In Search of the Booker Revolution

Ryan W. Scott

Available at: https://works.bepress.com/ryan_scott/1/
In 2005, the Supreme Court in United States v. Booker rendered the United States Sentencing Guidelines advisory. Arriving after eighteen years of complex and mandatory sentencing rules, the decision initially was heralded as revolutionary, both by critics and defenders of the federal Guidelines. But subsequent reports by the Sentencing Commission have shown few signs of a Booker revolution, revealing surprisingly minor changes.

The existing research on post-Booker sentencing is incomplete, however, because it has not examined the response of individual judges to the decision. That omission is critical, given that the reduction of inter-judge disparity was the central purpose of the Guidelines. Studying sentencing patterns by individual judges is notoriously difficult because the Commission does not disclose the identity of the sentencing judge when releasing case records. But one district court—the District of Massachusetts—has adopted a unique policy that makes key sentencing documents available to the public, allowing the analysis of judge-specific data.

This Article offers the first empirical evidence of individual judges’ responses to Booker, drawing on a dataset of sentences from 2002 to 2008 by judges in Boston who share a common case pool. An analysis of those sentences suggests a modest but clear increase in inter-judge disparity since Booker. The strength of the relationship between the identity of the sentencing judge and sentence length has increased, by some measures 60-80% above pre-Booker levels. The identity of the judge also has become a stronger predictor of sentencing relative to the Guidelines: some judges now sentence outside the guideline range more frequently, and to a greater extent, than their colleagues. Although it is difficult to know whether similar trends have played out in other districts, the Boston data suggest that judges’ disagreements about the Guidelines have a greater effect on sentencing outcomes since Booker.

* Associate Professor, Indiana University Maurer School of Law, Bloomington. The author would like to thank the judges of the United States District Court for the District of Massachusetts for adopting the public-access policy that made this Article possible. Thanks in particular to two judges of that court, Nancy Gertner and William Young, for their assistance and encouragement. Thanks as well to participants in the spring 2009 Yale Law School Sentencing Workshop, and to Amy Baron-Evans, Craig Bradley, Samuel Bray, Paul Cassell, Ken Dau-Schmidt, Paul Hofer, Robert Lawless, Andrew Martin, Michael McConnell, Ben Roin, Larry Solum, Michelle Spak, and Sandra Guerra Thompson for their comments on earlier drafts.
INTRODUCTION

In 2005, the Supreme Court in United States v. Booker\(^1\) rendered the United States Sentencing Guidelines “effectively advisory,” leaving judges free to impose any reasonable sentence consistent with the broad purposes of punishment outlined by Congress.\(^2\) After eighteen years in which judges treated the Guidelines as mandatory, the decision was expected to revolutionize federal sentencing. Critics of the Guidelines greeted Booker with enthusiasm, while defenders warned that the decision would plunge the courts into “pre-guideline chaos.”\(^3\) It was widely predicted that Booker would mark “the end of federal criminal sentencing as we know it.”\(^4\)

Today, signs of the Booker revolution are hard to find. Data released by the United States Sentencing Commission suggest that little has changed since the decision. Average sentence length has continued to increase, even for drug offenses, and the rate at which judges depart from the guideline sentencing range has increased only slightly. Most scholars and commentators describe the changes as surprisingly modest.\(^5\) Far from a revolution, Booker now looks like a non-event.

The existing research on post-Booker sentencing is incomplete, however, because it does not examine how individual judges have responded to the decision. That omission is critical, given that the reduction of inter-judge disparity was the central purpose of the Guidelines. As Attorney General Eric Holder explained in a June 2009 speech on sentencing policy, “the full impact of recent trends in sentencing jurisprudence is still unclear” and “these developments should be monitored carefully.”\(^6\) Central to that effort, Holder argued, is an assessment of whether post-Booker sentencing practices “show an increase in unwarranted sentencing disparities” based on “differences in judicial philosophy among judges working in the same courthouse.”\(^7\)

This Article addresses that critical gap in the research, offering the first empirical account of changes in sentencing patterns among individual judges in the wake of Booker. Studying sentencing by individual judges is notoriously difficult because the Commission refuses to disclose the identity of the sentencing judge when releasing case records. But one district court—the District of Massachusetts—has adopted a unique policy that makes key sentencing documents available to the public. Matching those documents with data released by the Commission allows the creation of a large dataset of sentences, from 2002 to 2008 by judges in Boston who share a common case pool.

An analysis of those sentences suggests, by two measures, a modest but clear increase in inter-judge disparity since Booker. First, the identity of the sentencing judge has become a stronger predictor of sentence length since the decision. For the relevant set of sentences, the “judge effect”—the percentage of variance in sentence length

\(^1\) 543 U.S. 220 (2005).
\(^2\) Id. at 245 (Breyer, J., writing for the Court).
\(^7\) Id. Holder also cited unwarranted regional disparity as an area for further research. Id.
explained by the identity of the judge—increased in the eighteen months after Booker. Although it weakened during the following eighteen months, it has grown sharply stronger since the Supreme Court’s decisions in Kimbrough v. United States\textsuperscript{8} and Gall v. United States,\textsuperscript{9} and now stands 60-80% above pre-Booker levels. The changes are modest, given that the rate of variance explained by the sentencing judge remains fairly small. But they suggest that judges’ general tendency toward lenience or severity plays a greater role in determining the length of prison sentences after Booker.

Second, the identity of the sentencing judge has become a stronger predictor of sentencing relative to the Guidelines. Some judges now sentence outside the guideline sentencing range much more frequently, and to a greater extent, than their colleagues. Since Booker, some judges in Boston continue to sentence below the guideline range at the same rate as before the decision, as little as 10% of the time. But the rate of below-range sentencing for other judges has tripled or quadrupled since Booker, with several judges sentencing below the Guidelines more than 40% or even 50% of the time. Statistical comparisons confirm that, since Booker, the identity of the sentencing judge has become a stronger predictor of the difference between the sentence imposed and the guideline range. These changes suggest that Booker has exacerbated a distinct form of inter-judge disparity, driven by disagreements about how often the Guidelines recommend an appropriate sentence.

These conclusions are necessarily tentative. As with any study of a single district court, there is a risk that the results are not representative of sentencing trends nationwide. And because inter-judge disparity is but one factor among many to consider in evaluating a sentencing system, nothing about these findings compels any judgment about whether Booker, on balance, has improved or worsened federal sentencing. Yet the Article explores an important dimension of post-Booker sentencing and supplies new information to help make those judgments. Drawing on the best evidence available, it offers an unprecedented look at how individual judges have responded to Booker.

The Article proceeds in four parts. Part I traces the recent history of federal sentencing, briefly describing the time periods of particular interest in studying post-Booker sentencing. It also summarizes the principal criticisms of the Guidelines regime. Readers familiar with Booker and the scholarly debate about the Guidelines can skim Part I. Part II sets forth the Article’s data and methods, which build on “natural experiment” studies that compared pre- and post-Guidelines sentencing. It details the case matching and selection process, describes two measures of inter-judge disparity, and proposes a subset of “discretionary sentences” that deserve particular attention from researchers. Part III reports the results, presenting descriptive and statistical comparisons of average sentence length and Guidelines sentencing patterns across time periods. Part IV suggests several implications of the Article’s findings.

I. A RECENT HISTORY OF FEDERAL SENTENCING

A search for the Booker revolution must begin by tracing the recent history of federal sentencing. Several recent stages of development in sentencing law are of particular interest: (1) the era of mandatory Sentencing Guidelines, beginning in 1987; (2) a brief window in 2003 and 2004 in which the PROTECT Act made the Guidelines

\textsuperscript{8}552 U.S. 85, 128 S. Ct. 558 (2007).
\textsuperscript{9}552 U.S. 38, 128 S. Ct. 586 (2007).
even stronger; (3) the 18-month period following Booker, in which district courts began to treat the Guidelines as “effectively advisory”; (4) the subsequent 18-month period, in which courts of appeals worked out some details of “reasonableness” review; and (5) the period following Kimbrough and Gall, which appear to authorize sentencing courts to reject the policy judgments of the Commission.

A. Prologue: Indeterminate Sentencing

Until the early 1980s, criminal sentencing in the federal system was “indeterminate.” Federal judges enjoyed almost entirely unfettered discretion in choosing the type and severity of sentence.10 Criminal statutes generally designated high maximum penalties and no minimum penalties, leaving judges free to impose a term of probation or imprisonment of any length within a broad range.11 Judges were under no obligation to give reasons for the sentence imposed,12 and appellate review of sentencing decisions was virtually nonexistent.13 Sentencing was also indeterminate in the sense that parole boards enjoyed broad discretion to reduce sentences after they were imposed.14 In the federal system, once a prisoner had served at least one-third of the sentence imposed by the judge, the parole board in its discretion could grant early release15 —again, without providing reasons and without any opportunity for meaningful review.16

The indeterminate sentencing regime reflected the dominant theory of punishment during the early twentieth century: the rehabilitation of offenders.17 Judges and prison officials were encouraged to tailor their decisions to the unique circumstances of the case and the needs of the individual offender, imposing punishment no greater than necessary to ensure that the offender was reformed.18 The federal system thus called upon judges and parole boards to make sentencing decisions “almost like a doctor or social worker exercising clinical judgment.”19

---

11 Id. at 11. The federal bank robbery statute, for example, provided that an offender “shall be fined not more than $5,000 or imprisoned not more than twenty years, or both.” Bank Robbery Act of 1934, Pub. L. No. 73-235, 48 Stat. 783 (codified as amended at 18 U.S.C. § 2113); see Jerome v. United States, 318 U.S. 101, 101-02 (1943).
13 Stith & Cabranes, supra note 10, at 9 & n.3.
14 See Act to Parole United States Prisoners, ch. 387, 36 Stat. 819 (June 25, 1910) (establishing and defining the powers of parole boards for federal penitentiaries).
15 Stith & Cabranes, supra note 10, at 19.
17 Mistretta v. United States, 488 U.S. 361, 364 (1989) (“Both indeterminate sentencing and parole were based on concepts of the offender’s possible, indeed probable, rehabilitation, a view that it was realistic to attempt to rehabilitate the inmate and thereby to minimize the risk that he would resume criminal activity upon his return to society.”); see also Kate Stith & Steve Y. Koh, The Politics of Sentencing Reform: The Legislative History of the Federal Sentencing Guidelines, 28 WAKE FOREST L. REV. 223, 227 (1993).
18 See Williams v. New York, 337 U.S. 241, 249-50 (1949) (describing the prevailing “belief that by careful study of the lives and personalities of convicted offenders many could be less severely punished and restored sooner to complete freedom and useful citizenship”).
In the space of a decade, however, rehabilitation fell from favor as the principal justifying aim of punishment in the United States. Skepticism about rehabilitation mounted as research revealed that few programs aimed at rehabilitating offenders were succeeding, and these findings led to a widespread (if oversimplified) belief that “nothing works.” By the early 1980s, Congress had concluded that “[w]e know too little about human behavior to be able to rehabilitate individuals on a routine basis or even to determine accurately whether or when a particular person has been rehabilitated.”

At the same time, the discretion enjoyed by sentencing judges and parole boards came under fire. Judge Marvin Frankel, the most influential critic of indeterminate sentencing in the 1970s, argued that “[t]he almost unchecked and sweeping powers we give to judges in the fashioning of sentences are terrifying and intolerable for a society that professes devotion to the rule of law.” The notion that sentences must be “individualized” was, in Frankel’s view, “prima facie at war with such concepts, at least as fundamental, as equality, objectivity, and consistency in the law.” These criticisms of indeterminate sentencing found support not only in anecdotal reports—some judges had developed a reputation as harsh or lenient at sentencing—but in early simulation studies that found wide disparity in the sentences chosen by different judges presented with identical case facts.

B. Mandatory Guidelines, 1987-2003

1. Inter-Judge Disparity and the Sentencing Guidelines

Following more than a decade of debate, Congress enacted the Sentencing Reform Act of 1984. Co-sponsored by Democratic Senator Ted Kennedy and Republican Senator Strom Thurmond, the principal purpose of the Act was to reduce inter-judge disparity in sentencing. Congress concluded that, too often, similarly

---


24 Id. at 9.


situated offenders received unjustifiably disparate sentences, solely because of the preferences and biases of the sentencing judge. Different constituencies in Congress emphasized different aspects of the problem. Liberals expressed concern that indeterminate sentencing allowed race discrimination to flourish, while “tough on crime” conservatives worried that too many judges were unduly lenient. But there was remarkable bipartisan agreement that unfettered discretion had resulted in an intolerable level of inter-judge sentencing disparity.

To reduce inter-judge disparity, the Act created the United States Sentencing Commission, “an independent commission in the judicial branch of the United States.” The Act directed the Commission to promulgate guidelines for use by sentencing courts in making virtually all important sentencing decisions, including whether to impose a term of imprisonment, the length of the sentence, terms of supervised release, and whether to impose consecutive or concurrent sentences. Congress specified that, in establishing guidelines for prison sentences, the Commission generally must use narrow ranges in which the maximum does not exceed the minimum by more than 25%. Judges were bound to follow the guidelines except in two circumstances: (1) on the government’s motion, based on a defendant’s substantial assistance to authorities; and (2) in “rare” cases in which the court found aggravating or mitigating circumstances “of a kind, or to a degree, not adequately taken into consideration by the Sentencing Commission.” The Act compelled judges to state the reasons for each sentence in open outcomes caused solely by the sentencing judge, and not by legitimate differences between offenses and offenders. See James M. Anderson et al., Measuring Interjudge Sentencing Disparity: Before and After the Federal Sentencing Guidelines, 42 J.L. & ECON. 271, 274 (1999) (defining “disparity” as “solely that variation caused by the identity of the decision maker”). What counts as a “legitimate” difference between cases justifying a higher or lower sentence is, of course, heavily contested and dependent on some underlying theory of punishment. See Michael M. O’Hear, The Original Intent of Uniformity in Federal Sentencing, 74 U. CIN. L. REV. 749, 749-50 (2006); Kevin Cole, The Empty Idea of Sentencing Disparity, 91 Nw. U.L. REV. 1336, 1337 (1997). But Congress in enacting the Sentencing Reform Act was convinced that the preferences and idiosyncrasies of judges frequently produced different sentencing outcomes that could not be explained by relevant characteristics of the offense and offender. See 18 U.S.C. § 3553(a)(6) (2008) (directing judges to impose sentences so as to avoid “unwarranted sentence disparities among defendants with similar records who have been found guilty of similar conduct”); S. Rep. No. 98-225, supra note 20, at 45 (“Sentencing disparities that are not justified by differences among offenses or offenders are unfair both to offenders and to the public.”).

