February 23, 2010

Inter-Judge Sentencing Disparity After Booker: A First Look

Ryan W. Scott
INTER-JUDGE SENTENCING DISPARITY AFTER BOOKER: A FIRST LOOK

Ryan W. Scott*

ABSTRACT
A central purpose of the Sentencing Reform Act was to reduce inter-judge sentencing disparity, driven not by legitimate differences between offenders and offense conduct, but by the philosophy, politics, or biases of the sentencing judge. The federal Sentencing Guidelines, despite their well-recognized deficiencies, succeeded in reducing that form of unwarranted disparity. But in a series of decisions from 2005 to 2007, the Supreme Court rendered the Guidelines advisory (Booker), set a highly deferential standard for appellate review (Gall), and explicitly authorized judges to reject the policy judgments of the Sentencing Commission (Kimbrough). Since then, the Commission has received extensive anecdotal reports of a surge in inter-judge disparity at sentencing.

This Article provides the first empirical evidence of inter-judge sentencing disparity since the Supreme Court upended federal sentencing, drawing on an original new dataset of sentences from the District of Massachusetts—the only district court that makes key sentencing documents available to the public. The data indicate a clear increase in inter-judge sentencing disparity, both in sentence length and in guideline sentencing patterns. Since Booker, Kimbrough, and Gall, the effect of the judge on sentence length has more than doubled in strength. In cases not subject to a mandatory minimum, the difference between the court’s more lenient and more severe judges translates into an average of more than two years in prison. The decisions also have altered guideline sentencing patterns. Some “business as usual” judges continue to sentence below the guideline range at essentially the same rate as before Booker, while other “free at last” judges now sentence below the guideline range at triple or quadruple their pre-Booker levels.

In explaining the spike in inter-judge sentencing disparity, the Article casts doubt on the conventional theories that persistent within-guideline sentencing is the product of inertia, fear of reversal, anchoring effects, strategic behavior, or simple laziness. Instead, it proposes that some judges actually agree with the Guidelines or consciously choose to impose within-range sentences for institutional reasons.

This article is available for download at http://ssrn.com/abstract=1446744.

* 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.
# TABLE OF CONTENTS

INTRODUCTION .............................................................................................................. 1  
I. A BRIEF HISTORY OF FEDERAL SENTENCING REFORM .............................................. 3  
  A. Inter-Judge Sentencing Disparity Before Booker .................................................... 3  
     1. The Sentencing Reform Act of 1984 ................................................................ 3  
     3. PROTECT Act (2003) ........................................................................................ 7  
     B. The Booker Revolution, 2005-2007 .................................................................. 8  
        1. Booker, Kimbrough, and Gall ........................................................................ 8  
        2. Average Sentence Length and Guideline Sentencing ....................................... 10  
        3. Inter-Judge Sentencing Disparity .................................................................. 14  
II. THE EMPIRICAL STUDY OF INTER-JUDGE SENTENCING DISPARITY .................. 15  
  A. Data and Methods .................................................................................................. 15  
     1. Judge-Specific Data .......................................................................................... 15  
     2. Natural Experiment Method ............................................................................. 16  
     3. Measures of Inter-Judge Disparity .................................................................... 17  
     4. Why Massachusetts? ......................................................................................... 19  
  B. Results ..................................................................................................................... 21  
     1. Sentence Length ................................................................................................ 22  
     2. Guideline Sentencing Patterns .......................................................................... 25  
III. IMPLICATIONS ......................................................................................................... 31  
  A. Conventional Explanations for Within-Range Sentencing ..................................... 32  
     1. Inertia ................................................................................................................ 32  
     2. Risk-Aversion .................................................................................................... 33  
     3. Anchoring, Strategic Behavior, and Laziness ..................................................... 34  
  B. Alternative Explanations for Within-Range Sentencing ....................................... 35  
     1. Agreement with the Guidelines ......................................................................... 35  
     2. Institutional Considerations .............................................................................. 38  
CONCLUSION ................................................................................................................. 39  
APPENDIX ....................................................................................................................... 41  
  A. Methodological Details .......................................................................................... 41  
     1. Period Selection ............................................................................................... 41  
     2. Case Matching .................................................................................................. 42  
     3. Random Distribution ....................................................................................... 44  
     4. Discretionary Sentences ................................................................................... 45  
  B. Detailed Results ..................................................................................................... 46  
     1. Regression Models ............................................................................................ 46  
     2. Alternative Time Periods ................................................................................... 51
INTRODUCTION

A central purpose of the Sentencing Reform Act of 1984 was to reduce inter-judge sentencing disparity. Congress was concerned that similarly situated defendants were receiving widely divergent sentences based on the philosophy, politics, and biases of the sentencing judge. The federal Sentencing Guidelines, promulgated by the United States Sentencing Commission, were designed to minimize that form of unwarranted disparity by designating a mandatory sentencing range, applicable to all judges, based on the circumstances of the offense and characteristics of the offender.

But in a series of decisions from 2005 to 2007, the Supreme Court upended the federal sentencing regime. In *United States v. Booker*, the Court resolved a constitutional defect in the design of the Guidelines by rendering them “effectively advisory,” leaving judges free to impose any reasonable sentence consistent with the broad purposes of punishment outlined by Congress. Three years later, in *Gall v. United States*, the Court directed appellate courts to review sentencing decisions under a “deferential abuse-of-discretion standard.” And on the same day, in *Kimbrough v. United States*, the Court indicated that district courts are now free to sentence outside the guideline range “based solely on policy considerations, including disagreements with the Guidelines.”

In the wake of those decisions, the Commission has received extensive anecdotal reports of a surge in inter-judge sentencing disparity. Attorney General Eric Holder, in a June 2009 speech on sentencing policy, issued a call for research into 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.” Prosecutors around the country echoed those concerns at the Commission’s 2009-2010 regional hearings. Patrick Fitzgerald, the U.S. Attorney for the Northern District of Illinois, warned that *Booker* has “re-introduced into federal sentencing both substantial district-to-district variations and substantial judge-to-judge variations.” Prosecutors have reported a similar spike in inter-judge disparity in “nearly all districts” in the Ninth Circuit. Frank Bowman calls the Supreme Court’s decisions a

---

2. *Id.* at 245 (Breyer, J., writing for the Court).
4. *Id.* at 52-53.
6. *Id.* at 101 (internal quotation marks omitted).
8. *See infra* notes 107-111 and accompanying text.
“debacle,” and warns that in white-collar cases, “we’re back to a pre-guidelines era” marked by “disparity and the potential for disparity.”

Those reports, if accurate, deserve urgent attention because they implicate Congress’s core objective in reforming federal sentencing. To date, however, the evidence has been strictly anecdotal. This Article addresses a critical gap in the research, offering the first empirical account of inter-judge sentencing disparity since the Supreme Court’s shake-up of federal sentencing. It does so by drawing on an original new dataset of sentences from the District of Massachusetts, the only district that makes key sentencing documents available to the public. Those records allow, for the first time, a study of how individual judges have responded to the federal sentencing revolution.

Analysis of those sentences reveals a clear increase in inter-judge disparity, both in sentence length and in guideline sentencing patterns. Following the Supreme Court’s decisions in *Booker*, *Kimbrough*, and *Gall*, the effect of the judge on sentence length has more than doubled in strength. In cases not governed by a mandatory minimum, the court’s three most lenient judges have imposed average sentences of 25.5 months or less, while its two most severe judges have imposed average sentences of 51.4 months or more. That stark difference translates to an average of more than two years in prison, depending on which of those judges is assigned the case.

Similarly, the Boston data reveal that some judges have taken advantage of their enhanced discretion to depart from the guidelines to a far greater extent than others. Two judges (call them “business as usual” judges) continue to impose below-guideline sentences at essentially the same rate as before *Booker*, as little as 16% of the time. But four other judges (call them “free at last” judges) now sentence below the guideline range at triple or quadruple their pre-*Booker* rates, as much as 53% of the time. In addition, the effect of the judge on how far sentences fall from the guideline range has more than doubled in the wake of *Booker*, *Kimbrough*, and *Gall*.

These results tend to corroborate the anecdotal reports of an increase in inter-judge sentencing disparity. Yet they 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 to consider in evaluating a sentencing system, the results do not compel any judgment about whether the Supreme Court’s decisions, on balance, have improved or worsened federal sentencing. Nonetheless, the Boston data offer an unprecedented look at how individual judges have responded to the Supreme Court’s decisions.

The Article proceeds in three parts. Part I explains the importance of inter-judge sentencing disparity to Congress’s reform efforts and describes the trio of Supreme Court decisions that reshaped federal sentencing between 2005 and 2007. Despite anecdotal reports of a surge in inter-judge disparity, neither the Commission nor other researchers have examined the effects of *Booker*, *Kimbrough*, and *Gall* on the sentencing patterns of individual judges.

---

13 *See infra* note 158 and accompanying text.
14 *See infra* Figure 5 and accompanying text.
15 *See infra* notes 161-165 and accompanying text.
16 *See infra* Table 3 and accompanying text.
Part II of the Article reports the empirical study. Part II.A describes the Article’s unique dataset of sentences linked to individual judges. It also summarizes the Article’s methods, which build on “natural experiment” studies of inter-judge disparity after the promulgation of the Guidelines. Part II.B reports the results of the study. Details of the data and methods, as well as full reports of the regression models, appear in the Appendix.

Part III considers possible explanations for the Article’s key finding of a spike in inter-judge sentencing disparity. It casts doubt on the conventional theories that persistent within-guideline sentencing is the product of inertia, fear of reversal, “anchoring” effects, strategic behavior, or simple laziness. Instead, it proposes two alternative explanations: some judges might actually agree with the Guidelines, or may elect to impose within-range sentences for institutional reasons.

I. A BRIEF HISTORY OF FEDERAL SENTENCING REFORM

Before describing the nuts and bolts of the empirical study, a brief history of federal sentencing reform is needed, both to demonstrate the importance of inter-judge disparity to sentencing reform, and to describe the Supreme Court decisions that radically altered federal sentencing law from 2005 to 2007.

A. Inter-Judge Sentencing Disparity Before Booker

1. The Sentencing Reform Act of 1984

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.17 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.18 Judges were under no obligation to give reasons for the sentence imposed,19 and appellate review of sentencing decisions was virtually nonexistent.20 The theory was that judges should “individualize” sentences to serve the rehabilitative needs of criminal defendants, “almost like a doctor or social worker exercising clinical judgment.”21

In practice, however, indeterminate sentencing gave judges so much discretion that criminal defendants faced starkly different levels of punishment depending on which judge happened to draw the case. Judge Marvin Frankel, the most influential critic of

---

18 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).
20 STITH & CABRANES, supra note 17, at 9 & n.3.
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.”22 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.”23 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.24 Following more than a decade of debate, Congress enacted the Sentencing Reform Act of 1984.25

A principal purpose of the Act was to reduce inter-judge disparity in sentencing.26 Congress concluded that, too often, similarly situated offenders received unjustifiably disparate sentences, solely because of the preferences and biases of the sentencing judge.27 To be sure, different constituencies in Congress emphasized different aspects of the problem. Democrats expressed concern that indeterminate sentencing allowed race discrimination to flourish, while “tough on crime” Republicans frequently worried that too many judges were unduly lenient.28 But there was remarkable bipartisan agreement

---

23 Id. at 9.
27 U.S.S.G. § 1A.1, intro to comment., pt. A, ¶ 2 (Congress “sought uniformity in sentencing by narrowing the wide disparity in sentences imposed by different federal courts for similar criminal conduct”).
28 STITH & CABRANES, supra note 17, at 38-48.
that unfettered discretion had resulted in an intolerable level of inter-judge sentencing disparity. 29

To reduce inter-judge disparity, the Act created the United States Sentencing Commission, “an independent commission in the judicial branch of the United States.” 30 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. 31 It provided that guidelines and amendments adopted by the Commission must be submitted to Congress for a period of review, but unless “modified or disapproved” by Congress would go into effect automatically. 32 Judges were bound to follow the Guidelines except in two circumstances (known as “departures”): (1) on the government’s motion, based on a defendant’s substantial assistance to authorities; 33 and (2) in “rare” cases 34 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.” 35 The Act compelled judges to state the reasons for each sentence in open court, and to issue a written statement of reasons in any case where the sentence fell outside the guideline range. 36 It also provided for appellate review of sentencing range calculations and for review of sentences outside the guideline range for abuse of discretion. 37


The Commission promulgated the first federal Sentencing Guidelines in 1987, and the mandatory-Guidelines regime remained essentially intact for 18 years. During that time, the Guidelines provoked strident opposition, particularly among scholars, the defense bar, and district court judges. A chorus of critics assailed the Guidelines for their severity, 38 for their inflexibility, 39 and for transferring too much power to prosecutors making charging and plea bargaining decisions. 40

29 The Act was co-sponsored by strange bedfellows in the Senate, Ted Kennedy and Strom Thurmond.
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.
36 18 U.S.C. § 3553(c)(1)-(2).
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.
Among the Guidelines’ many failures, however, reducing inter-judge disparity was a bright spot. In the late 1990s, several studies provided strong evidence that the Guidelines had reduced inter-judge sentencing disparity, 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 a random case-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 disparity in average sentences to the judge, and reductions in the rate of disparity to the Guidelines.

