ARTICLES

INTER-JUDGE SENTENCING DISPARITY AFTER *BOOKER*: A FIRST LOOK

Ryan W. Scott*

*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 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.*

* 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 the court, Nancy Gertner and William Young, for their assistance and encouragement. Thanks as well to participants in the Yale Law School Sentencing Workshop, to faculty workshop participants at the Louisiana State University Law Center, and to Amy Baron-Evans, Craig Bradley, Samuel Bray, Brian Broughman, Paul Cassell, Ken Dau-Schmidt, Paul Hofer, Robert Lawless, Leandra Lederman, Andrew Martin, Michael McConnell, Marc Miller, Ben Roin, Larry Solum, Michelle Spak, David Stras, and Sandra Guerra Thompson for their comments on earlier drafts.

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' recommendations or consciously choose to impose within-range sentences for institutional reasons.

INTRODUCTION....3
I. A BRIEF HISTORY OF FEDERAL SENTENCING REFORM....6
A. *Inter-Judge Sentencing Disparity Before* Booker....6
1. *The Sentencing Reform Act of 1984*....6
2. *Mandatory Sentencing Guidelines (1987-2004)*....9
3. *PROTECT Act (2003)*....11
B. *The* Booker *Revolution, 2005-2007*....13
1. Booker, Kimbrough, *and* Gall....13
2. *Average sentence length and guideline sentencing*....14
3. *Inter-judge sentencing disparity*....19
II. THE EMPIRICAL STUDY OF INTER-JUDGE SENTENCING DISPARITY....21
A. *Data and Methods*....21
1. *Judge-specific data*....21
2. *Natural experiment method*....24
3. *Measures of inter-judge disparity*....25
4. *Why Massachusetts?*....27
B. *Results*....30
1. *Sentence length*....30
2. *Guideline sentencing patterns*....35
III. IMPLICATIONS....41
A. *Conventional Explanations for Within-Range Sentencing*....42
1. *Inertia*....42
2. *Risk aversion*....44
3. *Anchoring*....45
4. *Strategic behavior*....46
5. *Laziness*....46
B. *Alternative Explanations for Within-Range Sentencing*....47
1. *Agreement with the Guidelines' recommendations*....47
2. *Institutional considerations*....50
CONCLUSION....52
APPENDIX....53
A. *Methodological Details*....53
1. *Period selection*....53
2. *Case matching*....54
3. *Random distribution*....56
4. *Discretionary sentences*....57
B. *Detailed Results*....58
1. *Regression models*....58
2. *Alternative time periods*....63

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*,[1] 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.[2] Three years later, in *Gall v. United States*,[3] the Court directed appellate courts to review sentencing decisions under a "deferential abuse-of-discretion standard."[4] And on the same day, in *Kimbrough v. United States*,[5] 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."[6]

In the wake of those decisions, the Commission has received extensive anecdotal reports of a surge in inter-judge sentencing disparity. The Department of Justice reported in a June 2010 memorandum that "[m]ore and more, we are receiving reports from our prosecutors that in many federal courts, a defendant's sentence will largely be determined by the judicial assignment of the case; i.e., which judge in the courthouse will conduct the sentencing."[7] 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."[8] Prosecutors around the country echoed those concerns at the Commission's 2009-2010 re-

1. 543 U.S. 220 (2005).

2. *Id.* at 245 (Breyer, J., delivering the opinion of the Court in part).

3. 552 U.S. 38 (2007).

4. *Id.* at 52-53.

5. 552 U.S. 85 (2007).

6. *Id.* at 101 (internal quotation marks omitted).

7. Memorandum from Jonathan J. Wroblewski, Dir., Office of Policy & Legislation, U.S. Dep't of Justice, to Hon. William K. Sessions III, U.S. Sentencing Comm'n 2 (June 28, 2010) [hereinafter Wroblewski Memorandum], *available at* http://sentencing.typepad.com/files/annual_letter_2010_final_062810.pdf.

8. Eric Holder, U.S. Att'y Gen., Remarks for the Charles Hamilton Houston Institute for Race and Justice and Congressional Black Caucus Symposium: Rethinking Federal Sentencing Policy, 25th Anniversary of the Sentencing Reform Act (June 24, 2009) (transcript available at http://www.usdoj.gov/ag/speeches/2009/ag-speech-0906241.html).

gional hearings.[9] 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."[10] Prosecutors have reported a similar spike in inter-judge disparity in "nearly all districts" in the Ninth Circuit.[11] Frank Bowman calls the Supreme Court's decisions a "debacle,"[12] and warns that in white-collar cases, "we're back to a pre-guidelines era" marked by "disparity and the most potential for disparity."[13]

Those reports, if accurate, deserve urgent attention because they implicate Congress's core objective in reforming federal sentencing. Inter-judge sentencing disparity, in the view of sentencing reformers, offends important rule-of-law principles, erodes respect for the courts, and undermines the deterrent effect of the criminal law. Congress, if it wishes, has several options available to address the problem by altering the Sentencing Guidelines to resolve the constitutional defects identified by the Supreme Court.

To date, however, the evidence of an uptick in inter-judge disparity 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. The 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 doubled in strength.[14] 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

9. *See infra* notes 113-17 and accompanying text.

10. Patrick J. Fitzgerald, U.S. Att'y, N. Dist. Ill., Statement Before the U.S. Sentencing Commission in the Regional Hearing on the State of Federal Sentencing 3 (Sept. 10, 2009) [hereinafter Fitzgerald Statement] (transcript available at http://www.ussc.gov/AGENDAS/20090909/Fitzgeraldtestimony.pdf).

11. Karin J. Immergut, U.S. Att'y, Dist. Or., Statement Before the U.S. Sentencing Commission in the Regional Hearing on the State of Federal Sentencing 12 (May 27, 2009) [hereinafter Immergut Statement] (transcript available at http://www.ussc.gov/AGENDAS/20090527/Immergut_testimony.pdf).

12. Frank O. Bowman, III, *Debacle: How the Supreme Court Has Mangled American Sentencing Law and How It Might yet Be Mended*, 77 U. CHI. L. REV. 367, 368 (2010).

13. Amir Efrati, *Looser Rules on Sentencing Stir Concerns About Equity*, WALL ST. J., Nov. 5, 2009, at A15.

14. *See infra* text accompanying note 175.

those judges is assigned the case.[15]

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.[16] 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*.[17]

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.

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' recommendations, or may elect to impose within-range sentences for institutional reasons.

15. *See infra* Figure 5 and accompanying text.

16. *See infra* notes 180-82 and accompanying text.

17. *See infra* Table 3 and accompanying text.

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.[18] 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.[19] Judges were under no obligation to give reasons for the sentence imposed,[20] and appellate review of sentencing decisions was virtually nonexistent.[21] 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."[22]

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. For prominent scholars, the evidence of "inter-judge sentencing disparity"—differences in sentencing outcomes caused by the judge, rather than by legitimate differences between offenses and offenders[23]—was overwhelming.[24] Many judges had developed a

18. KATE STITH & JOSÉ A. CABRANES, FEAR OF JUDGING: SENTENCING GUIDELINES IN THE FEDERAL COURTS 9-11 (1998). Parole boards added an additional layer of indeterminacy to federal sentences. The Sentencing Reform Act of 1984 abolished parole in the federal system.

19. *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, § 2(a), 48 Stat. 783, 783 (current version at 18 U.S.C. § 2113 (2006)); *see also* Jerome v. United States, 318 U.S. 101, 101-02 (1943).

20. Kevin R. Reitz, *Sentencing*, *in* THE HANDBOOK OF CRIME AND PUNISHMENT 542, 543 (Michael Tonry ed., 1998).

21. STITH & CABRANES, *supra* note 18, at 9 & 197 n.3.

22. United States v. Mueffelman, 327 F. Supp. 2d 79, 83 (D. Mass. 2004) (Gertner, J.); *see also* Douglas A. Berman, *Conceptualizing* Booker, 38 ARIZ. ST. L.J. 387, 389 (2006).

23. *See* James M. Anderson et al., *Measuring Interjudge Sentencing Disparity: Before and After the Federal Sentencing Guidelines*, 42 J.L. & ECON. 271, 274 (1999) (defining "disparity" as "solely that variation caused by the identity of the decision maker"). What counts as a "legitimate" difference between cases justifying a higher or lower sentence is, of course, heavily contested and dependent on some underlying theory of punishment. *See* Kevin Cole, *The Empty Idea of Sentencing Disparity*, 91 NW. U. L. REV. 1336, 1337 (1997);

reputation as especially harsh or lenient at sentencing, and numerous simulation studies found wide disparity in the sentences chosen by different judges presented with identical case facts.[25]

Reformers saw inter-judge disparity as problematic for several reasons. One was that inter-judge disparity threatens core rule-of-law principles. Judge Marvin Frankel, the most influential critic of indeterminate sentencing in the 1970s, called judges' unchecked discretion at sentencing "terrifying and intolerable for a society that professes devotion to the rule of law."[26] 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."[27] Although sentencing decisions properly take into account a wide range of facts and considerations, no one defends the proposition that sentencing outcomes should depend on the judge's politics, personality, or biases. Another was that inter-judge disparity erodes confidence in the courts by creating the appearance of unfairness and arbitrariness. As the Department of Justice recently reiterated, disparities between judges over time "breed disrespect" for courts, threatening the effectiveness of the criminal justice system.[28] During the 1970s, for example, federal corrections officials called sentencing disparity one of the "major causes of prison riots" because it fueled anger and resentment among prisoners.[29] Similarly, inter-judge disparity was seen as rendering the level of punishment less certain and predictable, undermining the deterrent effect of the criminal law.[30]

Michael M. O'Hear, *The Original Intent of Uniformity in Federal Sentencing*, 74 U. CIN. L. REV. 749, 749-50 (2006). But Congress concluded that inter-judge disparity, driven by judicial preferences and biases rather than offense and offender characteristics, is unwarranted. *See* 18 U.S.C. § 3553(a)(6) (2006) (directing judges to impose sentences so as "to avoid unwarranted sentence disparities among defendants with similar records who have been found guilty of similar conduct"); S. REP. NO. 98-225, at 45 (1983) ("Sentencing disparities that are not justified by differences among offenses or offenders are unfair both to offenders and to the public.").

24. *E.g.*, Norval Morris, *Towards Principled Sentencing*, 37 MD. L. REV. 267, 274 (1977).

25. *See* ANTHONY PARTRIDGE & WILLIAM B. ELDRIDGE, FED. JUDICIAL CTR., THE SECOND CIRCUIT SENTENCING STUDY: A REPORT TO THE JUDGES OF THE SECOND CIRCUIT 36 (1974); Kevin Clancy et al., *Sentence Decisionmaking: The Logic of Sentence Decisions and the Extent and Sources of Sentence Disparity*, 72 J. CRIM. L. & CRIMINOLOGY 524, 525-26 (1981); Shari Seidman Diamond & Hans Zeisel, *Sentencing Councils: A Study of Sentence Disparity and Its Reduction*, 43 U. CHI. L. REV. 109, 119-24 (1975) (analyzing the recommendations of "sentencing councils" in which panels of judges not assigned to a case would review the file and choose a sentence independently, then consult with the sentencing judge).

26. MARVIN E. FRANKEL, CRIMINAL SENTENCES: LAW WITHOUT ORDER 5 (1973).

27. *Id.* at 10.

28. Wroblewski Memorandum, *supra* note 7, at 2.

29. Diamond & Zeisel, *supra* note 25, at 110-11 (quoting J. BENNETT, OF PRISONS AND JUSTICE, S. DOC. NO. 88-70, at 319 (1964)).

30. Stephen J. Schulhofer & Ilene H. Nagel, *Negotiated Pleas Under the Federal Sentencing Guidelines: The First Fifteen Months*, 27 AM. CRIM. L. REV. 231, 237 (1989)

Following more than a decade of debate, Congress enacted the Sentencing Reform Act of 1984.[31] A principal purpose of the Act was to reduce inter-judge disparity in sentencing.[32] Congress concluded that, too often, similarly situated offenders received unjustifiably disparate sentences, solely because of the preferences and biases of the judge assigned to the case.[33] 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.[34] But there was remarkable bipartisan agreement that unfettered discretion had resulted in an intolerable level of inter-judge sentencing disparity.[35]

To reduce inter-judge disparity, the Act created the United States Sentencing Commission, "an independent commission in the judicial branch of the United States."[36] 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.[37] It provided that guidelines and amendments adopted by the Commission must be submitted to Congress for a period of review; unless they were "modified or disapproved" by Congress, they would go into effect automatically.[38] 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;[39] and (2) in "exceptional

("These disparities not only fostered undue optimism among offenders who hoped to 'beat the rap,' they also undermined deterrence and crime control objectives.").