32 Id. § 994(b)(2). Thus, a guideline range of 100-125 months would comply with the statute (25/100 = 25%), but a range of 75-100 months would not (25/75 = 33.3%). One consequence of this requirement was that the Commission could grant district courts greater discretion, in absolute terms, as the guideline maximum increases.
33 U.S.S.G. § 5K1.1; see 28 U.S.C. § 994(m).
34 U.S.S.G. § 5K2.0.
35 18 U.S.C. § 3553(b)(1). This was one of the provisions excised by the remedial opinion in Booker. See infra Part I.B.
court, and to issue a written statement of reasons in any case where the sentence fell outside the guideline range.\textsuperscript{36} It also provided for appellate review of sentencing range calculations and for review of sentences outside the guideline range for abuse of discretion.\textsuperscript{37}

The Commission promulgated the first federal Sentencing Guidelines in 1987, and soon after began to evaluate their performance. Early efforts to measure changes in inter-judge disparity using “matched case” techniques proved unsatisfactory. A 1991 study by the Commission, for example, compared a subset of cases from 1985 and 1989-1990 that shared particular facts that the Commission had incorporated into the Guidelines calculations.\textsuperscript{38} The study found a reduction in disparity, but the non-random method of selecting cases made it difficult to draw any meaningful conclusions from the results. It showed, for example, that the post-Guidelines sentences were significantly closer to the guideline range, but it could not demonstrate that the greater disparity in pre-Guidelines sentences was attributable to the judge, rather than to legitimate sentencing considerations not captured by the Guidelines or used as selection criteria.\textsuperscript{39} Difficulties in obtaining comparable data for the pre- and post-Guidelines periods frustrated many similar matched-case studies and statistical analyses.\textsuperscript{40}

In the late 1990s, however, several studies provided strong evidence that the Guidelines had reduced inter-judge sentencing disparities, at least to a modest degree. These studies used a “natural experiment” technique that focused on districts in which judges received case assignments from a common case pool using random assignment system. Each study measured inter-judge sentence disparity in two time periods, before and after the Guidelines went into effect. On the assumption that the distribution of cases was random in each period, they attributed any reduction in disparity in average sentences to the judge.\textsuperscript{41}

The two most prominent large-scale studies each found a measurable reduction in inter-judge sentencing disparity. The first, authored by James Anderson, Jeffrey Kling, and Kate Stith (“the Anderson-Kling-Stith study”), examined a sample of cases from approximately 25 cities nationwide in which the case distribution system was deemed sufficiently random.\textsuperscript{42} The study concluded that “Congress successfully achieved [its] goal” of “reducing interjudge nominal sentencing disparity,” finding that in 1986-1987 the estimated expected difference in the average length of sentence imposed by any two judges was 16\% to 18\%, and that under the Guidelines in 1988-1993 that figure had fallen to 8\% to 13\%.\textsuperscript{43}

\begin{flushright}
\textsuperscript{36} 18 U.S.C. § 3553(c)(1)-(2).
\textsuperscript{37} Id. § 3742(a)-(b). It was not until 1996 that the Supreme Court clarified the standard of appellate review as “abuse of discretion.” Koon v. United States, 518 U.S. 81 (1996). The Act’s appellate review provision was excised in Booker. See infra Part I.B.
\textsuperscript{39} Anderson et al., supra note 27, at 280.
\textsuperscript{40} Paul J. Hofer et al., The Effect of the Federal Sentencing Guidelines on Inter-Judge Sentencing Disparity, 90 J. CRIM. L. & CRIMINOLOGY 239, 244, 268-69, 274-76 (1999).
\textsuperscript{41} Anderson et al., supra note 27, at 291; Hofer et al., supra note 40, at 282.
\textsuperscript{42} Id. at 290 (Table 2).
\textsuperscript{43} Id. at 303.
\end{flushright}
The second, by Paul Hofer, Kevin Blackwell, and Barry Ruback (“the Hofer study”), compared a sample of cases from cities with a random case distribution system in two time periods, 1984-1985 and 1994-1995. Based on sentences by judges who remained on the bench during both periods, drawn from nine cities, the study found that the identity of the sentencing judge accounted for 2.32% of variation in sentences in the first period and 1.24% in the second, a reduction “almost by half under the guidelines.”

Using a larger sample from 41 cities in which the composition of the court had changed between periods, the study found larger reductions for most offense types—for drug offenses from 7.47% to 4.55%, and for firearm offenses from 18.08% to 14.00%—but increases in inter-judge disparity for immigration and robbery offenses. The authors concluded that, despite the fairly small percentage of variance attributable to judges in either period, the Guidelines had achieved “modest success” in reducing inter-judge disparity.

These studies, and other similar efforts by Joel Waldfogel and Abigail Payne, offer the best available evidence of the effect of the Guidelines on inter-judge sentencing disparity. Yet the authors of the studies readily acknowledge several limitations. One is that the studies do not measure the extent to which other sources of disparity, such as greater prosecutorial discretion, may have increased as a result of the Guidelines. Another is that they could not disentangle the effects of the Guidelines from the effects of other simultaneous changes in sentencing, such as the enactment of mandatory minimum sentences for drug offenses. But perhaps the greatest limitation, discussed at greater length below, is that they measure only disparity in average sentence length. That approach measures a judge’s “across-the-board” leniency or severity, but does not capture other important forms of variation between judges, like variation that depends on particular offense or offender characteristics.

2. Criticisms of the Guidelines

Whatever their accomplishments, the Guidelines have encountered strident opposition, particularly among scholars and district court judges. A complete catalogue of criticisms leveled at the federal sentencing guidelines in the last 20 years would be prohibitively lengthy. Brief discussion is warranted, however, for three of the most

---

44 Hofer et al., supra note 40, at 284.
45 Id. at 287. The percentages reported are derived from r-squared, a regression statistic that measures the fraction of variation in a dependent variable that is explained by the independent variable(s).
46 Id. at 203-04.
47 Id. at 298.
49 See Hofer et al., supra note 40, at 299-302; Anderson et al., supra note 27, at 302.
50 See Anderson et al., supra note 27, at 299.
51 See infra notes 237-239 and accompanying text.
52 For Proust.
prevalent criticisms: that the Guidelines are too harsh, that they are too inflexible, and that they have improperly shifted sentencing authority to prosecutors.

a. Harshness

By far the most frequent criticism is that the Guidelines are too harsh, and the argument comes in two varieties. One is that the Guidelines are too harsh in relative terms. Opponents note that average sentences in the federal system are now much longer compared with the pre-Guidelines era, especially for drug offenses. During the first seven years of the Guidelines, average prison sentences increased by 22 months (approximately 79%) for all offenses, and by 30 months (approximately 64%) for drug offenses. Meanwhile, from 1985 to 2006, the federal prison population swelled from approximately 40,000 inmates to more than 190,000, an increase of 375%. Sentences under the federal Guidelines also significantly exceed those under state guidelines and in other Western nations. The number and variety of such comparisons is impressive, even if their persuasive force ultimately depends on the premise that levels of punishment were correct in the early 1980s, or are now correct in state systems or in Europe.

Another contention is that the Guidelines are too harsh in absolute terms. A litany of judges, appointed by Presidents from both parties, has denounced the Guidelines for their severity, and the view that the federal Guidelines are too harsh has become something close to a consensus among scholars and commentators. Nonetheless, Congress and public opinion have generally supported the Guidelines’ stiff punishments. A national opinion survey by Paul Rossi and Richard Berk suggests that, with a few important exceptions, the Guidelines generally track societal norms and expectations about the level of punishment deserved for federal offenses.

---

53 STITH & CABRANES, supra note 10, at 63. The increase was still more dramatic considering that many prisoners served less than their full sentence under the previous indeterminate sentencing regime. The increase in estimated time served increased by 30 months (or 230%) for all offenses and 43 months (or 195%) for drug offenses. Id.
59 Id. at 1020-21; see PETER H. ROSSI & RICHARD A. BERK, JUST PUNISHMENTS: FEDERAL GUIDELINES AND PUBLIC VIEWS COMPARED 207-08 (1997); Ilene Nagel, Writing the Federal Guidelines, in ROSSI & BERK,
The sentiment that the Guidelines are too harsh exacerbates other objections to the Guidelines, described below. When judges perceive sentencing ranges as too severe, they feel pressure to find reasons to depart, fueling criticisms that the Guidelines are inflexible. Likewise, to the extent that the Guidelines have shifted power to prosecutors, high sentencing ranges enhance that power by giving the government greater leverage in plea negotiations. Severity thus animates or reinforces almost all other criticisms of the Guidelines regime.60

Claims that levels of punishment are too harsh depend on some underlying normative framework for determining “just” or “optimal” sentences,61 and the debate between competing justifications of punishment is outside the scope of this Article. But regardless of its normative basis, the harshness critique begins from the premise that, if the Guidelines were eliminated or weakened, judges would do a better job of selecting just or optimal sentences. Freed from the strictures of mandatory sentencing ranges, the argument goes, judges would do a better job than the Guidelines of recognizing the circumstances in which more lenient sentence would be just or optimal.

b. Inflexibility

For many judges, the most frustrating aspect of the Guidelines is their rigidity. The Guidelines, by design, specify a narrow sentencing range and tightly control the permissible grounds for departure. Although styled as “guidelines,” the Guidelines thus have operated more like mandatory rules.62

In fairness, it is possible to imagine sentencing regimes more “mandatory” and rule-like than the Guidelines.63 The Guidelines prescribe sentencing ranges, not precise sentences, and they allow for departure in circumstances not taken into account by the Commission. And some of the federal system’s most glaring injustices stem not from the Guidelines, but rather from statutory mandatory minimums, which are even more inflexible.64 But devising a more restrictive system than the federal Guidelines requires imagination, because none has ever been attempted in the United States.65

The inflexibility criticism implicates the familiar distinction between legal rules and standards.66 Sentencing rules announce clear and predictable outcomes based on the presence or absence of specific factors, and in theory do not require any discretionary choices by the judge. Yet because they imperfectly implement their underlying purposes,

---

61 Cassell, supra note 58, at 1020.
64 See Cassell, supra note 58, at 1045 (“[I]t is striking how many [Guidelines] ‘horror stories’ stem from mandatory minimums in general and from narcotics mandatory minimums in particular.”).
65 See Reitz, supra note 63, at 171 (concluding that the post-Booker Guidelines “remain as restrictive of judicial sentencing discretion as any system in the United States”).
rules are frequently both overinclusive and underinclusive, particularly in unusual circumstances not anticipated by the rulemaker.\textsuperscript{67} A host of judges, of all political stripes, have protested that strict application of the mandatory Sentencing Guidelines regime produces unjust results in too many cases.\textsuperscript{68} More flexible sentencing standards, they have argued, would call upon judges to apply the underlying purposes of punishment directly, achieving just results by considering and weighing all relevant factors and making a discretionary judgment.\textsuperscript{69}

The inflexibility critique depends on the assumption that, if the Guidelines were flexible rather than mandatory, judges could recognize when application of the Guidelines is inconsistent with the underlying purposes of sentencing and make case-specific corrections to do justice for individual defendants.\textsuperscript{70} In that sense, it differs from the harshness critique, and from other arguments grounded in the content of the Guidelines’ rules. If the Guidelines are too harsh, then the Commission may have simply chosen the wrong rules. For example, according to some underlying theory of punishment, the Commission may have chosen the wrong rules by generally prohibiting the consideration of age, drug addiction, employment history, childhood poverty, and other offender characteristics as a basis for downward departure.\textsuperscript{71} The inflexibility critique differs from those claims. It generally accepts the rules and objects only to particular circumstances in which the rules produce unjust results, according to some agreed-upon purposes of sentencing.

c.  Increased Prosecutorial Power

Another major theme in criticisms of the Guidelines is the transfer of power from judges to prosecutors. By setting detailed sentencing rules that were binding on judges, the Guidelines allowed prosecutors, through the exercise of their charging power, to exert considerable control over the guideline sentencing range.\textsuperscript{72}

The Commission sought to counteract a transfer of power to prosecutors by requiring that sentences take into account “real offense” factors not alleged in the indictment or proven to the jury.\textsuperscript{73} It further directed probation officers to ascertain the


\textsuperscript{70} See Douglas A. Berman, \textit{A Common Law for This Age of Federal Sentencing: The Opportunity and Need for Judicial Lawmaking}, 11 STAN. L. & PUB. POL’Y REV. 93, 93 n.6 (1999) (collecting authorities criticizing the Guidelines for “inappropriately restricting the discretion and flexibility that sentencing judges need to achieve justice in individual cases”).

\textsuperscript{71} Bowman, \textit{supra} note 60, at 168; see U.S.S.G. Ch. 4, Part H.

\textsuperscript{72} Kate Stith, \textit{The Arc of the Pendulum: Judges, Prosecutors, and the Exercise of Discretion}, 117 YALE L.J. 1420, 1430 (2008).

\textsuperscript{73} See U.S.S.G. §§ 1B1.1 to 1.9.
real facts of the case, beyond what the prosecutor and defense attorney had presented at trial or in the plea agreement. The Department of Justice instructed line prosecutors to press “the most serious, readily provable offense or offenses consistent with the defendant’s conduct” and to stipulate to facts only if they accurately reflected the defendant’s conduct.

None of these efforts, however, prevented a shift in sentencing discretion to the government. Under the Guidelines, prosecutors have considerable discretion to enter an agreement (a “charge bargain”) in which the government presses only lesser charges with a lower guideline sentencing range. And, if the prosecutor agrees, the government and defense counsel can strike a different sort of agreement (a “fact bargain”) whereby the guilty plea stipulates to an abbreviated set of facts, withholding “real offense” facts to avoid triggering particular adjustments under the Guidelines. In the face of charge and fact bargaining, probation officers have little independent evidence or incentive to dig deeper, particularly if it would disrupt a consummated plea agreement. Judges might never learn of the real facts or might collude with the parties to ignore them to avoid imposing a harsh Guidelines sentence. Centralized oversight of line prosecutors proved difficult given the sheer number of charging and plea decisions and the limited resources of Main Justice.

Because “real offense” facts are not systematically and reliably recorded, the frequency of charge and fact bargaining is exceedingly difficult to measure. One study, approved by the Department of Justice, examined sentencing practices immediately following the promulgation of the Guidelines in the late 1980s and early 1990s. It found evidence of guideline “manipulation” in 20% to 35% of cases. The authors also found evidence of stark differences in bargaining practices between U.S. Attorneys’ offices in three districts. Surveys of probation officers indicate widespread belief that fact bargaining is prevalent and plea agreements are incomplete or inaccurate. Although the available evidence is now dated, it indicates that the ability of prosecutors to selectively

---

74 Stith & Cabranes, supra note 10, at 85-91.
76 Stith, supra note 72, at 1448.
79 Anderson et al., supra note 27, at 280.
80 Schulhofer & Nagel, supra note 77, at 1290.
82 See David Yellen, Probation Officers Look at Plea Bargaining, and Do Not Like What They See, 8 FED. SENT. REPORTER 339, 339 (1996) (describing a survey in which 40% of probation officers reported that the guideline calculations in plea agreements do not “accurately and completely reflect all aspects of the case” in a majority of cases); Letter from Francesca D. Bowman, Chair, Probation Officers Advisory Group, to Richard P. Conaboy, Chairman, U.S. Sentencing Comm’n (Jan. 30, 1996), reprinted in 8 FED. SENT. REPORTER 1, 303-04 (1996).
83 See Marc L. Miller, Domination & Dissatisfaction: Prosecutors as Sentencers, 56 STAN. L. REV. 1211, 1230 (2004) (noting that “[t]his research in the late 1980s and early 1990s has not been repeated, and,
press charges and disclose facts, coupled with the marked increase in average sentence length under the Guidelines, has given prosecutors considerably more leverage in plea negotiations and influence over sentencing outcomes.

C. PROTECT Act: 2003-2004

Despite fifteen years of vigorous criticism, Congress voted in 2003 to make the Guidelines even tougher and less flexible. Effective May 1, 2003, Congress enacted a package of sentencing provisions as part of the Prosecutorial Remedies and Other Tools to End the Exploitation of Children Today Act (“PROTECT Act”).84 Championed by Representative Tom Feeney and dubbed the “Feeney Amendment,” the provisions responded to concerns in Congress and the Department of Justice about the prevalence of downward departures from the Guidelines.85 At the time, reports by the Commission showed strong growth in the rate of downward departures between 1991 and 2001, from 5.8% of all sentences to 18.1%.86 (The Commission later realized that the 2001 rate was incorrect.87)

Among other changes, the PROTECT Act (1) tightened the standard of appellate review for non-guideline sentences, replacing the “abuse of discretion” standard with de novo review;88 (2) directed the Commission to amend the Guidelines “to ensure that the incidence of downward departures are [sic] substantially reduced”;89 (3) prohibited the Commission from recognizing new permissible grounds for downward departure for two years;90 and (4) directed the Department of Justice to resist downward departures “not supported by the facts and the law.”91 The PROTECT Act also directly amended the Guidelines by adding specific upward adjustments for sex offenders and child pornography cases.92 Attorney General John Ashcroft implemented the PROTECT Act’s requirements by issuing a memorandum prohibiting fact bargaining and instructing prosecutors to “charge and pursue the most serious, readily provable offense or offenses that are supported by the facts of the case.”93

The PROTECT Act sentencing provisions drew strong criticism from scholars, judges, interest groups, and the defense bar.94 Responding to an earlier version of the Act that would have eliminated all Guidelines grounds for downward departure, Chief Justice

without the blessing of the Department of Justice, it would be enormously difficult or impossible to reconstruct”).