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. 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%.

The second, by Paul Hofer of the Sentencing Commission and two colleagues (“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.

41 See 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); Anderson et al., supra note 26, at 303.
42 Anderson et al., supra note 26, at 291; Hofer et al., supra note 41, at 282.
43 Anderson et al., supra note 26, at 290 (Table 2).
44 Id. at 303.
45 Hofer et al., supra note 41, at 284.
46 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).
47 Id. at 203-04.
48 Id. at 298.
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. A second 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. A third 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.


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”). Championed by Representative Tom Feeney and dubbed the “Feeney Amendments,” the provisions responded to concerns in Congress and the Department of Justice about the prevalence of downward departures from the Guidelines. 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%. (The Commission later realized that the 2001 rate was incorrect.)

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; (2) directed the Commission to amend the Guidelines “to ensure that the incidence of downward departures are [sic] substantially reduced”; (3) prohibited the Commission from recognizing new permissible grounds for downward departure for two

---


50 See Hofer et al., supra note 41, at 299-302; Anderson et al., supra note 26, at 302.

51 See Anderson et al., supra note 26, at 299.

52 See infra notes 137-138 and accompanying text.


55 Stith, supra note 40, 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 39, at 1228.

56 See infra notes 65-66 and accompanying text.

57 PROTECT Act § 401(d), 117 Stat. at 670.

58 Id. § 401(m)(2)(A).
years; and (4) directed the Department of Justice to resist downward departures “not supported by the facts and the law.” The PROTECT Act also directly amended the Guidelines by adding specific upward adjustments for sex offenders and child pornography cases.

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

In hindsight, it is clear that reports of an epidemic of judge-initiated downward departures were exaggerated. 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.

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.


1. Booker, Kimbrough, and Gall

In January 2005, the Supreme Court held in United States v. Booker that the Sentencing Reform Act violated the Sixth Amendment right to trial by jury. The Court’s fractured decision consisted of two majority opinions. One opinion, written by

---

59 Id. § 401(j)(2).
60 Id. § 401(l)(1).
61 Id. § 401(i).
68 Id. at 226-27, 243-44.
69 The case prompted six separate opinions, including two principal majorities and two principal dissents. Id. at 225.
Justice Stevens, extended the rule of *Apprendi v. New Jersey*\(^\text{70}\) and *Blakely v. Washington*\(^\text{71}\) to the federal Sentencing Guidelines. The Court held that, because the Guidelines permitted judges to find facts that trigger a sentence above the otherwise-applicable guideline maximum, they intruded upon the province of the jury.\(^\text{72}\)

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.\(^\text{73}\) Excising those provisions, the Court explained, “makes the Guidelines effectively advisory.”\(^\text{74}\) Judges must continue to calculate the applicable sentencing range, the Court explained, but need only “consider” it, along with the factors identified in Section 3553(a), in imposing a sentence.\(^\text{75}\)

Two subsequent decisions, issued on the same day in December 2007, clarified the role of appellate courts reviewing sentences for “reasonableness” and left no doubt that *Booker* had dramatically expanded the discretion of district courts at sentencing.\(^\text{76}\) In *Gall v. United States*,\(^\text{77}\) the Court held that courts of appeals 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.\(^\text{78}\) Instead, appellate courts must apply a “deferential abuse-of-discretion standard,” according due respect to “the district court’s decision that the § 3553(a) factors, on a whole, justify the extent of the variance [from the Guidelines].”\(^\text{79}\)

Simultaneously in *Kimbrough v. United States*,\(^\text{80}\) the Court held that judges, in applying the now-advisory Guidelines, are free to reject the Guidelines’ 100-to-1 ratio that treats one gram of crack cocaine as equivalent to 100 grams of powder cocaine.\(^\text{81}\) In reaching that conclusion, the Court relied upon—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.”\(^\text{82}\) Although it suggested that “closer review may be in order” in those

\(^{70}\) 530 U.S. 466 (2000).
\(^{71}\) 542 U.S. 296 (2004).
\(^{72}\) *Booker*, 543 U.S. at 244 (Stevens, J., writing for the Court).
\(^{73}\) *Booker*, 543 U.S. at 245 (Breyer, J., writing for the Court).
\(^{74}\) Id.
\(^{75}\) *Id.* at 245, 259-60.
\(^{77}\) 552 U.S. 38 (2007).
\(^{78}\) *Id.* at 47-48.
\(^{79}\) *Id.* at 52.
\(^{80}\) 552 U.S. 85 (2007).
\(^{81}\) *Id.* at 109-10.
\(^{82}\) *Id.* at 101 (internal quotation marks and alteration omitted).
circumstances, the Court left little doubt that judges now enjoy the freedom to categorically reject the Commission’s judgments about sentencing policy.

2. Average Sentence Length and Guideline Sentencing

Longtime critics of the Guidelines greeted Booker with enthusiasm, but the decision did not prompt immediate changes in sentencing outcomes. Average sentence length actually increased for several years after Booker, even for drug trafficking offenses. The rate of below-guideline sentencing jumped, but quickly leveled out, and the change was hardly “earth-shattering.” Many commentators lamented that, far from ushering in a revolution, the decision turned out to be a dud.

It would be premature to pronounce the Supreme Court’s sentencing decisions a non-event. Recent data from the Commission suggest that Kimbrough and Gall have, after a long delay, prompted meaningful changes in sentencing outcomes. As shown in Figure 1, average sentence length has reversed course, decreasing after Kimbrough and Gall:

---

83 Id. at 109.
84 See United States v. Herrera-Zuniga, 571 F.3d 568, 584-85 (6th Cir. 2009) (interpreting Kimbrough as recognizing “the broad authority of sentencing judges” to “categorically reject the sentencing range prescribed by the Guidelines”).
Average sentence length nationwide rose from 50.1 months in fiscal year 2004, immediately before *Booker*, to 51.8 months in fiscal years 2006 and 2007. But since *Kimbrough* and *Gall*, average sentence length has fallen to 47.2 months, the lowest level since 2002. Similarly, sentences for drug trafficking offenses rose from 81.3 months in fiscal year 2004, before *Booker*, to 83.2 months after *Booker* in fiscal year 2007. Drug trafficking sentences declined after *Kimbrough* and *Gall*, however, decreasing to 77.9 months in FY 2009. They are now less severe than before *Booker*.

Another measure is sentencing relative to the guideline range. Figure 2 shows the rate of above-range and below-range sentencing among all judges nationwide from fiscal years 2003-2009:

---

*Figure 1: Average Sentences Nationwide, by Fiscal Year*[^89]

The rate of below-range sentencing more than doubled after Booker in fiscal year 2005, from 5.5% to 13.0%, but retreated to 12.0% by fiscal year 2007, not much higher than the 8.6% rate under the mandatory Guidelines in 2002-2003. After Kimbrough and Gall, however, the increase in below-range sentencing has resumed, reaching 15.9% in fiscal year 2009. In the last quarter of 2009, the rate of below-range sentencing hit 16.1%. By way of comparison, the rate of below-range sentencing is creeping closer to the (incorrectly reported) 18.1% rate that prompted Congress to intervene in the PROTECT Act. The percentage of above-range sentences also has more than doubled, from 0.8% before Booker to 1.8% after Kimbrough and Gall.

Still, the changes in sentencing outcomes since Booker have fallen far short of the fundamental change many scholars expected. Average sentence length stands at approximately 2000-2003 levels, while drug trafficking sentences remain substantially above 2002-2003 levels. Within-range and government-sponsored sentences continue to account for more than 80% of sentences in the federal system.

Why has the response to Booker been relatively modest? The conventional wisdom, reflecting impatience with the pace of change, has focused on several explanations: inertia, risk-aversion, anchoring, strategic behavior, and laziness.

The most common conventional explanation for the slow response to Booker 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

---


92 See supra note 55 and accompanying text.
in 1987 and the *Booker* decision in 2005. It should not be surprising, the argument goes, that 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.

A second explanation is risk-aversion, among judges worried about 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 proposed 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 rational 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 if the Guidelines continue to exert a powerful influence despite being advisory.

A fourth is strategic behavior. Judges who celebrated *Booker*, eager to safeguard the sentencing discretion they had gained, took a “go slow” approach to reduce the risk of

---

93 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.

94 See, e.g., Judge Nancy Gertner, *Supporting Advisory Guidelines*, 3 HARV. L. POL’Y REV. 261, 270 (2009) (describing continued guideline sentencing as the result of “the habits ingrained during twenty years of mandatory Guideline sentencing,” and noting that “after the SRA, judges were trained only in the Guidelines”); Stith, supra note 40, 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).

95 United States v. Booker, 543 U.S. 220, 245 (Breyer, J., writing for the Court).


97 127 S. Ct. at 2462-65.


102 Stith, supra note 40, at 1496; Gertner, supra note 98, at 138.
interbranch retaliation. On this theory, judges secretly desire to flout the Guidelines more often, but have restrained themselves to avoid provoking Congress.

A final explanation is laziness. Some commentators have suggested, rather uncharitably, that judges find it easier to impose within-range sentences because it requires “less time in thought and less stress.” As one judge put it, a judge “who wants to be a lazy judge, will be able to do it very easily” by staying within the Guidelines.

3. Inter-Judge Sentencing Disparity

As commentators have puzzled over the fairly modest changes in sentencing outcomes, anecdotal reports from around the country have warned of a surge in inter-judge sentencing disparity in the wake of Booker, Kimbrough, and Gall. At its regional hearings in 2009-2010, the Commission heard extensive testimony from prosecutors that sentencing outcomes increasingly depend on which judge is assigned to the case. The U.S. Attorney for the Northern District of Illinois, Patrick Fitzgerald, told the Commission that Booker has “re-introduced into federal sentencing both substantial district-to-district variations and substantial judge-to-judge variations.” A survey of prosecutors in the Ninth Circuit revealed that “nearly all districts emphasize the wide variation seen between different judges within their districts.” In the Eastern District of New York, it appears “the range of variation between judges in [the same] courthouse has grown” since Booker. In Oregon, “sentencing tendencies have always been somewhat unique to each individual judge, but the differences since Booker have become more pronounced.”

Those claims are exceedingly difficult to evaluate because changes in inter-judge sentencing disparity are almost impossible to detect. Consistent with its longstanding

---

104 Weinstein, supra note 105, 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.
108 Karin J. Immergut, Testimony before the U.S. Sentencing Commission Regional Hearing in Palo Alto, California, at 12 (May 27, 2009), available at http://www.ussc.gov/AGENDAS/20090527/Immergut_testimony.pdf; see id. at 2 (warning that “the signs point to increasing sentencing disparity – including disparity based on differing judicial philosophies among judges working in the same courthouse”).
110 Immergut, supra note 10, at 6 (reporting that some judges “continue to follow the advisory guideline sentence in the majority of cases” while “other judges routinely decline to impose a guideline sentence”).
policy, the Commission has reported only *aggregate* data on post-*Booker* sentencing trends. Neither the Commission nor any independent researcher has examined how *individual judges* have responded to *Booker, Kimbrough,* and *Gall.*

That is a critical omission, and it has not gone unnoticed. Because a central goal of the Sentencing Reform Act was the reduction of *inter-judge* sentencing disparity, judge-specific data are needed to determine the extent to which *Booker* has advanced or undermined Congress’s objectives. Attorney General Eric Holder, in a June 2009 speech marking the 25th anniversary of the Sentencing Reform Act, called 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.” Existing research by the Commission does not permit such an assessment, leaving an important gap in our understanding of federal sentencing patterns.

**II. THE EMPIRICAL STUDY OF INTER-JUDGE SENTENCING DISPARITY**

This Article provides the first hard evidence of inter-judge sentencing disparity after the Supreme Court’s decisions in *Booker, Kimbrough,* and *Gall.* It overcomes the primary challenge in studying federal sentencing patterns—the lack of data that include the identity of the sentencing judge—by drawing on a unique new dataset of more than 2,200 cases from the District of Massachusetts, the lone federal district court that publicizes critical sentencing documents. Those records afford a rare opportunity to test how *Booker* has affected inter-judge sentencing disparity, both in sentence length and in guideline sentencing patterns.