31. Pub. L. No. 98-473, 98 Stat. 1837 (codified as amended in scattered sections of 18 and 28 U.S.C.).

32. U.S. SENTENCING COMM'N, SENTENCING GUIDELINES AND POLICY STATEMENTS 1.2 (1987); Stephen Breyer, *The Federal Sentencing Guidelines and the Key Compromises upon Which They Rest*, 17 HOFSTRA L. REV. 1, 4 (1988); *see also* Susan R. Klein & Jordan M. Steiker, *The Search for Equality in Criminal Sentencing*, 2002 SUP. CT. REV. 223, 232-33 (describing "the reduction of unwarranted disparity in sentencing" as "Congress's stated goal" in sentencing reform); Ilene H. Nagel, *Structuring Sentencing Discretion: The New Federal Sentencing Guidelines*, 80 J. CRIM. L. & CRIMINOLOGY 883, 895-99 (1990).

33. U.S. SENTENCING GUIDELINES MANUAL § 1A1.3 introductory cmt. (1987) ("Congress sought reasonable uniformity in sentencing by narrowing the wide disparity in sentences imposed for similar criminal offenses committed by similar offenders.").

34. STITH & CABRANES, *supra* note 18, at 38-48.

35. The Act was co-sponsored by strange bedfellows in the Senate: Ted Kennedy and Strom Thurmond. *Id.* at 38-39.

36. 28 U.S.C. § 991(a) (2006); *see also id.* §§ 994, 995(a)(1). The Commission's composition and location "in the judicial branch" are unusual, and scores of federal courts struck down the Act as unconstitutional before the Supreme Court rejected a separation of powers challenge in *Mistretta v. United States*, 488 U.S. 361, 380-412 (1989).

37. 28 U.S.C. § 994(a) (2006).

38. *Id.* § 994(p).

39. U.S. SENTENCING GUIDELINES MANUAL § 5K1.1 (2009); *see also* 28 U.S.C.

case[s]"[40] 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."[41] 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.[42] It also provided for appellate review of sentencing range calculations and for review of sentences outside the guideline range for abuse of discretion.[43]

2. *Mandatory Sentencing Guidelines (1987-2004)*

The Commission promulgated the first Federal Sentencing Guidelines in 1987, and the mandatory guidelines regime remained essentially intact for eighteen 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,[44] for their inflexibility,[45] and for transferring too much power to prosecutors making charging and plea bargaining decisions.[46]

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.[47] 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-

§ 994(n) (2006).

40. U.S. SENTENCING GUIDELINES MANUAL § 5K2.0 (2009).

41. 18 U.S.C. § 3553(b)(1) (2006). This was one of the provisions excised by the remedial opinion in *Booker*. *See infra* Part I.B.

42. 18 U.S.C. § 3553(c)(1)-(2) (2006).

43. *Id.* § 3742(a)-(b). It was not until 1996 that the Supreme Court clarified that the standard of appellate review was "abuse of discretion." *See* Koon v. United States, 518 U.S. 81, 92-100 (1996). The Act's appellate review provision was excised in *Booker*. *See infra* Part I.B.

44. *See* STITH & CABRANES, *supra* note 18, at 59-64; Paul G. Cassell, *Too Severe? A Defense of the Federal Sentencing Guidelines (and a Critique of Federal Mandatory Minimums)*, 56 STAN. L. REV. 1017, 1018 (2004).

45. *See, e.g.*, Vincent L. Broderick, *The Importance of Flexibility in Sentencing*, 78 JUDICATURE 182, 182 (1995); Daniel J. Freed, *Federal Sentencing in the Wake of Guidelines: Unacceptable Limits on the Discretion of Sentencers*, 101 YALE L.J. 1681, 1719-20, 1725-27 (1992); Gerald Heaney, *No End to Disparity*, 28 AM. CRIM. L. REV. 161 (1991); Marc L. Miller, *Domination & Dissatisfaction: Prosecutors as Sentencers*, 56 STAN. L. REV. 1211, 1236 (2004); Daniel Zlotnick, *The Future of Federal Sentencing Policy: Learning Lessons from Republican Judicial Appointees*, 79 U. COLO. L. REV. 1, 27-28 (2008).

46. Kate Stith, *The Arc of the Pendulum: Judges, Prosecutors, and the Exercise of Discretion*, 117 YALE L.J. 1420, 1430 (2008).

47. *See* Anderson et al., *supra* note 23, at 303; Paul J. Hofer et al., *The Effect of the Federal Sentencing Guidelines on Inter-Judge Sentencing Disparity*, 90 J. CRIM. L. & CRIMINOLOGY 239, 241, 291, 296 (1999).

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.[48]

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, examined a sample of cases from approximately twenty-five district offices nationwide in which the case distribution system was deemed sufficiently random.[49] 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%.[50]

The second, by Paul Hofer of the Sentencing Commission and two colleagues, compared a sample of cases from cities with a random case distribution system in two time periods, 1984-1985 and 1994-1995.[51] 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."[52] Using a larger sample from forty-one 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.[53] 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.[54]

These studies, and other similar efforts by Joel Waldfogel and Abigail Payne,[55] offer the best available evidence of the effect of the Guidelines on in-

48. Anderson et al., *supra* note 23, at 291; Hofer et al., *supra* note 47, at 282.

49. Anderson et al., *supra* note 23, at 290 tbl.2.

50. *Id.* at 303.

51. Hofer et al., *supra* note 47, at 284.

52. *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).

53. *Id.* at 293-94.

54. *Id.* at 298.

55. *See, e.g.*, A. Abigail Payne, *Does Inter Judge Disparity Really Matter? An Analysis of the Effects of Sentencing Reforms in Three Federal District Courts*, 17 INT'L REV. L. & ECON. 337 (1997) (using data from 1980 to 1991 for select types of cases in three federal district courts); Joel Waldfogel, *Aggregate Inter-Judge Disparity in Federal Sentencing: Evidence from Three Districts*, 4 FED. SENT'G REP. 151 (1991) (using data from three different district courts from 1984 to 1987); Joel Waldfogel, *Does Inter-Judge Disparity Justify Empirically Based Sentencing Guidelines?*, 18 INT'L REV. L. & ECON. 293 (1998) [hereinaf-

ter-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.[56] 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.[57] 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.[58]

3. *PROTECT Act (2003)*

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 of 2003 (PROTECT Act).[59] Championed by Representative Tom Feeney and dubbed the "Feeney Amendment," the provisions responded to concerns in Congress and the Department of Justice about the prevalence of downward departures from the Guidelines.[60] 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%.[61] The Commission later realized that the 2001 rate was incorrect.[62]

Among other changes, the PROTECT Act (1) tightened the standard of appellate review for nonguideline sentences, replacing the abuse of discretion standard with de novo review;[63] (2) directed the Commission to amend the

ter Waldfogel, *Empirically Based Sentencing*] (employing a natural experiment regression analysis using data from ten judges in San Francisco from 1984 to 1987).

56. *See* Anderson et al., *supra* note 23, at 302; Hofer et al., *supra* note 47, at 299-302.

57. *See* Anderson et al., *supra* note 23, at 299.

58. *See infra* notes 154-55 and accompanying text.

59. Pub. L. No. 108-21, § 401, 117 Stat. 650, 667-76 (codified as amended in scattered sections of 18, 28, and 42 U.S.C.).

60. *See* H.R. REP. NO. 108-66, at 58 (2003) (Conf. Rep.) (announcing an intention to address "the longstanding problem of downward departures from the Federal Sentencing Guidelines"); Stith, *supra* note 46, at 1465.

61. U.S. SENTENCING COMM'N, DOWNWARD DEPARTURES FROM THE SENTENCING GUIDELINES, at iv-v, 59-60 (2003); Stith, *supra* note 46, 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 also* Miller, *supra* note 45, at 1228 fig.1.

62. *See infra* notes 71-72 and accompanying text.

63. PROTECT Act § 401(d)(2).

Guidelines "to ensure that the incidence of downward departures are [sic] substantially reduced";[64] (3) prohibited the Commission from recognizing new permissible grounds for downward departure for two years;[65] and (4) directed the Department of Justice to resist downward departures "not supported by the facts and the law."[66] The PROTECT Act also directly amended the Guidelines by adding specific upward adjustments for sex offenders and child pornography cases.[67]

The PROTECT Act sentencing provisions drew strong criticism from scholars, judges, interest groups, and the defense bar.[68] Responding to an earlier version of the Act that would have eliminated *all* grounds for downward departure in the Guidelines, 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."[69] 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.[70]

In hindsight, it is clear that reports of an epidemic of judge-initiated downward departures were exaggerated.[71] 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.[72]

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.

64. *Id.* § 401(m)(2)(A).

65. *Id.* § 401(j)(2).

66. *Id.* § 401(*l*)(1)(A).

67. *Id.* § 401(i).

68. *See* Noelle Tsigounis Valentine, Note, *An Exploration of the Feeney Amendment: The Legislation that Prompted the Supreme Court to Undo Twenty Years of Sentencing Reform*, 55 SYRACUSE L. REV. 619, 628-29 (2005).

69. Alan Vinegrad, *The New Federal Sentencing Law*, 15 FED. SENT'G REP. 310, 313 (2003).

70. *See* Letter from Leonidas Ralph Mecham, Sec'y, Judicial Conference of the U.S., to Senator Orrin G. Hatch, Chairman, Comm. on the Judiciary 3 (Apr. 3, 2003), *available at* http://www.nacdl.org/public.nsf/2cdd02b415ea3a64852566d6000daa79/departures/$FILE/judconf_feeney.pdf.

71. Max Schanzenbach, *Have Federal Judges Changed Their Sentencing Practices? The Shaky Empirical Foundations of the Feeney Amendment*, 2 J. EMPIRICAL L. STUD. 1, 1 (2005); Stith, *supra* note 46, at 1464-65.

72. U.S. SENTENCING COMM'N, *supra* note 61, at iv.

B. *The* Booker *Revolution, 2005-2007*

1. Booker*,* Kimbrough*, and* Gall

In January 2005, the Supreme Court held in *United States v. Booker*[73] that the Sentencing Reform Act violated the Sixth Amendment right to trial by jury.[74] The Court's fractured decision consisted of two majority opinions.[75] One opinion, written by Justice Stevens, extended the rule of *Apprendi v. New Jersey*[76] and *Blakely v. Washington*[77] 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.[78]

The second majority 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.[79] Excising those provisions, the Court explained, "makes the Guidelines effectively advisory."[80] Judges must continue to calculate the applicable sentencing range, the Court explained, but need only "consider" it, along with the factors identified in § 3553(a), in imposing a sentence.[81]

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.[82] In *Gall v. United States*,[83] 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 sen-

73. 543 U.S. 220 (2005).

74. *Id.* at 226-27, 243-44.

75. The case prompted six separate opinions, including two principal majorities and two principal dissents. *Id.* at 225.

76. 530 U.S. 466 (2000).

77. 542 U.S. 296 (2004).

78. *Booker*, 543 U.S. at 244 (Stevens, J., delivering the opinion of the Court in part).

79. *Booker*, 543 U.S. at 245 (Breyer, J., delivering the opinion of the Court in part).

80. *Id.*

81. *Id.* at 245, 259-60.

82. *See* Ryan Scott Reynolds, Note, *Equal Justice Under Law: Post*-Booker*, Should Federal Judges Be Able to Depart from the Federal Sentencing Guidelines to Remedy Disparity Between Codefendants' Sentences?*, 109 COLUM. L. REV. 538, 560, 563-64 (2009) (noting that both cases expanded district courts' discretion and that some courts of appeals have responded by reconsidering their treatment of particular sentencing factors); *The Supreme Court, 2007 Term—Leading Cases*, 122 HARV. L. REV. 276, 333 (2008) (asserting that *Kimbrough* and *Gall* "appear to loosen the hold of the Guidelines").

83. 552 U.S. 38 (2007).

tence.[84] 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]."[85]

Simultaneously in *Kimbrough v. United States*,[86] 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 one hundred grams of powder cocaine.[87] 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."[88] Although it suggested that "closer review may be in order" in those circumstances,[89] the Court left little doubt that judges now enjoy the freedom to categorically reject the Commission's judgments about sentencing policy.[90]

2. *Average sentence length and guideline sentencing*

Longtime critics of the Guidelines greeted *Booker* with enthusiasm,[91] but the decision did not prompt immediate changes in sentencing outcomes. Average sentence length actually increased for several years after *Booker*,[92] even for drug trafficking offenses. The rate of below-guideline sentencing jumped, but quickly leveled out, and the change was hardly "earth-shattering."[93] Many

84. *Id.* at 47.

85. *Id.* at 51-52.

86. 552 U.S. 85 (2007).

87. *Id.* at 109-10.

88. *Id.* at 101-02 (quoting Brief for Respondent at 16, *Kimbrough*, 552 U.S. 85 (No. 06-6330)) (internal quotation marks and alteration omitted).

89. *Id.* at 109.

90. *See, e.g.*, 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" (quoting Spears v. United States, 129 S. Ct. 840, 844 (2009))).

91. For a collection of initial reactions, see Erik Luna & Barton Poulson, *Restorative Justice in Federal Sentencing: An Unexpected Benefit of* Booker*?*, 37 MCGEORGE L. REV. 787, 787-88 (2006). Two federal judges in the District of Massachusetts publicly praised *Booker* shortly after it was announced. *See* Shelley Murphy, *Two Boston Jurists Hail Return of Discretion*, BOSTON GLOBE, Jan. 13, 2005, at A20.