84 Pub L. No. 108-21, § 401(h), 117 Stat 650, 672.
86 Stith, supra note 85, at 1465 (describing the numbers before Congress in 2003 as “powerful,” showing “persistent increases in the rate of noncooperation downward departures during the 1990s—especially after the Koon decision was handed down in 1996”); see Miller, supra note 83, at 1228.
87 See infra notes 97-98 and accompanying text.
88 PROTECT Act § 401(d), 117 Stat. at 670.
89 Id. § 401(m)(2)(A).
90 Id. § 401(j)(2).
91 Id. § 401(l)(1).
92 Id. § 401(i).
94 See Noelle Tsigounis Valentine, Note, An Exploration of the Feeney Amendment: The Legislation that Prompted the Supreme Court to Undo Twenty Years of Sentencing Reform, 55 SYRACUSE L. REV. 619, 628-29 (2005).
William Rehnquist warned Congress that the bill “would do serious harm to the basic structure of the sentencing guideline system and would seriously impair the ability of courts to impose just and reasonable sentences.”\footnote{Alan Vinegrad, The New Federal Sentencing Law, 15 FED. SENT. REPORTER 310, 313 (2003).} The Judicial Conference of the United States took exception to the allegation that judges were driving up the rate of downward departures, noting that most of the increase was concentrated in southwestern border districts where the justice system faced “crisis” conditions.\footnote{See Letter of Leonidas Ralph Meacham to Senator Orrin Hatch, Apr. 3, 2003, at 3, available at http://www.uscourts.gov/judiciary2003/feeneyamendment.pdf.}

In hindsight, it is clear that reports of an epidemic of judge-initiated downward departures were exaggerated.\footnote{Max Schanzenbach, Have Federal Judges Changed Their Sentencing Practices? The Shaky Empirical Foundations of the Feeney Amendment, 2 J. EMPIRICAL L. STUDIES 1, 28-29 (2005); Stith, supra note 85, at 1464-65.} In response to the PROTECT Act, the Commission revealed that approximately 40% of the sentences it had reported as judge-initiated downward departures in fiscal year 2001 were in fact government sponsored, typically due to a plea agreement or “fast track” program.\footnote{U.S. SENTENCING COMM’N, DOWNWARD DEPARTURES FROM THE SENTENCING GUIDELINES iv-v, 59-60 (Oct. 2003).}

Nonetheless, the PROTECT Act greatly curtailed judges’ discretion to depart from the Guidelines. It resulted in changes to the Guidelines themselves that narrowed the permissible circumstances for departure. And because it toughened the standard of review, judges concerned about reversal on appeal had strong incentives to impose within-range sentences.


In 2005, the Supreme Court held in United States v. Booker\footnote{543 U.S. 220 (2005).} that the Sentencing Reform Act violated the Sixth Amendment right to trial by jury.\footnote{Id. at 226-27, 243-44.} That conclusion, and the Court’s surprising decision to remedy the constitutional defect by rendering the Guidelines “advisory,”\footnote{Id. at 227, 245.} appeared to revolutionize federal sentencing. Based on the early years of post-Booker sentencing, however, many observers concluded that the Booker revolution never materialized.

1. The Booker Revolution

On January 12, 2005, the Supreme Court issued a fractured decision in United States v. Booker\footnote{The case prompted six separate opinions, including two principal majorities and two principal dissents.} consisting of two majority opinions.\footnote{Id. at 225.} One opinion, written by Justice Stevens, extended the rule of Apprendi v. New Jersey\footnote{530 U.S. 466 (2000).} and Blakely v. Washington\footnote{542 U.S. 296 (2004).} to the federal Sentencing Guidelines. It reiterated that, under the Sixth Amendment, “[a]ny fact (other than a prior conviction) which is necessary to support a sentence exceeding the maximum authorized by the facts established by a plea of guilty or a jury verdict must be
admitted by the defendant or proved to a jury beyond a reasonable doubt.” The Court noted that the Guidelines set a mandatory sentencing range based on the facts established by the plea or verdict, and that departures were unavailable in most cases. Because a judge could impose a sentence above the guideline range only upon finding some fact—an aggravating circumstance “of a kind, or to a degree, not adequately taken into consideration by the Sentencing Commission”—the Guidelines regime violated the Sixth Amendment. In light of the Court’s previous Sixth Amendment cases, this result came as no surprise.

Another opinion, written by Justice Breyer, held that the proper remedy for the Sixth Amendment violation was to sever two provisions of the Sentencing Reform Act that made the Guidelines mandatory. Excising those provisions, the Court explained, “makes the Guidelines effectively advisory.” Henceforth judges must continue to calculate the applicable sentencing range, but need only “consider” it, along with the factors identified in Section 3553(a), in imposing a sentence. Those changes cured the constitutional defect because all nine Justices agreed that, within statutory limits, fully discretionary sentencing based on judge-found facts poses no Sixth Amendment problem. Unlike the constitutional holding, this remedial holding came as a shock. None of the parties or amici had proposed it.

Opponents of the Guidelines greeted Booker with enthusiasm. Judges who chafed at the mandatory Guidelines regime saw the decision as “a kind of Emancipation Proclamation.” Judge Jack Weinstein described himself as “elated” and predicted that

---

105 Booker, 543 U.S. at 244 (Stevens, J., writing for the Court).
106 Id. at 234.
109 Booker, 543 U.S. at 245 (Breyer, J., writing for the Court).
110 Id.
111 Id. at 245, 259-60.
112 Id. at 245.
113 Id. at 233 (Stevens, J., writing for the Court) (“If the Guidelines as currently written could be read as merely advisory provisions that recommended, rather than required, the selection of particular sentences in response to differing sets of facts, their use would not implicate the Sixth Amendment.”).
116 Bowman, supra note 5, at 280; see Hon. Nancy Gertner, Sentencing Reform: When Everyone Behaves Badly, 57 ME. L. REV. 569, 579 (2005) (describing some judges who cheered Booker as a “free at last” moment). Two federal judges in the District of Massachusetts publicly praised the decision shortly after it
“most judges will be, too.”

Scholars who had long criticized the Guidelines recognized the decision a momentous opportunity to fix federal sentencing. The defense bar was “ecstatic.”

Supporters of the Guidelines, on the other hand, despaired. The Justice Department warned that because of Booker, “the risk increases that sentences across the country will become wildly inconsistent.” Senator Charles Grassley called the decision “disappointing” and worried that “without required sentencing guidelines, the criminal courts could revert to pre-guideline chaos.” Representative Feeney, who had sponsored the PROTECT Act amendments, decried the decision as “an egregious overreach into Congress’ constitutional power.”

In both groups, however, the conventional wisdom saw Booker as revolutionary. For better or worse, many observers believed that the decision marked “the end of federal criminal sentencing as we know it.”

2. The Booker Dud

In the years since Booker, however, the conventional wisdom has shifted. Aggregated data released by the Sentencing Commission suggest that little has changed as a result of the decision.

One measure by which Booker has disappointed is average sentence length. In its final report on the effects of Booker, the Commission concluded that “[t]he severity of sentences imposed has not changed substantially.” In fact, according to Commission
statistics reproduced in Figure 1, average sentence length has slightly increased since Booker, both overall and for drug trafficking offenses:

![Figure 1: Average Sentences Nationwide, by Fiscal Year](image)

Average sentences nationwide rose to 59.3 months by fiscal year 2007, compared with 56 months in early fiscal year 2003 and 57 months in 2003-2004 under the PROTECT Act. Sentences for drug trafficking rose from 83 months under the PROTECT Act to 85.6 months in fiscal year 2007. The percentage of defendants who receive a prison sentence, as opposed to a sentence of probation or a fine, also has inched upward since Booker, from 85.9% to 88.6%.

Nor have aggregate sentencing patterns relative to the Guidelines changed dramatically since Booker. Figure 2 shows the rate of above-range, below-range, and

---

126 FINAL REPORT, supra note 125, at 71 (Table 3); U.S. SENTENCING COMMISSION, SOURCEBOOK OF FEDERAL SENTENCING STATISTICS 2003-2008, Appx. B, National Data (2003-2008). The FY 2003 period ends on the effective date of the PROTECT Act, and the PROTECT Act period ends on June 24, 2004, when Blakely was decided. FINAL REPORT, supra note 125, at 71 (Table 3). Because the Commission’s annual sourcebooks report immigration offenses in a slightly different manner than the Final Report, the averages for immigration offenses in the first two periods reflect FY 2003 and FY 2004 pre-Blakely figures, respectively.

127 FINAL REPORT, supra note 125, at 71 (Table 3) (reporting figures for the first calendar year after Booker).
within-range and government-sponsored sentencing among all judges nationwide from fiscal years 2003-2008.

![Figure 2: Guideline Sentencing, by Time Period](image)

Although the rate of below-range sentencing after *Booker* more than doubled (from 5.5% to 12.0%) compared with the PROTECT Act period, that figure is not much higher than the rate (of 8.6%) under the mandatory Guidelines in 2002-2003. The percentage of above-range sentences also has nearly doubled (from 0.8% to 1.5%), but remains quite low. Within-range and government-sponsored sentences continue to account for more than 85% of sentences in the federal system. As Max Schanzenbach and Emerson Tiller have observed, the changes in Guidelines sentencing patterns since *Booker* are noticeable, but hardly “earth-shattering.” By way of comparison, the rate of below-range sentencing remains well below the (incorrectly reported) 18.1% rate that prompted Congress to intervene in the PROTECT Act.

---


130 See supra note 86 and accompanying text.
Based on the aggregate data in the Commission’s reports, the conventional wisdom about the effects of Booker has changed. Far from ushering in a revolution, the decision now looks like a dud. Frank Bowman calls the effects of Booker “strikingly modest.” Douglas Berman concludes that “the Booker decision does not appear to have radically transformed either basic practices or typical outcomes in the federal sentencing system” and finds the state of post-Booker sentencing “[d]isappointingly . . . Guidelines-centric.” Sandra Guerra Thompson calls the cumulative result of Booker “much ado about not very much” and “a major disappointment” for those who had hoped the decision would bring about improvements. Judge Nancy Gertner of the District of Massachusetts, channeling Yogi Berra, calls post-Booker sentencing “déjà vu all over again.” Judge Graham Mullen likens the Guidelines to a “vampire buried without a stake . . . resurrected to stalk our jurisprudence once more.”

Several explanations have been proposed for the missing Booker revolution. One is inertia. Three-quarters of district court judges in active status, and more than half of all sitting district court judges, were appointed between the effective date of the Guidelines in 1987 and the Booker decision in 2005. It would not be surprising for these judges, who have spent their entire careers treating the Guidelines as mandatory, to continue to follow them in the great majority of cases even though they are now advisory. Some

131 Schanzenbach & Tiller, supra note 129, at 739 (noting that “most observers” believe “the fundamentals of sentencing changed little post-Booker”); see also Zlotnick, supra note 68, at 15; Timothy J. Droske, Correcting Native American Sentencing Disparity Post-Booker, 91 MARQ. L. REV. 723, 774 (2008); Kevin P. McCormick, Untangling the Capricious Effects of Market Loss in Securities Fraud Sentencing, 82 TUL. L. REV. 1145, 1161-62 (2008); D. Michael Fisher, 46 DUQ. L. REV. 65, 77-78 (2007) (“While the change is noticeable, it does not reflect the fear of some post-Booker commentators that judges, now invested with a new kind of discretion, would ignore the Guidelines and sentence defendants however they saw fit.”); Michael M. O’Hear, The Duty to Avoid Disparity: Implementing 18 U.S.C. § 3553(a)(6) After Booker, 37 McGeorge L. REV. 627, 645 (2006); Jeffrey S. Hurd, Federal Sentencing and the Uncertain Future of Reasonableness Review, 84 DEN. U. L. REV. 835, 860 (2007); Frank O. Bowman, III, ‘Tis a Gift to be Simple: A Model Reform of the Federal Sentencing Guidelines, 18 FED. SENT. REPORTER 301 (2006). Schanzenbach and Tiller express misgivings about this conventional wisdom, leaving open “the possibility that sentences in fact did change post-Booker, but such changes are masked by the interplay between district and circuit courts.” Schanzenbach & Tiller, supra note 129, at 740. In Part III, below, I suggest another factor that may have masked the effects of Booker: the unavailability of judge-specific data.

132 Bowman, supra note 5, at 319.

133 Berman, supra note 129, at 349-51.


138 See Federal Judicial Center, Federal Judges Biographical Database, at www.fjc.gov/history/home.nsf (last visited Oct. 11, 2008). There are 1,016 sitting federal district court judges, including 651 judges in active status. Of them, 593 judges (58%), including 506 in active status (78%), were appointed between the effective date of the first Sentencing Guidelines on November 1, 1987 and the Booker decision on January 12, 2005.

139 Stith, supra note 72, at 1496-97 (concluding that “the gravitational pull of the Guidelines on the pendulum of sentencing practice remains strong” based, in part, on the “reluctan[ce]” of “incumbent sentencing decision makers” who were obliged to follow the Guidelines for two decades).
commentators have also suggested, less charitably, that judges find it easier to impose within-range sentences because it requires “less time in thought and less stress.”

Another explanation is fear of reversal. The Guidelines are now advisory, but sentences remain subject to appellate review for “reasonableness.” In *Rita v. United States*, the Supreme Court held that courts of appeals may presume that a within-Guidelines sentence is reasonable. A judge anxious to avoid having a sentence vacated on appeal therefore has an incentive to stay within the Guidelines.

A third explanation is “anchoring,” the well-documented cognitive error in which decisionmakers begin with an initial value, even one that is irrational, and fail to make sufficient adjustments. One study has shown, in an experimental setting, that starting values provided to a person choosing a sentence may influence the final result, even if the test subject knows that the initial value is arbitrary. Presumably sentencing guidelines, which judges know to be non-arbitrary, will have an even stronger influence. Because the Court has emphasized that the Guidelines continue to serve as “the starting point and the initial benchmark” for every federal sentence, it should not be surprising that the Guidelines continue to exert a powerful influence despite being advisory.

A fourth explanation is tactical: judges have taken a “go slow” approach to safeguard the sentencing discretion they gained in *Booker*. On this theory, judges secretly desire to flout the Guidelines more often, but have restrained themselves to avoid provoking Congress.

### E. Kimbrough and Gall: 2007-2008

On the same day in December 2007, the Supreme Court issued two decisions that reinforced district courts’ sentencing discretion under *Booker*. In *Kimbrough v. United States*, the Court held that judges, in applying the now-advisory Guidelines, are free to reject the 100-to-1 ratio that treats one gram of crack cocaine as equivalent to 100 grams

---

141 United States v. Booker, 543 U.S. 220, 245 (Breyer, J., writing for the Court).
143 127 S. Ct. at 2462-65.
150 Weinstein, *supra* note 140, at 211. On this theory, the change in party control of Congress in 2006 and the White House in 2008 could embolden district court judges to depart more frequently.
of powder cocaine.\textsuperscript{152} In reaching that conclusion, the Court emphasized—and seemed to endorse—the government’s concession that “as a general matter, courts may vary from Guidelines ranges based solely on policy considerations, including disagreements with the Guidelines.”\textsuperscript{153} Although it suggested that “closer review may be in order” in those circumstances,\textsuperscript{154} the Court implied that after Booker judges enjoy the freedom to categorically reject the Commission’s judgments about sentencing policy.\textsuperscript{155}

In \textit{Gall v. United States},\textsuperscript{156} the Court held that appellate courts conducting reasonableness review may not insist upon “extraordinary” circumstances to justify a sentence outside the guideline range, and rejected the use of a “rigid mathematical formula” to determine the strength of the justifications required for the particular sentence.\textsuperscript{157} Describing appellate review after Booker, the Court stressed that courts of appeals must apply “an abuse-of-discretion standard,” according “due deference to the district court’s decision that the § 3553(a) factors, on a whole, justify the extent of the variance.”\textsuperscript{158}

By rejecting appellate rules designed to constrain sentencing courts, \textit{Kimbrough} and \textit{Gall} marked an important expansion of district courts’ discretion under Booker.\textsuperscript{159} Recognizing the potential implications, some scholars have described the decisions as aftershocks to the \textit{Booker} earthquake.\textsuperscript{160} Likewise, the defense bar welcomed the decisions as “a potential sea change in federal sentencing” and a serious threat to the “‘guidelines-light’ approach” that has prevailed since \textit{Booker}.\textsuperscript{161}

Consistent with those predictions, recent reports by the Commission suggest that sentencing outcomes—seemingly frozen in place despite the light and heat of \textit{Booker}—have begun to thaw since \textit{Kimbrough} and \textit{Gall}. In FY 2008, for the first time in over a decade, average sentence length declined slightly, from 59.3 months to 56.9 months.\textsuperscript{162} In addition, the rate of below-guideline sentencing inched upward for the first time since \textit{Booker}, from 12.0\% in FY 2007,\textsuperscript{163} to 13.8\% in FY 2008 after \textit{Kimbrough} and \textit{Gall}.\textsuperscript{164} Although the changes are small, and have not yet drawn a significant reaction from the

\begin{itemize}
\item\textsuperscript{152} 128 S. Ct. at 575.
\item\textsuperscript{153} \textit{Id.} at 570 (internal quotation marks and alteration omitted).
\item\textsuperscript{154} \textit{Id.} at 575.
\item\textsuperscript{155} \textit{See} United States v. Herrera-Zuniga, 571 F.3d 568, 584-85 (6th Cir. 2009) (interpreting \textit{Kimbrough} as recognizing “the broad authority of sentencing judges” to “‘categorically’ reject the sentencing range prescribed by the Guidelines”).
\item\textsuperscript{156} 552 U.S. 38, 128 S. Ct. 586 (2007).
\item\textsuperscript{157} 128 S. Ct. at 595.
\item\textsuperscript{158} \textit{Id.} at 597.
\item\textsuperscript{160} \textit{See}, e.g., Stephanos Bibas & Susan Klein, \textit{The Sixth Amendment and Criminal Sentencing}, 30 CARDOZO L. REV. 775, 775 (2008).
\item\textsuperscript{161} Christopher R. Clifton, \textit{Champion}, April 2009, at 53, 53.
\item\textsuperscript{162} \textit{See supra} Figure 1 \& note 126.
\item\textsuperscript{163} \textit{See supra} Figure 2 \& note 128.
\item\textsuperscript{164} U.S. SENTENCING COMMISSION, FINAL POST-KIMBROUGH/GALL DATA REPORT, Table 1 (2008).
\end{itemize}
Commission or commentators, preliminary data from FY 2009 suggest that both trends are accelerating.\footnote{U.S. SENTENCING COMMISSION, Preliminary Quarterly Data Report, Second Quarter FY 2009, Tables 1, 19 Figure C (2009) (reporting that the rate of below-range sentencing has climbed to 15.3\% and that average sentence length has continued to decline).}

** * * * *

Despite extensive reports by the Commission, the existing research on post-
Booker sentencing contains no discussion of how individual judges have responded to the decision. In keeping with its longstanding practice, the Commission has disclosed only aggregate data on post-Booker sentencing trends, based on sentences imposed by all judges nationwide, or all judges within circuits or districts. Neither the Commission nor any independent researcher has examined whether individual judges have responded to the Booker revolution in similar or different ways.

