposure to gravity can affect the comparative study of the courts in even wider settings, for example also with regard to the division of geographic jurisdiction. Scholarship analyzing the differences across courts in rural and urban areas found smaller rural or suburban courts to have higher sentencing levels than large urban courts, and such differences were ascribed to caseload pressures and mainly to the courts communities' ties and norms.[65] However, rural areas generally also face lower crime levels than large urban areas, and thus relative comparison across cases can lead rural judges to judge the same cases more harshly than urban judges.

While there could be good arguments in favor of local courts serving their local communities, including through adjustments in their sentencing levels, there are also reasons why awareness of such a phenomenon is important, and such impact is not necessarily always desirable. The local districts benefiting from harsher punishment levels and increased deterrence are not the ones internalizing the costs associated with such benefits, since the cost of imprisonment is born at the state (for state prisons) or county (for county jails) level. Parole boards in prisons were documented to be susceptible to relative judgments bias as well – and to be less likely to release prisoners serving "short terms" in comparison to the prison population, even if such punishments were considered harsh by the judges imposing them and even when the trial judges recommended such prisoners will be released after serving their minimum sentence only.[66] With prison overpopulation becoming an increasingly prevalent issue in the US, and Supreme Court decisions in several states ordering releases of prisoners from overcrowded prisons, possible misalignments in the sentencing of offenders across district courthouses within the county or the state might be a factor that should be taken into account in release decisions.

---

[65] *See, e.g.,* Brian D. Johnson, *Contextual Disparities in Guidelines Departures: Courtroom Social Contexts, Guidelines Compliance, and Extralegal Disparities in Criminal Sentencing*, 43 CRIMINOLOGY 761 (2005); Jeffery T. Ulmer and Brian Johnson, *Sentencing in Context: A Multilevel Analysis*, 42 (1) CRIMINOLOGY 137 (2004); Paula Kautt, Location, Location, Location: Interdistrict and Intercircuit Variation in Sentencing Outcomes for Federal Drug-trafficking Offenses, 19 *JUSTICE QUARTERLY* 633 (2002); James Eisenstein, Roy Flemming et al., *The Contours of Justice: Communities and Their Courts* (1988).

[66] See, Emerson, *supra* note 11, 433-434, quoting Caleb Foote "The Sentencing Function," in ANNUAL CHIEF JUSTICE WARREN CONFERENCE ON ADVOCACY IN THE UNITED STATES, A PROGRAM FOR PRISON REFORM: THE FINAL REPORT (1972). Foote explains that since incarceration was rarely used as a punishment in marijuana cases, in the first few cases sentenced to imprisonment judges viewed it as an extremely harsh sentence and recommended offenders will be released after the minimum time served. The parole board at the Adult Authority in California however, was dealing with a population of offenders most of whom serve sentences of at least 30 months. In comparison to the general prison population, the lower sentences imposed in marijuana cases seemed too lenient to warrant early releases, not only after the minimum time served (usually 6-12 months) but even after 18 months.

VI. CONCLUSION

Judicial professionalism is a hallmark of the courts system. The fact that judges routinely hear many and varying types of cases is widely perceived as enhancing judicial competence and experience, granting judges more knowledge and a wider perspective on cases. Aside from the virtues of judicial experience, however, the paper identifies an additional effect of the wider perspective judges have – such perspective depends on the scope of cases in judicial caseloads.

Legal judgments are affected by the relative comparison of a particular case to the other cases in the caseload before the judge. As a result, when judicial caseloads contain cases of different gravity, the evaluation of severity of a particular case and the punitive outcome in that case will be affected by how that case fares in comparison to the other cases in the caseload. The results of the paper provide evidence that judges that are exposed to graver cases in their caseloads sentence defendants significantly more leniently than judges whose caseloads contain only a limited scope of gravity. The results also indicate that such differences are substantial not only in magnitude but in their lasting effect as well. When gravity levels in judicial caseloads change, sentencing patterns are updated accordingly, but such updating is gradual and takes a long time to implement. Initial judicial exposure to different gravity levels through caseloads, it seems, impacts not only the immediate comparison across cases, but also the formation of more general judicial sentencing practices that judges continue to apply in future cases.