**A. Data and Methods**

1. 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, 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 does not provide a code or number that would permit an analysis of sentencing patterns among judges anonymously. With the exception of studies by the Commission and its staff, the Anderson-Kling-Stith study marks the only

---

112 See supra notes 26-28 and accompanying text.
113 Holder, supra note 7.
115 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)).
116 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.
time in over 25 years that scholars have received permission to study case records that identify the sentencing judge.\textsuperscript{117}

This study overcomes that obstacle by drawing on a unique new dataset of more than 2,200 sentences from the District of Massachusetts. The data were gathered using a method, pioneered by Max Schanzbach and Emerson Tiller,\textsuperscript{118} that matches publicly available docket information with corresponding information in the Commission’s case records. Changes in the Commission’s data-disclosure practices in 2004 make the case-matching method far less effective for cases decided after Booker.\textsuperscript{119} But sentencing documents disclosed by the District of Massachusetts—and no other federal court—make it possible to generate a rich dataset of post-Booker sentences from that district.

By special vote of the court in 2001, the District of Massachusetts makes a case document called the “Statement of Reasons” public, and available online, for every criminal sentence, unless the presiding judge orders it sealed.\textsuperscript{120} The Statement of Reasons, which must be completed and submitted to the Commission for every sentence, reports a host of details about the sentence, including the offender’s offense level, criminal history category, guideline range, any statutory minimum sentence, and the basis for any departure.\textsuperscript{121} Those additional data points greatly improve the efficiency and reliability of the case-matching process. The district’s extraordinary policy, which apparently defies a contrary policy statement by the Judicial Conference,\textsuperscript{122} 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.”\textsuperscript{123}

Aided by the information in Statements of Reasons, the case-matching method proved highly effective. Based on docket information for cases in the district’s Boston division, I generated a dataset of 2,659 sentences imposed between October 1, 2001 and September 30, 2008.\textsuperscript{124}

2. Natural Experiment Method

Building on previous studies of inter-judge sentencing disparity, this study employs a “natural experiment” method. Because all federal judges are equally bound by the Supreme Court’s decisions in Booker, Kimbrough, and Gall, there is no control group

\footnotesize
\textsuperscript{117} See Anderson et al., supra note 26, at 287.
\textsuperscript{118} Schanzenbach & Tiller, supra note 87, at 729-30.
\textsuperscript{119} Specifically, the Commission no longer reports the date of sentencing, but instead reports only the month and year, greatly increasing the chance that multiple cases in the Commission’s data will match publicly available docket information for a given case. See infra note 227 and accompanying text.
\textsuperscript{121} The documents are also a gold mine of qualitative data. 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.
\textsuperscript{124} Details of the case-matching technique are set forth in the Appendix, infra notes 220-229 and accompanying text.
of judges unaffected by recent changes in sentencing law. Researchers can capture changes in inter-judge disparity over time, however, by taking before-and-after measurements from a group of judges who share a common case pool and a random case-assignment system.\textsuperscript{125} Assuming each judge hears a sufficient number of cases, and the distribution of cases is truly random, then \textit{average} sentencing outcomes for each judge should be the same. Inter-judge variation in average outcomes is properly attributed to the judge, rather than case-specific considerations, because the average reflects a random cross-section of the common case pool.

Accordingly, two types of sentences were excluded from the initial set. First, to ensure a sufficient number of cases per judge to draw reliable conclusions from average sentencing outcomes, judges who did not satisfy minimum caseload requirements were excluded.\textsuperscript{126} Second, to ensure that sentencing outcomes were the product of random distribution, the dataset was narrowed to judges sitting in Boston who drew their cases from the shared Boston case wheel. The court’s rules provide for distribution of cases “by lot” within the division that includes Boston,\textsuperscript{127} and statistical tests indicate that cases were indeed distributed randomly.\textsuperscript{128}

The result is a large dataset of 2,262 sentences imposed by 10 judges, all in active status, who served side-by-side in Boston continuously from 2001 to 2008.\textsuperscript{129} Judges included in the study had between 175 and 264 sentences during that period, an average 226 sentences per judge. The dataset is not a sample of sentences during that time, but accounts for more than 90\% of sentences matching the selection criteria.\textsuperscript{130}

An important assumption of the natural-experiment method is that changes in sentencing outcomes are exogenous, caused by developments in sentencing law rather than on-the-ground factors in Boston. An analysis of the mixture of cases in the Boston pool does not suggest any meaningful change in the type of offenders sentenced during the relevant time period.\textsuperscript{131}

3. Measures of Inter-Judge Disparity

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

The most basic measure of sentencing outcomes is sentence length. The Hofer, Anderson-Kling-Stith, and Waldfogel studies measured sentencing outcomes using a single metric: average prison term, in months.\textsuperscript{132} This study uses the same measure.\textsuperscript{133}

\textsuperscript{125} Anderson et al., supra note 26, at 291; Hofer et al., supra note 41, at 282.
\textsuperscript{126} Specifically, sentences were excluded if the sentencing judge was on pace to impose fewer than 25 sentences in a two-year period. \textit{Cf.} Anderson et al., supra note 26, at 287 (using a cutoff of 30 cases, including jurisdictional transfers and acquittals, in a two-year period).
\textsuperscript{127} D. Mass. Local Rule 40.1(B)(3).
\textsuperscript{128} See infra notes 230-233 and accompanying text.
\textsuperscript{129} For a detailed breakdown of the sentence count for each judge, see infra Table A2.
\textsuperscript{130} See infra note 226 & Tale A1 and accompanying text.
\textsuperscript{131} See infra notes Table A3 and accompanying text.
\textsuperscript{132} See Hofer et al., supra note 41, at 307-08 (technical appendix); Anderson et al., supra note 26, at 281; Waldfogel, \textit{Empirically Based Sentencing Guidelines}, supra note 49, at 294. Consistent with the
Linear regression models can analyze inter-judge disparity in sentence length by calculating the percentage of variance in sentence length explained by the judge assigned to the case.\textsuperscript{134}

To capture changes in inter-judge disparity over time, this study performs that analysis during three time periods:

1. **Pre-Booker**: October 1, 2001 – June 23, 2004 (\(\approx 33\) months)
2. **Post-Booker**: January 12, 2005 – June 30, 2006 (\(\approx 35\) months)
3. **Kimbrough/Gall**: December 10, 2007 – September 30, 2008 (\(\approx 10\) months)\textsuperscript{135}

Changes over time in the percentage of variance explained by the judge indicate increases or decreases in inter-judge sentencing disparity.

In addition, this study examines sentence length in the subset of cases not subject to a mandatory minimum sentence. As previous researchers have recognized, mandatory minimums may interfere with accurate assessment of inter-judge sentencing disparity by creating the illusion of inter-judge consistency. To guard against that risk, the Hofer study recommended that future researchers “exclude cases where mandatory minimum statutes truncate the [sentencing] range.”\textsuperscript{136} This study follows that recommendation by separately analyzing cases not governed by a mandatory minimum, which account for 66.4\% of sentences in the dataset.

A second measure of sentencing outcomes is sentencing relative to the guideline range. Disparity in average sentence length provides an incomplete picture because it captures only judges’ general tendency toward leniency or severity, sometimes called the “primary judge effect.”\textsuperscript{137} It does not capture other forms of inter-judge disparity linked to particular offense or offender characteristics.\textsuperscript{138} This study therefore supplements that measure by analyzing inter-judge disparity in guideline sentencing. Using the Guidelines as a reference point does not suggest or assume that the guideline sentencing range is “correct” or just.\textsuperscript{139} But it measures a distinct form of inter-judge disparity, driven by differences in judges’ reactions to the Guidelines themselves. Indeed, anecdotal reports from prosecutors have focused on this form of disparity, warning that some judges

---

\textsuperscript{133} Sentence length is measured as a term of imprisonment in months. Following the Sentencing Commission, a sentence of probation is coded as zero months of imprisonment.

\textsuperscript{134} See infra notes 244-246 and accompanying text.

\textsuperscript{135} Because the Commission has not yet released sentencing data for fiscal year 2009 and beyond, the Kimbrough/Gall period, of necessity, is shorter than the other periods. For a full discussion of period selection issues, see infra, notes 214-219 and accompanying text.

\textsuperscript{136} Hofer et al., supra note 41, at 275 n.103.

\textsuperscript{137} Id. at 240-41; see STITH & CABRANES, supra note 17, at 119.

\textsuperscript{138} Hofer et al., supra note 41, at 240-41; see id. at 297 (calling judge-to-judge disparity in average sentence length “the tip of the disparity iceberg”); STITH & CABRANES, supra note 17, at 119 (acknowledging “[a] possibility that comparing each judge’s average sentence masks considerable variability within each set of sentences”).

\textsuperscript{139} See Bowman, supra note 88, at 296.
routinely sentence within the guideline range, while others routinely sentence below the range.\textsuperscript{140}

The study analyzes guideline sentencing outcomes in part through a straightforward description of changes in sentencing patterns over time. It also analyzes \textit{how far}, on average, each judge sentences from the Guidelines by calculating average distance from the guideline range.\textsuperscript{141} Again, linear regression models can determine the percentage of variance in that metric explained by the judge. Because guideline sentencing patterns are highly sensitive to changes in the law governing departures, this study performs both types of analysis during five time periods:

1. \textit{Mandatory Guidelines}: October 1, 2001 – April 30, 2003 ($\approx 19$ months) 
2. \textit{PROTECT Act}: May 1, 2003 – June 23, 2004 ($\approx 14$ months) 
4. \textit{Post-Booker II}: July 1, 2006 – December 9, 2007 ($\approx 17$ months) 
5. Kimbrough/Gall: December 10, 2007 – September 30, 2008 ($\approx 10$ months)\textsuperscript{142}

As a final measure of guideline sentencing patterns, the study examines a subset of sentences (call them “discretionary sentences”) in which judges were free, as a legal and practical matter, to sentence below the guideline range. The documents from the District of Massachusetts show that in a surprising number of cases—almost 20\% of the Boston sentences in the dataset—judges did not have the option of imposing a below-range sentence. Sometimes a statutory mandatory minimum makes it unlawful to sentence below the guideline minimum. Sometimes, by the time of sentencing, the defendant has already served a term in custody within the guideline range. And sometimes the guideline sentencing range includes a sentence of probation, making a below-range sentence effectively impossible.\textsuperscript{143}

Those constraints suggest that the high rate of within-range sentencing that has continued since \textit{Booker} is partially misleading, the product of legal and practical obstacles rather than continued fealty to the Guidelines. But they also suggest that there exists a narrower class of discretionary sentences, of special interest to researchers studying inter-judge disparity, in which judges had the full range of guideline sentencing options available. Thus, in its review of guideline sentencing outcomes, the study conducts a separate analysis of discretionary sentences.

4. Why Massachusetts?

This study depends on data from the District of Massachusetts, and it is no accident that this particular court makes its sentencing documents available to the public. The judges of the District of Massachusetts take a special interest in sentencing; indeed, several are well-respected as sentencing experts. Judges Nancy Gertner,\textsuperscript{144} William

\textsuperscript{140} See supra notes 107-111 and accompanying text.
\textsuperscript{141} Average distance from the guideline range is calculated using all of the judge’s sentences, treating within-range sentences as zero months. See infra note 246 and accompanying text.
\textsuperscript{142} For an explanation of the cutoff dates for each period, see infra notes 214-218 and accompanying text.
\textsuperscript{143} See infra notes 234-241 and accompanying text.
\textsuperscript{144} See, e.g., Judge Nancy Gertner, From Omnipotence to Impotence: American Judges and Sentencing, 4 OHIO ST. CRIM. L.J. 523 (2007) [hereinafter Gertner, Omnipotence to Impotence]; Judge Nancy Gertner,
Lessons from Criminal Trials and Sentencing

One Judge’s Perspective

Moreover, the same qualities that led the court to approve its disclosure policy might make it dissimilar from other courts. 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. Also, average sentences in Massachusetts are slightly higher than sentences nationwide, and the rate of below-guideline sentencing in Massachusetts is higher than the rate nationwide. Those differences set the District of Massachusetts 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, it avoids the risk that inter-district disparity in prosecutorial practices might be mistaken for inter-judge disparity. In Massachusetts, the Criminal Division


See, e.g., William G. Young, An Open Letter to U.S. District Judges, 50 FED. L. REV 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.