That is a critical omission, and it has not gone unnoticed. Because the central goal of the Sentencing Reform Act was the reduction of inter-judge sentencing disparity,\footnote{See supra notes 27-29 and accompanying text.} judge-specific data are needed to determine the extent to which Booker has advanced or undermined Congress’s objectives. The Obama Justice Department has acknowledged as much, calling for an assessment of whether post-Booker sentencing practices “show an increase in unwarranted sentencing disparities” based on “differences in judicial philosophy among judges working in the same courthouse.”\footnote{Holder, supra note 6.} Existing research by the Commission does not permit such an assessment, leaving a crucial gap in our understanding of the Booker revolution.

II. DATA AND METHODS

This Article provides the first hard evidence of how individual judges have responded to Booker. It overcomes the primary challenge in studying federal sentencing patterns—the lack of data that include the identity of the sentencing judge—by building a unique dataset of cases from the District of Massachusetts, the lone federal district court that makes public the Statement of Reasons for criminal sentences. Although there are obvious drawbacks to focusing on a single district court, the Massachusetts data are the most comprehensive available, and there are also advantages to focusing on sentences drawn from a common pool of cases. The data thus afford a rare opportunity to test how Booker has affected inter-judge sentencing disparity.

A. Data

1. The Challenge of Gathering Judge-Specific Data

The most frustrating obstacle to the study of federal sentencing is the unavailability of data that include the identity of the sentencing judge. Despite its statutory responsibilities for collecting and disseminating information about federal sentencing,\footnote{See 28 U.S.C. § 995(a)(12)-(16).} the Commission removes all judge-identifying information from the data it...
releases to judges, scholars, and the public. The Commission not only withholds the name of the sentencing judge, but refuses to provide a code or number that would permit an analysis of sentencing patterns among judges anonymously.

That policy has been roundly criticized by scholars. Its ostensible purpose is to prevent the release of defendant-identifying information—a laudable goal in an age of web sites like “Who’s A Rat” that make it easier for criminal enterprises to detect and retaliate against witnesses who cooperate with authorities. As discussed below, however, that risk is unavoidable, as other case information can be used to link most defendants to the Commission’s records. Concealing the identity of the judge does nothing to advance the privacy interests of defendants, and much to frustrate valuable research. I join the chorus of scholars calling for the Commission to promote transparency and facilitate the study of federal sentencing by releasing case records that include judge identifiers.

But with the Commission’s policy in place for more than 20 years, scholars have struck upon several methods of working around the problem. One is to obtain special dispensation for particular projects. Both of the large-scale studies of post-Guidelines sentencing, described above, used nationwide data that included judge identifying information. The Anderson-Kling-Stith study used a special extract of data from the Administrative Office of the United States Courts, using codes rather than judges’ names and with strict conditions on the use of the data. The Hofer study had full access to data from the Commission, presumably because two of the authors were senior

---

169 U.S. SENTENCING COMM’N, GUIDE TO PUBLICATIONS & RESOURCES 2007-2008 45 (2007), available at http://www.ussc.gov/publicat/Cat2005.pdf (“Pursuant to the policy on public access to Sentencing Commission documents and data, all case and defendant identifiers have been removed from the data.” (internal citation omitted)).

170 The Feeney Amendment authorized Congress or the Justice Department to request data that include the identity of the sentencing judge, but did not provide for public dissemination of that information. See Prosecutorial Remedies and Other Tools to End the Exploitation of Children Today (PROTECT) Act of 2003, Pub. L. No. 108-21, § 401(h), 117 Stat. 650, 672.

171 See Public Access to Sentencing Commission Documents and Data, 54 Fed. Reg. 51,279, 51,282 (Dec. 13, 1989) (“No information that will identify an individual defendant or other person identified in the sentencing information will be disclosed to persons or entities outside of the Commission without the express permission of the court for which the information was prepared.”).

172 See Who’s A Rat – About Us, at http://www.whosarat.com/aboutus.php. The site “DOES NOT CONDONE VIOLENCE AGAINST INFORMANTS OR LAW ENFORCEMENT OFFICERS,” but quickly adds that it “takes no responsibility for” and “exercises no control over” its members. Id.

173 See infra notes 181-183 and accompanying text.


175 See Steven L. Chanenson, Write On!, 115 YALE L.J. POCKET PART 146, 147 (2006); Schanzenbach & Tiller, supra note 129, at 741-42; Marc L. Miller, Sentencing Reform “Reform” through Sentencing Information Systems, in THE FUTURE OF IMPRISONMENT 121, 146-48 (Michael Tonry ed. 2004); Miller, supra note 174, at 1356 & n.19.

176 See supra notes 42-47 and accompanying text.

177 Anderson et al., supra note 27, at 287.
researchers for the Commission at the time.\textsuperscript{178} This level of access fully solves the problem, but it is extremely rare. With the exception of studies by the Commission and its staff, the Anderson-Kling-Stith study marks the only time in over 20 years that scholars have received such blanket permission to study case records that identify the sentencing judge.

Another method is to start from scratch, examining and hand-coding sentencing records to create a database. This approach generates useful data, but it is far from ideal. Because coding cases is a labor-intensive process (occupying roughly 30 full-time staff members at the Commission), studies using this technique have focused on a single city or a small group of cities.\textsuperscript{179} The hand-coding process also carries an unavoidable risk of data entry error. And, worst of all, the information available is limited. Researchers can access the “Order and Judgment” for the case, which provides basic information about the sentence imposed. But the Judicial Conference has voted to make the “Statement of Reasons” for federal sentences non-public, and therefore unavailable at the courthouse or on PACER (Public Access to Court Electronic Records), the web-based system that provides access to most federal district court filings.\textsuperscript{180} Without the Statement of Reasons, researchers have no way to compare the sentence imposed with the guideline range, or to determine why the judge chose the particular sentence. That information is critical to the analysis of sentencing under the Guidelines.

A third method, developed by Max Schanzenbach and Emerson Tiller, uses docket information available on PACER to match cases in the Commission’s database. As part of a study of the influence of judges’ party affiliation on sentencing decisions, Schanzenbach and Tiller ran nationwide searches for cases filed on twenty random dates during three judicial terms from 1999 to 2002.\textsuperscript{181} They used docket information for those cases to match records in PACER with records released by the Commission. In comparing cases, they relied principally on the date and length of the sentence, but also (when necessary) the amount of any fine, the offense type, and the Hispanic ethnicity of the defendant.\textsuperscript{182} They successfully matched about 80% of sentences returned in their searches, yielding a dataset of 2,265 cases.\textsuperscript{183} One advantage of this method is that, although the matching process is labor-intensive, the coding process is straightforward. A researcher need only add a judge identifier to records released by the Commission, avoiding the need to duplicate the extensive data entry performed by the Commission and reducing the risk of coding error.

\textsuperscript{178} Hofer et al., supra note 40, at 239 n.*.
\textsuperscript{179} See, e.g., Waldfogel, \textit{Empirically Based Sentencing Guidelines}, supra note 48, at 294 (data from 10 judges in San Francisco); Payne, supra note 48, at 337 (data limited to certain types of cases in three federal courts); Waldfogel, \textit{Aggregate Inter-Judge Disparity}, supra note 48, at 151 (data from three district courts). Waldfogel and Payne are to be commended for their diligence in gathering data in the 1980s and early 1990s. Before the federal courts provided online access to most case documents, collecting these data required an in-person visit to the courthouse. Schanzenbach & Tiller, supra note 129, at 728.
\textsuperscript{181} Schanzenbach & Tiller, supra note 129, at 729-30.
\textsuperscript{182} \textit{Id.} at 729. Each of those data points ordinarily appears in the criminal docket, with the exception of ethnicity. Schanzenbach and Tiller presumably determined ethnicity by asking whether the defendant had a Hispanic-sounding name.
\textsuperscript{183} See \textit{id.} at 730.
Although Emerson and Tiller’s method holds promise, several challenges arise when deploying it to study post-Booker sentencing. First, PACER permits searches for cases filed during specified time periods, but not cases sentenced during specified time periods. Criminal cases vary greatly in length, with some proceeding to sentencing in a few months and others dragging on for years, so filing date is of little help in predicting the date of sentencing. Second, beginning in fiscal year 2004, the Commission stopped disclosing the exact date of sentencing in the records it releases to the public; it now reports only the month and year. That change expands the pool of potential matches for every case on PACER, making the matching procedure more time-consuming and less likely to succeed. Third, sampling particular filing dates yields only a tiny fraction of all sentences in a given year or other time period. Schanzenbach and Tiller’s dataset, for example, appears to have captured only 1.25% of sentences reported in 1999-2002. That makes comparisons between individual judges difficult because of the low number of cases per judge and distortions caused by multiple-defendant cases.

2. Why Massachusetts?

This Article uses a modified version of Schanzenbach and Tiller’s method, drawing upon data from the District of Massachusetts. By special vote of the court in 2001, the District of Massachusetts makes the Statement of Reasons public, and available online, for every criminal sentence, unless the presiding judge orders it sealed. This extraordinary policy, which apparently defies a contrary policy statement by the Judicial Conference, reflects the court’s commitment to greater openness and transparency in sentencing decisions. As former Chief Judge William Young has observed, “[t]he District of Massachusetts is a shining exception to the prevailing secrecy about sentencing.”

184 Office of Policy Analysis, U.S. Sentencing Comm’n, Variable Codebook 4 (Oct. 18, 2006) (beginning with the 2004 Codebook, “[t]he day part of the date has been removed”).

185 There were 179,973 sentences nationwide during that period. See U.S. Sentencing Comm’n, 2000 Sourcebook of Federal Sentencing Statistics Appx. B (National); U.S. Sentencing Comm’n, 2001 Sourcebook of Federal Sentencing Statistics Appx. B (National); U.S. Sentencing Comm’n, 2002 Sourcebook of Federal Sentencing Statistics Appx. B (National). The counts are not entirely comparable because Schanzenbach and Tiller searched for cases filed on specific dates in 1999-2002, whereas the Commission reports cases sentenced during those years. Schanzenbach & Tiller, supra note 129, at 729. Schanzenbach and Tiller were studying the effects of a judge’s party affiliation, not the identity of the judge or changes over time, so this percentage was of no consequence in their study.

186 Most judges in Schanzenbach and Tiller’s dataset had only one sentence, see Schanzenbach & Tiller, supra note 129, at 732 & n.53, and a single sentence is obviously an insufficient basis for drawing conclusions about the sentencing patterns of an individual judge. Because cases often have multiple defendants, some judges in Schanzenbach and Tiller’s dataset had multiple observations. Id. at 732 n.53. But codefendants are often charged with similar offenses, as in cases involving large drug-trafficking conspiracies. Thus, even when the filing-date method captures multiple observations for the same judge, those observations are not necessarily representative of the judge’s sentences as a whole.


188 See Judicial Conference Policy Statement, supra note 180.

Access to the Statement of Reasons provides two kinds of valuable information. First, the documents themselves are a gold mine. Many judges attach transcripts from the sentencing hearing or write narrative descriptions of their reasons, offering a rare glimpse of how judges are sentencing—on a day-to-day basis in ordinary, unreported cases—after Booker. That information makes Massachusetts uniquely well-suited for qualitative study. Second, when matching docket reports to records released by the Commission, the Statement of Reasons provides additional data points that can resolve ambiguities between potential matches. The high match rate increases the number of cases available for analysis, making Massachusetts uniquely well-suited for quantitative study.

The judges of the District of Massachusetts are also well-respected on sentencing issues. Judges Nancy Gertner, William Young, and Patti Saris have all written scholarly articles on sentencing. It is no coincidence that Massachusetts alone has decided to make its Statements of Reasons public. Many judges of the court have a special interest in sentencing and hope to facilitate public debate.

As with any study of a single district, the results may not be representative of sentencing nationwide. In Massachusetts, for example, the same qualities that led the court to approve its unusual disclosure policy and to generate a body of sentencing scholarship might make it dissimilar from other courts in relevant ways. Massachusetts is also one of the nation’s most politically Democratic states, although the district’s 15-member bench is split roughly evenly, with eight Democrats and seven Republicans presently sitting.

190 The offense level, criminal history category, and sentencing range under the Guidelines, along with the defendant’s year of birth, proved particularly helpful in matching records. None of that information appears on the criminal docket on PACER.


192 See, e.g., William G. Young, An Open Letter to U.S. District Judges, 50 FED. LAW 30 (July 2003). Judge Young’s remarkable 177-page decision in Green not only anticipated the invalidation of the Guidelines on Sixth Amendment grounds, but contains one of the most comprehensive critiques of the Guidelines ever assembled.


194 Hofer et al., supra note 40, at 279.

195 N.Y. TIMES, 2008 Presidential Race: Massachusetts, Oct. 10, 2008 (noting that over the last ten Presidential elections, Massachusetts has been the most solidly Democratic state in the country).

196 See Federal Judges Biographical Database, supra note 138. Party control at the state or local level should not raise serious concerns about representativeness because federal judges are appointed by the President, making it difficult for local political machines to capture the bench. In any case, a study of Boston judges does not involve any greater risk of party effects than past studies of San Francisco, see Waldfogel, Empirically Based Sentencing Guidelines, supra note 48, at 294, or New York City and Philadelphia, see Payne, supra note 48, at 337. Even the large-scale national studies have ensured a random distribution of cases by limiting their dataset to cities where several judges shared a single case wheel, which necessarily oversamples sentences in locations with disproportionately Democratic populations.
A comparison between sentencing trends nationwide and in the District of Massachusetts indicates potentially important differences. Figures 3a and 3b show average sentences and sentencing relative to the Guidelines, respectively, for fiscal years 2003-2008:

![Figure 3a: Average Sentences by Fiscal Year](image)

![Figure 3b: Guideline Sentencing by Fiscal Year](image)

Perhaps surprisingly, average sentences in Massachusetts are slightly higher than sentences nationwide. But the court also has imposed below-Guideline sentences more frequently than other courts, particularly immediately after Booker in fiscal year 2005. The gap between the national and Massachusetts figures for guideline sentencing is partially attributable to “fast track” programs for immigration offenses, which account for 7.4% of sentences nationwide. Those programs boost the nationwide rate of government-sponsored sentences relative to districts, like Massachusetts, that have no

---


198 See U.S.S.G. § 5K3.1 (Early Disposition Program Departure). Fast-track programs ease a crushing burden on courts and prosecutors in border districts, but they are controversial because they must be authorized by the Attorney General and are not available in all districts, injecting obvious regional disparity into sentencing outcomes.
fast-track authority. Nonetheless, the rate of below-range sentencing in Massachusetts sets it apart from other district courts, potentially undermining its representativeness.