92. U.S. SENTENCING COMM'N, FINAL REPORT ON THE IMPACT OF *UNITED STATES V. BOOKER* ON FEDERAL SENTENCING, at vii (2006) [hereinafter FINAL REPORT], *available at* http://www.ussc.gov/booker_report/Booker_Report.pdf (documenting a modest increase in average sentence length and concluding that "[t]he severity of sentences imposed has not changed substantially").

93. Max M. Schanzenbach & Emerson H. Tiller, *Reviewing the Sentencing Guidelines: Judicial Politics, Empirical Evidence, and Reform*, 75 U. CHI. L. REV. 715, 739 (2008); *see also* Douglas A. Berman, *Tweaking* Booker*: Advisory Guidelines in the Federal System*, 43 HOUS. L. REV. 341, 349 (2006) ("[D]ata on post-*Booker* sentencing outcomes re-

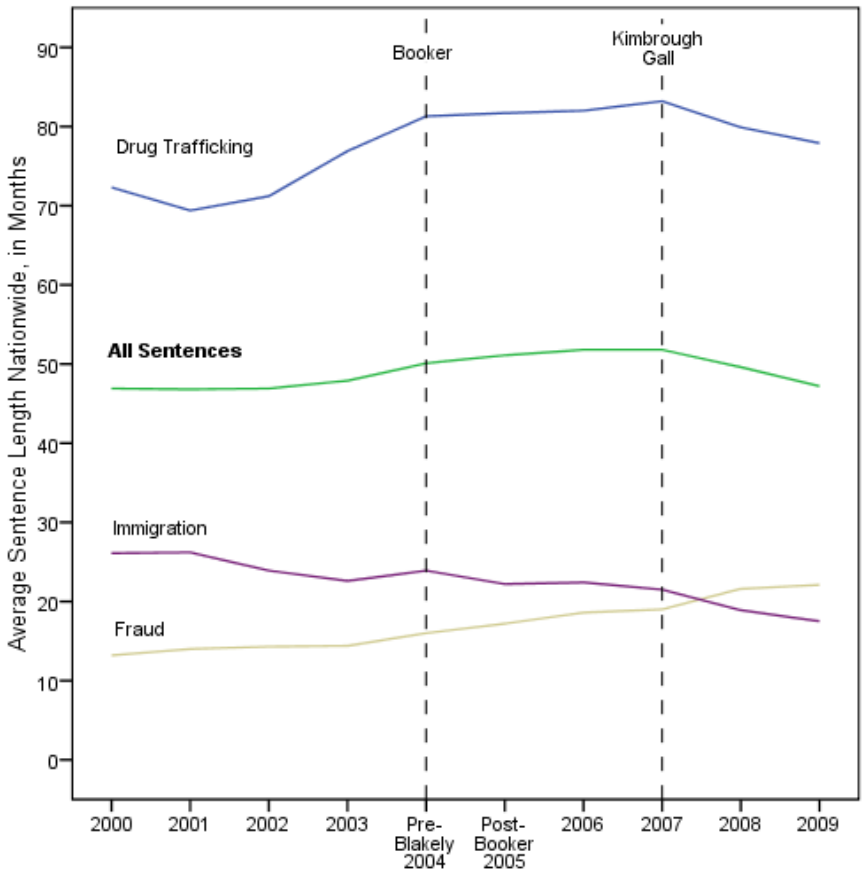
commentators lamented that, far from ushering in a revolution, the decision turned out to be a dud.[94]

It would be premature to pronounce the Supreme Court's sentencing decisions a nonevent. 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*.

leased by the Commission reveal only relatively small changes in the patterns of sentencing outcomes.").

94. *See* Schanzenbach & Tiller, *supra* note 93, at 739 (noting that "most observers" believe "the fundamentals of sentencing changed little post-*Booker*"); *see also* Berman, *supra* note 93, at 348 ("[T]he *Booker* decision does not appear to have radically transformed either basic practices or typical outcomes in the federal sentencing system."); Frank O. Bowman, III, *The Year of Jubilee . . . or Maybe Not: Some Preliminary Observations About the Operation of the Federal Sentencing System After* Booker, 43 HOUS. L. REV. 279, 319 (2006) (calling the changes "strikingly modest"); D. Michael Fisher, *Striking a Balance: The Need To Temper Judicial Discretion Against a Background of Legislative Interest in Federal Sentencing*, 46 DUQ. L. REV. 65, 77-78 (2007) ("While the change is noticeable, it does not reflect the fear of some post-*Booker* commentators that judges, now invested with a new kind of discretion, would ignore the Guidelines and sentence defendants however they saw fit."); Jeffrey S. Hurd, *Federal Sentencing and the Uncertain Future of Reasonableness Review*, 84 DEN. U. L. REV. 835, 860 (2007); Michael M. O'Hear, *The Duty To Avoid Disparity: Implementing 18 U.S.C. § 3553(a)(6) After* Booker, 37 MCGEORGE L. REV. 627, 645 (2006); Zlotnick, *supra* note 45, at 15.

FIGURE 1
Average Sentences Nationwide, by Fiscal Year[95]

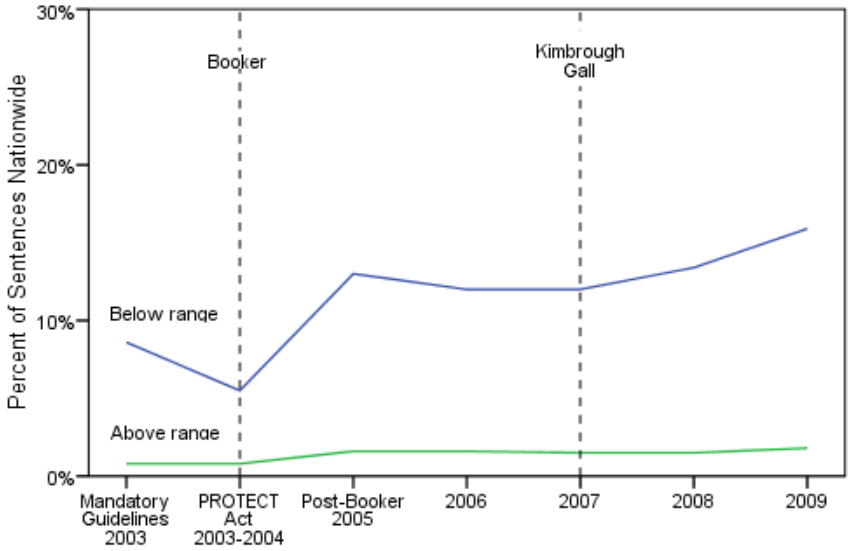


Average sentence length nationwide rose from 50.1 months in fiscal year 2004, immediately before *Booker*, to 51.8 months in fiscal years 2006 and 2007. But since *Kimbrough* and *Gall*, average sentence length has fallen to 46.8 months. Similarly, sentences for drug trafficking offenses rose from 81.3 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 fiscal year 2009. They are now less severe than before *Booker*.

95. The data for this Figure comes from table 13 in the 2000-2009 editions of the U.S. Sentencing Commission's *Sourcebook of Federal Sentencing Statistics* and table 19 in the 2006-2009 editions of the Commission's *4th Quarter Preliminary Data Reports*. The fiscal year 2004 period ends on June 24, 2004, the date of the Supreme Court's decision in *Blakely v. Washington*, 542 U.S. 296 (2004). FINAL REPORT, *supra* note 92, at 71 tbl.3.

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 2
Guideline Sentencing, by Fiscal Year[96]



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, compared with 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. Preliminary data for the first half of fiscal year 2010 indicate that the rate of below-range sentencing has jumped to 16.9%.[97] That means the rate of below-range sentencing is creeping close to the (incorrectly reported) 18.1% rate that prompted Congress to intervene with the PROTECT Act.[98] The percentage of above-range sentences also has more than doubled, from 0.8% before *Booker* to 1.8% after *Kimbrough* and *Gall*.

96. The data for this Figure come from figure G and table N in the 2000-2009 editions of the U.S. Sentencing Commission's *Sourcebook of Federal Sentencing Statistics* and table 1 in the 2006-2009 editions of the Commission's *4th Quarter Preliminary Data Reports*. *See also* FINAL REPORT, *supra* note 92, app. E-1. Percentages for all post-*Booker* periods combine traditional departures with nonguideline sentences based on the § 3553(a) factors (sometimes called "variances").

97. U.S. SENTENCING COMM'N, PRELIMINARY QUARTERLY DATA REPORT 1 tbl.1 (2010), *available at* http://www.ussc.gov/sc_cases/USSC_2010_Quarter_Report_2nd.pdf.

98. *See supra* note 61 and accompanying text.

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 in 1987 and the *Booker* decision in 2005.[99] It should not be surprising, the argument goes, that judges who have spent their entire careers treating the Guidelines as mandatory continue to follow them in the great majority of cases even though they are now advisory.[100]

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."[101] In *Rita v. United States*,[102] the Supreme Court held that courts of appeals may presume that a within-guideline sentence is reasonable.[103] A judge anxious to avoid having a sentence vacated on appeal therefore has an incentive to stay within the Guidelines.[104]

99. *See Biographical Directory of Federal Judges*, FED. JUD. CENTER, http://www.fjc.gov/history/home.nsf/page/judges.html (last visited Oct. 11, 2008). There are 1016 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.

100. *See, e.g.*, 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 46, 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).

101. United States v. Booker, 543 U.S. 220, 263-64 (2004) (Breyer, J., delivering the opinion of the Court in part).

102. 551 U.S. 338 (2007).

103. *Id.* at 347.

104. Nancy Gertner, *What Yogi Berra Teaches About Post-*Booker *Sentencing*, 115 YALE L.J. POCKET PART 137, 140 (2006), http://www.thepocketpart.org/images/pdfs/50.pdf (describing decisions of appellate courts that reinforce the Guidelines and reporting that "[d]istrict judges have gotten the message"); Jack King, *Up, Down or Lazy? Panelists Discuss Federal Sentencing After* Rita, CHAMPION, Sept.-Oct. 2007, at 8-9; *see also* Kevin R. Reitz, *The Enforceability of Sentencing Guidelines*, 58 STAN. L. REV. 155, 171 (2005) (concluding that the post-*Booker* Guidelines "remain as restrictive of judicial sentencing discretion as any system in the United States"). The incentive against departure from the Guide-

A third proposed explanation is "anchoring," the well-documented cognitive error in which decision makers begin with an initial value, even one that is irrational, and fail to make rational adjustments.[105] 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.[106] Presumably sentencing guidelines, which judges know to be nonarbitrary, 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,[107] it should not be surprising if the Guidelines continue to exert a powerful influence despite being advisory.[108]

A fourth proposed explanation is strategic behavior. The idea is that judges, eager to safeguard the sentencing discretion they gained in *Booker*, have taken a "go slow" approach to reduce the risk of interbranch retaliation.[109] On this theory, judges secretly desire to sentence outside the Guidelines more often, but have restrained themselves to avoid provoking Congress.[110]

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."[111] 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.[112]

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*,

lines was at least as strong before *Booker*. *See* Stephanos Bibas, *The Feeney Amendment and the Continuing Rise of Prosecutorial Power to Plea Bargain*, 94 J. CRIM. L. & CRIMINOLOGY 295, 302 (2004).

105. Gertner, *supra* note 100, at 270; *see also* Cass R. Sunstein, *Behavioral Analysis of Law*, 64 U. CHI. L. REV. 1175, 1188 (1997); Amos Tversky & Daniel Kahneman, *Judgment Under Uncertainty: Heuristics and Biases*, 185 SCIENCE 1124, 1128-30 (1974).

106. *See* Birte Englich et al., *Playing Dice with Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making*, 32 PERSONALITY & SOC. PSYCHOL. BULL. 188, 194 (2006).

107. Gall v. United States, 552 U.S. 38, 49 (2007).

108. Gertner, *supra* note 104, at 138; Stith, *supra* note 46, at 1496.

109. *See* Zlotnick, *supra* note 45, at 14-15; Daniel A. Chatham, Note, *Playing with Post-*Booker *Fire: The Dangers of Increased Judicial Discretion in Federal White-Collar Sentencing*, 32 J. CORP. L. 619, 637-38 (2007) (recommending this approach).

110. *Cf.* Jack B. Weinstein, *The Role of Judges in a Government of, by, and for the People: Notes for the Fifty-Eighth Cardozo Lecture*, 30 CARDOZO L. REV. 1, 211 (2008). 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.