Hofer et al., supra note 41, at 279.

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).

See Federal Judges Biographical Database, supra note 93. 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 Waldfoegel, Empirically Based Sentencing Guidelines, supra note 49, at 294, or New York City and Philadelphia, see Payne, supra note 49, 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.

U.S. SENTENCING COMM’N, 2002-2008 SOURCEBOOK OF FEDERAL SENTENCING STATISTICS Appx. B (National, D. Mass.) (2002-2008). 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. 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 programs have the effect of boosting the nationwide rate of government-sponsored sentences compared with districts, like Massachusetts, that have no fast-track authority.

of a single U.S. Attorney’s office charges and prosecutes virtually all federal cases.\textsuperscript{152} Second, it avoids the risk that inter-region disparity in the types of offenses committed or prosecuted might be mistaken for inter-judge disparity by comparing judges who share a common case pool. This study focuses on a core group of judges who drew cases at random from a common pool in Boston. Third, it avoids concerns about inter-circuit disparity caused by differences in appellate courts. In the wake of Booker, regional courts of appeals split on a number of questions concerning reasonableness review,\textsuperscript{153} and Schanzenbach and Tiller have found that the partisan alignment of circuit courts can affect sentence length and the likelihood of departure.\textsuperscript{154} Examining a single district ensures that all judges being studied were bound to follow the same circuit precedent, subject to review by the same mix of appellate judges.

Of course, this study’s findings about the sentencing patterns in Boston do not necessarily explain sentencing patterns in far-flung cities nationwide. It offers a first look at inter-judge disparity after Booker, by no means the final word. Nonetheless, the Massachusetts documents offer researchers unparalleled access to judge-specific sentencing data, and therefore the best available evidence of how sentencing by individual judges has changed in the wake of Booker, Gall, and Kimbrough.

B. Results

Analysis of the Boston data reveals a clear increase in inter-judge sentencing disparity, both in sentence length and in guideline sentencing patterns. The effect of the judge on sentence length has doubled in strength since Kimbrough and Gall. And in their guideline sentencing patterns, judges have responded in starkly different ways to Booker, with some following a “free at last” pattern and others a “business as usual” pattern.

\textsuperscript{152} See United States Attorney’s Office District of Massachusetts, Divisions, at http://www.usdoj.gov/usa/mag/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.


\textsuperscript{154} See Schanzenbach & Tiller, supra note 87, 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)
1. Sentence Length

Among Boston judges as a whole, average sentence length has increased since *Booker*. Figure 3 shows the increase, both for all sentences and for sentences not governed by a statutory mandatory minimum.\(^\text{155}\)

**Figure 3: Average Sentence Length, Boston Judges**

Average sentence length climbed from 47.8 months before *Booker*, to 58.3 months in the years following *Booker*, to 63.7 months after *Kimbrough* and *Gall*. Excluding cases subject to a mandatory minimum, the increase is more gradual, from 30.8 months before *Booker*, to 33.7 months after *Booker*, to 35.5 months after *Kimbrough* and *Gall*.

But average sentence length for the district as a whole masks significant variation among individual judges. Figure 4a shows the distribution of average sentence length for each judge as it has changed over time—before *Booker*, after *Booker*, and after *Kimbrough* and *Gall*. Each dot represents the average sentence for a single judge. Figure 4b shows the same distribution, but leaves high and low values unshaded to make it easier to see how the remaining dots are clustered:

---

\(^{155}\) 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.
Although the difference between the highest and lowest averages remains essentially unchanged between periods, the distribution of averages has widened compared to the pre-Booker period. After Kimbrough and Gall, in particular, two clusters of judges are readily apparent: one cluster following the trend toward higher sentences with averages around 70 months, and another cluster splitting off with averages around 45 months.

Statistical analysis confirms that the effect of the judge on sentence length has grown stronger since Kimbrough and Gall. Table 1 reports the results:

<table>
<thead>
<tr>
<th></th>
<th>% Variance Explained</th>
<th>Avg. Variance Explained</th>
<th>Model Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Booker</td>
<td>2.9%</td>
<td>10.7 months</td>
<td>.001*</td>
</tr>
<tr>
<td>Post-Booker</td>
<td>2.5%</td>
<td>10.9 months</td>
<td>.003*</td>
</tr>
<tr>
<td>Kimbrough/Gall</td>
<td>6.1%</td>
<td>15.5 months</td>
<td>.044*</td>
</tr>
</tbody>
</table>

* Significant at the .05 level

Table 1: Summary of Linear Regression Models Sentence Length

---

156 For details of these regression models for sentence length, see infra Table A4.
For each period, the “% Variance Explained” column reports the percentage of variance in sentence length explained by which judge was assigned to the case. The “Avg. Variance Explained” column converts that percentage into actual months of variance explained, as an average for all sentences. The “Model Significance” column reports the statistical significance of the model.\(^{157}\)

For the full set of sentences, the regression models indicate a delayed reaction, but ultimately a sharp uptick in inter-judge sentencing disparity since *Booker*. In the years before the decision, the percentage of variance in sentence length explained by the identity of the judge stood at 2.9%. Immediately after *Booker*, the rate actually declined slightly to 2.5%. But in the *Kimbrough/Gall* period, it rose sharply to 6.1%. That means the effect of the judge on sentence length is now more than twice as strong as in the three years before *Booker*.\(^{158}\)

The increase in inter-judge disparity is even clearer in cases not governed by a mandatory minimum sentence. As previous researchers have noted, mandatory minimums affect sentence length for all judges and, as a result, may mask changes in inter-judge disparity. Figure 5 shows the pre-*Booker* and post-*Booker* distribution of average sentences for cases not subject to a mandatory minimum:

\(^{157}\) For discussion of the regression models generally, see infra notes 244-246 and accompanying text.

\(^{158}\) For a discussion of period-selection issues, see infra notes 214-219 and accompanying text.
For cases not subject to a mandatory minimum, the trend is unmistakable. The distribution of average sentences among judges has grown substantially wider since *Booker*: from a total spread of 15 months before *Booker*, to almost 30 months after *Booker*, to almost 40 months in the wake of *Kimbrough* and *Gall*.

The stark differences between judges have real consequences for criminal defendants. Before *Booker*, regardless of the judge, a defendant in Boston not facing a mandatory minimum could expect that the judge’s average sentence would fall between 25.9 months and 40.2 months. Today, after *Kimbrough* and *Gall*, three judges on the court are imposing average sentences of 25.5 months or less, while two other judges on the court are imposing average sentences of 51.4 months or more. That is an average difference of more than two years in prison, depending on which judge is assigned to the case.

Again, statistical analysis confirms that, for sentences not subject to a mandatory minimum, the relationship between the identity of the judge and the length of the sentence has grown stronger since *Booker*. Table 2 reports the results:

<table>
<thead>
<tr>
<th></th>
<th>% Variance Explained</th>
<th>Avg. Variance Explained</th>
<th>Model Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Booker</td>
<td>1.4%</td>
<td>4.9 months</td>
<td>.368</td>
</tr>
<tr>
<td>Post-Booker</td>
<td>3.1%</td>
<td>8.0 months</td>
<td>.021*</td>
</tr>
<tr>
<td><em>Kimbrough/Gall</em></td>
<td>8.0%</td>
<td>10.3 months</td>
<td>.180</td>
</tr>
</tbody>
</table>

* Significant at the .05 level

**Table 2: Summary of Linear Regression Models**

**Sentence Length, Excluding Mandatory Minimums**

For sentences not governed by a mandatory minimum, in the pre-*Booker* period the rate of variance in sentence length explained by the identity of the judge was very small, just 1.4%, and the relationship was not statistically significant. After *Booker*, however, the rate more than doubled to 3.1% and the identity of the judge became a statistically significant predictor of sentence length. Since *Kimbrough* and *Gall*, the rate of variance explained has increased again to 8.0%—more than five times pre-*Booker* levels—although the model for that period is not yet statistically significant.

2. Guideline Sentencing Patterns

Similarly, analysis of guideline sentencing patterns since *Booker* indicates that, consistent with anecdotal reports from around the country, there has been a spike in inter-judge disparity. Some Boston judges have embraced their newfound discretion to depart

---

159 For details of these regression models, see *infra* Table A5.
160 The fact that the model for non-mandatory minimum sentences in the *Kimbrough/Gall* period is not significant reinforces that the results are preliminary, but does not undermine their persuasiveness. Statistical significance is highly sensitive to sample size, and testing a subset of cases obviously shrinks the sample. Because the *Kimbrough/Gall* period of necessity has about one-third as many cases as the other periods, and the relationship is strongly positive, the lack of statistical significance is a reason for caution, but not a reason to discount the results.
from the guideline range more enthusiastically than others. Consider the below-range sentencing patterns of four judges, A, B, C, and D:

- Sentences by Judge A closely track the pattern for the district as a whole. Under the Mandatory Guidelines, 19.6% of Judge A’s sentences fell below the guideline range. Under the PROTECT Act, that figure fell sharply to 7.7%. But after *Booker*, it rebounded to well above pre-*Booker* levels at over 35%, and has remained at those levels continuously for more than four years.

- 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%) followed by 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.

- 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 after *Booker*, from 10.5% to 10.0%, and has remained stable throughout the other periods as well (13.3%, 19.1% and 16.1%).

- Sentences by Judge D fit a “return to form” pattern. 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. Recently it has returned to 38.6% and 34.6% in the two most recent periods.

---

161 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 “black list” of disfavored judges. See Letter of Leonidas Ralph Mecham, supra note 64, 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 that judge-identifying information can enable valuable research without targeting individual judges.

162 See Gertner, *Omnipotence to Impotence*, supra note 144, at 579 (using the phrase “free at last” to describe the reaction to *Booker* among some district court judges).

163 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*.

164 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.

165 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
Figures 6a-6d show the sentencing patterns of Judges A, B, C, and D, overlaid on the average sentencing pattern for the district as a whole:

- Figure 6a: Guideline Sentencing, Judge A
- Figure 6b: Guideline Sentencing, Judge B
- Figure 6c: Guideline Sentencing, Judge C
- Figure 6d: Guideline Sentencing, Judge D

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.
These disparate patterns suggest that reports by the Commission fail to capture important differences in the way that individual judges have responded to *Booker*. Judge B imposed sentences below the guideline range in about 11% of cases before *Booker*, but in the most recent period has imposed below-range sentences in about 53% of cases. Judge C, by contrast, also sentenced below the guideline range in approximately 10% of cases before *Booker*, but most recently has imposed below-range sentences in only 16% of cases. That is a stark difference in post-*Booker* sentencing behavior, and tends to corroborate anecdotal reports of a surge in inter-judge sentencing disparity.

Statistical analysis of how far, on average, each judge has sentenced from the guideline range confirms an increase in inter-judge disparity in guideline sentencing. Figure 7a shows the distribution of average distance from the guideline range, with each dot representing the average distance for one judge:

![Figure 7a: Distribution in Average Distance from Guideline Range](image)

Under the mandatory Guidelines in 2002-2003, average distance from the guideline range was tightly clustered within a range of 4.5 months. Perhaps surprisingly, the spread increased under the PROTECT Act, covering 8.2 months. But after *Booker*, the distribution has widened dramatically and grown broader in every period. In the most recent period, following *Kimbrough* and *Gall*, average distances from the Guidelines span 20.0 months, ranging from 4.2 months to a remarkable 24.2 months.

As expected, the trend is even more pronounced for “discretionary” sentences in which the sentencing judge was free, as a legal and practical matter, to sentence outside
the guideline range. Figure 7b shows the distribution of average distance from the guideline range for the subset of discretionary sentences:

Under the mandatory Guidelines, average distance from the guideline range was clustered within a range of 5.6 months. Under the PROTECT Act the range increased to 12.4 months. Since then, the distribution has widened to 16.5 months, then 23.8 months, and most recently 23.4 months.

For criminal defendants in the 80% of cases where the judge has full discretion to sentence outside the guideline range, the difference between judges has serious consequences. Under the mandatory Guidelines in 2002-2003, regardless of the judge assigned to the case, a criminal defendant could expect an average sentence 7.8 months or less from the Guidelines. Today, in the wake of *Kimbrough* and *Gall*, three judges in Boston continue to sentence on average 6.1 months or less from the guideline range. But a different group of three Boston judges sentences, on average, 24.6 months or more from the guideline range. That is an average difference of more than a year and a half in prison, depending on the judge.