On the other hand, there are a number of advantages—other than the unique trove of data—to focusing on judges in a single district when studying inter-judge sentencing disparity. First, this approach avoids the risk that inter-district disparity in prosecutorial practices might be mistaken for inter-judge disparity. This is an important advantage, given the well-developed criticism that the Guidelines have shifted power over sentencing to prosecutors. In Massachusetts, the Criminal Division of a single U.S. Attorney’s office charges and prosecutes virtually all federal cases.

Second, it allows a comparison of judges who share a common case pool, avoiding the risk that inter-region disparity in the types of offenses committed or prosecuted might be mistaken for inter-judge disparity. The District of Massachusetts has small offices in Springfield and Worcester, but almost all of the court’s judges sit in Boston and participate in the random case distribution system for the court’s Eastern Division. This analysis employs a “natural experiment” method that depends on a random distribution of cases and therefore focuses on the Boston case pool.

Third, it avoids concerns about inter-circuit disparity caused by differences in appellate precedent. In the wake of Booker, circuit courts split on a number of questions concerning reasonableness review, and their divergent approaches are thought to have influenced sentencing decisions in the district courts. Also, the composition of courts of appeals may itself influence sentencing decisions. Schanzenbach and Tiller’s research suggests that the partisan alignment of some circuit courts can affect sentence length and the likelihood of departure. Examining a single district ensures that all judges in the

---


201 See supra notes 72-83 and accompanying text.

202 See United States Attorney’s Office District of Massachusetts, Divisions, at http://www.usdoj.gov/usoa/ma/divisions.html. It is possible that prosecutors and defense attorneys in the district change their charging and plea bargaining practices in response to the judge assigned to the case, based on the judge’s reputation. Because such changes reflect an assessment of the judge, rather than differences between prosecutors or between defense attorneys, they are properly treated as sources of inter-judge disparity.


204 See Gall v. United States, 128 S. Ct. 586 (2007) (rejecting the conclusion of some courts of appeals that a significant variance from the Guidelines requires an extraordinary justification); Rita v. United States, 127 S. Ct. 2456 (2007) (affirming the decision of some courts of appeals to apply a presumption of reasonableness when reviewing within-range sentences on appeal).


206 Schanzenbach & Tiller, supra note 129, at 735. For foundational research on the influence of party affiliation on courts of appeals, see CASS R. SUNSTEIN ET AL., ARE JUDGES POLITICAL? AN EMPIRICAL INVESTIGATION OF THE FEDERAL JUDICIARY (2006).
dataset were bound to follow the same circuit precedent, subject to review by the same mix of appellate judges.

This Article examines only sentences in the District of Massachusetts. Its findings about the sentencing patterns of judges in Boston cannot explain changes in Boise, Baltimore, Birmingham, or Bangor. Nonetheless, the District of Massachusetts offers researchers unparalleled access to pre- and post-


Booker sentencing information. It therefore offers the best available evidence of how sentencing by individual judges has changed since Booker.

B. Methods

1. Case Matching and Selection

To facilitate comparisons between pre- and post-


Booker sentencing, this Article examines sentences in five time periods:

1. Mandatory Guidelines: October 1, 2001 – April 30, 2003 (∼ 19 months)
2. PROTECT Act: May 1, 2003 – June 23, 2004 (∼ 14 months)
3. Post-
Booker I: January 12, 2005 – June 30, 2006 (∼ 18 months)
4. Post-
Booker II: July 1, 2006 – December 9, 2007 (∼ 17 months)

The Mandatory Guidelines period begins on the first day of fiscal year 2002. The PROTECT Act period begins on the effective date of the Act and ends on the date Blakely was decided. Because of the chaos that followed that decision, the interregnum between Blakely and Booker is ignored. Post-


Booker sentences are divided into two periods. One (“Post-
Booker I”) captures approximately 18 months of sentencing immediately following Booker, beginning on the date of the decision and ending on June 30, 2006, the midpoint of calendar year 2006. The next (“Post-
Booker II”) captures the following 17 months of sentencing, leading up to the Supreme Court’s decisions in Kimbrough and Gall. The final period (“Kimbrough/Gall”) begins on the date of those decisions and ends on September 30, 2008, the last day of the most recent fiscal year for which Sentencing Commission data are available.

These cutoff dates were selected with two objectives in mind: to isolate distinct phases in the recent history of federal sentencing and, where possible, to create periods of approximately equal length, around 18 months. Unfortunately, two periods—the PROTECT Act period (14 months) and the Kimbrough/Gall period (10 months)—necessarily cover fewer months because of the timing of the Supreme Court’s decisions.

Using Schanzenbach and Tiller’s matching technique as a starting point, I matched case dockets on PACER with electronic case records released by the Commission. Because a judge-specific analysis of post-


Booker sentencing requires a

...

207 The Commission’s post-


Booker reports have ignored the period between Blakely and Booker as well. See, e.g., FINAL REPORT, supra note 125.

208 At its Data and Research Conference in May 2009, the Commission distributed flash drives containing the full set of sentencing data files through fiscal year 2008. The release of FY2007 and FY2008 data ahead of the ordinary schedule was unexpected, and a valuable benefit for participants.
sufficient number of sentences per judge, I sought to maximize the number of cases matched in each period. I therefore did not restrict the search to a sample of dates, but instead searched for every criminal case filed in the Boston office between January 1, 2000 and June 30, 2008.\textsuperscript{209} I restricted the search to Boston cases because an important assumption of the “natural experiment” approach is that cases were randomly distributed in each time period.\textsuperscript{210} Cases filed in Springfield and Worcester are not assigned using the same distribution “deck” as Boston cases, and are almost always assigned to the judges who have chambers in those cities, such as Judge Ponsor (Springfield) and Judge Saylor (Worcester).\textsuperscript{211}

The initial search yielded around 5,000 cases, which included dismissals, jurisdictional transfers, or acquittals that did not result in a sentence. For cases in which a sentence was imposed, I first attempted to find a match in the Commission’s database using information in the docket sheet. When the docket provided insufficient information—a common occurrence for fiscal years 2004 and later—information from the Statement of Reasons was used to narrow the list of potential matches. This method proved highly reliable: less than 0.4% of sentences could not be matched because of multiple similar sentences in the Commission’s data.\textsuperscript{212} Like Schanzenbach and Tiller, however, I encountered a surprising number of sentences, about 8.5% of those in the initial search, that did not look similar to any of the Commission’s records.\textsuperscript{213} I echo their concern that this is a significant amount of missing data.\textsuperscript{214}

The process resulted in 2,659 matched cases, more than 90% of the Boston sentences in the Commission’s files. Table 1 lists the number and percentage of cases in the Commission’s data that were successfully matched, by fiscal year:

<table>
<thead>
<tr>
<th>Fiscal Year</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Boston Cases</td>
<td>497</td>
<td>479</td>
<td>315</td>
<td>306</td>
<td>418</td>
<td>460</td>
<td>403</td>
<td>2,878</td>
</tr>
<tr>
<td>Matched Cases</td>
<td>445</td>
<td>433</td>
<td>292</td>
<td>273</td>
<td>387</td>
<td>430</td>
<td>369</td>
<td>2,629</td>
</tr>
<tr>
<td>% of Cases Matched</td>
<td>89.5%</td>
<td>90.4%</td>
<td>92.7%</td>
<td>89.2%</td>
<td>92.6%</td>
<td>93.5%</td>
<td>91.6%</td>
<td>91.3%</td>
</tr>
</tbody>
</table>

Table 1: Matched Cases, by Fiscal Year\textsuperscript{215}

\textsuperscript{209} PACER’s “Reports” tool allows searches by Case Type, including criminal cases. I included pending and terminated defendants, but excluded cases involving fugitive defendants. I also conducted targeted searches for cases with earlier filing dates that were “closed” during fiscal year 2002, to ensure a comparable percentage of matched cases in each year being studied.

\textsuperscript{210} See infra notes 227-232 and accompanying text.

\textsuperscript{211} See D. Mass. Local Rule 40.1(B)(3), (C) (establishing separate random case-distribution systems for each division).

\textsuperscript{212} Cf. Schanzenbach & Tiller, \textit{supra} note 129, at 730 (reporting that only 3% of sentences could not be matched using the docket sheet alone, mostly in immigration cases).

\textsuperscript{213} See \textit{id}. at 730.

\textsuperscript{214} \textit{Id.}

\textsuperscript{215} Fiscal years 2004 and 2005 include fewer sentences because they exclude sentences imposed between \textit{Blakely} and \textit{Booker}. Boston cases were identified using the Commission’s parole office code, except that cases without any parole office code were included.
The Commission’s datasets include only initial sentences imposed, not sentences imposed on remand following appeal. In Massachusetts, as in the rest of the nation, sentencing appeals are relatively few and usually unsuccessful. In fiscal year 2006, for example, there were 516 sentences in the District of Massachusetts, but just 41 sentencing appeals (an appeal rate of approximately 7.9%), and only 9 of those appeals resulted in remand (a success rate of 22.0%, for an effective reversal rate of 1.7%). Two-thirds of successful appeals were brought by defendants, not the government.

Two categories of sentences from the pool of matched cases were excluded. First, to ensure a sufficient number of cases per judge to draw reliable conclusions about sentencing behavior, sentences by judges on pace to impose fewer than 25 sentences over a two-year period were excluded. Because this Article examines sentences over a seven-year period, some judges have a sufficient caseload during some periods, but not in others. The Article principally relies on sentences by a “core” set of judges who imposed a sufficient number of sentences in all five periods. In a few instances, it also reports results for an “expanded” set that includes additional judges who imposed a sufficient number of sentences in at least two periods.

Second, to ensure that the data were the product of random distribution, the sentences of judges who principally sat in Springfield or Worcester were removed, even if those judges also handled a small number of Boston cases. The court’s local rules do not provide for random distribution to judges outside the division in which the case was filed. Senior judges were not automatically excluded, but were subject to the same minimum-caseload requirement as other judges. As it happens, no senior judge on the

---

216 The Commission does record and track data concerning resentencing, but does not release those data to the public. Interview with Paul Hofer (May 8, 2009).
218 See id. Tables 55, 56, 56A & n.2.
219 See id. Tables 56, 56A.
220 The Anderson-Kling-Stith study excluded judges who did not have an average of 30 criminal cases per two-year period, including dismissals, transfers, and acquittals. See Anderson et al., supra note 27, at 287. Because the Commission’s data include only cases that resulted in a sentence, I used a slightly lower cutoff, excluding judges with an average of fewer than 25 sentences per two-year period.
221 Nine judges had a sufficient number of sentences in all periods, and therefore qualified as “core” judges. A tenth judge, Judge F, also was treated as a “core” judge because the judge imposed a sufficient number of sentences in every period except the Kimbrough/Gall period, the most recent and shortest period. Judge F’s sentences during the Kimbrough/Gall period were ignored.
222 The “expanded” set of judges includes three additional judges who imposed a sufficient number of sentences in two periods. To avoid distortions created by a small number of cases in a particular period, the expanded set excludes sentences by those judges in any period where they did not impose a sufficient number of sentences.
223 See D. Mass. Local Rule 40.1(B)(3). The court’s rules do allow for the transfer of cases between divisions “for good cause,” D. Mass. Local Rule 40.1(F), but the clerk’s office reports that such transfers are infrequent.
224 Although senior judges have the option of “limit[ing] the category of cases or types of alleged criminal offenses” they accept, D. Mass. Local Rule 40.1(B)(3), the clerk of court reports that senior judges in the District of Massachusetts historically have not exercised that option. Interview with Sarah Allison Thornton, Clerk of the United States District Court for the District of Massachusetts, Oct. 14, 2008. Because senior judges must accept assignments by lot, like their colleagues, their participation remains fully random, even if they handle fewer cases overall. See D. Mass. Local Rule 40.1(B)(3).
court carried a sufficient caseload during all periods to qualify as a “core” judge, although some sentences by senior judges appear in the “expanded” set of cases.  

Table 2 lists the number of sentences per “core” judge, by period:

<table>
<thead>
<tr>
<th>Mandate Guidelines</th>
<th>PROTECT Act</th>
<th>Post-Booker I</th>
<th>Post-Booker II</th>
<th>Kimbrough/Gall</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge A</td>
<td>56</td>
<td>39</td>
<td>68</td>
<td>48</td>
<td>17</td>
</tr>
<tr>
<td>Judge B</td>
<td>54</td>
<td>38</td>
<td>36</td>
<td>47</td>
<td>36</td>
</tr>
<tr>
<td>Judge C</td>
<td>60</td>
<td>38</td>
<td>64</td>
<td>64</td>
<td>31</td>
</tr>
<tr>
<td>Judge D</td>
<td>52</td>
<td>36</td>
<td>50</td>
<td>41</td>
<td>26</td>
</tr>
<tr>
<td>Judge E</td>
<td>82</td>
<td>45</td>
<td>55</td>
<td>49</td>
<td>33</td>
</tr>
<tr>
<td>Judge F</td>
<td>40</td>
<td>34</td>
<td>44</td>
<td>57</td>
<td>--</td>
</tr>
<tr>
<td>Judge G</td>
<td>65</td>
<td>38</td>
<td>30</td>
<td>53</td>
<td>34</td>
</tr>
<tr>
<td>Judge H</td>
<td>79</td>
<td>46</td>
<td>45</td>
<td>47</td>
<td>23</td>
</tr>
<tr>
<td>Judge I</td>
<td>54</td>
<td>38</td>
<td>41</td>
<td>44</td>
<td>36</td>
</tr>
<tr>
<td>Judge J</td>
<td>61</td>
<td>42</td>
<td>56</td>
<td>68</td>
<td>22</td>
</tr>
<tr>
<td>Total</td>
<td>603</td>
<td>394</td>
<td>489</td>
<td>518</td>
<td>258</td>
</tr>
</tbody>
</table>

Table 2: Sentence Count for Core Judges

An important assumption of the natural experiment method is a random distribution of cases—and, by extension, of sentences—to judges in the dataset. The court’s local rules provide for assignment “by lot,” with separate categories of cases based on the anticipated length of the proceedings. The clerk’s office reports that case assignments are indeed randomized, using a computerized “deck” that differentiates between divisions and case categories.

Following the Hofer and Waldfogel studies, chi-square analyses were conducted to test the randomness of sentence assignment using

---

225 The Anderson-Kling-Stith study did not exclude senior judges, but it did exclude judges who did not hear a full caseload, on the assumption that such judges “were unlikely to have fully participated in the randomization.” Anderson et al., supra note 27, at 287. For a nationwide study where the actual practice among judges cannot be determined easily, that assumption is justified. In this study, however, the court rules and historical practice indicate that senior judges in the dataset did receive a fully random caseload.

226 In this table and throughout the article, I use letters rather than names to identify judges. Identifying judges by name is unnecessary because Congress saw inter-judge disparity as a concern regardless of which particular judges reached inconsistent results. I also suspect that the Administrative Office’s reticence about releasing judge identifying information reflects concern that it could be used to build a “blacklist” of disfavored judges. See Letter of Leonidas Ralph Mecham, supra note 96, at 3 (urging Congress not to direct the release of judge identifying information on the ground that it may lead to “unfair criticism” of judges based on “isolated cases”). Although I see no reason why federal judges who enjoy life tenure cannot withstand criticism—even “unfair” criticism—of their decisions, I hope this Article illustrates how judge-identifying information can enable valuable research without targeting individual judges.

227 Because cases, rather than sentences, are randomly assigned, differences between judges as to the rate of transfer, dismissal, or acquittal could undermine the assumption that sentences are randomly distributed. As the Hofer study noted, however, the available research shows that although “differences in dismissal and acquittal rates do exist among judges . . . they are negligible and appear unlikely to explain differences in sentences.” Hofer et al., supra note 40, at 283 n.121.

228 D. Mass. Local Rule 40.1(B)(3).

several case attributes that cannot easily be changed after filing: the defendant’s race, gender, age, and education. All four tests supported the conclusion that the distribution was random for the dataset as a whole. The gender, age, and education tests further supported the conclusion that the distribution was random in each period.