The most important implication of this phenomenon is that rules for case assignment across courts or judges are not merely operational in nature, but can influence the substantive outcomes reached by judges in particular cases. In light of the variance in court models, the increased movement toward specialization in the courts, and the great discretion given to local court administration in cases assignments, such consequences might be especially concerning. More broadly, from the policymaker's perspective, as specialization becomes increasingly prevalent in the legal system, connecting divergence in outcomes to institutional capacity and learning the reasons for any such influences can serve as an important tool in implementing future policies more suitable to achieving their goals. While as a practical matter, complete consistency of judgments is an unattainable ideal, when the sources of bias are connected to institutional capacity they might be easier to address by policymakers and architects of the courts than when disparities are assumed to be based on preferences.

Table 4. Summary Statistics

Variable	Obs.	Mean	Std. Dev.	Min	Max
Sentence	1838	7.95	9.95	0	60
Deviation	1817	-1.78	8.32	-120	50
Deviation (winsorized)	1817	-1.70	6.24	-24	23.5
Conformity Above	1838	.044	.205	0	1
Conformity High	1838	.141	.349	0	1
Conformity Low	1838	.201	.401	0	1
Conformity Below	1838	.055	.229	0	1
Offense Gravity Score (OGS)	1838	4.14	2.29	1	14
Prior Record Score (PRS)	1838	2.08	2.10	0	6
Mandatory Minimum	1838	.944	4.64	0	60
Trial	1508	.102	.301	0	1
White Defendant	1838	.501	.500	0	1
African-American Defendant	1838	.395	.489	0	1
Hispanic Defendant	1838	.063	.243	0	1
Female Defendant	1837	.135	.342	0	1

Summary statistics are for the regression sample, including only offenses eligible for discretionary incarceration for the $1 \leq T \leq 59$ net hearing days period, for judges identified as having "low" or "high" *Exposure*.

Table 5. Sentencing outcomes (in months)

	Sentence				
	(1)	(2)	(3)	(4)	(5)
Exposure ("High"=1)	-2.173*	-1.797*	-1.904**	-1.939*	-2.049**
	(0.925)	(0.652)	(0.623)	(0.694)	(0.688)
	[0.030]	[0.013]	[0.006]	[0.012]	[0.008]
Time (net days)	-0.018	-0.009	-0.008	-0.013	-0.011
	(0.019)	(0.010)	(0.009)	(0.010)	(0.010)
	[0.347]	[0.350]	[0.424]	[0.221]	[0.263]
Exposure*Time	0.061+	0.043*	0.044*	0.049**	0.052**
	(0.031)	(0.017)	(0.016)	(0.017)	(0.017)
	[0.062]	[0.018]	[0.012]	[0.009]	[0.007]
Offense Gravity Score (OGS)		3.016***	3.000***	2.825***	2.809***
		(0.172)	(0.172)	(0.239)	(0.238)
		[0.000]	[0.000]	[0.000]	[0.000]
Prior Record Score (PRS)		1.247***	1.207***	1.105***	1.076***
		(0.182)	(0.186)	(0.197)	(0.204)
		[0.000]	[0.000]	[0.000]	[0.000]
Mandatory Minimum (months)				0.488***	0.486***
				(0.051)	(0.051)
				[0.000]	[0.000]
Trial				2.006***	2.014***
				(0.502)	(0.495)
				[0.001]	[0.001]
Defendant Characteristics	No	No	Yes	No	Yes
Offense Type	No	Yes	Yes	Yes	Yes
R-Square	0.106	0.537	0.541	0.581	0.584
N	1838	1838	1837	1508	1507