Statistical analysis reinforces that inter-judge disparity in distance from the guideline range has increased since *Booker*. Table 3 reports the results of linear

---

166 See supra notes 234-242 and accompanying text (defining “discretionary” sentences and explaining why they are of special relevance in measuring inter-judge disparity in guideline sentencing).
regression models calculating the percentage of variance explained by the judge, both for all sentences and for discretionary sentences:

<table>
<thead>
<tr>
<th></th>
<th>% Variance Explained</th>
<th>Avg. Variance Explained</th>
<th>Model Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All Sentences</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mandatory Guidelines</td>
<td>1.0%</td>
<td>1.4 months</td>
<td>.847</td>
</tr>
<tr>
<td>PROTECT Act</td>
<td>2.4%</td>
<td>2.7 months</td>
<td>.482</td>
</tr>
<tr>
<td>Post-Booker I</td>
<td>3.6%</td>
<td>3.9 months</td>
<td>.089†</td>
</tr>
<tr>
<td>Post-Booker II</td>
<td>3.7%</td>
<td>4.8 months</td>
<td>.048*</td>
</tr>
<tr>
<td>Kimbrough/Gall</td>
<td>6.6%</td>
<td>7.1 months</td>
<td>.073†</td>
</tr>
<tr>
<td><strong>Discretionary Sentences</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mandatory Guidelines</td>
<td>1.3%</td>
<td>1.8 months</td>
<td>.853</td>
</tr>
<tr>
<td>PROTECT Act</td>
<td>3.6%</td>
<td>3.7 months</td>
<td>.384</td>
</tr>
<tr>
<td>Post-Booker I</td>
<td>4.5%</td>
<td>4.9 months</td>
<td>.105</td>
</tr>
<tr>
<td>Post-Booker II</td>
<td>5.1%</td>
<td>6.2 months</td>
<td>.038*</td>
</tr>
<tr>
<td>Kimbrough/Gall</td>
<td>9.4%</td>
<td>9.1 months</td>
<td>.037*</td>
</tr>
</tbody>
</table>

* Significant at the .05 level
† Significant at the .10 level

Table 3: Summary of Linear Regression Models
Distance from Guideline Range

The models confirm 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 during the first eighteen months after Booker and 6.6% (more than double pre-Booker levels) since Kimbrough and Gall. As expected, the trend is even stronger for discretionary sentences, with the identity of the judge explaining 9.4% of the variance (nearly triple pre-Booker levels) since Kimbrough and Gall.

Together, these measures of sentencing outcomes in Boston tend to corroborate anecdotal reports of a surge in inter-judge sentencing disparity. Since the Supreme Court’s decisions in Booker, Kimbrough, and Gall, the effect of the judge on sentence length has more than doubled in strength. In cases not subject to a mandatory minimum, the court’s three most lenient judges are imposing average sentences of 25.5 months or less, while its two most severe judges are imposing average sentences of 51.4 months or more, resulting in an average difference of more than two years in prison depending on which judge is assigned the case. Similarly, the effect of the judge on how far sentences fall from the guideline range has more than doubled. In Boston, some judges continue to impose below-guideline sentences at essentially the same rate as before Booker, as little

---

167 Hofer et al., supra note 41, 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.

168 For details of these regression models, see infra Tables A6 & A7.
as 16% of the time, while other judges now sentence below the guideline range at triple or quadruple their pre-Booker levels, as much as 53% of the time.

III. IMPLICATIONS

The results of the empirical study, showing a spike in inter-judge sentencing disparity after *Booker*, *Gall*, and *Kimbrough*, come as unwelcome news. Although the study examines only one district court, its findings tend to corroborate anecdotal evidence from around the country warning of greater judge-to-judge disparity in sentencing outcomes. If the same trends have played out in other districts, they would mark a step backward from Congress’s goal of reducing inter-judge sentencing disparity.

It is true that, despite the uptick in inter-judge disparity, the effect of the judge remains relatively modest. Even after *Kimbrough* and *Gall*, the judge accounts for 6.1% of variation in sentence length (8.0% in cases not subject to a mandatory minimum), and 6.4% of variation in distance from the guideline range (9.1% for discretionary sentences). Yet both the strength of the effect and size of the change are larger than those reported in the Hofer study, suggesting that *Booker*, *Kimbrough*, and *Gall* may have altered inter-judge disparity to a degree comparable to the original Guidelines. Moreover, as the Anderson-Kling-Stith study 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.” Although it is too early to despair a return to “pre-guideline chaos,” the preliminary evidence—from the only district court in which this sort of study is possible—is discouraging.

It also bears emphasis that inter-judge sentencing disparity is but one consideration among many in evaluating the federal sentencing system. It is entirely possible to conclude that *Booker*, *Kimbrough*, and *Gall* have improved federal sentencing, on balance, by allowing judges greater flexibility to reject unjust guidelines and impose just sentences. And there are other urgent priorities for federal sentencing reform, including the elimination of the 100:1 crack-powder ratio in drug sentencing, reevaluation of mandatory minimum sentences, and confronting unwarranted disparity created by prosecutor charging and bargaining practices. Nonetheless, reducing inter-judge sentencing disparity was one of Congress’s primary goals in the Sentencing Reform Act, and evidence of backsliding ought to be taken seriously.

What explains the uptick in inter-judge sentencing disparity? Specifically, why have “business as usual” judges continued to impose so many sentences within the guideline range? The Boston data tend to undermine the conventional explanations that within-guideline sentencing is the product of inertia, risk-aversion, anchoring, strategic

---

169 See *supra* notes 107-111 and accompanying text.
170 See *supra* Tables 1 & 2 and accompanying text.
171 See *Hofer et al.* *supra* note 41, at 287; see also *Waldfogel, Empirically Based Sentencing Guidelines, supra* note 49, at 294.
172 Anderson et al., *supra* note 26, at 294.
174 In addition to calling for research on inter-judge sentencing disparity, Attorney General Holder has convened a department-wide Sentencing and Corrections Working Group to consider those issues. *Holder, supra* note 7.
behavior, or laziness. Instead, I propose two possible explanations that have received surprisingly little attention. Some judges might actually agree with the Guidelines’ sentencing recommendations more often than their colleagues. And some judges might choose to impose within-range sentences for institutional reasons, such as deference to the Commission or a belief that the Guidelines carry democratic legitimacy.

A. Conventional Explanations for Within-Range Sentencing

As discussed above, the modest initial response to Booker has prompted extensive speculation about why so many judges, freed from the shackles of the mandatory Guidelines regime, have continued to impose within-guideline sentences more than 80% of the time. The conventional wisdom points to five factors: (1) inertia among 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” caused by the requirement to use the guideline range as a starting point; (4) strategic behavior by judges anxious to avoid provoking Congress; and (5) simple laziness. The Boston data, however, tend to undermine many of those explanations.

1. Inertia

The first, most common explanation of judges’ unexpectedly mild 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, on this theory, that a generation of judges that has “grown up” with the Guidelines would cling to them even though they have become advisory.

The Boston data, however, provide little support for that hypothesis. If the inertia theory is correct, then judges appointed before 1987, who tasted freedom under the pre-Guidelines regime, should respond more swiftly and enthusiastically than their younger colleagues to the return of their sentencing discretion. It happens that, 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 when the Guidelines were mandatory. As shown in Figures 8a and 8b, however, the difference between those groups is negligible:

---

175 See supra notes 93-104 and accompanying text.
176 See supra notes 93-105 and accompanying text.
From 2002 to 2008, there has been no meaningful difference in average sentence length based on whether a judge was appointed before the effective date of the Guidelines in 1987. If anything, it was judges appointed after 1987, not their more senior colleagues, whose sentences changed most noticeably after *Booker*, jumping from 46.0 months to 60.6 months.\(^{177}\)

Nor do patterns in guideline sentencing suggest a continuing “inertia” effect. It appears that pre-1987 judges responded more quickly in the immediate aftermath of *Booker*, with a higher rate of below-range sentencing than post-1987 judges during the first eighteen months (33.8% compared with 21.4%). But the difference has already evaporated. For more than two and a half years, both groups have sentenced below the guideline range around 30% of the time.\(^{178}\)

2. Risk-Aversion

A second conventional explanation 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

\(^{177}\) Pre-1987 service by the sentencing judge was not a statistically significant predictor of sentence length in any of the three periods.

\(^{178}\) Pre-1987 service by the sentencing judge was not a statistically significant predictor of how far a sentence falls from the guideline range during any period.
for the Supreme Court and courts of appeals to work out the details of that standard.\textsuperscript{179} Some judges are more risk-averse than others, and on this theory, the most risk-averse judges may cling to the guideline range to avoid reversal and resentencing.

The risk-aversion theory assumes that all judges wish to sentence outside the guideline range to the same extent, but that some judges are more timid than others in the face of uncertainty. If the theory is sound, then decisions like Kimbrough and Gall should reduce inter-judge disparity in guideline sentencing because they simultaneously reduce uncertainty and reduce the risk of reversal. By making clear in Gall that the appellate standard is extremely deferential,\textsuperscript{180} and by indicating in Kimbrough that judges may categorically reject the Sentencing Commission’s policy judgments,\textsuperscript{181} the Court made clear to even the most risk-averse judges that they are free to sentence outside the guideline range with confidence.

But in Boston, Kimbrough and Gall did not reduce inter-judge disparity. To the contrary, differences between judges grew even more acute. The effect of the judge on sentence length reached double pre-Booker levels, while the effect on distance from the guidelines strengthened to triple pre-Booker levels.\textsuperscript{182} It seems entirely plausible that risk aversion affected sentencing decisions in the years immediately after Booker.\textsuperscript{183} But the threat of appellate review does not offer a satisfying explanation of why inter-judge sentencing disparity has persisted after Kimbrough and Gall, when the threat became greatly diminished.

3. Anchoring, Strategic Behavior, and Laziness

The remaining conventional explanations—the “anchoring” effect of the Guidelines,\textsuperscript{184} strategic behavior by judges to avoid provoking Congress,\textsuperscript{185} and simple laziness\textsuperscript{186}—are more difficult to evaluate. Judges do not readily disclose their motives and character flaws, and may not even be aware of their cognitive biases. In theory, any of those factors could affect some judges more than others, and therefore partially explain the spike in inter-judge disparity in sentencing outcomes.

But none is terribly persuasive. The anchoring effect of the guideline range must compete with other anchors supplied by prosecutors, defense counsel, the probation office, victim impact testimony, and the statutory minimum. Judges do not approach sentencing with a single number in mind, but with multiple numbers from various sources. And most judges have imposed sentence in hundreds of cases, “dropping anchor” each time and potentially reducing the influence of the Guidelines act as a source of cognitive error.

\textsuperscript{179} See supra notes 95-98 and accompanying text.
\textsuperscript{180} Gall v. United States, 552 U.S. 38, 51-51 (2007).
\textsuperscript{182} See supra Table 7 and accompanying text.
\textsuperscript{183} See Gertner, supra note 98, at 138 (concluding, in the period between Booker and the decisions in Kimbrough and Gall, that appellate courts were closely policing sentences on appeal and that “[d]istrict judges have gotten the message”).
\textsuperscript{184} See supra notes 99-102 and accompanying text
\textsuperscript{185} See supra notes 103-104 and accompanying text.
\textsuperscript{186} See supra notes 105-106 and accompanying text.
Strategic behavior by judges, designed to avoid provoking Congress, is certainly plausible. A rich empirical literature has documented other forms of strategic behavior by judges. Yet the sheer number of district court judges and sentencing decisions makes this sort of strategy difficult to execute. No individual judge could prevent the more than 600 other judges nationwide from sentencing in a manner that agitates Congress. Further, in any particular case, the strategic benefits of remaining within the guideline range would seem tiny compared with the costs to the criminal defendant. Moreover, the risk of congressional retaliation substantially subsided in 2006 when Democrats took control of Congress. Yet inter-judge sentencing disparity has continued to increase.

Judicial laziness is a particularly unpersuasive explanation. A few judges, themselves critics of the Guidelines, have floated the rather self-congratulatory theory that their colleagues cling to guideline range because they prefer “less work and less stress” and are unwilling to approach the task of sentencing with sufficient intellectual rigor. Although judges—like the rest of us—undoubtedly consider their personal time constraints in approaching their work, the upheaval of Booker, Kimbrough, and Gall forced judges to approach sentencing decisions with unusual caution and seriousness. It is difficult to imagine that a vast segment of the federal bench, despite being duty-bound to impose sentences consistent with § 3553(a), is mechanically following the Guidelines just to avoid thinking too much.