111. *Id.*

112. King, *supra* note 104, at 9 (quoting Judge Myron Thompson).

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 "reintroduced into federal sentencing both substantial district-to-district variations and substantial judge-to-judge variations."[113] A survey of other districts in the Ninth Circuit revealed that "nearly all emphasize the wide variation seen between different judges within their districts."[114] In the Eastern District of New York, it appears "the range of variation *between* judges in [the same] courthouse has grown" since *Booker*.[115] In Oregon, "sentencing tendencies have always been somewhat unique to each individual judge, but the differences since *Booker* have become more pronounced."[116] In the Eastern District of Virginia, some judges "are more inclined to use the freedom granted by *Booker* and its progeny" than others.[117] Summing up its position in a July 2010 memorandum, the Department of Justice reported that "[m]ore and more, we are receiving reports from our prosecutors that in many federal courts, a defendant's sentence will largely be determined by . . . which judge in the courthouse will conduct the sentencing."[118]

Those claims, if accurate, deserve urgent attention. Congress's central objective in reforming federal sentencing was to reduce inter-judge disparity,[119] and on that score the mandatory Guidelines can claim success. If *Booker* has compromised that progress, Congress could take corrective action. Consistent with the Court's Sixth Amendment cases, Congress has several options for avoiding the constitutional problem.[120] The most straightforward, charted by Justice Stevens in his dissent from the remedial opinion in *Booker*, is to "*Blakely*-ize" the Guidelines by affording criminal defendants a jury trial right with respect to aggravating factors that increase the otherwise-applicable guideline maximum.[121] Some state guideline systems have followed that approach,[122]

113. Fitzgerald Statement, *supra* note 10, at 3.

114. Immergut Statement, *supra* note 11, at 12 (warning that "the signs point to increasing sentencing disparity—including disparity based on differing judicial philosophies among judges working in the same courthouse").

115. Benton J. Campbell, U.S. Att'y, E. Dist. N.Y., Statement Before the U.S. Sentencing Commission in the Regional Hearing on the State of Federal Sentencing 8 (July 9, 2009) (transcript available at http://www.ussc.gov/AGENDAS/20090709/Campbell_testimony.pdf).

116. Immergut Statement, *supra* note 11, 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").

117. Dana Boente, U.S. Att'y, E. Dist. Va., Statement Before the U.S. Sentencing Commission in the Regional Hearing on the State of Federal Sentencing 3 (July 9, 2009) (transcript available at http://www.ussc.gov/AGENDAS/20090709/Boente_testimony.pdf).

118. Wroblewski Memorandum, *supra* note 7, at 2.

119. *See supra* notes 32-35 and accompanying text.

120. For an overview of proposed reforms, see Berman, *supra* note 93, at 356-71.

121. *Id.* at 364-65 (discussing proposals to "*Blakely*-ize" the Guidelines); *see also* Unit-

and the American Law Institute is poised to recommend it as part of the new Model Penal Code provisions on sentencing.[123]

Yet claims of an uptick in inter-judge sentencing disparity are exceedingly difficult to evaluate because the changes are almost impossible to detect. Consistent with its longstanding 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 twenty-fifth 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."[124] 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 2200 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

ed States v. Booker, 543 U.S. 220, 276-79, 284-85 (2005) (Stevens, J., dissenting in part).

122. *E.g.*, KAN. STAT. ANN. § 21-4716(b) (West 2010).

123. *See* MODEL PENAL CODE: SENTENCING § 7.07B(2) (Tentative Draft No. 1, 2007).

124. Holder, *supra* note 8.

federal sentencing,[125] the Commission removes all judge-identifying information from the data it releases to judges, scholars, and the public.[126] The Commission not only withholds the name of the sentencing judge, but does not provide a code or number that would permit analysis of judges' sentencing patterns while keeping their identities confidential.[127] With the exception of studies by the Commission and its staff, the Anderson-Kling-Stith study marks the only time in over twenty-five years that scholars have received permission to study case records that identify the sentencing judges.[128]

The Commission's policy can be traced to the Judicial Conference of the United States, which on behalf of federal judges extracted a promise of secrecy from the Commission as a condition of supplying basic sentencing records.[129] Ostensibly, the purpose of the policy is to prevent the release of defendant information, but sensitive personal data are already removed from the Commission's data releases. More likely, federal judges simply wish to shield themselves from criticism—an astonishing expectation for public officials who enjoy life tenure. The policy has been roundly criticized by scholars,[130] and even some judges.[131] I join the chorus calling for the Commission to promote transparency and facilitate the study of federal sentencing by releasing sentencing data that include judge-specific information.[132]

This study overcomes that obstacle by drawing on a unique new dataset of more than 2200 sentences from the District of Massachusetts. The data were

125. *See* 28 U.S.C. § 995(a)(12)-(16) (2006).

126. U.S. SENTENCING COMM'N, GUIDE TO PUBLICATIONS & RESOURCES 2007-2008, at 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)).

127. 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* 28 U.S.C. § 994(w)(3)-(4) (2006).

128. *See* Anderson et al., *supra* note 23, at 287.

129. *See* Public Access to Sentencing Commission Documents and Data, 54 Fed. Reg. 51,279, 51,282 (Dec. 13, 1989).

130. Mark H. Bergstrom & Joseph S. Mistick, *The Pennsylvania Experience: Public Release of Judge-Specific Sentencing Data*, 16 FED. SENT'G REP. 57, 63 (2003) (noting that Pennsylvania now releases judge-identifying information and that "[m]any of the negative outcomes predicted during the development of the policy have not materialized"); Marc L. Miller, *A Map of Sentencing and a Compass for Judges: Sentencing Information Systems, Transparency, and the Next Generation of Reform*, 105 COLUM. L. REV. 1351, 1356 n.19, 1385 (2005).

131. *See, e.g.*, Richard G. Kopf, *A Brief and Modest Proposal*, SENT'G L. & POL'Y (July 28, 2010, 5:19 PM), http://sentencing.typepad.com/sentencing_law_and_policy/2010/07/a-brief-and-modest-proposal-an-original-essay-from-us-district-judeg-richard-kopf-.html.

132. *See* Miller, *supra* note 130, at 1356 & n.19; Marc L. Miller, *Sentencing Reform "Reform" Through Sentencing Information Systems*, *in* THE FUTURE OF IMPRISONMENT 121, 146-48 (Michael Tonry ed., 2004); Schanzenbach & Tiller, *supra* note 93, at 741-42; Steven L. Chanenson, *Write On!*, 115 YALE L.J. POCKET PART 146, 147 (2006), http://www.thepocketpart.org/images/pdfs/1.pdf.

gathered using a method, pioneered by Max Schanzenbach and Emerson Tiller,[133] 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*.[134] 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 now makes public a case document called the "Statement of Reasons." This document is available online for every criminal sentence, unless the presiding judge orders it sealed.[135] 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, and guideline range, as well as any statutory minimum sentence, and the basis for any departure.[136] 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,[137] 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."[138]

Aided by the information in the 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 2659 sentences imposed between October 1, 2001, and September 30, 2008.[139]

133. Schanzenbach & Tiller, *supra* note 93, at 729-30.

134. 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 242.

135. United States v. Green, 346 F. Supp. 2d 259, 277 n.66 (D. Mass. 2004) (Young, C.J.) (citing Minutes of the Court Meeting of the District of Massachusetts 4 (Sept. 4, 2001)). Judges may order the Statement of Reasons sealed for case-specific reasons, *id.*, such as the protection of a defendant who cooperated with authorities. In practice, judges rarely order the Statement of Reasons sealed, minimizing the risk of selection bias. I encountered fewer than five cases (out of more than 2200 total coded) in which the Statement of Reasons was unavailable.

136. 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*.

137. *See* JUDICIAL CONFERENCE OF THE U.S., REPORT OF THE PROCEEDINGS OF THE JUDICIAL CONFERENCE OF THE UNITED STATES 17 (2001).

138. United States v. Kandirakis, 441 F. Supp. 2d 282, 332 n.76 (D. Mass. 2006) (Young, J.).

139. Details of the case-matching technique are set forth in Part A.2 of the Appendix.

Throughout the Article, I use letters rather than names to identify judges. Identifying judges by name is unnecessary because inter-judge disparity is a concern regardless of which particular judges reached inconsistent results.[140] Also, the Administrative Office's reticence to release judge-identifying information reflects concerns that individual judges will be subject to "unfair criticism" based on "isolated cases."[141] Although I see no reason why federal judges who enjoy life tenure cannot withstand criticism—even "unfair" criticism—of their decisions, this Article illustrates that judge-identifying information can enable valuable research without targeting individual judges.

2. *Natural experiment method*

Building on previous studies of inter-judge sentencing disparity, this study employs a natural experiment method. Unlike a controlled experiment, in which researchers themselves change a condition to study its effects, a natural experiment examines the effects of exogenous changes that occur in the world without any prompting by researchers, such as the enactment of a new law or a series of Supreme Court decisions. Because all federal judges are equally bound by the Supreme Court's decisions in *Booker*, *Kimbrough*, and *Gall*, there is no control group 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 random case-assignment system and a common case pool.[142] Assuming each judge hears a sufficient number of cases, and the distribution of cases is truly random, then *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.[143] 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 di-

140. *See supra* notes 26-30 and accompanying text.

141. *See* Letter from Leonidas Ralph Mecham to Senator Orrin G. Hatch, *supra* note 70, at 3.

142. Anderson et al., *supra* note 23, at 290-91; Hofer et al., *supra* note 47, at 282.

143. Specifically, sentences were excluded if the sentencing judge was on pace to impose fewer than twenty-five sentences in a two-year period. *Cf.* Anderson et al., *supra* note 23, at 288 (using a cutoff of thirty cases, including jurisdictional transfers and acquittals, in a two-year period).

vision that includes Boston,[144] and statistical tests indicate that cases were indeed distributed randomly.[145]

The result is a large dataset of 2262 sentences imposed by ten judges, all in active status, who served side-by-side in Boston continuously from 2001 to 2008.[146] Judges included in the study had between 175 and 264 sentences during that period, an average of 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.[147]

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.[148]

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.[149] This study uses the same measure.[150] 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.[151]

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 (≈ 33 months)
2. Post-*Booker*: January 12, 2005 - December 9, 2007 (≈ 35 months)

144. D. MASS. LOCAL R. 40.1(B)(3) (2008).

145. *See infra* notes 245-48 and accompanying text.

146. For a detailed breakdown of the sentence count for each judge, see Table A2.

147. *See infra* note 241 & Table A1 and accompanying text.

148. *See infra* Table A3 and accompanying text.

149. *See* Anderson et al., *supra* note 23, at 281; Hofer et al., *supra* note 47, technical app. at 307-09; Waldfogel, *Empirically Based Sentencing*, *supra* note 55, at 294. Consistent with the Sentencing Commission's convention, sentences of probation are coded as zero months of imprisonment. *See* Hofer et al., *supra* note 47, technical app. at 307-09; *see also* STITH & CABRANES, *supra* note 18, at 62 (following the same convention).

150. 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.

151. *See infra* notes 259-61 and accompanying text.

3. *Kimbrough*/*Gall*: December 10, 2007 - September 30, 2008 (≈ 10 months)[152]

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."[153] This study follows that recommendation by separately analyzing cases not governed by a mandatory minimum, which account for 66.7% 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."[154] It does not capture other forms of inter-judge disparity linked to particular offense or offender characteristics.[155] 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.[156] 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 routinely sentence within the guideline range, while others routinely sentence below the range.[157]

The study analyzes guideline sentencing outcomes in part through a straightforward description of changes in sentencing patterns over time. It also analyzes *how far*, on average, each judge sentences from the Guidelines by calculating average distance from the guideline range.[158] Again, linear regression

152. 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 232-37 and accompanying text.

153. Hofer et al., *supra* note 47, at 275 n.103.

154. *Id.* at 240-41; *see also* STITH & CABRANES, *supra* note 18, at 119.

155. Hofer et al., *supra* note 47, at 240-41; *see also* STITH & CABRANES, *supra* note 18, at 119 (acknowledging "[a] possibility that comparing each judge's average sentence masks considerable variability within each set of sentences"); Hofer et al., *supra* note 47, at 296 (calling judge-to-judge disparity in average sentence length "the tip of the disparity iceberg").

156. *See* Bowman, *supra* note 94, at 296.

157. *See supra* notes 113-17 and accompanying text.

158. 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 261 and accompanying text.

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. Mandatory Guidelines: October 1, 2001 - April 30, 2003 (≈ 19 months)
2. PROTECT Act: May 1, 2003 - June 23, 2004 (≈ 14 months)
3. Post-*Booker* I: January 12, 2005 - June 30, 2006 (≈ 18 months)
4. Post-*Booker* II: July 1, 2006 - December 9, 2007 (≈ 17 months)
5. *Kimbrough*/*Gall*: December 10, 2007 - September 30, 2008 (≈ 10 months)[159]

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.[160]

Those constraints suggest that the high rate of within-range sentencing that has continued since *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,[161] William Young,[162] and Patti Saris[163] have written scholarly

159. For an explanation of the cutoff dates for each period, see *infra* notes 232-37 and accompanying text.

160. *See infra* notes 249-56 and accompanying text.

161. *See, e.g.*, Nancy Gertner, *Circumventing Juries, Undermining Justice: Lessons from Criminal Trials and Sentencing*, 32 SUFFOLK U. L. REV. 419 (1999); Nancy Gertner, *From Omnipotence to Impotence: American Judges and Sentencing*, 4 OHIO ST. CRIM. L.J.

articles on sentencing issues. Their public-access policy reflects a laudable commitment to transparency and public debate on federal sentencing.