All Models include district fixed effects, year fixed effects and district by year fixed effects.
Robust clustered standard errors in parentheses. P-values in brackets.
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 6. Sentencing relative to the standard guidelines range (in months)

	Deviation				
	(1)	(2)	(3)	(4)	(5)
Exposure ("High"=1)	-1.683*	-1.707*	-1.839*	-1.550*	-1.731*
	(0.695)	(0.708)	(0.677)	(0.693)	(0.704)
	[0.026]	[0.026]	[0.014]	[0.038]	[0.024]
Time (net days)	-0.015	-0.015	-0.012	-0.015	-0.013
	(0.015)	(0.015)	(0.015)	(0.012)	(0.012)
	[0.338]	[0.355]	[0.416]	[0.238]	[0.285]
Exposure*Time	0.032	0.037+	0.040+	0.028	0.032
	(0.019)	(0.021)	(0.021)	(0.021)	(0.022)
	[0.109]	[0.095]	[0.078]	[0.202]	[0.170]
Offense Gravity Score (OGS)		-1.076***	-1.097***	-1.070**	-1.094**
		(0.277)	(0.280)	(0.320)	(0.325)
		[0.001]	[0.001]	[0.003]	[0.003]
Prior Record Score (PRS)		-0.898***	-0.927***	-1.046***	-1.077***
		(0.166)	(0.175)	(0.163)	(0.179)
		[0.000]	[0.000]	[0.000]	[0.000]
Mandatory Minimum (months)				0.067	0.068
				(0.067)	(0.067)
				[0.327]	[0.327]
Trial				1.440*	1.483*
				(0.596)	(0.586)
				[0.026]	[0.020]
Defendant Characteristics	No	No	Yes	No	Yes
Offense Type	No	Yes	Yes	Yes	Yes
R-Square	0.053	0.153	0.160	0.157	0.164
N	1817	1817	1816	1487	1486

All Models include district fixed effects, year fixed effects and district by year fixed effects.
Robust clustered standard errors in parentheses. P-values in brackets.
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 7. Sentencing in harsher guidelines categories

	Above Departure[1]					*Above Departure or Aggravated Range*				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Exposure ("High"=1)	-0.0242**	-0.0145*	-0.0146*	-0.0169*	-0.0165*	-0.0827***	-0.0674**	-0.0681**	-0.0723**	-0.0734**
	(0.008)	(0.006)	(0.007)	(0.007)	(0.007)	(0.024)	(0.021)	(0.021)	(0.026)	(0.027)
	[0.004]	[0.022]	[0.026]	[0.017]	[0.017]	[0.001]	[0.001]	[0.001]	[0.005]	[0.006]
Time (net days)	-0.0006**	-0.0003*	-0.0003*	-0.0004+	-0.0004+	-0.0017**	-0.0012*	-0.0012*	-0.0014*	-0.0014*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)
	[0.003]	[0.023]	[0.031]	[0.058]	[0.070]	[0.009]	[0.016]	[0.019]	[0.029]	[0.033]
Exposure*Time	0.0008***	0.0005**	0.0005**	0.0005*	0.0005*	0.0018*	0.0016**	0.0016**	0.0019**	0.0020**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
	[0.000]	[0.001]	[0.002]	[0.023]	[0.023]	[0.015]	[0.005]	[0.006]	[0.009]	[0.009]
Offense Gravity Score (OGS)		-0.0008	-0.0009	-0.0005	-0.0005		-0.0186***	-0.0187***	-0.0152***	-0.0153***
		(0.001)	(0.001)	(0.001)	(0.001)		(0.004)	(0.004)	(0.004)	(0.004)
		[0.587]	[0.543]	[0.737]	[0.710]		[0.000]	[0.000]	[0.000]	[0.000]
Prior Record Score (PRS)		-0.0080***	-0.0082***	-0.0075***	-0.0075***		-0.0452***	-0.0452***	-0.0426***	-0.0426***
		(0.001)	(0.001)	(0.001)	(0.001)		(0.003)	(0.003)	(0.003)	(0.003)
		[0.000]	[0.000]	[0.000]	[0.000]		[0.000]	[0.000]	[0.000]	[0.000]
Mandatory Minimum (months)				-0.0001	-0.0001				-0.0023	-0.0023
				(0.000)	(0.000)				(0.003)	(0.003)
				[0.762]	[0.782]				[0.398]	[0.380]
Trial				0.0064	0.0055				0.0453	0.0441
				(0.006)	(0.006)				(0.032)	(0.032)
				[0.249]	[0.333]				[0.163]	[0.173]
Defendant Characteristics	No	No	Yes	No	Yes	No	No	Yes	No	Yes
Offense Type	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Pseudo R-Square	0.073	0.181	0.188	0.206	0.215	0.059	0.207	0.208	0.211	0.212
N	1690	1690	1689	1362	1361	1835	1835	1834	1495	1494