B. Alternative Explanations for Within-Range Sentencing

As alternatives to the conventional account, I propose two other explanations for the persistent high rate of sentences within the guideline range. One is that some judges actually agree, on the merits, with the Guidelines’ recommendations about the appropriate level of punishment. Another is that some judges, more than their colleagues, defer to the Guidelines for institutional reasons. Surprisingly, both of these possibilities have been largely ignored by the legal literature.

1. Agreement with the Guidelines

The most promising explanation for many judges’ continued fidelity to the Guidelines is also the simplest. Some judges might actually agree with the Guidelines to a greater extent than their colleagues. The fact that some judges have followed a


189 See supra notes 105-106 and accompanying text.

“business as usual” pattern of guideline sentencing,\textsuperscript{191} even as the degree of freedom they enjoy at sentencing has dramatically expanded, suggests that they simply do not wish to alter their pre-\textit{Booker} sentencing practices. That fundamental difference of opinion between “business as usual” and “free at last” judges results in increased inter-judge sentencing disparity.

Remarkably, however, the scholarly literature has essentially ignored that possibility. The enormous body of pre-\textit{Booker} literature criticizing the federal Guidelines frequently created the impression of uniform opposition among judges.\textsuperscript{192} Judges themselves commented that “[t]oday almost all federal judges agree that these guidelines need substantial change, if not complete rejection.”\textsuperscript{193} News reports described hostility so pervasive that “[m]any judges regard as a traitor any colleague who serves on the U.S. Sentencing Commission.”\textsuperscript{194} A few judges even resigned from the bench or refused to hear drug cases in protest of the Guidelines regime.\textsuperscript{195} Scholars cheerfully generalized that “everybody loves to hate the Federal Sentencing Guidelines.”\textsuperscript{196} Even members of Congress took notice that judges seemed to “hate the sentencing guidelines.”\textsuperscript{197}

Surveys of federal judges, however, have documented a persistent split in opinion. The most recent survey, a 2003 project by the Sentencing Commission, asked district court judges to rate various aspects of the Guidelines on a scale from 1 to 6, with 1 being the worst possible rating and 6 being the best.\textsuperscript{198} Table 4 excerpts some of the results:

\textsuperscript{191} See supra note 164 and accompanying text.
\textsuperscript{194} Naftali Bendavid, \textit{Breyer’s Role as Sentencing Pioneer Still Rankles}, LEGAL TIMES, May 16, 1994, at 7.
As 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. For five of the nine specific purposes of punishment in § 3553(a), a majority of judges rated the Guidelines a 5 or 6. And for the remaining purposes of punishment, a substantial minority of judges gave the Guidelines a high rating. Asked how often the Guidelines prescribe a just punishment, 37.0% gave them a 5 or 6. Asked whether the Guidelines offer sufficient flexibility at sentencing, 24.4% of district court judges gave them a 5 or 6.

Older surveys also confirm the existence of a sizeable contingent of judges who support the Guidelines. A 1993 survey by the ABA Journal caused a stir when it found that 45% of federal judges believed the Guidelines should be “scraped.” Similarly, a Federal Judicial Center survey from 1996 found that 73% of district court judges believed mandatory guidelines were not “necessary to direct the sentencing process.” The converse of those findings is that about half of district court judges believed some form of sentencing guidelines should be preserved, and 27% believed mandatory guidelines were necessary. Although those judges were in the minority, they still formed a large and stable portion of the federal bench.

The survey results thus suggest a simple explanation for increasing inter-judge disparity after Booker, Kimbrough, and Gall. For almost twenty years, a substantial contingent of federal judges has quietly agreed with the Guidelines. And now that judges must directly apply the purposes of punishment in § 3553(a), inter-judge disagreements about the performance of the Guidelines have direct consequences for criminal defendants. Judges who generally believe the Guidelines perform poorly are free to routinely reject them, while judges who generally believe the Guidelines perform well...

---

199 See id. at 12, 15, 24.
200 Id. at 24.
201 Id. at 2-3.
202 Id. at 12.
203 Id. at 15.
remain free to routinely follow them. The result is a spike in judge-to-judge sentencing disparity.

2. Institutional Considerations

Another possible explanation for the persistence of within-range sentencing is that some judges, more than their colleagues, find institutional reasons for deference to the Commission persuasive. In particular, they may be persuaded by the competence of the Commission relative to individual judges, or by the democratic legitimacy of the Guidelines.

First, concerns about institutional competence may persuade some judges to stick close to the Guidelines. The Commission is designed to serve as an expert body, with trained staff, a dedicated research arm, and exhaustive data concerning federal sentencing practices. Judges may choose to accord strong respect to the Commission’s recommendations based on doubts about their own competence, as individual judges, to make systemic judgments about sentencing policy. They may be reluctant, for example, to “individualize” sentences based on their own predictions about the future dangerousness of individual defendants. As one judge put it in a post-Booker opinion, “unlike Congress or the Commission, we judges lack the institutional capacity (and frankly, the personal competence) to set up and then enforce one new, well-chosen, theoretically coherent, national standard.”

Second, concerns about institutional legitimacy may lead some judges to accord strong respect to the Guidelines. Congress reserves to itself the power to review, modify, and reject changes to the Guidelines. Judges may conclude, based on Congress’s stamp of approval, that the Guidelines carry democratic legitimacy, making it generally inappropriate for judges to substitute their own policy and punishment values for those embodied in the Guidelines.


211 United States v. Cage, 451 F.3d 585, 593 (10th Cir. 2006) (describing the Guidelines as “an expression of popular political will about sentencing”); Wilson, 350 F. Supp. 2d at 915; Wanning, 354 F. Supp. 2d at 1062 n.9; Bibas et al., supra note 206, at 1388 (“Most importantly, Congress has democratic legitimacy; courts do not.”).
No doubt institutional considerations like these interact with other judgments at sentencing. In a case where the advisory guideline sentence strikes the judge as grossly unjust, institutional respect for the Guidelines likely makes little difference. But in a case where the advisory guideline range seems just a little too high, and not grossly excessive, the judge’s assessment of the institutional strengths of the Commission and the democratic legitimacy of the Guidelines may make the difference between a within-range or below-range sentence.

To be sure, critics of the Guidelines have vigorously challenged these institutional justifications for guideline sentencing. Scholars have noted, for example, that the Commission does not behave like most expert administrative agencies, frequently failing to marshal evidence or even provide a reasoned explanation in support of its judgments. They have also raised serious questions about whether the unique structure and composition of the Commission actually advances, rather than undermines, its legitimacy. Many commentators have urged that the Commission’s work, in general and in a host of specific instances, should not be entitled to deference. But not all judges will find those arguments persuasive. The point is not that the institutional reasons for deferring to the Commission are correct, but that they supply a basis for the kind of good-faith disagreement that can fuel inter-judge sentencing disparity.

We cannot know for certain why many judges, contrary to expectations, have continued to impose within-guideline sentences at a high rate. The Boston data do not foreclose the possibility that the conventional explanations—habit, anxiety, cognitive error, strategic behavior, and laziness—might play some role. But it is important to remember Occam’s razor. Simpler explanations should not be neglected. Some judges might actually agree with the Guidelines, or find institutional reasons for deference to the Guidelines compelling.

CONCLUSION

Consistent with anecdotal reports from around the country, the first empirical study of individual judges’ response to Booker, Kimbrough, and Gall reports a spike in inter-judge sentencing disparity. Among judges in Boston, the effect of the judge on sentence length has more than doubled in strength. So has the effect of the judge on how far sentences fall from the guideline range. A clear split has emerged between “free at last” judges, whose rate of below-range sentencing has tripled or quadrupled to as high as 53%, and “business as usual” judges, whose rate of below-range sentencing has hardly changed since Booker and remains as low as 16%. The consequences for criminal defendants are significant. In cases not governed by a mandatory minimum, drawing one of the court’s more severe judges, rather than its more lenient judges, means an average difference of more than two years in prison.

212 Luby, supra note 192, at 1202 (“The Commission . . .rarely justifies its guidelines, consistently avoids on-the-record decisionmaking, and operates unencumbered by the procedural safeguards that ensure the political legitimacy of other administrative agencies.”); Kate Stith & José A. Cabranes, Judging Under the Federal Sentencing Guidelines, 91 Nw. U. L. Rev. 1247, 1270-71 (1997) (noting that “the Sentencing Commission almost never explains the reason behind a particular Guidelines rule” characterizing the Guidelines as a “compilation of administrative diktats”).

These findings are necessarily tentative. They reveal how judges in Boston have responded to *Booker*, *Kimbrough*, and *Gall*, but they may not be representative of sentencing trends nationwide. And, of course, inter-judge disparity is just one factor to consider in reforming the federal sentencing system. It is entirely possible to conclude that, despite the spike in inter-judge disparity, *Booker* on balance represents a step in the right direction.

But the advantages of *Booker* were immediately obvious. Greater flexibility has allowed sentencing judges to reject sentences they see as excessive and to do justice for individual offenders much more frequently. The systemic consequences for inter-judge uniformity, by contrast, have been more difficult to assess and slower to develop. This Article thus offers a critical first look at how *Booker*, *Kimbrough*, and *Gall* have affected one of Congress’s top sentencing reform priorities.
APPENDIX

This Appendix provides additional details concerning the methodology and results of the empirical study.

A. Methodological Details

1. Period Selection

The study examines sentences between fiscal years 2002 and 2008, the last year for which data are available. In evaluating inter-judge disparity in sentence length, the study divides that period into three time periods. The Pre-Booker period begins on October 1, 2001, the first day of fiscal year 2002, and ends on the date of the Supreme Court’s decision in Blakely v. Washington.\(^\text{214}\) Because of the chaos that followed that decision, the interregnum between Blakely and Booker is ignored.\(^\text{215}\) The Post-Booker period extends almost three years, from the date of the Booker decision until December 9, 2007. The Kimbrough/Gall period begins on December 10, 2007, the date of those decisions, and ends on September 30, 2008, the last day of the fiscal year.\(^\text{216}\)

In evaluating inter-judge disparity in guideline sentencing, the study subdivides the pre-Booker and post-Booker periods to create five periods. The Pre-Booker period is divided into a Mandatory Guidelines period and a PROTECT Act period, with the PROTECT Act period beginning on May 1, 2004, the effective date of the Act. The Post-Booker period is divided into two periods, “Post-Booker I” and “Post-Booker II,” with the latter period beginning on July 1, 2006.

These cutoff dates were selected with two competing objectives in mind: to create periods large enough to ensure a sufficient number of cases per judge, but small enough to capture relevant changes in sentencing law. Because sentence length only indirectly reflects guideline sentencing patterns, longer periods allow for a larger number of sentences without ignoring potentially relevant legal changes.\(^\text{217}\) For guideline sentencing, however, the PROTECT Act marks a critical change because it was explicitly designed to reduce the number of downward departures.\(^\text{218}\) The study therefore divides the Pre-Booker period to separate the effects of the PROTECT Act, and divides the Post-Booker period to create periods of roughly equal length. For the sake of completeness, the Appendix also discusses how alternative time periods would affect the regression models.\(^\text{219}\)

\(^{215}\) The Commission’s post-Booker reports have ignored the period between Blakely and Booker as well. See, e.g., FINAL REPORT, supra note 86.
\(^{216}\) 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.
\(^{217}\) The Kimbrough/Gall period is shorter than the other periods because no data are available for fiscal year 2009.
\(^{218}\) See supra notes 53-61 and accompanying text.
\(^{219}\) See infra Tables A4-A5 and accompanying text.
2. Case Matching

The Sentencing Commission has a longstanding policy of withholding judge-identifying information.\(^{220}\) It has been roundly criticized by scholars,\(^{221}\) and I join the chorus calling for the Commission to promote transparency and facilitate the study of federal sentencing by releasing case records that include judge identifiers.\(^{222}\)

Max Schanzenbach and Emerson Tiller have developed a work-around, however, that uses docket information available on PACER (Public Access to Court Electronic Records), 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.\(^{223}\) 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.\(^{224}\) They successfully matched about 80% of sentences returned in their searches.\(^{225}\)

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. The search extended to every criminal case filed in the Boston office between January 1, 2000 and June 30, 2008.\(^{226}\) 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\(^{227}\)—information from the Statement of Reasons was used to narrow the list of

---

\(^{220}\) Its ostensible purpose is to prevent the release of defendant-identifying information. *See Public Access to Sentencing Commission Documents and Data*, 54 Fed. Reg. 51,279, 51,282 (Dec. 13, 1989). But that risk is unavoidable because a determined researcher could use other case information to link virtually any individual defendant to the Commission’s records.


\(^{223}\) Schanzenbach & Tiller, *supra* note 87, at 729-30.