Chi-square analysis based on the defendant’s race supported the conclusion that the distribution was random in the Mandatory Guidelines, PROTECT Act, and Kimbrough/Gall periods. For the two Post-Booker periods, however, the race of the defendant was not demonstrably independent of the identity of the sentencing judge. The likely culprit is drug conspiracy cases, which frequently involve multiple defendants of the same race. The Hofer study encountered similar difficulties with using chi-square tests based on race for large cities, and in light of the results for other attributes, the results for race do not undermine the premise that sentences were distributed randomly.

Another important assumption of this natural experiment is that changes in sentencing outcomes from 2002 to 2008 are exogenous, caused by Booker and related developments in sentencing law rather than on-the-ground factors in Boston. If, for example, the mix of cases prosecuted in the District of Massachusetts changed during those years, we could not confidently attribute any changes in sentencing to Booker. As Table 3 shows, however, the composition of the case pool for Boston judges has not meaningfully changed from period to period:

<table>
<thead>
<tr>
<th></th>
<th>Drug Trafficking</th>
<th>Fraud</th>
<th>Immigration</th>
<th>Firearms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandatory Guidelines</td>
<td>41.2%</td>
<td>13.9%</td>
<td>10.8%</td>
<td>7.3%</td>
</tr>
<tr>
<td>PROTECT Act</td>
<td>38.4%</td>
<td>12.6%</td>
<td>13.0%</td>
<td>11.1%</td>
</tr>
<tr>
<td>Post-Booker I</td>
<td>42.2%</td>
<td>11.0%</td>
<td>12.0%</td>
<td>9.1%</td>
</tr>
<tr>
<td>Post-Booker II</td>
<td>37.6%</td>
<td>15.3%</td>
<td>9.5%</td>
<td>11.1%</td>
</tr>
<tr>
<td>Kimbrough/Gall</td>
<td>43.5%</td>
<td>14.4%</td>
<td>11.2%</td>
<td>13.7%</td>
</tr>
<tr>
<td>All Periods</td>
<td>40.4%</td>
<td>13.5%</td>
<td>11.2%</td>
<td>9.9%</td>
</tr>
</tbody>
</table>

Table 3: Percent of Cases for Each Offense Type, by Period

Together, the four largest primary offense types in Boston—drug trafficking, fraud, immigration, and firearms—account for about 75% of the case pool. The percentage of

---

230 Because chi-square analysis depends on a minimum number of cases per cell, the race variable (the Commission’s NEWRACE) was limited to white, black, and Hispanic offenders, omitting the “other” category. Similarly, the education variable (NEWEDUC) omitted the “college graduate” category, which applied to too few defendants. The Commission’s Age variable was coded into three categories: age 18-29, age 30-39, and age 40 and over.

231 Chi-square tests on age uncovered no significant relationship in any period. Tests on education uncovered no significant relationship in any period except Kimbrough/Gall, and that result likely was affected by the smaller population of cases. Tests on gender uncovered no significant relationship in any period except Post-Booker II. Given the results for gender in adjacent periods and for the dataset as a whole, that result does not call into question the premise that the distribution of cases was random.

232 Hofer et al., supra note 40, at 320 (technical appendix).

cases of each type shifts slightly from period to period, but there are no trends in composition of the case pool that might account for differences in sentencing outcomes. Accordingly, although the “natural experiment” lacks a control group (such as a group of federal judges somehow exempt from *Booker*), nothing in the available data suggests that Boston judges have confronted a meaningfully different mixture of cases since *Booker*.

2. Measures of Sentencing by Individual Judges

In studies of sentencing by individual judges, a foundational design question is how to measure sentencing outcomes. Previous natural experiment studies have relied exclusively on average sentence length. This Article supplements that measure by also examining sentencing relative to the sentencing range under the Guidelines.

The Hofer, Anderson-Kling-Stith, and Waldfogel studies measured sentencing outcomes using a single metric: average prison term, in months. The studies recognized that mandatory minimum sentences, a number of which were enacted contemporaneous with the Guidelines, might conceal changes in inter-judge disparity by forcing sentences higher for all judges. To guard against that risk, the Hofer study recommended that researchers “exclude cases where mandatory minimum statutes truncate the [sentencing] range.” Government-sponsored sentence reductions for “substantial assistance” or other cooperation present a similar risk by lowering sentences for all judges. Accordingly, in examining sentence length, this Article separately considers cases that did not involve a mandatory minimum or government-sponsored sentence reduction.

For studies of sentences that predate the Sentencing Reform Act, a focus on average sentence length made sense because no alternative benchmark was available. But the approach has drawbacks. Average sentence length captures only a judge’s across-the-board leniency or severity, sometimes called the “primary judge effect.” It does not capture other forms of inter-judge disparity linked to particular offense or offender characteristics.

To illustrate, suppose Judge X sentences drug offenses harshly but white-collar offenses leniently. Judge Y does exactly the opposite, sentencing white-collar offenses harshly and drug offenses leniently. It is entirely possible for Judges X and Y to have similar average sentence length, with high sentences for one set of offenses offsetting low sentences for the other. Yet a defendant convicted of a drug or white-collar offense stands to receive starkly different results based solely on the identity of the sentencing judge, a clear instance of unwarranted disparity. As Kate Stith and Judge José Cabranes have explained, there is “[a] possibility that comparing each judge’s average sentence

---

234 See Hofer et al., *supra* note 40, at 307-08 (technical appendix); Anderson et al., *supra* note 27, at 281; Waldfogel, *Empirically Based Sentencing Guidelines, supra* note 48, at 294. Consistent with the Sentencing Commission’s convention, sentences of probation are coded as zero months of imprisonment. Hofer et al., *supra* note 40, at 307-08 (technical appendix); *see also* STITH & CABRANES, supra note 10, at 62 (following the same convention).

235 Hofer et al., supra note 40, at 275 n.103.

236 See 18 U.S.C. § 3553(e); U.S.S.G. § 5K1.1.

237 Id. at 240-41.
masks considerable variability within each set of sentences.\textsuperscript{238} Likewise, the authors of the Hofer study acknowledge that by examining differences in average sentence length, they had focused on “the tip of the disparity iceberg.”\textsuperscript{239}

To supplement average sentence length, studies of post-\textit{Booker} sentencing should also compare sentences to the range recommended under the Sentencing Guidelines. The Guidelines are designed to take into account a dizzying number of facts concerning the circumstances of the offense and the history and characteristics of the defendant. Assuming a random distribution of cases, those facts will be randomly distributed among judges. Calculating differences in average sentence length relative to the Guidelines therefore captures another form of potential inter-judge disparity, driven not by general harshness or leniency but by judges’ conclusions about how often the advisory Guidelines recommend an appropriate sentence.

The practical difficulties that confronted early matched-case studies of the Guidelines do not arise when using the Guidelines as a point of comparison in studies of post-\textit{Booker} sentencing. Judges have calculated the Guidelines sentencing range continuously in cases before and after \textit{Booker}, ensuring that data are available throughout the relative time periods.\textsuperscript{240} Although the Guidelines are regularly amended by the Commission, they seldom change dramatically from year to year.\textsuperscript{241} The Guidelines also have proved important to policymakers. The Commission regularly reports sentencing patterns relative to the Guidelines, both nationwide and district-by-district.\textsuperscript{242} And the perceived rate of downward departures from the Guidelines prompted Congress to intervene by passing the PROTECT Act.\textsuperscript{243}

Using the Guidelines as a reference point does not suggest or assume that the guideline sentencing range is “correct” or just.\textsuperscript{244} If all judges strictly followed the Guidelines, they might well reach wrong results in a broad set of cases, according to some underlying theory of punishment.\textsuperscript{245} But if some judges almost always sentence within the guideline range, while other judges almost never sentence within the guideline range, an important form of inter-judge disparity would be present regardless of which judges are right. Guidelines sentencing patterns thus allow researchers to study a distinct form of inter-judge sentencing disparity.

\textsuperscript{238} \textsc{Stith} \& \textsc{Cabrines}, \textit{supra} note 10, at 119. The FJC study of Second Circuit judges found, for example, that although the judges sentenced at roughly the same level on average, different judges “had very different opinions about which particular cases deserved more severe or lenient punishment,” meaning that “the relative similarity in the overall averages masked a greater disparity at the individual case level.” Hofer et al., \textit{supra} note 40, at 251.

\textsuperscript{239} Hofer et al., \textit{supra} note 40, at 297.

\textsuperscript{240} United States v. Booker, 543 U.S. 220, 245 (2005) (Breyer, J., writing for the Court) (holding that judges must continue to calculate and consider the Guidelines).

\textsuperscript{241} The most notable changes during the period of this Article were the Commission’s extensive revisions pursuant to the PROTECT Act. \textit{See supra} notes 84-92 and accompanying text. Given the importance of those changes, the Article examines the Mandatory Guidelines and PROTECT Act periods separately.


\textsuperscript{243} \textit{See supra} notes 84-92 and accompanying text.

\textsuperscript{244} \textit{See Bowman, supra} note 5, at 296.

3. Discretionary Sentences

The Statements of Reasons for cases in the dataset also underscore a previously
underappreciated point that is relevant to future research on guideline sentencing. For
legal and practical reasons, judges frequently do not have the option of sentencing outside
the guideline range. In over 20% of cases in Boston, the judge was effectively
constrained from sentencing either above or below the Guidelines. Excluding those cases
yields a slightly narrower subset of sentences—call them “discretionary sentences”—of
particular interest when measuring inter-judge disparity in sentencing under the
Guidelines.

In reviewing the District of Massachusetts documents and coding thousands of
cases, several recurring constraints became apparent. First, a statutory mandatory
minimum sometimes prevents judges from imposing a below-range sentence. By
operation of the Guidelines, whenever a mandatory minimum exceeds the guideline
minimum, then the bottom end of the guideline range effectively shifts upward. Thus,
if the sentencing range under the Guidelines is 51-63 months, but the statutory minimum
is 60 months, then the sentencing range becomes 60-63 months. A judge who imposes
a 60-month sentence under those circumstances has imposed a within-range sentence, but
had no option to impose a below-range sentence. In this dataset, a statutory mandatory
minimum made it impossible for the judge to impose a below-range sentence in 6.4% of
cases.

Second, the time a defendant already has spent in custody sometimes prevents
judges from imposing a below-range sentence. In the federal system, defendants may
receive credit for time served in official detention prior to the date the sentence
commences, and it is common for a judge to impose a sentence of “time served,”
allowing the defendant to be released immediately. In some cases, the time served by the
defendant at the time of sentencing exceeds the guideline minimum. When that happens,
the judge cannot impose a sentence below the guideline range because the defendant
already has served a within-range sentence. In this dataset, a term of time served
prevented a below-range sentence in 4.2% of cases.

Third, notwithstanding their reputation for severity, the Guidelines often
recommend a sentence of probation as an appropriate punishment. For sentences with a
guideline range of 0-6 months, in Zone A of the sentencing table, a sentence of probation
is a within-range sentence. For sentences with a guideline range of 6-12 months, in

---

246 See id. § 5G1.1(c)(2). If the statutory minimum exceeds both the guideline minimum and the guideline
maximum, then the statutory minimum becomes the guideline sentence. Id. § 5G1.1(b).
247 See id. § 5G1.1 Commentary.
248 Missing data prevented the coding of constraints for 1.2% of cases in the dataset. Percentages reported
for each constraint are based on the remaining cases.
250 Although not a legal constraint, the federal judges with whom I have spoken cannot imagine
circumstances in which a judge would impose a sentence of less than time served, implying that the prior
detention was unlawful. See Interview with Paul Cassell (Oct. 10, 2008); cf. U.S.S.G. § 1B1.10
Application Note 3 (prohibiting the reduction of a sentence “below time served” following a downward
amendment to the Guidelines).
251 To the extent that the “time served” constraint overlapped with other constraints, the case was coded as
“time served.”
Zone B of the sentencing table, a sentence of probation qualifies as a within-range sentence if it includes some conditions of intermittent, community, or home confinement.253 Judges who take advantage of these options can impose a term of probation without sentencing below the guideline range. In this dataset, the Guidelines recommended a sentence of 0-6 months in 5.1% of cases, and a sentence in Zone B in 3.9% of cases.

Together, these constraints were present in 19.6% of cases and accounted for almost one-third of within-Guidelines sentences. Yet the existing literature on federal sentencing has almost entirely ignored the role that they play in limiting the discretion of district courts.254 As noted above, in response to the PROTECT Act, judges (rightly) objected to criticisms about the high rate of downward departures. Many of those departures were in fact sponsored by the government, as a product of “crisis” conditions in border districts, fundamentally changing the nature of the sentencing decision.255 So too, after Booker, judges can object to criticisms that they remain too wedded to the Guidelines. Many within-range sentences can be traced to statutory and practical constraints that limit their options.

An implication of these constraints is that there exists a narrower class of cases, just under 80% in this dataset, in which the court is free to sentence above or below the Guidelines, without practical or legal obstacles. Such “discretionary sentences” ought to be of particular interest in studies of inter-judge disparity because, in theory, they are the only cases in which such disparity would be fully apparent.

III. RESULTS

The Massachusetts data suggest two conclusions about sentencing by individual judges in the wake of Booker. First, the identity of the sentencing judge has become a stronger predictor of sentence length since the decision. Second, the identity of the sentencing judge has become a stronger predictor of sentencing relative to the Guidelines. Some judges now sentence outside the guideline sentencing range more frequently, and to a greater extent, than their colleagues.

A. Average Sentence Length

Among Boston judges, as among judges nationwide, average sentence length has increased since Booker. Table 4 shows the increase, both for all sentences and for sentences not governed by a statutory mandatory minimum.256

253 Id. § 5B1.1(a)(2).
254 The Hofer study recognized the importance of one of these constraints, recommending that future researchers exclude cases in which a mandatory minimum truncates the guideline range. Hofer et al., supra note 40, at 275 n.103.
255 See supra notes 96-98 and accompanying text.
256 Cases were treated as having no mandatory minimum if the court sentenced below the otherwise-applicable minimum based on the statutory “safety valve,” 18 U.S.C. § 3553(f), or a government “substantial assistance” motion, see 18 U.S.C. § 3553(e); U.S.S.G. § 5K1.1.
At first glance, the distribution of average sentence length among individual judges suggests little change since Booker. Figure 5 shows average sentence lengths for each core judge by period, with each dot representing the average for a single judge:

Although the difference between the judge with the highest average and the judge with the lowest average was larger in the Post-Booker I period (56.8 months) than in the Mandatory Guidelines (39.2 months) and PROTECT Act periods (46.2 months), it declined to pre-Booker levels in the two most recent periods. And it is difficult to detect any stable change in the pattern of distribution of average sentence length following Booker.

A statistical analysis reveals that, in the full set of sentences, the relationship between the identity of the sentencing judge and sentence length has followed an uneven course since Booker. Linear regression models for each period, using dummy indicators

Table 4: Average Sentence Length, Core Boston Judges

<table>
<thead>
<tr>
<th></th>
<th>All Sentences</th>
<th>No Mandatory Minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandatory Guidelines</td>
<td>48.0 months</td>
<td>31.4 months</td>
</tr>
<tr>
<td>PROTECT Act</td>
<td>49.9 months</td>
<td>31.1 months</td>
</tr>
<tr>
<td>Post-Booker I</td>
<td>57.5 months</td>
<td>32.8 months</td>
</tr>
<tr>
<td>Post-Booker II</td>
<td>58.9 months</td>
<td>34.8 months</td>
</tr>
<tr>
<td>Kimbrough/Gall</td>
<td>63.9 months</td>
<td>35.0 months</td>
</tr>
</tbody>
</table>

Figure 5: Average Sentence Distribution
for judges as independent variables, can measure the association between the identity of the judge and sentence length.\textsuperscript{257} R-squared can be used to measure the strength of the relationship, and to facilitate comparisons between periods.\textsuperscript{258} As in the Hofer study, the percentage of variance explained by the model is converted into actual months, as an average across all sentences for all judges.\textsuperscript{259} Table 4 reports the results:

<table>
<thead>
<tr>
<th>R-Squared</th>
<th>Variance Explained</th>
<th>Model Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandatory Guidelines</td>
<td>4.2%</td>
<td>13.0 months</td>
</tr>
<tr>
<td>PROTECT Act</td>
<td>4.7%</td>
<td>13.3 months</td>
</tr>
<tr>
<td>Post-Booker I</td>
<td>3.8%</td>
<td>14.1 months</td>
</tr>
<tr>
<td>Post-Booker II</td>
<td>2.2%</td>
<td>9.7 months</td>
</tr>
<tr>
<td>Kimbrough/Gall</td>
<td>6.1%</td>
<td>15.5 months</td>
</tr>
</tbody>
</table>

* Not significant at the .05 level

Table 5: Linear Regression Models
Identity of Judge and Sentence Length

The models suggest a mixed bag of changes since Booker. Before the decision, the percentage of variance in sentence length explained by the identity of the judge stood at 4.7%. In the eighteen months after Booker the rate fell to 3.8%. In the next eighteen months it further declined to 2.2%, an average of 3.6 months per sentence less than its pre-Booker level, and the identity of the judge was not a statistically significant predictor of sentence length. But in the Kimbrough/Gall period, the rate rose sharply to 6.1%, about 30-45% higher than pre-Booker levels depending on whether the point of comparison is the Mandatory Guidelines or PROTECT Act period. The seesaw trend makes it difficult to draw firm conclusions based on the full set of sentences.