As with any study of a single district, however, the results may not be representative of sentencing nationwide.[164] 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 liberal and Democratic states,[165] although at the time of the study the district's fifteen-member bench was split roughly evenly, with eight Democrats and seven Republicans presently sitting.[166] Also, average sentences in Massachusetts are slightly longer than sentences nationwide, and the rate of below-guideline sentencing in Massachusetts is higher than the rate nationwide.[167] Those differences set the District of

523 (2007) [hereinafter Gertner, *Omnipotence to Impotence*]; Nancy Gertner, Rita *Needs* Gall—*How To Make the Sentencing Guidelines Advisory*, 85 DEN. U. L. REV. 63 (2007); Gertner, *supra* note 100; Gertner, *supra* note 104; Nancy Gertner, *Women Offenders and the Sentencing Guidelines*, 14 YALE J.L. & FEMINISM 291 (2002); Nancy Gertner, *Federal Sentencing Guidelines: A View from the Bench*, HUM. RTS., Spring 2002, at 6.

162. *See, e.g.*, William G. Young, *An Open Letter to U.S. District Judges*, 50 FED. LAW., July 2003, at 30. Judge Young's remarkable 177-page decision in *United States v. Green*, 346 F. Supp. 2d 259 (D. Mass. 2004), 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.

163. *See* Patti B. Saris, *Below the Radar Screens: Have the Sentencing Guidelines Eliminated Disparity? One Judge's Perspective*, 30 SUFFOLK U. L. REV. 1027 (1997).

164. Hofer et al., *supra* note 47, at 279.

165. *2008 Presidential Race: Massachusetts*, N.Y. TIMES, Nov. 4, 2008, http://elections.nytimes.com/2008/president/states/massachusetts (noting that over the last ten Presidential elections, Massachusetts has been the most solidly Democratic state in the country); Lydia Saad, *Political Ideology: "Conservative" Label Prevails in the South*, GALLUP (Aug. 14, 2009), http://www.gallup.com/poll/122333/political-ideology-conservative-label-prevails-south.aspx (using Gallup poll results to show that Massachusetts is the most liberal state in the nation, trailing only the District of Columbia).

166. *See Biographical Directory of Federal Judges*, *supra* note 99. Party affiliation is an imperfect proxy for ideology, and Republican judges in a politically liberal state like Massachusetts probably skew more liberal than their Republican colleagues nationwide. Nonetheless, a study of Boston judges does not involve any greater risk of party and ideology effects than past studies of San Francisco, *see* Waldfogel, *Empirically Based Sentencing*, *supra* note 55, at 294, or New York City and Philadelphia, *see* Payne, *supra* note 55, at 337. Even the large-scale national studies have ensured a random distribution of cases by limiting their data set to district offices where multiple judges shared a single case wheel, which necessarily oversamples sentences in cities with disproportionately liberal and Democratic populations.

167. The data supporting this comparison can be found in appendix B of the 2002-2008 editions of the U.S. Sentencing Commission's *Sourcebook of Federal Sentencing Statistics*. The gap between the national and Massachusetts figures for guideline sentencing is partially attributable to "fast-track" programs for immigration offenses, *see* U.S. SENTENCING GUIDELINES MANUAL § 5K3.1 (2009) (authorizing departures for early disposition programs), which account for 7.9% of sentences nationwide, according to the fiscal year 2008 data. 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 out-

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.[168] In Massachusetts, the Criminal Division of a single U.S. Attorney's office charges and prosecutes virtually all federal cases.[169] 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,[170] and Schanzenbach and Tiller have found that the partisan alignment of circuit courts can affect sentence length and the likelihood of departure.[171] 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*, but by no means the final word. Nonetheless, the Massachusetts documents offer researchers unparalleled access to judge-specific sentencing data, and therefore the best available evi-

comes. 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.

168. *See* Ilene H. Nagel & Stephen J. Schulhofer, *A Tale of Three Cities: An Empirical Study of Charging and Bargaining Practices Under the Federal Sentencing Guidelines*, 66 S. CAL. L. REV. 501, 552-58 (1992) (documenting inter-district disparities driven by prosecutor and defense practices); Daniel Richman, *Federal Sentencing in 2007: The Supreme Court Holds—The Center Doesn't*, 117 YALE L.J. 1374, 1403-06 (2008) (discussing inter-district disparities driven by differences among federal prosecutors' offices).

169. *See Divisions*, U.S. ATT'Y'S OFF. DISTRICT MASS. (last visited Sept. 3, 2010), http://www.usdoj.gov/usao/ma/divisons.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.

170. *See* Gall v. United States, 552 U.S. 38, 47 (2007) (rejecting the conclusion of some courts of appeals that a significant variance from the Guidelines requires an extraordinary justification); Rita v. United States, 551 U.S. 338, 347 (2007) (affirming the decision of some courts of appeals to apply a presumption of reasonableness when reviewing within-range sentences on appeal).

171. *See* Schanzenbach & Tiller, *supra* note 93, 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 ANALYSIS OF THE FEDERAL JUDICIARY (2006).

dence of how sentencing by individual judges has changed in the wake of *Booker*, *Gall*, and *Kimbrough*.

B. *Results*

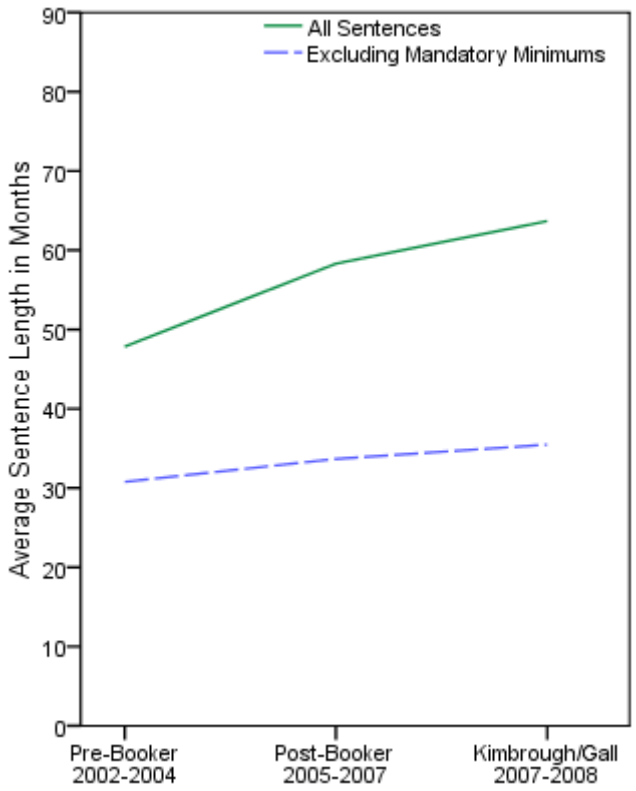
Analysis of the Boston data reveals a clear increase in inter-judge sentencing disparity, both in sentence length and in guideline sentencing patterns. The effect of the judge on sentence length has doubled in strength since *Kimbrough* and *Gall*. And in their guideline sentencing patterns, judges have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

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.[172]

172. 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) (2006), or a government "substantial assistance" motion, *see* 18 U.S.C. § 3553(e) (2006); U.S. SENTENCING GUIDELINES MANUAL § 5K1.1 (2009).

FIGURE 3
Average Sentence Length, Boston Judges



Average sentence length climbed from 47.7 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:

FIGURE 4A
Average Sentence Distribution, Including Mandatory Minimums

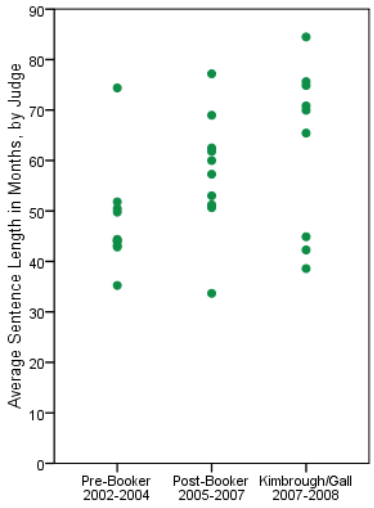
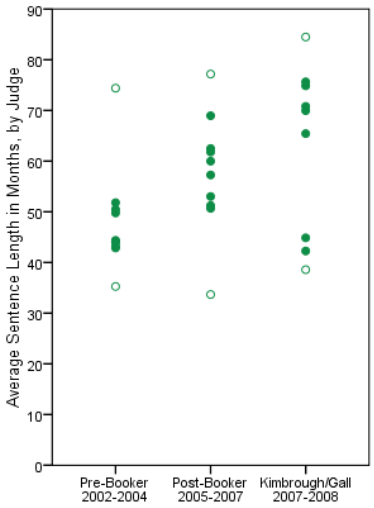


FIGURE 4B
Average Sentence Distribution, Including Mandatory Minimums (High and Low Values Unshaded)



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 1
Summary of Linear Regression Models, Sentence Length[173]

	% Variance Explained	Avg. Variance Explained	Model Significance
Pre-*Booker*	3.1%	10.8 months	.001*
Post-*Booker*	2.5%	10.9 months	.003*
Kimbrough/*Gall*	6.1%	15.5 months	.044*

*Note: * Significant at the .05 level*

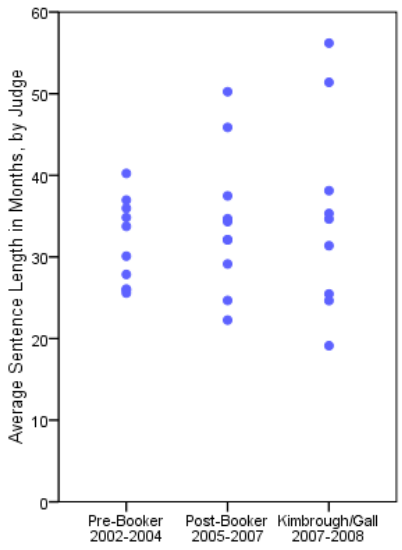
173. For details of these regression models for sentence length, see Table A4.

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.[174]

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 3.1%. 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 twice as strong as in the three years before *Booker*.[175]

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:

FIGURE 5
Average Sentence Distribution,
Excluding Mandatory Minimums



174. For discussion of the regression models generally, see *infra* notes 259-61 and accompanying text.

175. For a discussion of period-selection issues, see Appendix A.1.

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 25.6 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 2
Summary of Linear Regression Models, Sentence Length
Excluding Mandatory Minimums[176]

	% Variance Explained	Avg. Variance Explained	Model Significance
Pre-*Booker*	1.4%	4.9 months	.368
Post-*Booker*	3.1%	8.0 months	.021*
Kimbrough/*Gall*	8.0%	10.3 months	.180

*Note: * Significant at the .05 level*

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. For sentences since *Kimbrough* and *Gall*, the model is not statistically significant, so it is not possible to exclude the possibility that the relationship was the product of chance.[177] The limited data for that period suggest, howev-

176. For details of these regression models, see Table A5.

177. The fact that the model for the *Kimbrough*/*Gall* period is not significant reinforces the need for caution in interpreting the results for cases not governed by a mandatory minimum. Statistical significance is highly sensitive to sample size, and the *Kimbrough*/*Gall* period necessarily has about one-third as many cases as the other periods, even before excluding mandatory minimums. Although the relationship in the *Kimbrough*/*Gall* period is

er, that the rate of variance explained has increased further to 8.0%—more than five times pre-*Booker* levels.

2. *Guideline sentencing patterns*

Similarly, analysis of guideline sentencing patterns after *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 new-found discretion to depart from the guideline range more enthusiastically than others. Consider the below-range sentencing patterns of four judges, A, B, C, and D.[178]

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"[179] 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 (41.7%, 36.2%, and 52.8%). Judge B's rate of below-range sentencing has more than quadrupled.[180]

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 10.9%, and has remained stable throughout the other periods as well (13.3%, 18.8%, and 16.1%).[181]

---

strongly positive, the model falls well short of statistical significance.

178. This Article uses letters, rather than names, to identify judges. *See supra* text accompanying notes 140-41.

179. *See* Gertner, *supra* note 104, at 137-38 (using the phrase "free at last" to describe the reaction to *Booker* among some district court judges).

180. 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 32.7% in the two post-*Booker* periods, and 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 34.1% and 31.6% in the two post-*Booker* periods. So has Judge G's rate of below-range sentencing, which went from 13.9% in the Mandatory Guidelines period, to 10.5% in the PROTECT Act period, to 33.3%, 34.0%, and 32.4% in the three periods since *Booker*.

181. 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 (24.4%, 25.5%, and 17.4%). Sentences by Judge I seemed to fit this pattern during the eighteen months after *Booker* with a rate of 22.0%, 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 47.7% in the Post-*Booker* II period, and stands at 38.9% in the *Kimbrough*/*Gall* period.