(1) For above departures – results include 18 judges from 9 matches, since in one match neither judge sentenced any cases above the guidelines range.
All Models include district fixed effects, year fixed effects and district by year fixed effects.
Logit coefficients are reported as marginal effects, evaluated at the sample mean.
Robust clustered standard errors in parentheses. P-values in brackets. + p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 8. Sentencing in lenient guidelines categories

	Below Departure					*Below Departure or Mitigated Range*				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Exposure ("High"=1)	0.0230	0.0116	0.0122	0.0197*	0.0207*	0.0181	0.0105	0.0113	0.0136	0.0144
	(0.021)	(0.007)	(0.008)	(0.008)	(0.009)	(0.039)	(0.023)	(0.022)	(0.030)	(0.030)
	[0.280]	[0.121]	[0.108]	[0.019]	[0.021]	[0.641]	[0.654]	[0.614]	[0.646]	[0.627]
Time (net days)	0.0001	0.0001	0.0001	0.0003+	0.0003+	0.0006	0.0005	0.0005	0.0008	0.0008
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)
	[0.726]	[0.399]	[0.399]	[0.082]	[0.086]	[0.476]	[0.248]	[0.262]	[0.173]	[0.167]
Exposure*Time	-0.0002	-0.0002	-0.0002	-0.0005*	-0.0005*	0.0005	0.0001	0.0001	0.0001	0.0001
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
	[0.635]	[0.140]	[0.124]	[0.012]	[0.010]	[0.603]	[0.904]	[0.853]	[0.896]	[0.858]
Offense Gravity Score (OGS)		0.0065***	0.0064***	0.0084***	0.0082***		0.0337***	0.0342***	0.0408***	0.0414***
		(0.001)	(0.001)	(0.002)	(0.002)		(0.004)	(0.004)	(0.004)	(0.005)
		[0.000]	[0.000]	[0.000]	[0.000]		[0.000]	[0.000]	[0.000]	[0.000]
Prior Record Score (PRS)		0.0089***	0.0091***	0.0115***	0.0116***		0.0387***	0.0395***	0.0460***	0.0464***
		(0.001)	(0.001)	(0.001)	(0.002)		(0.005)	(0.005)	(0.005)	(0.005)
		[0.000]	[0.000]	[0.000]	[0.000]		[0.000]	[0.000]	[0.000]	[0.000]
Mandatory Minimum (months)				-0.0005*	-0.0005+				-0.0052***	-0.0051***
				(0.000)	(0.000)				(0.001)	(0.001)
				[0.049]	[0.062]				[0.000]	[0.000]
Trial				-0.0115*	-0.0110*				-0.0360	-0.0380+
				(0.005)	(0.005)				(0.022)	(0.022)
				[0.023]	[0.024]				[0.110]	[0.089]
Defendant Characteristics	No	No	Yes	No	Yes	No	No	Yes	No	Yes
Offense Type	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Pseudo R-Square	0.071	0.267	0.272	0.270	0.278	0.148	0.411	0.415	0.416	0.419
N	1793	1793	1792	1486	1485	1802	1802	1801	1495	1494