\(^{224}\) *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.

\(^{225}\) *See id.* at 730.

\(^{226}\) 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.

\(^{227}\) The Commission made date matching much more difficult because, beginning in 2004, case records no longer include the exact date of sentencing, but only the month and year. *OFFICE OF POLICY ANALYSIS,*
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.

The process resulted in 2,659 matched cases, more than 90% of the Boston
sentences in the Commission’s files. Table A1 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 A1: Matched Cases, by Fiscal Year

Table A2 lists the final number of sentences for each core judge, by period:

<table>
<thead>
<tr>
<th>Judge</th>
<th>Mandatory 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>
<td>228</td>
</tr>
<tr>
<td>Judge B</td>
<td>54</td>
<td>38</td>
<td>36</td>
<td>47</td>
<td>36</td>
<td>211</td>
</tr>
<tr>
<td>Judge C</td>
<td>60</td>
<td>38</td>
<td>64</td>
<td>64</td>
<td>31</td>
<td>257</td>
</tr>
<tr>
<td>Judge D</td>
<td>52</td>
<td>36</td>
<td>50</td>
<td>41</td>
<td>26</td>
<td>205</td>
</tr>
<tr>
<td>Judge E</td>
<td>82</td>
<td>45</td>
<td>55</td>
<td>49</td>
<td>33</td>
<td>264</td>
</tr>
<tr>
<td>Judge F</td>
<td>40</td>
<td>34</td>
<td>44</td>
<td>57</td>
<td>--</td>
<td>175</td>
</tr>
<tr>
<td>Judge G</td>
<td>65</td>
<td>38</td>
<td>30</td>
<td>53</td>
<td>34</td>
<td>220</td>
</tr>
<tr>
<td>Judge H</td>
<td>79</td>
<td>46</td>
<td>45</td>
<td>47</td>
<td>23</td>
<td>240</td>
</tr>
<tr>
<td>Judge I</td>
<td>54</td>
<td>38</td>
<td>41</td>
<td>44</td>
<td>36</td>
<td>213</td>
</tr>
<tr>
<td>Judge J</td>
<td>61</td>
<td>42</td>
<td>56</td>
<td>68</td>
<td>22</td>
<td>249</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>603</td>
<td>394</td>
<td>489</td>
<td>518</td>
<td>258</td>
<td>2,262</td>
</tr>
</tbody>
</table>

Table A2: Sentence Count for Judges

---

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”).

228 Cf. Schanzenbach & Tiller, supra note 87, at 730 (reporting that only 3% of sentences could not be matched using the docket sheet alone, mostly in immigration cases). 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. I echo their concern that this is a significant amount of missing data. See id. at 730.

229 Fiscal years 2004 and 2005 include fewer sentences because they exclude sentences imposed between Blakely and Booker. Boston cases were identified using the Commission’s parole office code, except that cases without any parole office code were included.
3. Random Distribution

Following the Hofer and Waldfogel studies, chi-square analyses were conducted to test the randomness of sentence assignment, using 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. But for the two Post-Booker periods, 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. As Table A3 shows, however, the composition of the case pool for Boston judges has not meaningfully changed from period to period:

<table>
<thead>
<tr>
<th>Periods</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 A3: Percent of Cases for Each Offense Type, by Period

---

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 41, at 320 (technical appendix).

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 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.

4. Discretionary Sentences

In evaluating guideline sentencing patterns, this study draws a distinction between “discretionary” sentences and sentences in which the judge, for legal or practical reasons, lacked the ability to sentence below the guideline range. Several recurring constraints became apparent in the course of coding thousands of case records from the District of Massachusetts.

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. For example, 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 the Boston 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. It is common, in such cases, for a judge to impose a sentence of “time served,” allowing the defendant to be released immediately. If the time served by the defendant at the time of sentencing exceeds the guideline minimum, 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, a sentence of probation is a within-range sentence. For sentences with a guideline range of 6-12 months, a sentence of probation qualifies as a

---

234 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).

235 See id. § 5G1.1 Commentary.

236 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.


238 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, which would imply 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).

239 To the extent that the “time served” constraint overlapped with other constraints, the case was coded as “time served.”

within-range sentence if it includes some conditions of intermittent, community, or home confinement. 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 probation in 9.0% 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. Many within-range sentences can be traced to statutory and practical constraints that limit their options.

B. Detailed Results

1. Regression Models

The study reports the results of ordinary least squares (OLS) regression models, using a separate linear model for each period. Dummy indicators for each judge served as independent variables. R-squared was used to measure the percentage of variance in the dependent variable explained by the identity of the sentencing judge. As in the Hofer study, the percentage of variance explained by the model is then converted into actual months, as an average across all sentences for all judges.

The two independent variables examined in the sets of regression models are sentence length and distance from the guideline range. Sentence length is measured in months of imprisonment, with a sentence of probation treated as zero months. Distance from the guideline range measures how far each sentence falls from the guideline range. Specifically, for above-range sentences, the distance was calculated as the difference between the sentence imposed and the guideline maximum. For below-range sentences, the distance was calculated as the difference between the sentence imposed and the guideline minimum. For within-range sentences, the distance was coded as zero.

The following pages provide detailed results for the regression models. The regression model for each period is described in a separate column. The dummy

---

241 Id. § 5B1.1(a)(2).
242 See Hofer et al., supra note 41, at 275 n.103 (recognizing the distorting effect of mandatory minimums).
244 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. 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, 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.
245 Hofer et al., supra note 41, 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.
246 Distance from the guideline range should always be either zero or a positive number, and a handful of cases were omitted due to logic problems, likely because the total sentence reflected consecutive sentences but the judge or the 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, in calculating distance from the guideline range, I treated a guideline minimum or maximum of life imprisonment as 470 months as well.
variables for judges, Judge4 through Judge14, appear in separate rows. The dummy
variables are nonsequential because some judges were excluded to ensure a random
distribution, and consistent with ordinary coding practices for categorical variables, one
judge (Judge 3) was omitted. Each cell reports the coefficient and, below it in
parentheses, the standard error.

Table A4 reports detailed regression results for sentence length:

<table>
<thead>
<tr>
<th></th>
<th>Pre-Booker</th>
<th>Post-Booker</th>
<th>Kimbrough/Gall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>74.38*</td>
<td>59.99*</td>
<td>84.47*</td>
</tr>
<tr>
<td></td>
<td>(5.50)</td>
<td>(6.74)</td>
<td>(10.76)</td>
</tr>
<tr>
<td>Judge4</td>
<td>–31.19*</td>
<td>–2.72</td>
<td>–14.54</td>
</tr>
<tr>
<td></td>
<td>(8.41)</td>
<td>(9.28)</td>
<td>(18.45)</td>
</tr>
<tr>
<td>Judge5</td>
<td>–24.61*</td>
<td>17.17†</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(9.06)</td>
<td>(9.60)</td>
<td></td>
</tr>
<tr>
<td>Judge6</td>
<td>–30.37*</td>
<td>–9.32</td>
<td>–42.20*</td>
</tr>
<tr>
<td></td>
<td>(8.22)</td>
<td>(9.14)</td>
<td>(17.01)</td>
</tr>
<tr>
<td>Judge7</td>
<td>–30.01*</td>
<td>–8.76</td>
<td>–19.06</td>
</tr>
<tr>
<td></td>
<td>(8.48)</td>
<td>(10.12)</td>
<td>(14.90)</td>
</tr>
<tr>
<td>Judge9</td>
<td>–22.57*</td>
<td>8.97</td>
<td>–13.66</td>
</tr>
<tr>
<td></td>
<td>(8.33)</td>
<td>(9.07)</td>
<td>(15.46)</td>
</tr>
<tr>
<td>Judge10</td>
<td>–39.13*</td>
<td>2.49</td>
<td>–9.63</td>
</tr>
<tr>
<td></td>
<td>(8.60)</td>
<td>(9.87)</td>
<td>(16.21)</td>
</tr>
<tr>
<td>Judge11</td>
<td>–23.90*</td>
<td>–26.32*</td>
<td>–39.60*</td>
</tr>
<tr>
<td></td>
<td>(8.48)</td>
<td>(10.05)</td>
<td>(14.90)</td>
</tr>
<tr>
<td>Judge13</td>
<td>–31.51*</td>
<td>–6.98</td>
<td>–8.85</td>
</tr>
<tr>
<td></td>
<td>(8.22)</td>
<td>(10.12)</td>
<td>(15.11)</td>
</tr>
<tr>
<td>Judge14</td>
<td>–30.28*</td>
<td>1.81</td>
<td>–45.90*</td>
</tr>
<tr>
<td></td>
<td>(7.81)</td>
<td>(9.84)</td>
<td>(16.79)</td>
</tr>
<tr>
<td>Number of cases</td>
<td>996</td>
<td>1006</td>
<td>257</td>
</tr>
<tr>
<td>R²</td>
<td>.029</td>
<td>.025</td>
<td>.061</td>
</tr>
<tr>
<td>Significance</td>
<td>.001*</td>
<td>.003*</td>
<td>.044*</td>
</tr>
</tbody>
</table>

* Significant at the .05 level
† Significant at the .10 level

Table A4: Linear Regression Model Results
Sentence Length, all Sentences
Table A5 reports detailed regression results for sentence length, for the subset of cases not governed by a mandatory minimum:

<table>
<thead>
<tr>
<th></th>
<th>Pre-Booker</th>
<th>Post-Booker</th>
<th>Kimbrough/Gall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>36.96*</td>
<td>34.68*</td>
<td>56.20*</td>
</tr>
<tr>
<td></td>
<td>(4.99)</td>
<td>(5.61)</td>
<td>(10.06)</td>
</tr>
<tr>
<td>Judge4</td>
<td>–6.86</td>
<td>–5.54</td>
<td>–37.08*</td>
</tr>
<tr>
<td></td>
<td>(6.92)</td>
<td>(7.90)</td>
<td>(16.30)</td>
</tr>
<tr>
<td>Judge5</td>
<td>3.28</td>
<td>15.57†</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(7.44)</td>
<td>(8.17)</td>
<td></td>
</tr>
<tr>
<td>Judge6</td>
<td>–11.40</td>
<td>–10.01</td>
<td>–24.82†</td>
</tr>
<tr>
<td></td>
<td>(7.06)</td>
<td>(7.55)</td>
<td>(13.55)</td>
</tr>
<tr>
<td>Judge7</td>
<td>–2.12</td>
<td>–2.59</td>
<td>–30.73*</td>
</tr>
<tr>
<td></td>
<td>(7.34)</td>
<td>(8.28)</td>
<td>(13.37)</td>
</tr>
<tr>
<td>Judge9</td>
<td>–3.21</td>
<td>–2.57</td>
<td>–18.08</td>
</tr>
<tr>
<td></td>
<td>(6.99)</td>
<td>(7.71)</td>
<td>(13.37)</td>
</tr>
<tr>
<td>Judge10</td>
<td>–10.89</td>
<td>11.19</td>
<td>–4.81</td>
</tr>
<tr>
<td></td>
<td>(6.97)</td>
<td>(8.37)</td>
<td>(14.23)</td>
</tr>
<tr>
<td>Judge11</td>
<td>–0.99</td>
<td>–12.43</td>
<td>–31.55*</td>
</tr>
<tr>
<td></td>
<td>(7.28)</td>
<td>(8.24)</td>
<td>(13.06)</td>
</tr>
<tr>
<td>Judge13</td>
<td>–9.10</td>
<td>–0.36</td>
<td>–21.56</td>
</tr>
<tr>
<td></td>
<td>(6.97)</td>
<td>(8.06)</td>
<td>(13.06)</td>
</tr>
<tr>
<td>Judge14</td>
<td>–11.02</td>
<td>2.80</td>
<td>–20.88</td>
</tr>
<tr>
<td></td>
<td>(7.44)</td>
<td>(7.93)</td>
<td>(12.80)</td>
</tr>
<tr>
<td>Number of cases</td>
<td>720</td>
<td>631</td>
<td>142</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.014</td>
<td>.031</td>
<td>.080</td>
</tr>
<tr>
<td>Significance</td>
<td>.368</td>
<td>.021*</td>
<td>.180</td>
</tr>
</tbody>
</table>

* Significant at the .05 level
† Significant at the .10 level

Table A5: Linear Regression Model Results
Sentence Length, Excluding Mandatory Minimums
Table A6 reports detailed regression results for distance from the guideline range:

<table>
<thead>
<tr>
<th></th>
<th>Mandatory Guidelines</th>
<th>PROTECT Act</th>
<th>Post-Booker I</th>
<th>Post-Booker II</th>
<th>Kimbrough Gall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.66 (1.76)</td>
<td>3.77 (2.77)</td>
<td>8.48* (3.14)</td>
<td>9.88* (3.78)</td>
<td>4.16 (5.41)</td>
</tr>
<tr>
<td>Judge4</td>
<td>2.39 (2.60)</td>
<td>-1.97 (4.06)</td>
<td>5.09 (4.14)</td>
<td>2.29 (5.26)</td>
<td>12.78 (8.66)</td>
</tr>
<tr>
<td>Judge5</td>
<td>0.88 (2.82)</td>
<td>-1.95 (4.09)</td>
<td>5.13 (4.56)</td>
<td>-1.34 (5.13)</td>
<td></td>
</tr>
<tr>
<td>Judge6</td>
<td>3.27 (2.57)</td>
<td>0.18 (3.89)</td>
<td>0.18 (4.24)</td>
<td>-1.90 (5.11)</td>
<td>6.67 (8.36)</td>
</tr>
<tr>
<td>Judge7</td>
<td>2.95 (2.84)</td>
<td>0.79 (4.12)</td>
<td>3.13 (4.74)</td>
<td>1.36 (5.29)</td>
<td>20.00* (7.22)</td>
</tr>
<tr>
<td>Judge9</td>
<td>3.98 (2.69)</td>
<td>6.20 (4.06)</td>
<td>-6.63 (4.14)</td>
<td>-5.04 (5.11)</td>
<td>10.80 (7.91)</td>
</tr>
<tr>
<td>Judge10</td>
<td>3.76 (2.74)</td>
<td>-1.83 (4.03)</td>
<td>-4.13 (4.45)</td>
<td>14.19* (5.60)</td>
<td>8.34 (7.73)</td>
</tr>
<tr>
<td>Judge11</td>
<td>4.57† (2.72)</td>
<td>3.65 (4.12)</td>
<td>-2.09 (4.70)</td>
<td>5.00 (5.56)</td>
<td>15.54* (7.38)</td>
</tr>
<tr>
<td>Judge13</td>
<td>3.26 (2.63)</td>
<td>-0.71 (4.09)</td>
<td>2.14 (5.08)</td>
<td>-1.29 (5.18)</td>
<td>1.14 (7.33)</td>
</tr>
<tr>
<td>Judge14</td>
<td>3.41 (2.46)</td>
<td>4.16 (3.80)</td>
<td>-0.65 (4.53)</td>
<td>-3.53 (5.38)</td>
<td>0.58 (8.24)</td>
</tr>
<tr>
<td>Number of cases</td>
<td>490</td>
<td>355</td>
<td>418</td>
<td>452</td>
<td>214</td>
</tr>
<tr>
<td>R²</td>
<td>.010</td>
<td>.024</td>
<td>.036</td>
<td>.037</td>
<td>.066</td>
</tr>
<tr>
<td>Significance</td>
<td>.847</td>
<td>.482</td>
<td>.089†</td>
<td>.048*</td>
<td>.073†</td>
</tr>
</tbody>
</table>

* Significant at the .05 level
† Significant at the .10 level

Table A6: Linear Regression Model Results
Distance from the Guideline Range, All Sentences
Finally, Table A7 reports detailed regression results for distance from the guideline range, using only “discretionary” sentences:

<table>
<thead>
<tr>
<th></th>
<th>Mandatory Guidelines</th>
<th>PROTECT Act</th>
<th>Post-Booker I</th>
<th>Post-Booker II</th>
<th>Kimbrough Gall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>2.24 (2.29)</td>
<td>4.90 (3.58)</td>
<td>9.53* (3.78)</td>
<td>11.81* (4.57)</td>
<td>5.20 (6.51)</td>
</tr>
<tr>
<td>Judge4</td>
<td>3.60 (3.43)</td>
<td>–2.36 (5.37)</td>
<td>9.33† (5.19)</td>
<td>4.66 (6.56)</td>
<td>19.47† (10.93)</td>
</tr>
<tr>
<td>Judge5</td>
<td>0.58 (3.46)</td>
<td>–2.59 (5.26)</td>
<td>5.68 (5.43)</td>
<td>–1.44</td>
<td></td>
</tr>
<tr>
<td>Judge6</td>
<td>4.41 (3.34)</td>
<td>0.20 (5.03)</td>
<td>2.41 (5.31)</td>
<td>–0.59</td>
<td>7.80 (9.94)</td>
</tr>
<tr>
<td>Judge7</td>
<td>3.40 (3.58)</td>
<td>1.44 (5.44)</td>
<td>5.08 (5.84)</td>
<td>3.07 (6.56)</td>
<td>23.43* (8.59)</td>
</tr>
<tr>
<td>Judge9</td>
<td>4.57 (3.38)</td>
<td>9.84† (5.44)</td>
<td>–7.09 (5.13)</td>
<td>–5.66</td>
<td>13.08 (9.46)</td>
</tr>
<tr>
<td>Judge10</td>
<td>4.82 (3.52)</td>
<td>–2.23 (5.26)</td>
<td>–3.23 (5.67)</td>
<td>18.08* (6.84)</td>
<td>10.59 (9.33)</td>
</tr>
<tr>
<td>Judge11</td>
<td>5.86 (3.46)</td>
<td>6.40 (5.58)</td>
<td>1.89 (6.44)</td>
<td>9.37 (7.06)</td>
<td>22.01* (9.10)</td>
</tr>
<tr>
<td>Judge13</td>
<td>4.53 (3.43)</td>
<td>–0.21 (5.26)</td>
<td>3.62 (6.23)</td>
<td>–1.28</td>
<td>0.92 (8.66)</td>
</tr>
<tr>
<td>Judge14</td>
<td>3.61 (3.09)</td>
<td>4.05 (4.77)</td>
<td>0.28 (5.48)</td>
<td>–2.91</td>
<td>0.80 (9.94)</td>
</tr>
<tr>
<td>Number of cases</td>
<td>383</td>
<td>268</td>
<td>318</td>
<td>348</td>
<td>171</td>
</tr>
<tr>
<td>R²</td>
<td>.013</td>
<td>.036</td>
<td>.045</td>
<td>.051</td>
<td>.094</td>
</tr>
<tr>
<td>Significance</td>
<td>.835</td>
<td>.384</td>
<td>.105</td>
<td>.038*</td>
<td>.037*</td>
</tr>
</tbody>
</table>

* Significant at the .05 level
† Significant at the .10 level

Table A7: Linear Regression Model Results
Distance from the Guideline Range, Discretionary Sentences
2. Alternative Time Periods

For the sake of completeness, the tables below summarize alternative regression models based on slightly different methods of dividing the 2002-2008 sentences into periods.

As discussed above, a five-period division is preferable for guideline sentencing outcomes. The PROTECT Act in 2003 was explicitly intended to reduce the rate of downward departures from the Guidelines, and changes in appellate precedent in the years following Booker are thought to have influenced district courts. Combining pre-Booker and post-Booker sentences into a single period therefore risks missing the effects of relevant legal changes.

Nonetheless, Table A8 reports the results for regression models using the same three-period division used for sentence length:

<table>
<thead>
<tr>
<th>Identity of Judge and Distance from Guideline Range, Three Periods</th>
<th>% Variance Explained</th>
<th>Avg. Variance Explained</th>
<th>Model Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Sentences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Booker</td>
<td>1.2%</td>
<td>1.7 months</td>
<td>.338</td>
</tr>
<tr>
<td>Post-Booker</td>
<td>1.7%</td>
<td>3.0 months</td>
<td>.095†</td>
</tr>
<tr>
<td>Kimbrough/Gall</td>
<td>6.6%</td>
<td>7.1 months</td>
<td>.073†</td>
</tr>
<tr>
<td>Discretionary Sentences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Booker</td>
<td>1.5%</td>
<td>2.1 months</td>
<td>.346</td>
</tr>
<tr>
<td>Post-Booker</td>
<td>2.7%</td>
<td>4.2 months</td>
<td>.031*</td>
</tr>
<tr>
<td>Kimbrough/Gall</td>
<td>9.4%</td>
<td>9.1 months</td>
<td>.037*</td>
</tr>
</tbody>
</table>

* Significant at the .05 level
† Significant at the .10 level

Table A8: Linear Regression Models

The results do not differ in any meaningful way from the models reported in the text. Both sets of models indicate that, before Booker, the relationship between the identity of the judge and distance from the guideline range was weak and not statistically significant. Both time periods also indicate that the strength of the judge effect increased after Booker, and increased sharply again after Kimbrough and Gall. If anything, the three-period models suggest an even more dramatic shift, with the strength of the relationship in the most recent period more than five times pre-Booker levels for all sentences, and more than six times pre-Booker levels for discretionary sentences.

For sentence length, as discussed above, a three-period division is preferable. Longer periods benefit the analysis in two ways: (1) by ensuring a larger number of cases per judge, which is central to the reliability of a natural experiment; and (2) by increasing the chances of identifying a statistically significant judge effect, since statistical significance is highly sensitive to sample size. And because neither the PROTECT Act

---

See supra Table 3 and accompanying text.
nor any court decision in the years immediately following *Booker* directly affected sentence length, the three-period division does not omit any potentially material legal changes.  

Nonetheless, Table A9 reports the results for regression models using the same five-period division used for guideline sentencing outcomes:

<table>
<thead>
<tr>
<th></th>
<th>% Variance Explained</th>
<th>Avg. Variance Explained</th>
<th>Model Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Sentences</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mandatory Guidelines</td>
<td>4.2%</td>
<td>13.0 months</td>
<td>.002*</td>
</tr>
<tr>
<td>PROTECT Act</td>
<td>4.7%</td>
<td>13.3 months</td>
<td>.028*</td>
</tr>
<tr>
<td>Post-Booker I</td>
<td>3.8%</td>
<td>14.1 months</td>
<td>.028*</td>
</tr>
<tr>
<td>Post-Booker II</td>
<td>2.2%</td>
<td>9.7 months</td>
<td>.264</td>
</tr>
<tr>
<td>Kimbrough/Gall</td>
<td>6.1%</td>
<td>15.5 months</td>
<td>.044*</td>
</tr>
<tr>
<td>Excluding Mandatory Minimums</td>
<td></td>
<td></td>
<td></td>
</tr>
<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>

* Significant at the .05 level  
† Significant at the .10 level

Table A9: Linear Regression Models  
Identity of Judge and Sentence Length, Five Periods

As expected, the shorter periods introduce greater uncertainty by making more of the models nonsignificant. Like the three-period models, these models indicate that the strength of the judge effect has increased after *Booker*. Each set, however, sends somewhat conflicting signals.

For all sentences, the models reveal a stronger judge effect in the Mandatory Guidelines (4.2%) and PROTECT Act periods (4.7%) individually than in the combined pre-*Booker* period (2.9%).  

They also show a substantial dip in the judge effect during the second 18-month period after *Booker*. The bottom-line finding remains the same: after *Kimbrough* and *Gall*, the judge effect is statistically significant and stronger than in any pre-*Booker* period. But the change appears more modest, 30-45% above pre-*Booker* levels rather than double pre-*Booker* levels.

For sentences not subject to a mandatory minimum, three of the five models are not statistically significant, including models for the PROTECT Act period and two post-*Booker* periods. That makes comparisons hazardous. The general trend appears similar: the judge effect grew stronger in the 18 months after *Booker* than during any previous period, and grew even stronger (although not yet statistically significant) since

---

248 *See supra* notes 217-218 and accompanying text.  
249 *See supra* Table 1 and accompanying text.
Kimbrough and Gall. But again, the change appears more modest, 60-80% above pre-
Booker levels rather than several times pre-Booker levels.²⁵⁰

These alternative models reveal that, to some extent, the change in inter-judge
sentencing disparity depends on the point of reference. Although shorter periods
introduce considerable noise, comparing a narrow slice of pre-Booker sentences with a
narrow slice of post-Booker sentences can make the change in inter-judge sentencing
disparity appear smaller, or even disappear. The natural-experiment method, however,
depends for its reliability on a sufficient number of cases per judge. The Hofer,
Waldfogel, and Anderson-Stith-Kling studies used periods of at least two years, and as
many as six years.²⁵¹ The text therefore relies, where possible and appropriate in light of
the underlying legal changes, on longer period lengths.

²⁵⁰ See supra Table 2 and accompanying text.
²⁵¹ See Hofer et al., supra note 41, at 284 (two years); Waldfogel, Empirically Based Sentencing
Guidelines, supra note 49, at 295 (three years); Anderson et al., supra note 26, at 290-91 (two years before
the Guidelines, six years after).