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 41.5% and 34.6% in the two most recent periods.[182]

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

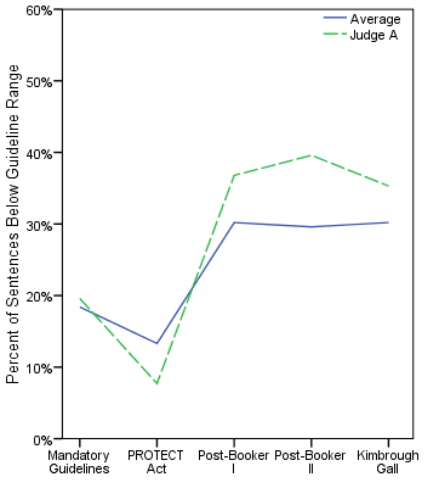
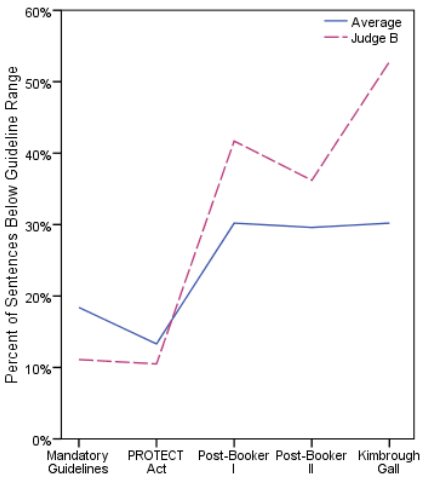


FIGURE 6B
Guideline Sentencing, Judge B



182. 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, nearly tripled to 32.1% in the Post-*Booker* I period, dropped again to 19.1% in the Post-*Booker* II period, and has nearly doubled again to 31.8% in the *Kimbrough*/*Gall* period.

FIGURE 6C
Guideline Sentencing,
Judge C

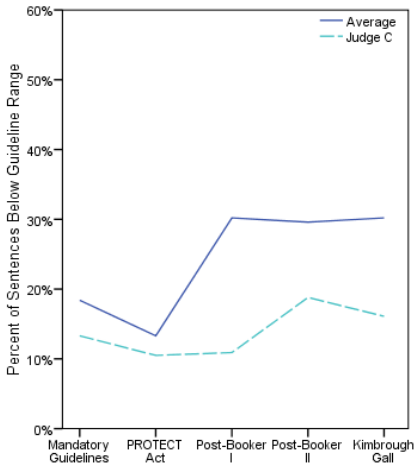
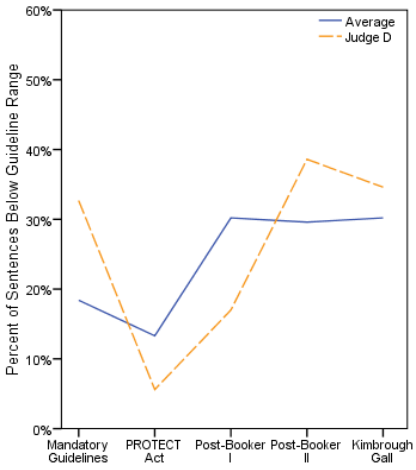


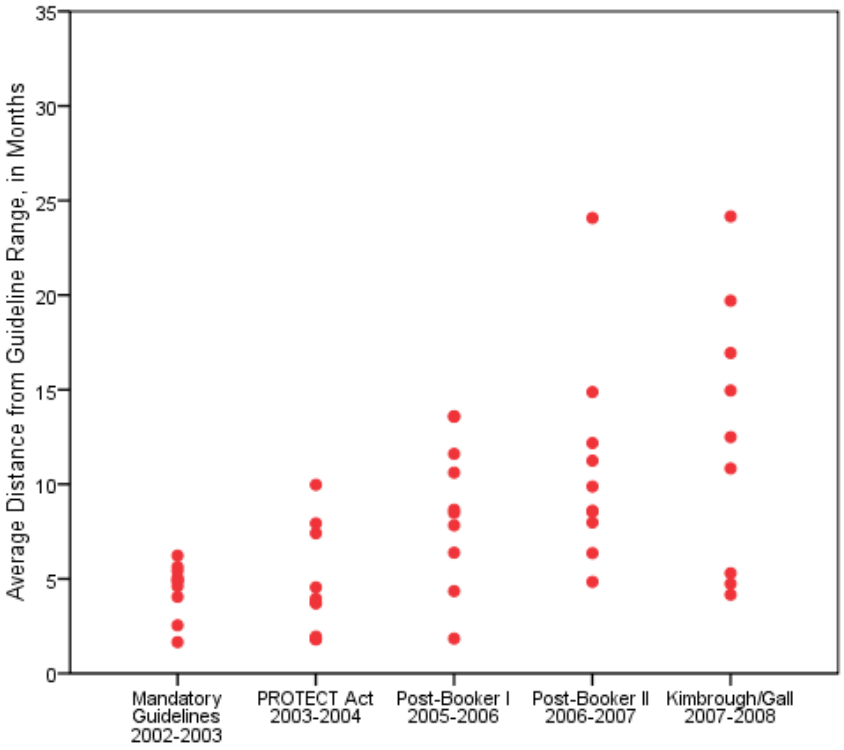
FIGURE 6D
Guideline Sentencing,
Judge D



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,
All Sentences

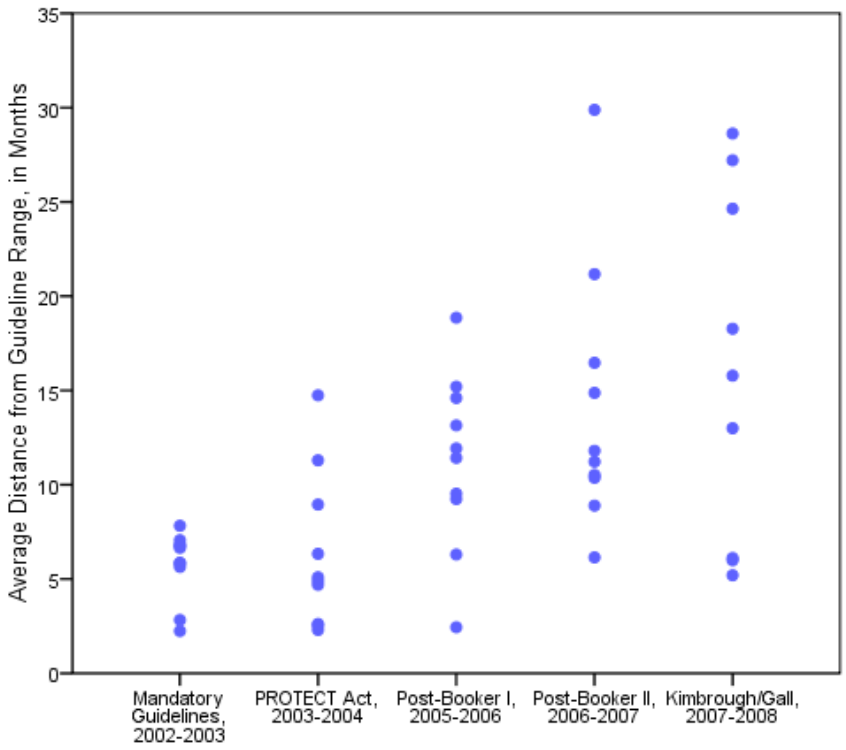


Under the mandatory Guidelines in 2002-2003, average distance from the guideline range was tightly clustered within a range of 4.5 months. Perhaps surprisingly, the spread increased under the PROTECT Act, covering 8.2 months. But after *Booker*, the distribution has widened dramatically and grown broader 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.[183] Figure 7b shows the distribution of average distance from the guideline range for the subset of discretionary sentences:

183. *See infra* Appendix A.4 (defining "discretionary" sentences and explaining why they are of special relevance in measuring inter-judge disparity in guideline sentencing).

FIGURE 7B
Distribution in Average Distance from Guideline Range,
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 within 7.8 months or less of 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 outside of 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*.[184] Table 3 reports the results of linear regression models calculating the percentage of variance explained by the judge, both for all sentences and for discretionary sentences:

TABLE 3
Summary of Linear Regression Models
Distance from Guideline Range[185]

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

*Note: * Significant at the .05 level*

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 sub-

184. *See* Hofer et al., *supra* note 47, 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.127.

185. For details of these regression models, see Tables A6 and A7.

ject 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 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 reinforce anecdotal evidence from around the country warning of greater judge-to-judge disparity in sentencing outcomes.[186] 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 6.6% of variation in distance from the guideline range (9.4% for discretionary sentences).[187] Yet both the strength of the effect and size of the change are larger than those reported in the Hofer study,[188] 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 inter-judge disparity itself is small or unimportant."[189] Although it is too early to despair a return to "pre-guideline chaos,"[190] 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 entire-

186. *See supra* Part I.B.3.

187. *See supra* Tables 1 & 2 and accompanying text.

188. *See* Hofer et al., *supra* note 47, at 287; *see also* Waldfogel, *Empirically Based Sentencing*, *supra* note 55, at 295.

189. Anderson et al., *supra* note 23, at 294 n.53.

190. That was what one Senator predicted in the immediate aftermath of *Booker*. Press Release, Senator Chuck Grassley, Supreme Court Decision on Sentencing Guidelines (Jan. 13, 2005), *available at* 2005 WLNR 2769009.

ly 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 reevaluating mandatory minimum sentences and confronting unwarranted disparity created by prosecutorial charging and bargaining practices.[191] 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, even as their "free at last" colleagues have begun to impose below-range sentences far more frequently? The Boston data tend to undermine some of the conventional explanations that within-guideline sentencing is the product of inertia, risk aversion, anchoring, strategic 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) cognitive "anchoring" errors caused by using the guideline range as a starting point; (4) strategic behavior by judges anxious to avoid provoking Congress; and (5) simple laziness.[192] The Boston data, however, tend to undermine several of those explanations.

1. *Inertia*

The first, most common explanation of judges' unexpectedly mild reaction to *Booker* was that most sitting judges have spent their entire careers imposing sentences under the Guidelines' framework. It should come as no surprise, on

191. 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. *See* Holder, *supra* note 8.

192. *See supra* notes 99-112 and accompanying text.

this account, that a generation of judges that has "grown up" with the Guidelines would cling to them even though they have become advisory.[193]

The Boston data allow a partial test of that theory. If inertia were the primary reason why judges impose sentences in the guideline range, then we might expect to see a difference in sentencing patterns between long-tenured judges and their more junior colleagues. Judges appointed before 1987, who tasted freedom under the pre-guidelines regime, might cast off the yoke of the Guidelines more readily because they have personal experience imposing sentences in a fully discretionary system. Judges appointed after 1987, by contrast, might experience greater inertia because they have never sentenced outside the mandatory guidelines regime. It happens that, of the ten core judges in Boston from 2002 to 2008, five joined the court before 1987, while the other five joined the court after 1987 while the Guidelines were mandatory.

A comparison of sentencing outcomes, however, does not suggest a greater "inertia effect" among more junior judges. As shown in Figures 8a and 8b the difference between groups is negligible:

FIGURE 8A
Average Sentences
Pre- and Post-1987 Judges

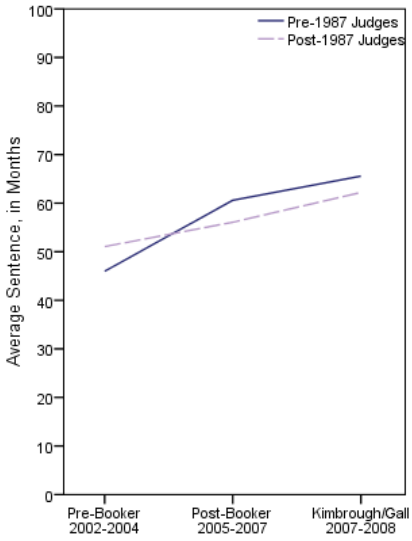
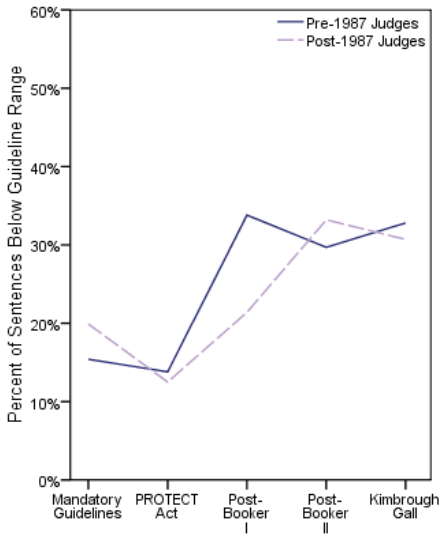


FIGURE 8B
Guideline Sentencing
Pre- and Post-1987 Judges



From 2002 to 2008, there has been no meaningful difference in average sentence length based on whether a judge was appointed before the effective

193. *See supra* notes 99-100 and accompanying text.

date of the Guidelines in 1987.[194] 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 (33.9% compared with 21.4%). But the difference quickly evaporated. For more than 2.5 years, both groups have sentenced below the guideline range around 30% of the time.[195]

Of course, these similarities between judges appointed before 1987 and their more junior colleagues do not rule out the possibility that inertia influenced the sentencing patterns of some judges. Eighteen years of mandatory Guidelines may leave habits deeply ingrained in anyone. Still, with no direct measure available, pre-1987 service is a credible proxy for judges' resistance to legal change in this context, and it has no apparent explanatory power. The experience in Boston suggests that more than inertia is at work.