All Models include district fixed effects, year fixed effects and district by year fixed effects.
Logit coefficients are reported as marginal effects, evaluated at the sample mean.
Robust clustered standard errors in parentheses. P-values in brackets. + p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 9. Sentencing outcomes excluding cases with mandatory minimum sentences

	Excluding cases with mandatory minimums equal to or exceeding the guidelines upper limit						*Excluding cases with mandatory minimums equal to or exceeding the guidelines lower limit*					
	OLS		LOGIT				OLS		LOGIT			
	Sentence (months)	*Deviation (months)*	*Above Departure*	*Above Departure or Aggravated Range*	*Below Departure or Mitigated Range*	*Below Departure*	*Sentence (months)*	*Deviation (months)*	*Above Departure*	*Above Departure or Aggravated Range*	*Below Departure or Mitigated Range*	*Below Departure*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Exposure	-2.318**	-2.060*	-0.0181*	-0.0646*	0.0235	0.0306*	-1.961*	-2.201*	-0.0116*	-0.0511+	0.0263	0.0387*
("High"=1)	(0.701)	(0.819)	(0.007)	(0.026)	(0.045)	(0.013)	(0.830)	(0.933)	(0.005)	(0.027)	(0.063)	(0.018)
	[0.004]	[0.021]	[0.013]	[0.014]	[0.598]	[0.022]	[0.029]	[0.029]	[0.027]	[0.061]	[0.677]	[0.028]
Time (net days)	-0.013	-0.018	-0.0004+	-0.0011+	0.0011	0.0005+	-0.003	-0.020	-0.0002	-0.0010+	0.0013	0.0005
	(0.010)	(0.013)	(0.000)	(0.001)	(0.001)	(0.000)	(0.010)	(0.014)	(0.000)	(0.001)	(0.001)	(0.000)
	[0.201]	[0.192]	[0.088]	[0.052]	[0.159]	[0.086]	[0.767]	[0.181]	[0.157]	[0.094]	[0.233]	[0.113]
Exposure*Time	0.055**	0.040	0.0005*	0.0018**	0.0001	-0.0007*	0.045*	0.041	0.0003+	0.0015*	0.0007	-0.0007*
	(0.017)	(0.026)	(0.000)	(0. 001)	(0.001)	(0.000)	(0.019)	(0.029)	(0.000)	(0.001)	(0.001)	(0.000)
	[0.005]	[0.135]	[0.026]	[0.009]	[0.901]	[0.012]	[0.028]	[0.178]	[0.057]	[0.029]	[0.630]	[0.044]
Offense Gravity	3.018***	-1.150**	-0.0007	-0.0146***	0.0612***	0.0114***	3.182***	-1.176**	-0.0005	-0.0150***	0.0749***	0.0136***
Score (OGS)	(0.246)	(0.366)	(0.001)	(0.003)	(0.006)	(0.002)	(0.248)	(0.392)	(0.001)	(0.003)	(0.008)	(0.003)
	[0.000]	[0.005]	[0.631]	[0.000]	[0.000]	[0.000]	[0.000]	[0.007]	[0.592]	[0.000]	[0.000]	[0.000]
Prior Record	1.288***	-1.141***	-0.0081***	-0.0388***	0.0687***	0.0160***	1.324***	-1.175***	-0.0054***	-0.0403***	0.0828***	0.0187***
Score (PRS)	(0.202)	(0.220)	(0.001)	(0.003)	(0.006)	(0.002)	(0.216)	(0.256)	(0.001)	(0.003)	(0.007)	(0.002)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Mandatory	0.292*	-0.108	-0.0025	-0.0012	-0.0056**	-0.0003	0.338*	-0.396***				
Minimum	(0.123)	(0.163)	(0.002)	(0.005)	(0.002)	(0.001)	(0.128)	(0.065)				
	[0.028]	[0.517]	[0.158]	[0.795]	[0.004]	[0.518]	[0.017]	[0.000]	No	No	No	No
Trial	2.051***	1.655**	0.0088	0.0564+	-0.0576+	-0.0156*	2.067**	1.675*	0.0039	0.0447	-0.0781+	-0.0205*
	(0.527)	(0.572)	(0.007)	(0.032)	(0.033)	(0.007)	(0.601)	(0.609)	(0.005)	(0.035)	(0.042)	(0.009)
	[0.001]	[0.009]	[0.219]	[0.077]	[0.083]	[0.024]	[0.003]	[0.013]	[0.394]	[0.206]	[0.066]	[0.018]
R-Square/												
Pseudo R-Square	0.534	0.168	0.216	0.2324	0.3807	0.2538	0.521	0.177	0.241	0.265	0.367	0.250
N	1312	1291	1182	1300	1300	1292	1137	1116	1039	1127	1127	1119