As previous researchers have noted, however, mandatory minimums and government-sponsored sentence reductions affect sentence length for all judges and, as a result, may mask changes in inter-judge disparity.\textsuperscript{260} An analysis of Boston sentences not governed by a mandatory minimum or government-sponsored sentence reduction reveals that the strength of the relationship between the identity of the judge and length of sentences has increased since Booker. Tables 6a and 6b report the results.

\textsuperscript{257} Sentence length is measured as a term of imprisonment in months. The Sentencing Commission codes a sentence of probation as a sentence of zero months.

\textsuperscript{258} In linear regression, the r-squared statistic is a value between 0 and 1 that describes the percentage of variance in the dependent variable that is explained by the independent variable. \textit{See generally} MICHAEL O. FINKELSTEIN & BRUCE LEVIN, STATISTICS FOR LAWYERS 345 (1990). For a discussion of some uses and limitations of r-squared, see David R. Stras & Ryan W. Scott, \textit{An Empirical Analysis of Life Tenure: A Response to Professors Calabresi and Lindgren}, 20 HARV. J. L. PUB. POL’Y 791, 817 (2007). The r-squared values here are shown as percentages.

\textsuperscript{259} Hofer et al., \textit{supra} note 40, at 287. Actual months of variance explained were determined by “(1) multiplying the total variance by the portion of the variance accounted for by judges, and (2) finding the square root of the result, thus translating the numbers back into absolute terms.” \textit{Id.} at 287 n.187.

\textsuperscript{260} \textit{See supra} notes 235-236 and accompanying text.
<table>
<thead>
<tr>
<th></th>
<th>R-Squared</th>
<th>Variance Explained</th>
<th>Model Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandatory Guidelines</td>
<td>5.0%</td>
<td>9.7 months</td>
<td>.008</td>
</tr>
<tr>
<td>PROTECT Act</td>
<td>4.4%</td>
<td>8.6 months</td>
<td>.195*</td>
</tr>
<tr>
<td>Post-Booker I</td>
<td>6.1%</td>
<td>10.7 months</td>
<td>.028</td>
</tr>
<tr>
<td>Post-Booker II</td>
<td>4.4%</td>
<td>10.0 months</td>
<td>.104*</td>
</tr>
<tr>
<td>Kimbrough/Gall</td>
<td>8.0%</td>
<td>10.3 months</td>
<td>.180*</td>
</tr>
</tbody>
</table>

* Not significant at the .05 level

**Table 6a: Linear Regression Models**

**Identity of Judge and Sentence Length, Excluding Mandatory Minimums**

<table>
<thead>
<tr>
<th></th>
<th>R-Squared</th>
<th>Variance Explained</th>
<th>Model Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandatory Guidelines</td>
<td>5.2%</td>
<td>10.1 months</td>
<td>.020</td>
</tr>
<tr>
<td>PROTECT Act</td>
<td>5.6%</td>
<td>9.8 months</td>
<td>.112*</td>
</tr>
<tr>
<td>Post-Booker I</td>
<td>7.9%</td>
<td>12.8 months</td>
<td>.011</td>
</tr>
<tr>
<td>Post-Booker II</td>
<td>5.3%</td>
<td>11.0 months</td>
<td>.075†</td>
</tr>
<tr>
<td>Kimbrough/Gall</td>
<td>9.3%</td>
<td>11.5 months</td>
<td>.191*</td>
</tr>
</tbody>
</table>

* Not significant at the .05 level
† Significant only at the .10 level

**Table 6b: Linear Regression Models**

**Further Excluding Government-Sponsored Sentence Reductions**

For sentences not governed by a mandatory minimum, the portion of variance in sentence length explained by the identity of the judge rose to 6.1% in the eighteen months after *Booker*, compared with 5.0% under the Mandatory Guidelines in 2002-2003 and 4.4% under the PROTECT Act in 2003-2004. That represents an increase of 20-40% over pre-*Booker* levels, depending on the point of comparison. Although the rate declined to PROTECT Act levels again in the following eighteen months, it has since increased to 8.0% in the *Kimbrough/Gall* period, roughly 60-80% above pre-*Booker* levels.

And the results are more pronounced if government-sponsored sentence reductions are excluded. The rate of variation in sentence length explained by the identity of the judge rose to 7.9% after *Booker*, compared with 5.2% under the Mandatory Guidelines in 2002-2003 and 5.6% under the PROTECT Act. That is a 40-50% increase over pre-*Booker* levels, representing an average of about 3 months per sentence. Again the rate declined to pre-*Booker* levels during the following eighteen months, but again it climbed to its highest levels in the *Kimbrough/Gall* period, reaching 9.3% or approximately 65-85% above pre-*Booker* levels.

Notably, for both sets of sentences not governed by a mandatory minimum, the regression models for the *Kimbrough/Gall* period do not reach the level of statistical significance. That reinforces that the results are preliminary, but does not undermine

---

261 The results are similar if, instead of excluding sentences governed by a mandatory minimum, sentence length is measured as the difference between the sentence and the mandatory minimum.
their persuasiveness. Statistical significance is highly sensitive to sample size, and testing a subset of cases obviously shrinks the sample. Particularly for the Kimbrough/Gall period, which of necessity has about half as many cases as the other periods, the lack of a statistically significant relationship is a reason for caution, but not a reason to discount the results.

Together, these measures of sentence length tell a similar story. In the first eighteen months after Booker, as judges in Boston first began exercising their broader discretion at sentencing, the identity of the judge became a stronger predictor of sentence length. That effect faded in the next eighteen months, as district court judges tested the limits of that discretion in dialogue with the court of appeals. But in the wake of Kimbrough and Gall, the “judge effect” has become stronger than it has been since at least 2001, by some measures 60-80% above pre-Booker levels.

B. Sentencing Relative to the Guidelines

1. Descriptive Comparison

The Boston data also allow a descriptive account of how judges’ sentencing patterns under the Guidelines have changed since Booker. Figure 6 shows the rate of sentencing below the guideline range for all judges in the dataset, by time period:

![Figure 6: Guideline Sentencing, by Time Period](image)

This is the sort of aggregated information that the Commission has reported in the past. It does not indicate how individual judges have responded to Booker.
Figures 6a-6d overlay the sentencing patterns of four Boston judges from the dataset, Judges A, B, C, and D. They illustrate how individual judges’ responses to *Booker* may differ from the average, and from the responses of other judges:

**Figure 6a: Guideline Sentencing, Judge A**

**Figure 6b: Guideline Sentencing, Judge B**

**Figure 6c: Guideline Sentencing, Judge C**

**Figure 6d: Guideline Sentencing, Judge D**
Sentences by Judge A roughly fit the guideline sentencing pattern for all Boston judges combined, with a noticeable decline in below-range sentences under the PROTECT Act, followed by an even larger increase after *Booker*. But each of the other judges’ sentencing patterns fits a distinctive shape, typical of others in the dataset:

- Sentences by Judge B fit a “free at last” pattern: a low rate of below-range sentencing in the two pre-*Booker* periods (11.1% and 10.5%) and a much higher rate in the three post-*Booker* periods (40.0%, 37.5%, and 52.8%). Judge B’s rate of below-range sentencing has more than quadrupled since *Booker*.263

- Sentences by Judge C fit a “business as usual” pattern, with very little change between periods. Judge C’s rate of below-range sentencing moved less than one-half of one percent between the PROTECT Act and Post-*Booker* I periods, from 10.5% to 10.0%, and remained stable in the Mandatory Guidelines (13.3%), Post-*Booker* II (19.1%), and *Kimbrough/Gall* (16.1%) periods as well.264

- Sentences by Judge D fit a “return to form” pattern, in which *Booker* effectively nullified the effects of the PROTECT Act. Judge D’s rate of below-range sentencing stood at 32.7% in the Mandatory Guidelines period, but plummeted to 5.6% under the PROTECT Act. After slowly increasing to 17.0% in the eighteen months after *Booker*, it has returned to 38.6% and 34.6% in the two most recent periods.265

These disparate patterns suggest that the aggregate data released by the Commission fail to capture important differences in the way that individual judges have responded to *Booker*. In some cases, the differences are quite stark. Figure 7 shows the

262 See Gertner, *supra* note 116, at 579 (using the phrase “free at last” to describe the reaction to *Booker* among some district court judges).
263 Sentences by Judges E, F, and G also fit this pattern. Judge E’s rate of below-range sentencing approximately tripled since *Booker*, from 7.3% in the Mandatory Guidelines period and 13.3% in the PROTECT Act period, to 34.0% and 33.3% in the two Post-*Booker* periods, before falling to 21.2% in the *Kimbrough/Gall* period. Judge F’s rate of below-range sentencing has more than doubled, from 15.0% in the Mandatory Guidelines period and 14.7% in the PROTECT Act period, to 32.5% and 32.8% in the two Post-*Booker* periods. So has Judge G’s rate of below-range sentencing, which went from 13.8% in the Mandatory Guidelines period, to 10.5% in the PROTECT Act period, to 34.5%, 33.3%, and 32.4% in the three periods since *Booker*.
264 Sentences by Judge H fit a similar pattern. Judge H’s below-range sentencing rates in the pre-*Booker* periods (16.5% and 23.9%) are very similar to those in the post-*Booker* periods (23.1%, 26.4%, and 17.4%). Sentences by Judge I seemed to fit this pattern during the eighteen months after *Booker*, with a rate of 22.5%, compared with 25.9% in the Mandatory Guidelines period and 21.1% under the PROTECT Act. But Judge I’s rate of below-range sentencing more than doubled to 46.7% in the Post-*Booker* II period and stands at 38.9% in the *Kimbrough/Gall* period.
265 The sentencing pattern of Judge J is unique and highly volatile. From a below-range sentencing rate of 24.6% in the Mandatory Guidelines period, it dropped to 11.9% under the PROTECT Act, more than tripled to 34.0% in the Post-*Booker* I period, dropped again to 18.3% in the Post-*Booker* II period, and has nearly doubled again to 31.8% in the *Kimbrough/Gall* period.
distribution of below-range sentencing rates, for the expanded set of judges, by time period:

![Figure 7: Distribution of Below-Range Sentencing Rates](image)

After a contraction under the PROTECT Act, the distribution in rates of below-range sentences has widened beyond pre-Booker levels and has steadily grown. Since Booker, some judges continue to sentence below the guideline range as low as 10% of the time, while others are sentencing below the guideline range over 40% or even 50% of the time. One judge, who is in senior status and is not a “core” judge because of a light caseload in most periods, sentenced below the guideline range a remarkable 71% of the time during the eighteen months after Booker.

2. Statistical Comparison

A statistical comparison confirms that inter-judge disparity in sentencing relative to the Guidelines has increased since Booker. The difference between the Guidelines sentencing range and the sentence imposed, using zero for within-range sentences, measures how far sentences fall from the guideline range.\(^{266}\) Linear regression models

\(^{266}\) For above-range sentences, the difference was calculated as the difference between the sentence imposed and the guideline maximum. For below-range sentences, the difference was calculated as the difference between the guideline minimum and the sentence imposed. Within-range sentences were coded as zero. The difference should always be positive, and a handful of cases were omitted because of logic problems likely caused because the total sentence reflected consecutive sentences but the judge or the
can describe the relationship between that difference and the identity of the judge. Again, r-squared measures the portion of variance explained by the model, and it has been converted into average months per sentence across all judges. Table 7 reports the results for all sentences and for discretionary sentences:

| Identity of Judge and Difference Between Sentence and Guidelines |
|---------------------|------------------|------------------|------------------|
|                      | R-Squared | Variance Explained | Model Significance |
| All Sentences        |           |                  |                  |
| Mandatory Guidelines | 1.0%      | 1.4 months        | .847*            |
| PROTECT Act          | 2.4%      | 2.7 months        | .482*            |
| Post-Booker I        | 3.6%      | 3.9 months        | .089†             |
| Post-Booker II       | 3.7%      | 4.8 months        | .048†             |
| Kimbrough/Gall       | 6.6%      | 7.1 months        | .073†             |
| Discretionary Sentences |        |                  |                  |
| Mandatory Guidelines | 1.3%      | 1.8 months        | .853*            |
| PROTECT Act          | 3.6%      | 3.7 months        | .384*            |
| Post-Booker I        | 4.5%      | 4.9 months        | .105*             |
| Post-Booker II       | 5.1%      | 6.2 months        | .038              |
| Kimbrough/Gall       | 9.4%      | 9.1 months        | .037              |

* Not significant at the .05 level
† Significant only at the .10 level

Table 7: Linear Regression Models

The models show that, before Booker, the identity of the sentencing judge bore a very small and nonsignificant relationship to the distance between the sentence imposed and the guideline range. Since Booker, however, the identity of the judge has become a statistically significant predictor of how far a sentence will fall from the Guidelines. And the relationship has grown steadily stronger, explaining 3.6% of variance for all sentences (50% above pre-Booker levels) during the first eighteen months after Booker and 6.6% (more than 150% above pre-Booker levels) since Kimbrough and Gall. As expected, the trend is even more pronounced for discretionary sentences, with the identity of the judge explaining 9.4% of the variance (more than 175% above pre-Booker levels) since Kimbrough and Gall.

Together, the trends in patterns of sentencing relative to the Guidelines indicate that, since Booker, the identity of the judge has become a stronger predictor of whether a defendant will receive a within-range sentence and of how far outside the guideline range the sentence will fall.

Commission recorded the guideline minimum and maximum for only one offense. The Commission codes a sentence life imprisonment as 470 months. See U.S. SENTENCING COMMISSION, VARIABLE CODEBOOK FOR INDIVIDUAL OFFENDERS 64 (2009). For consistency, guideline minimum or maximum sentences of life imprisonment were recoded as 470 months as well.

Hofer et al., supra note 40, at 287. Actual months of variance explained were determined by “(1) multiplying the total variance by the portion of the variance accounted for by judges, and (2) finding the square root of the result, thus translating the numbers back into absolute terms.” Id. at 287 n.187.
IV. IMPLICATIONS

Both measures of sentencing outcomes—sentence length and sentencing relative to the Guidelines—indicate a modest but clear increase in inter-judge disparity since Booker. The identity of the sentencing judge has become a stronger predictor of sentence length since the decision, by some measures 60-80% above pre-Booker levels. Although the trend has been uneven, the relationship has reached its strongest levels after Kimbrough and Gall. Those results suggest that, in Boston, judges’ overall tendency toward lenience or severity has a greater effect on sentencing outcomes since Booker.

At the same time, the identity of the sentencing judge has become a stronger predictor of sentencing relative to the Guidelines. Some judges now sentence outside the guideline range much more frequently, and to a greater extent, than their colleagues. Since Booker, some Boston judges continue to sentence below the guideline range as low as 10% of the time, while other judges’ rate below-range sentencing has tripled or quadrupled, exceeding 40% or even 50%. In addition, the strength of the relationship between the identity of the judge and how far sentences fall outside the guideline range has grown steadily, climbing 150-175% above pre-Booker levels. These changes suggest that Booker has exacerbated a distinct form of inter-judge disparity, driven by disagreements about how well the Guidelines perform in recommending sentences.

Both changes are unwelcome news. If Booker has produced similar trends in other districts, it would mark a step backward from Congress’s goal of reducing inter-judge disparity. Although it is still too early, and the changes too modest, to despair over a return to “pre-guideline chaos,” the preliminary evidence—from the only district court in which this sort of study is possible—is discouraging.