2. *Risk aversion*

A second conventional explanation for the generally modest effects of *Booker* is that some district court judges were anxious to avoid reversal on appeal. It took several years for the Supreme Court and appellate courts to work out the details of "reasonableness" review, and during that time district courts faced considerable uncertainty at the appellate level.[196] On the "risk aversion" theory, some courageous judges immediately took advantage of their newfound discretion to sentence outside the guideline range. But other judges, more risk-averse than their colleagues, have continued to cling to the guideline range to avoid the embarrassment and hassle of reversal and resentencing.

Again, the Boston data allow a partial test of that explanation. If risk aversion were the primary reason why many judges continue to impose sentences in the guideline range, then we might expect that decisions like *Kimbrough* and *Gall* would produce more consistent guideline sentencing patterns among members of the court. After all, *Kimbrough* and *Gall* reduced the risk of appellate reversal by announcing a highly deferential standard of review,[197] and by indicating that sentencing judges may categorically reject the Commission's policy judgments.[198] The Court thus made clear, even to the most risk-averse judges, that they are free to sentence outside the guideline range with confidence. If post-*Booker* differences in guideline sentencing patterns were attributable to different levels of risk aversion, rather than other factors, then elimi-

194. Pre-1987 service by the sentencing judge was not a statistically significant predictor of sentence length in any of the three periods.

195. Pre-1987 service by the sentencing judge was not a statistically significant predictor of how far a sentence fell from the guideline range during any period.

196. *See supra* notes 101-04 and accompanying text.

197. Gall v. United States, 552 U.S. 38, 51 (2007).

198. Kimbrough v. United States, 552 U.S. 85, 101-02 (2007).

nating the source of risk should allow later guideline sentencing patterns to come into closer alignment.

But in Boston, *Kimbrough* and *Gall* had the opposite effect: differences between judges' guideline sentencing patterns have grown even more acute. Differences between "free at last" and "business as usual" judges, for example, became more stark following *Kimbrough* and *Gall*. At the same time, the effect of the judge on sentence length reached double pre-*Booker* levels, while the effect of the judge on distance from the Guidelines strengthened to triple pre-*Booker* levels.[199] Far from allowing more timid judges to "catch up" to their more risk-tolerant colleagues, *Kimbrough* and *Gall* appear to have made inter-judge disparity worse.

No doubt risk aversion affected sentencing decisions in the years immediately after *Booker*.[200] But the fact that inter-judge sentencing disparity appears to have persisted, even after *Kimbrough* and *Gall* greatly reduced the threat of appellate reversal, suggests that risk aversion is an incomplete explanation.

3. *Anchoring*

Another conventional explanation why some judges continue to sentence within the Guidelines, grounded in behavioral psychology, is that the Guidelines cause a form of cognitive error known as "anchoring."[201] Crucially, the anchoring theory does not propose that judges *intentionally* rely on the Guidelines as a legitimate and persuasive source of sentencing advice. Rather, the theory is that judges *irrationally* assign too much weight to the guideline range, just because it offers some initial numbers. Research has shown that giving a sentencing official an initial value, even one that is known to be arbitrary, can influence the length of a sentence.[202]

The Boston data do not shed any light on the possibility that anchoring might account for some inter-judge sentencing disparity. Comparisons are impossible because judges do not readily disclose, and may not be aware of, their cognitive biases.

Still, the anchoring explanation seems strained because the Guidelines are *supposed* to serve as an anchor. The Supreme Court in *Gall* admonished sentencing courts that "the Guidelines should be the starting point and the initial benchmark" for every criminal sentence.[203] Not only is it rational for judges to

199. *See supra* Tables 1 & 3 and accompanying text.

200. *See* Gertner, *supra* note 104, at 140 (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").

201. *See supra* notes 105-08 and accompanying text.

202. Englich et al., *supra* note 106, at 194.

203. Gall v. United States, 552 U.S. 38, 49 (2007).

give consideration to the guideline range, but it is legally compelled. And as discussed below, rational people can disagree about how much weight to give the Guidelines, relative to other considerations.

Moreover, to the extent the guideline range operates as an irrational "anchor" just because it supplies some initial numbers, its effects likely are offset by other anchors tugging in different directions. In every criminal case, competing "starting point" numbers may be offered by defense counsel, prosecutors, the probation office, victim impact testimony, and the statutory sentencing range. Judges also may recall the numbers they selected in other cases, perhaps hundreds of cases in a long career. As a result, judges do not approach sentencing thinking about a single set of "anchor" numbers—the guideline minimum and maximum—but with multiple numbers from various sources.

4. *Strategic behavior*

Strategic behavior by judges is another conventional explanation for persistent within-range sentencing patterns in the wake of *Booker*.[204] On this account, although judges were delighted by *Booker* and eager to impose below-range sentences, they have taken a "go slow" approach to avoid provoking a response from Congress. In theory that explanation is entirely plausible, although necessarily speculative. A rich empirical literature has documented other forms of strategic behavior by judges,[205] and nothing in this study forecloses similar behavior in the sentencing context.

Yet in these circumstances, widespread strategic behavior seems unlikely because of challenges in coordinating a "go slow" strategy. The sheer number of district court judges and sentencing decisions makes this sort of strategy difficult to sustain. Individual district court judges must realize that nothing can prevent their more than 600 colleagues nationwide from sentencing in a manner that agitates Congress. Thus, faced with particular cases in which the guideline range seems unduly harsh, the strategic benefits of a within-range sentence would appear tiny compared with the human costs of an unjust sentence. Given the extent of persistent within-range sentencing since *Booker*, strategic behavior seems like a weak explanation.

5. *Laziness*

Judicial laziness, the final conventional explanation, is particularly unpersuasive. A few judges, themselves critics of the Guidelines, have floated the

204. *See supra* notes 109-10 and accompanying text.

205. *See, e.g.*, LEE EPSTEIN & JACK KNIGHT, THE CHOICES JUSTICES MAKE 9-18 (1998); THOMAS H. HAMMOND ET AL., STRATEGIC BEHAVIOR AND POLICY CHOICE ON THE U.S. SUPREME COURT 65-248 (2005); JEFFREY A. SEGAL & HAROLD J. SPAETH, THE SUPREME COURT AND THE ATTITUDINAL MODEL REVISITED 97-109 (2002).

rather self-congratulatory theory that their colleagues cling to the guideline range because they prefer "less work and less stress" and are unwilling to approach the task of sentencing with sufficient intellectual rigor.[206] Although judges—like the rest of us—undoubtedly consider their personal time constraints in approaching their work,[207] the upheaval of *Booker*, *Kimbrough*, and *Gall* forced judges to approach sentencing decisions with renewed 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 accounts, I propose two other explanations for the persistent high rate of sentences within the guideline range. One is that some judges consistently 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' recommendations*

The most promising explanation for many judges' continued fidelity to the Guidelines is also the simplest. Some judges might *actually agree* with the sentences recommended by the Guidelines more frequently than their colleagues. Although now free to exercise independent judgment, some judges consistently arrive at sentencing outcomes that match the Commission's recommendations. The fact that some judges have followed a "business as usual" pattern of guideline sentencing,[208] even as the degree of freedom they enjoy at sentencing has dramatically expanded, suggests that they simply do not wish to alter their pre-*Booker* sentencing practices. That fundamental difference of opinion between "business as usual" and "free at last" judges tends to produce inter-judge sentencing disparity.

Remarkably, however, the scholarly literature has essentially ignored that possibility. The enormous body of pre-*Booker* literature criticizing the Federal Guidelines frequently created the impression of *uniform* opposition among judges.[209] News reports described hostility so pervasive that "[m]any judges

206. *See supra* notes 111-12 and accompanying text.

207. Richard A. Posner, *What Do Judges and Justices Maximize? (The Same Thing Everybody Else Does)*, 3 SUP. CT. ECON. REV. 1, 31 (1993) (describing a "judicial utility function" that includes work time and leisure time).

208. *See supra* note 181 and accompanying text.

209. *See* Gertner, *Omnipotence to Impotence*, *supra* note 161, at 530, 539 (describing "robust judicial opposition to the Guidelines"); Gertner, *supra* note 100, at 267 (describing

regard as a traitor any colleague who serves on the U.S. Sentencing Commission."[210] A few judges even resigned from the bench or refused to hear drug cases in protest of the guidelines regime.[211] Scholars cheerfully generalized that "everybody loves to hate the Federal Sentencing Guidelines."[212] Even members of Congress took notice that judges seemed to "hate the sentencing guidelines."[213]

Surveys of federal judges, however, have documented a persistent split in opinion. The most recent, a 2010 survey by the Sentencing Commission, asked district court judges whether the guideline sentencing range was "generally appropriate," "too low," or "too high" for various categories of offenses.[214] Table 4a excerpts some of the results:

TABLE 4A
Results of 2010 Survey of District Court Judges[215]

Offense	Too High	Generally Appropriate	Too Low
Drug Trafficking—Methamphetamine	34%	60%	6%
Fraud	10%	65%	24%
Firearms	23%	70%	7%
Illegal Reentry into the U.S.	34%	55%	11%
Drug Trafficking—Crack Cocaine	70%	28%	2%
Child Pornography—Possession	70%	26%	3%

how district court judges "overwhelmingly opposed the Guidelines"); Joseph W. Luby, *Reining in the "Junior Varsity Congress": A Call for Meaningful Judicial Review of the Federal Sentencing Guidelines*, 77 WASH. U. L.Q. 1199, 1276 (1999) ("[T]he Commission has little legitimacy in the sentencing regime. Its Guidelines are reviled (even though tolerated) by lawyers, judges, and commentators alike."); José A. Cabranes, *Sentencing Guidelines: A Dismal Failure*, N.Y. L.J., Feb. 11, 1992, at 2 (claiming that "virtually everyone who is associated with the federal justice system" deems the Guidelines a "dismal failure").

210. Naftali Bendavid, *Judicial Traitor or Consensus Builder? Breyer's Role as Sentencing Pioneer Still Rankles*, LEGAL TIMES, May 16, 1994, at 7.

211. *Criticizing Sentencing Rules, U.S. Judge Resigns*, N.Y. TIMES, Sept. 30, 1990, § 1, at 22; Don J. DeBenedictis, *The Verdict Is In*, A.B.A. J., Oct. 1993, at 78.

212. Klein & Steiker, *supra* note 32, at 232-33; *see also* Stephanos Bibas, Blakely*'s Federal Aftermath*, 16 FED. SENT'G REP. 333, 342 (2004) ("To put it bluntly, many judges and others hate the Guidelines . . . ."); Julie R. O'Sullivan, *In Defense of the U.S. Sentencing Guidelines' Modified Real-Offense System*, 91 NW. U. L. REV. 1342, 1343-44 (1997) ("[J]udging from the scholarly commentary, virtually everyone loves to hate [the Guidelines].").

213. Blakely v. Washington *and the Future of the Sentencing Guidelines: Hearing Before the S. Comm. on the Judiciary*, 108th Cong. 3-4 (2004) (statement of Sen. Orrin Hatch, Chairman, S. Comm. on the Judiciary) (stating "I can understand why").

214. U.S. SENTENCING COMM'N, RESULTS OF SURVEY OF UNITED STATES DISTRICT JUDGES JANUARY 2010 THROUGH MARCH 2010 tbl.8 (2010).

215. *See id.*

For most drug trafficking offenses, as well as for firearms, fraud, and immigration offenses, a majority of judges agrees with the Commission that the guideline sentencing range is generally appropriate.[216] Even for the offenses that generate the most forceful criticism of the Commission, such as trafficking in crack cocaine and possession of child pornography, a substantial minority—more than 25% of judges—believes that the guideline range is generally appropriate. The survey also revealed that a substantial minority of judges agrees with the Commission that many offender characteristics are not "ordinarily relevant" in deciding whether to depart from the guideline range.[217]

Those disagreements among judges are not new. A 2003 survey 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.[218] Table 4b excerpts some of the results:

TABLE 4B
Results of 2003 Survey of District Court Judges[219]

1 or 2 (Worst)	3 or 4	5 or 6 (Best)
How well do the Guidelines achieve the purposes of sentencing in § 3553(a)?		
22.9%	38.6%	38.4%
How often do guideline sentences provide just punishment?		
20.2%	42.8%	37.0%
How often do the Guidelines offer sufficient flexibility at sentencing?		
45.0%	30.6%	24.4%

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.[220] For five of the nine specific purposes of punishment in § 3553(a), a majority of judges rated the Guidelines a 5 or 6.[221]

216. *Id.* For all drug types except crack cocaine, more than 50% of judges responded that the guideline range was generally appropriate. *See id.*

217. *See id.* at tbls.8 & 13. For example, although 41% of judges consider vocational skills "ordinarily relevant," 53% consider that factor "not ordinarily relevant," and 6% consider it "never relevant." *Id.*

218. *See* LINDA DRAZGA MAXFIELD, U.S. SENTENCING COMM'N, FINAL REPORT SURVEY OF ARTICLE III JUDGES ON THE FEDERAL SENTENCING GUIDELINES, at ES-1 (2003), *available at* http://www.ussc.gov/Judge_Survey/execsum.pdf.