All Models include district fixed effects, year fixed effects and district by year fixed effects. Logit coefficients are reported as marginal effects, evaluated at the sample mean. Robust clustered standard errors in parentheses. P-values in brackets. + p<0.10, * p<0.05, ** p<0.01, *** p<0.001

APPENDIX

Table A1. Robustness tests for different time periods

Panel A. Sentencing gap (OLS)

	Sentence (months)			*Deviation (months)*		
	T≤49	T≤54	T≤59	T≤49	T≤54	T≤59
	(1)	(2)	(3)	(4)	(5)	(6)
Exposure ("High"=1)	-2.353**	-1.920**	-2.049**	-1.984*	-1.563*	-1.731*
	(0.818)	(0.649)	(0.688)	(0.845)	(0.645)	(0.704)
	[0.010]	[0.008]	[0.008]	[0.030]	[0.025]	[0.024]
Time (net days)	0.003	-0.003	-0.011	-0.014	-0.017	-0.013
	(0.015)	(0.013)	(0.010)	(0.019)	(0.016)	(0.012)
	[0.867]	[0.825]	[0.263]	[0.463]	[0.298]	[0.285]
Exposure*Time	0.069*	0.048*	0.052**	0.049	0.029	0.032
	(0.030)	(0.020)	(0.017)	(0.032)	(0.023)	(0.022)
	[0.031]	[0.025]	[0.007]	[0.143]	[0.216]	[0.170]
Offense Gravity Score (OGS)	2.959***	2.866***	2.809***	-1.142**	-1.138**	-1.094**
	(0.249)	(0.229)	(0.238)	(0.393)	(0.350)	(0.325)
	[0.000]	[0.000]	[0.000]	[0.009]	[0.004]	[0.003]
Prior Record Score (PRS)	1.033***	1.023***	1.076***	-1.159***	-1.156***	-1.077***
	(0.226)	(0.213)	(0.204)	(0.229)	(0.204)	(0.179)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Mandatory Minimum (months)	0.473***	0.494***	0.486***	0.065	0.073	0.068
	(0.052)	(0.045)	(0.051)	(0.076)	(0.068)	(0.067)
	[0.000]	[0.000]	[0.000]	[0.405]	[0.297]	[0.327]
Trial	2.277**	1.867**	2.014***	1.408+	1.298*	1.483*
	(0.667)	(0.545)	(0.495)	(0.727)	(0.574)	(0.586)
	[0.003]	[0.003]	[0.001]	[0.068]	[0.036]	[0.020]
R-Square	0.585	0.589	0.584	0.163	0.167	0.164
N	1226	1364	1507	1209	1345	1486

All Models include defendant characteristics, offense type, district fixed effects, year fixed effects and district by year fixed effects.