But inter-judge disparity is only one factor relevant in evaluating a sentencing system. The Boston results therefore prompt several additional questions. Why has inter-judge disparity increased since Booker? And, in light of Booker’s potential to address longstanding criticisms of the Guidelines—harshness, inflexibility, and prosecutorial power—does the change really matter?

A. Explaining the Increase in Inter-Judge Disparity

The Boston data alone cannot provide any definitive explanation of why inter-judge disparity has increased since Booker. They can, however, assist in evaluating whether the same reasons offered for the modest aggregate effects of Booker might explain the significant effects of Booker for individual judges. As discussed above, the most frequently cited explanations are (1) force of habit, for a generation of judges that has always treated the Guidelines as mandatory; (2) fear of reversal, in the face of “reasonableness” review by courts of appeals; (3) “anchoring” and other cognitive errors; and (4) tactical concerns, for judges anxious to avoid provoking Congress.

The most common explanation for judges’ fairly muted reaction to Booker is that most sitting judges were appointed after 1987, and have spent their entire careers imposing sentences under the Guidelines framework. It should come as no surprise, the argument goes, that a generation of judges that has “grown up” with the Guidelines might cling to them now that they have become advisory.

---

268 Grassley Press Release, supra note 3.
269 See supra notes 138-150 and accompanying text.
270 See supra notes 138-140 and accompanying text.
The Boston dataset, however, provides no support for that hypothesis. Of the ten core judges in Boston from 2002-2008, five joined the court before 1987 and gained pre-Guidelines sentencing experience, while the other five joined the court after 1987 and have no comparable experience. As shown in Figures 8a and 8b, the difference between those groups is negligible:

Since *Booker*, judges appointed before 1987 have imposed sentences at levels only slightly lower than their junior colleagues, never more than 4.7 months on average. And it was judges appointed after 1987, not judges appointed before the original Guidelines, whose sentences changed most noticeably after *Booker*, jumping from 42.2 months to 60.0 months. 271 Although pre-1987 judges had a higher rate of below-range sentencing than post-1987 judges in the eighteen months after *Booker* (33.8% compared with 21.4%), that difference has disappeared in subsequent periods, with both groups sentencing below the guideline range around 30% of the time. 272 In Boston, judges who began their careers before 1987 have continued, in the wake of *Booker*, to sentence in essentially the same way as their colleagues.

Another explanation offered for the generally modest effects of *Booker* is that district court judges are anxious to avoid reversal on appeal. *Booker* directed appellate courts to review sentences for “reasonableness,” and it has taken several years for the

---

271 Pre-1987 service by the sentencing judge was a statistically significant predictor of sentence length only during the PROTECT Act period \((p = .020)\), and has little explanatory power \((r\text{-}squared = 1.4\%)\).

272 Pre-1987 service by the sentencing judge was a statistically significant predictor of a within-range sentence only during the Post-*Booker* I period \((p = .003)\), and was not a statistically significant predictor of how far a sentence falls from the guideline range during any period.
Supreme Court and courts of appeals to work out the details of that standard. In theory, concerns about appellate review could help to explain greater inter-judge disparity since *Booker* because some judges may be more risk-averse than others.

The Commission does not link sentence and appellate records, making it impossible to determine the appeal rate or reversal rate of individual judges using the Boston dataset. Still, the timing of the trends in sentence length among Boston judges lends some support to that theory. Since *Kimbrough* and *Gall*, high-profile cases in which the Supreme Court affirmed sentences as reasonable under *Booker*, inter-judge disparity in sentence length has increased sharply. That suggests, unsurprisingly, that Boston judges are sensitive to the risk of reversal on appeal.

The Boston data shed no light on the strength of other explanations, like the “anchoring” effect of the Guidelines or tactical concerns about provoking Congress. In concept, those factors may affect different judges in different ways, but the effect would be difficult to measure, even if judges were inclined to disclose their motives and cognitive biases.

But the trends identified in this Article, together with previous research on the federal judiciary, also suggest a more basic explanation for the increase in inter-judge disparity since *Booker*. Some judges might actually agree with the Guidelines to a greater extent than their colleagues. During the first eighteen months after *Booker*, when the scope of judges’ newfound sentencing discretion was uncertain, inter-judge disparity in sentence length increased in Boston. In the following eighteen months, after a string of decisions from the First Circuit imposing sharp limits on that discretion, the effect faded. But most recently, in the wake of *Kimbrough* and *Gall*, which repudiated the First Circuit’s approach and stressed the breadth of district courts’ discretion under *Booker*, it has roared back, reaching its highest levels since at least 2001. The fact that changes in inter-judge disparity have mirrored changes in the degree of freedom enjoyed by sentencing courts suggests that some judges do not wish to alter their pre-*Booker* sentencing practices as much as their colleagues.

That suggestion should not come as a surprise. Although the federal judiciary includes many vocal critics of the Guidelines, surveys by the Commission have demonstrated that a significant contingent of district court judges generally supports the Guidelines. In a 2003 survey by the Commission, judges were asked to rate various

273 See *supra* notes 141-144 and accompanying text.
274 Id.
275 See Gertner, *supra* note 136, at 138 (concluding, in the period before *Kimbrough* and *Gall*, that appellate courts were policing sentences on appeal and that “[d]istrict judges have gotten the message”);
276 See *supra* notes 145-149 and accompanying text.
277 See *supra* Tables 5a, 5b and accompanying text (reporting changes for sentences not governed by a mandatory minimum and for sentences not subject to a government-sponsored sentence reduction).
278 See, e.g., United States v. Pho, 433 F.3d 53, 61 (1st Cir. 2006) (emphasizing that “this newfound discretion, though broad, is not limitless” and holding that judges may not impose sentences based on “general disagreement” with policies adopted by the Commission), abrogated by *Kimbrough* v. United States, 552 U.S. 85 (2007); United States v. Smith, 445 F.3d 1, 7 (1st Cir. 2006); United States v. Zapete-Garcia, 447 F.3d 57, 60-61 (1st Cir. 2006); United States v. Thurston, 456 F.3d 211, 218-20 (1st Cir. 2006).
279 See *supra* Tables 5a, 5b and accompanying text.
280 *Kimbrough*, 128 S. Ct. at 566 n.4 (identifying *Pho* as one of the decisions adopting the position ultimately rejected by the Court).
aspects of the Guidelines, on a scale from 1 to 6.\textsuperscript{281} Asked how well the Guidelines achieve the purposes of sentencing in § 3553(a), 38.4\% of district court judges gave them a 5 or 6, compared with 22.9\% who gave them a 1 or 2.\textsuperscript{282} Asked whether the Guidelines offer sufficient flexibility at sentencing, 24.4\% of district court judges gave them a 5 or 6, compared with 45.0\% who gave them a 1 or 2.\textsuperscript{283} Similarly, in a 1996 survey by the Federal Judicial Center, more than 25\% of district court judges responded that mandatory sentencing guidelines are necessary.\textsuperscript{284} These judges may be in the minority, but they form a large and stable portion of the federal bench, and they were not clamoring for a \textit{Booker} revolution.

The notion that some judges actually agree with the Guidelines seems more plausible than the (rather uncharitable) suggestion that guideline sentencing is mostly the result of inertia, anxiety, cognitive error, and intellectual laziness. Judges might choose to accord strong respect to the Guidelines based on modesty about their own role, reluctance to hazard predictions about particular defendants, deference to the Commission as an expert body, commitment to the project of inter-judge sentencing uniformity, or belief that the Guidelines derive democratic legitimacy from their approval by Congress. Although those reasons have been vigorously challenged by opponents of the Guidelines, they supply an ample basis for the kind of good-faith disagreement that can fuel inter-judge sentencing disparity.

\textbf{B. Booker and Criticisms of the Guidelines}

Although the reduction of unwarranted disparity was Congress’s central goal in the Sentencing Reform Act, it is by no means the only factor relevant in evaluating a sentencing system. It is possible to conclude that, even if inter-judge disparity has increased since \textit{Booker}, the decision on balance has improved federal sentencing because the benefits of addressing the principal criticisms of the Guidelines—harshness, inflexibility, and prosecutorial power—outweigh the costs to Congress’s goal of inter-judge uniformity.\textsuperscript{285} The Boston data permit an assessment of how \textit{Booker} implicates each of those concerns.

To the surprise of many commentators, \textit{Booker} has coincided with an increase in sentence severity nationwide, even for drug offenses.\textsuperscript{286} In the Boston dataset, average sentence length rose from 49.9 months under the PROTECT Act to 57.5 months in the eighteen months after \textit{Booker}, and has climbed to 63.9 months since \textit{Kimbrough} and \textit{Gall}.\textsuperscript{287} Frank Bowman has proposed that, even though sentences have risen nationwide,

\textsuperscript{282} Id. at 24.
\textsuperscript{283} Id. at 15.
\textsuperscript{286} See supra Figure 1 (nationwide average sentences).
\textsuperscript{287} See supra Table 3.
Booker may have blunted an even stronger upward trend that otherwise would have continued. That scenario is plausible, although necessarily speculative.

The Boston data suggest another scenario: prosecutors may be fighting back against Booker—and winning—through charging and bargaining practices. As Figure 9 shows, the increase in sentence length has closely tracked concurrent increases in statutory minimum and guideline minimum sentences:

![Figure 9: Guideline and Statutory Minimum Sentences](image)

The average guideline minimum stood at 59.0 months before Booker, but jumped to 71.8 months in the eighteen months after the decision, and jumped again to 80.7 months after Kimbrough and Gall. Similarly, the average statutory minimum climbed from 15.5 months under the PROTECT Act to 22.8 months after Booker and again to 29.7 months after Kimbrough and Gall. The increase in statutory minimums reflects an increase in the portion of cases governed by a mandatory minimum, from 13.7% and 17.0% in the pre-Booker periods, to 23.7% and 25.4% in the post-Booker periods, and 31.3% since Kimbrough and Gall.

It is unlikely that the increase in statutory and guideline minimum sentences resulted solely from changes by Congress or the Commission. Although the PROTECT Act added a few new mandatory minimums and directed the Commission to increase certain base offense levels and upward adjustments, those provisions were focused on sex

---

288 Bowman, supra note 5, at 283-84.
289 For sentences in which the court imposed a sentence below the statutory minimum based on the “safety valve” provision or a government-sponsored departure, the statutory minimum was treated as zero. See supra note 256.
offenses and child pornography, which (along with prostitution) account for just 2.3% of cases in the Boston dataset. The Commission’s amendments to the Guidelines since Booker have increased offense levels and adjustments for a host of specific offenses, including identity theft, terrorism offenses, distributing anabolic steroids, and possessing semiautomatic weapons. But each of those changes affects only a small number of cases, likely insufficient to cause the broad increase in guideline minimums since Booker.

Instead, the changes in statutory and guideline minimum sentences could be driven by prosecutors. By selecting which charges to file against a defendant, prosecutors have control over whether a mandatory minimum sentence will apply, and in some instances how high the mandatory minimum will be. And by negotiating the details of the recitation of facts in a plea agreement, prosecutors often can influence the guideline range. Neither “real offense” sentencing under the Guidelines nor the 2003 Ashcroft Memorandum—which directs prosecutors to “charge and pursue the most serious, readily provable offense or offenses that are supported by the facts of the case”—have neutralized those prosecutorial advantages. It is possible that Booker led prosecutors to alter their bargaining behavior, using mandatory minimums and guideline sentencing factors more aggressively to counteract any reduction in sentence length by judges. If that proves to be the case nationwide, then the decision will have done little to curb prosecutorial power, one of the major criticisms of the Guidelines.

Whatever Booker’s effects on sentence severity and prosecutorial leverage, the decision unquestionably has addressed the criticism that the Guidelines were too inflexible. Since 2005, the Supreme Court has rendered the Guidelines advisory (Booker), created a highly deferential standard of appellate review (Gall), and apparently authorized judges to categorically reject the Commission’s judgments about sentencing policy (Kimbrough). Judges in Boston have taken advantage of the expansion of their discretion. Every judge in the dataset has imposed below-range sentences more frequently since Booker, and some judges have done so three or four times more often. By reducing the rigidity of the Guidelines, Booker alleviated a major source of frustration for judges.

At the same time, however, the increase in inter-judge disparity in Guidelines sentencing patterns tends to undermine an important assumption of the inflexibility critique. Commentators who fault the Guidelines for their rigidity do not seek flexibility for its own sake, but instead assume that, with greater latitude at sentencing, judges will recognize when application of the Guidelines is inconsistent with the purposes of punishment and make appropriate case-specific corrections. Yet in the Boston dataset, some judges after Booker saw a need for a below-range sentence at triple or quadruple

---

290 PROTECT Act § 401(d), 117 Stat. at 670.
293 See Stith, supra note 72, at 1448; Richman, supra note 78, at 1403-06; supra notes 76-82 and accompanying text.
294 See supra notes 62-71 and accompanying text.
295 See supra Figure 7, notes 262-263 and accompanying text.
296 See supra notes 70-71 and accompanying text.
the rate of their colleagues, despite drawing a random set of cases from an identical case pool. That suggests that judges simply disagree about how often the Guidelines recommend a just sentence—routinely, not just in specific cases—and that the greater flexibility of *Booker* has allowed those disagreements to drive sentencing outcomes more frequently.

Given *Booker*’s mixed scorecard in addressing criticisms of the Guidelines, an increase in inter-judge disparity is an important data point. To be sure, sentencing trends among judges Boston may not be playing out in the same way in all districts nationwide. And the magnitude of the “judge effect”—even assuming it has increased by 60-80% for sentence length or 150-175% for guideline sentencing—remains relatively small, accounting for under 10% of the variation in sentencing outcomes. Yet those figures are slightly larger than the results reported in the Hofer and Waldfogel studies. Moreover, as Anderson, Kling and Stith observed, the small fraction of variance explained by the identity of the sentencing judge “tells us that there are many additional factors that drive differences in sentences, but it does not lead us to conclude that interjudge disparity itself is small or unimportant.”

For obvious reasons, Congress and the Sentencing Commission would find it troubling if inter-judge disparity has increased nationwide. The Commission believes that the Guidelines recommend an appropriate sentence in all but “rare” cases, and Congress’s efforts to curtail downward departures in the PROTECT Act signal strong disapproval of high rates of below-range sentencing. News that some judges were sentencing outside the Guidelines range far more frequently than their colleagues, in 40% or even 50% of cases, might signal to Congress or the Commission that those judges too often reach unjust results. Congress also considers the reduction of inter-judge disparity a worthy goal in itself, and the animating purpose of the Guidelines.

Critics of the Guidelines, too, ought to find the trends in Boston unsettling. The most obvious benefit of *Booker*, greater flexibility for sentencing judges, was immediately obvious and has become even clearer since *Kimbrough* and *Gall*. But the decision’s potential costs—an increase in inter-judge disparity, perhaps with no reduction in severity or prosecutorial power—are less obvious and more difficult to measure. The experience in Boston thus provides valuable information for reevaluating federal sentencing after the *Booker* revolution.

**CONCLUSION**

As the first empirical account of changes in sentencing patterns among individual federal judges since *Booker*, this Article addresses a critical gap in the existing research. But its conclusions are necessarily tentative. Its results may not be representative of sentencing trends nationwide. It also focuses inter-judge sentencing disparity, when a comprehensive assessment of post-*Booker* sentencing must consider other major

---

297 See supra notes 194-198 and accompanying text.
299 Anderson et al., supra note 27, at 294.
300 U.S.S.G. § 5K2.0.
301 See supra notes 84-93 and accompanying text.
criticisms of the Guidelines regime, including sentence severity, inflexibility, and prosecutorial power.

Yet it offers an unprecedented glimpse at how individual judges have responded to Booker, and the results are generally discouraging. Among judges in Boston who share a common case pool, there has been a modest but clear increase in inter-judge disparity since Booker. The identity of the sentencing judge has become a stronger predictor of sentence length, particularly since the Supreme Court’s decisions in Kimbrough and Gall, by some measures rising 60-80% above pre-Booker levels. At the same time, the identity of the sentencing judge has become a stronger predictor of sentencing relative to the Guidelines. Some Boston judges now sentence outside the guideline range much more frequently, and to a greater extent, than their colleagues.

Even if similar trends have played out nationwide, they do not compel any judgment about whether Booker, on balance, has improved or worsened federal sentencing. But after four years of puzzling over the muted response to Booker, the Boston experience suggests a quiet revolution may be underway.