Robust clustered standard errors in parentheses. P-values in brackets. + p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table A1 (continued). Robustness tests for different time periods

Panel B. Sentencing in harsher guidelines categories (LOGIT)

	Above Departure			*Above Departure or Aggravated Range*		
	T≤49	T≤54	T≤59	T≤49	T≤54	T≤59
	(7)	(8)	(9)	(10)	(11)	(12)
Exposure ("High"=1)	-0.0332*	-0.0246*	-0.0165*	-0.0920*	-0.0813**	-0.0734**
	(0.013)	(0.010)	(0.007)	(0.038)	(0.030)	(0.027)
	[0.011]	[0.014]	[0.017]	[0.015]	[0.007]	[0.006]
Time (net days)	-0.0006	-0.0004	-0.0004+	-0.0013	-0.0014	-0.0014*
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)
	[0.116]	[0.111]	[0.070]	[0.146]	[0.104]	[0.033]
Exposure*Time	0.0011**	0.0008*	0.0005*	0.0026*	0.0023*	0.0020**
	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)
	[0.007]	[0.015]	[0.023]	[0.031]	[0.015]	[0.009]
Offense Gravity Score (OGS)	-0.0008	-0.0007	-0.0005	-0.0147***	-0.0148***	-0.0153***
	(0.002)	(0.002)	(0.001)	(0.004)	(0.004)	(0.004)
	[0.659]	[0.675]	[0.710]	[0.000]	[0.000]	[0.000]
Prior Record Score (PRS)	-0.0101***	-0.0089***	-0.0075***	-0.0467***	-0.0437***	-0.0426***
	(0.001)	(0.001)	(0.001)	(0.003)	(0.003)	(0.003)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Mandatory Minimum	0.0000	0.0000	-0.0001	-0.0025	-0.0028	-0.0023
	(0.000)	(0.000)	(0.000)	(0.003)	(0.004)	(0.003)
	[0.980]	[0.972]	[0.782]	[0.460]	[0.436]	[0.380]
Trial	0.0062	0.0071	0.0055	0.0457	0.0480	0.0441
	(0.007)	(0.007)	(0.006)	(0.051)	(0.043)	(0.032)
	[0.396]	[0.295]	[0.333]	[0.372]	[0.261]	[0.173]
Pseudo R-Square	0.215	0.215	0.215	0.220	0.220	0.212
N	1099	1213	1361	1218	1353	1494

All Models include defendant characteristics, offense type, district fixed effects, year fixed effects and district by year fixed effects. Logit coefficients are reported as marginal effects, evaluated at the sample mean.
Robust clustered standard errors in parentheses. P-values in brackets. + p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table A1 (continued). Robustness tests for different time periods

Panel C. Sentencing in lenient guidelines categories (LOGIT)

	Below Departure or Mitigated Range			*Below Departure*		
	T≤49	T≤54	T≤59	T≤49	T≤54	T≤59
	(13)	(14)	(15)	(16)	(17)	(18)
Exposure ("High"=1)	0.0114	0.0118	0.0144	0.0230**	0.0223*	0.0207*
	(0.028)	(0.030)	(0.030)	(0.008)	(0.010)	(0.009)
	[0.685]	[0.694]	[0.627]	[0.006]	[0.020]	[0.021]
Time (net days)	0.0004	0.0006	0.0008	0.0004*	0.0005*	0.0003+
	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)
	[0.496]	[0.324]	[0.167]	[0.028]	[0.023]	[0.086]
Exposure*Time	0.0002	0.0002	0.0001	-0.0006**	-0.0006*	-0.0005*
	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)
	[0.801]	[0.783]	[0.858]	[0.001]	[0.010]	[0.010]
Offense Gravity Score (OGS)	0.0371***	0.0393***	0.0414***	0.0084***	0.0098***	0.0082***
	(0.006)	(0.005)	(0.005)	(0.002)	(0.002)	(0.002)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Prior Record Score (PRS)	0.0448***	0.0466***	0.0464***	0.0106***	0.0135***	0.0116***
	(0.006)	(0.006)	(0.005)	(0.002)	(0.002)	(0.002)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Mandatory Minimum	-0.0043***	-0.0048***	-0.0051***	-0.0003*	-0.0004*	-0.0005+
	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)
	[0.000]	[0.000]	[0.000]	[0.030]	[0.045]	[0.062]
Trial	-0.0305+	-0.0305	-0.0380+	-0.0110*	-0.0114*	-0.0110*
	(0.018)	(0.021)	(0.022)	(0.005)	(0.005)	(0.005)
	[0.085]	[0.138]	[0.089]	[0.016]	[0.037]	[0.024]
Pseudo R-Square	0.443	0.437	0.419	0.316	0.280	0.278
N	1218	1353	1494	1128	1257	1485

All Models include defendant characteristics, offense type, district fixed effects, year fixed effects and district by year fixed effects. Logit coefficients are reported as marginal effects, evaluated at the sample mean.
Robust clustered standard errors in parentheses. P-values in brackets. + p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table A2. Robustness tests for "placebo" sample

	OLS		LOGIT			
	Sentence	*Deviation*	*Above Departure*	*Above Departure or Aggravated Range*	*Below Departure or Mitigated Range*	*Below Departure*
	(1)	(2)	(3)	(4)	(5)	(6)
Exposure ("High"=1)	0.108	0.153	0.0044	-0.0387	-0.0099	0.0004
	(0.573)	(0.823)	(0.012)	(0.025)	(0.020)	(0.011)
	[0.852]	[0.854]	[0.719]	[0.116]	[0.625]	[0.973]
Time (net days)	0.022+	0.036	0.0005*	0.0005	-0.0002	-0.0003
	(0.011)	(0.025)	(0.000)	(0.001)	(0.000)	(0.000)
	[0.061]	[0.161]	[0.024]	[0.488]	[0.477]	[0.229]
Exposure*Time	-0.026+	-0.025	-0.0004	0.0001	0.0001	-0.0003
	(0.014)	(0.022)	(0.000)	(0.001)	(0.001)	(0.000)
	[0.076]	[0.276]	[0.173]	[0.862]	[0.813]	[0.292]
Offense Gravity Score (OGS)	3.020***	-0.335	-0.0003	-0.0256***	0.0195***	0.0049***
	(0.200)	(0.295)	(0.002)	(0.006)	(0.002)	(0.001)
	[0.000]	[0.271]	[0.840]	[0.000]	[0.000]	[0.000]
Prior Record Score (PRS)	1.417***	-0.810***	-0.0074**	-0.0534***	0.0236***	0.0097***
	(0.146)	(0.146)	(0.003)	(0.006)	(0.002)	(0.001)
	[0.000]	[0.000]	[0.009]	[0.000]	[0.000]	[0.000]
Mandatory Minimum	0.467***	0.142	0.0008*	0.0030	-0.0038**	-0.0019+
	(0.051)	(0.084)	(0.000)	(0.002)	(0.001)	(0.001)
	[0.000]	[0.107]	[0.044]	[0.110]	[0.003]	[0.097]
Trial	3.854**	-0.019	0.0168	0.0319	-0.0275**	-0.0128+
	(1.277)	(1.208)	(0.025)	(0.042)	(0.009)	(0.007)
	[0.007]	[0.987]	[0.500]	[0.447]	[0.001]	[0.068]
R-Square / Pseudo R-Square	0.646	0.106	0.166	0.173	0.396	0.274
N	1733	1749	1486	1724	1643	1290

All Models include defendant characteristics, offense type, district fixed effects, year fixed effects and district by year fixed effects. Logit coefficients are reported as marginal effects, evaluated at the sample mean.
Robust clustered standard errors in parentheses. P-values in brackets.
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001