219. *Id.* at 12 exh. II-11, 15 exh. II-14, 24 exh. II-23, *available at* http://www.ussc.gov/Judge_Survey/jschap2.pdf.

220. *Id.* at 24 exh. II-23.

221. *Id.* at 2 exh. II-1, 3 exh. II-3.

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.[222] Asked whether the Guidelines offer sufficient flexibility at sentencing, 24.4% of district court judges gave them a 5 or 6.[223] 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 at least a decade, even before the shake-up of federal sentencing, a substantial contingent of federal judges has quietly agreed with the Commission that the Guidelines generally recommend appropriate sentences. Now that judges must directly apply the purposes of punishment in § 3553(a), those inter-judge disagreements 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 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 arguments about the competence of the Commission relative to individual judges, or about 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.[224] 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.[225] They may be reluctant, for example, to "individualize" sentences based on their own predictions about the future dangerousness of defendants.[226]

222. *Id.* at 12 exh. II-11.

223. *Id.* at 15 exh. II-14.

224. *See* Rita v. United States, 551 U.S. 338, 349-50 (2007); United States v. Wilson, 350 F. Supp. 2d 910, 914-15 (D. Utah 2005) (Cassell, J.). For a thoughtful defense of the Commission's work in the controversial context of the child pornography guidelines, see *United States v. Cunningham*, 680 F. Supp. 2d 844, 862-64 (N.D. Ohio 2010).

225. United States v. Wanning, 354 F. Supp. 2d 1056, 1062 n.9 (D. Neb. 2005); *see also* Stephanos Bibas et al., *Policing Politics at Sentencing*, 103 NW. U. L. REV. 1371, 1388 (2009) (stating that "courts lack the institutional competence to make systemic policy choices," but "Congress has established an agency, the Sentencing Commission, to collect data and the views of various constituencies in formulating policies and rules").

226. Kevin R. Reitz, *Sentencing Facts: Travesties of Real-Offense Sentencing*, 45 STAN. L. REV. 523, 553-55 (1993) (acknowledging a division of opinion among scholars, but

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."[227]

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.[228] 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.[229]

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 claims about competence and legitimacy. Scholars have noted, for example, that the Commission does not behave like most expert administrative agencies—it frequently fails to marshal evidence or even provide a reasoned explanation in support of its judgments.[230] They have also raised serious questions about whether the unique structure and composition of the Commission actually undermine, rather than advance, its legitimacy.[231] Many commentators

reviewing research showing that "despite our faith that we can spot those offenders most likely to recidivate, individualized predictions of future dangerousness are little better than a game of chance"); *see also* Erica Beecher-Monas & Edgar Garcia-Rill, *Danger at the Edge of Chaos: Predicting Violent Behavior in a Post-*Daubert *World*, 24 CARDOZO L. REV. 1845, 1845-47 (2003); Stephen J. Morse, *Blame and Danger: An Essay on Preventive Detention*, 76 B.U. L. REV. 113, 126 & n.39 (1996); Paul H. Robinson, Commentary, *Punishing Dangerousness: Cloaking Preventive Detention as Criminal Justice*, 114 HARV. L. REV. 1429, 1450 (2001).

227. United States v. Tabor, 365 F. Supp. 2d 1052, 1061 (D. Neb. 2005).

228. 28 U.S.C. § 994(p) (2006).

229. United States v. Cage, 451 F.3d 585, 593 (10th Cir. 2006) (describing the Guidelines as "an expression of popular political will about sentencing"); *Wanning*, 354 F. Supp. 2d at 1062 n.9; *Wilson*, 350 F. Supp. 2d at 915; Bibas et al., *supra* note 225, at 1388 ("Most importantly, Congress has democratic legitimacy; courts do not.").

230. Luby, *supra* note 209, 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," and characterizing the Guidelines as a "compilation of administrative *diktats*").

231. *See* Frank O. Bowman, III, *Mr. Madison Meets a Time Machine: The Political*

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' recommendations, 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' responses 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 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.

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.

Science of Federal Sentencing Reform, 58 STAN. L. REV. 235, 254-55 (2005).

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*.[232] Because of the chaos that followed that decision, the interregnum between *Blakely* and *Booker* is ignored.[233] 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.[234]

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.[235] For guideline sentencing, however, the PROTECT Act marks a critical change because it was explicitly designed to reduce the number

232. 542 U.S. 296 (2004).

233. The Commission's post-*Booker* reports have largely ignored the period between *Blakely* and *Booker* as well. *See, e.g.*, FINAL REPORT, *supra* note 92.

234. 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 fiscal years 2007 and 2008 data ahead of the ordinary schedule was unexpected, and a valuable benefit for participants.

235. The *Kimbrough*/*Gall* period is shorter than the other periods because no data are available for fiscal year 2009.

of downward departures.[236] 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.[237]

2. *Case matching*

To obtain judge-specific sentencing data, Max Schanzenbach and Emerson Tiller have developed a work-around 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.[238] 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) on the amount of any fine, the offense type, and the Hispanic ethnicity of the defendant.[239] They successfully matched about 80% of sentences returned in their searches.[240]

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.[241] The initial search yielded around 5000 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[242]—information

236. *See supra* notes 59-67 and accompanying text.

237. *See infra* Tables A4 & A5 and accompanying text.

238. Schanzenbach & Tiller, *supra* note 93, at 729-30.

239. *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.

240. *See id.* at 730.

241. 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.

242. 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. U.S. SENTENCING COMM'N, VARIABLE CODEBOOK FOR INDIVIDUAL OFFENDERS 63 (2009) (noting that after fiscal year 2003, "[t]he date on which the defendant was sentenced" is not available).

from the Statement of Reasons was used to narrow the list of potential matches. This method proved highly reliable: less than 0.4% of sentences could not be matched because of multiple similar sentences in the Commission's data.[243]

The process resulted in 2659 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 A1
Matched Cases, by Fiscal Year[244]

	2002	2003	2004	2005	2006	2007	2008	Total
All Boston Cases	497	479	315	306	418	460	403	2878
Matched Cases	445	433	292	273	387	430	369	2629
% of Cases Matched	89.5%	90.4%	92.7%	89.2%	92.6%	93.5%	91.6%	91.3%

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

TABLE A2
Sentence Count for Judges

	Mandatory Guidelines	PROTECT Act	Post-*Booker I*	Post-*Booker II*	*Kimbrough/ Gall*	Total
Judge A	56	39	68	48	17	228
Judge B	54	38	36	47	36	211
Judge C	60	38	64	64	31	257
Judge D	52	36	50	41	26	205
Judge E	82	45	55	49	33	264
Judge F	40	34	44	57	—	175
Judge G	65	38	30	53	34	220
Judge H	79	46	45	47	23	240
Judge I	54	38	41	44	36	213
Judge J	61	42	56	68	22	249
Total	603	394	489	518	258	2262

243. *Cf.* Schanzenbach & Tiller, *supra* note 93, 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.*

244. 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.[245] 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.[246]

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,[247] and in light of the results for other attributes, the results for race do not undermine the premise that sentences were distributed randomly.[248]

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 A3
Percent of Cases for Each Offense Type, by Period

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

245. 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.

246. 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.

247. Hofer et al., *supra* note 47, technical app. at 320.

248. *Cf.* Waldfogel, *Empirically Based Sentencing*, *supra* note 55, at 295 (relying exclusively on a test of randomness using gender).

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 would appear to account for changes in inter-judge sentencing disparity.

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.[249] 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.[250] 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 6.3% of cases.[251]

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.[252] 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.[253]

249. *See* U.S. SENTENCING GUIDELINES MANUAL § 5G1.1(c)(2) (2009). 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).

250. *See id.* § 5G1.1 cmt.

251. 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.

252. *See* 18 U.S.C. § 3585(b) (2006).

253. 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* Telephone Interview with Paul Cassell, Professor and Former District Court Judge for the District of Utah (Oct. 10, 2008); *cf.* U.S. SENTENCING GUIDELINES MANUAL § 1B1.10 cmt. n.3 (2009) (prohibiting the reduction of a sentence "below time served" following a downward amendment to the Guidelines).

In this dataset, a term of time served prevented a below-range sentence in 4.1% of cases.[254]

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.[255] For sentences with a guideline range of 6-12 months, a sentence of probation qualifies as a within-range sentence if it includes some conditions of intermittent, community, or home confinement.[256] 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 19.4% of cases and accounted for almost one-third of within-guideline sentences. Yet the existing literature on federal sentencing has almost entirely ignored the role that they play in limiting the discretion of district courts.[257] Many within-range sentences can be traced to statutory and practical constraints that limit the sentencing judge's 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.[258] *R*-squared was used to measure the percentage of variance in the dependent variable explained by the identity of the sentencing judge.[259] 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.[260]

254. To the extent that the "time served" constraint overlapped with other constraints, the case was coded as "time served."

255. U.S. SENTENCING GUIDELINES MANUAL § 5B1.1(a)(1) (2009).

256. *Id.* § 5B1.1(a)(2).

257. *See* Hofer et al., *supra* note 47, at 275 & nn.101-03 (recognizing the distorting effect of mandatory minimums).

258. *See* Waldfogel, *Empirically Based Sentencing*, *supra* note 55, at 295.

259. 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 & Lindgren*, 30 HARV. J.L. & PUB. POL'Y 791, 817 (2007). The *R*-squared values here are shown as percentages.

260. Hofer et al., *supra* note 47, 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

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.[261]

Neither sentence length nor distance from the guideline range is perfectly normally distributed, an assumption of OLS regression. By definition neither can have a negative value, and both have a large number of cases with a value of zero, resulting in a "censored" distribution. The Hofer study concluded that, notwithstanding those features, the distribution of sentence length is sufficiently normal to permit reliable regression models.[262] To account for the potential effects of the cutoff at zero, however, each model was reanalyzed using the Tobit estimation technique, which is designed for partially censored distributions.[263] The results were materially identical, for both dependent variables and for all periods, in calculating statistical significance and in approximating changes in goodness of fit.

The following pages provide detailed results for the regression models. The regression model for each period is described in a separate column. The dummy 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 (Judge3) was omitted. Each cell reports the coefficient and, below it in parentheses, the standard error.

absolute terms." *Id.* n.127.

261. 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 of life imprisonment as 470 months. *See* U.S. SENTENCING COMM'N, 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.

262. Hofer et al., *supra* note 47, at 312.

263. *See* JEFFREY M. WOOLDRIDGE, ECONOMETRIC ANALYSIS OF CROSS SECTION AND PANEL DATA 518-20 (2002).

Table A4 reports detailed regression results for sentence length:

TABLE A4
Linear Regression Model Results
Sentence Length, Including Mandatory Minimums

	Pre-*Booker*	Post-*Booker*	*Kimbrough*/*Gall*
Constant	74.38* (5.43)	59.99* (6.74)	84.47* (10.76)
Judge4	–31.19* (8.30)	–2.72 (9.28)	–14.54 (18.45)
Judge5	–24.61* (8.95)	17.17 (9.60)	—
Judge6	–30.37* (8.11)	–9.32 (9.14)	–42.20* (17.01)
Judge7	–30.01* (8.37)	–8.76 (10.12)	–19.06 (14.90)
Judge9	–22.57* (8.33)	8.97 (9.07)	–13.66 (15.46)
Judge10	–40.26* (8.48)	2.49 (9.87)	–9.63 (16.21)
Judge11	–23.90* (8.37)	–26.32* (10.05)	–39.60* (14.90)
Judge13	–31.51* (8.11)	–6.98 (10.12)	–8.85 (15.11)
Judge14	–30.28* (7.71)	1.81 (9.84)	–45.90* (16.79)
Number of cases	997	1007	258
R^2	.031	.025	.061
Significance	.001*	.003*	.044*

*Note: * Significant at the .05 level*

Table A5 reports detailed regression results for sentence length for the subset of cases not governed by a mandatory minimum:

TABLE A5
Linear Regression Model Results
Sentence Length, Excluding Mandatory Minimums

	Pre-*Booker*	Post-*Booker*	*Kimbrough*/*Gall*
Constant	36.96* (4.99)	34.68* (5.61)	56.20* (10.06)
Judge4	–6.86 (6.92)	–5.54 (7.90)	–37.08* (16.30)
Judge5	3.28 (7.44)	15.57 (8.17)	—
Judge6	–11.40 (7.06)	–10.01 (7.55)	–24.82 (13.55)
Judge7	–2.12 (7.34)	–2.59 (8.28)	–30.73* (13.37)
Judge9	–3.21 (6.99)	–2.57 (7.71)	–18.08 (13.37)
Judge10	–10.89 (6.97)	11.19 (8.37)	–4.81 (14.23)
Judge11	–0.99 (7.28)	–12.43 (8.24)	–31.55* (13.06)
Judge13	–9.10 (6.97)	–0.36 (8.06)	–21.56 (13.06)
Judge14	–11.02 (7.44)	2.80 (7.93)	–20.88 (12.80)
Number of cases	721	632	143
R^2	.014	.031	.080
Significance	.368	.021*	.180

*Note: * Significant at the .05 level*

Table A6 reports detailed regression results for distance from the guideline range:

Table A6
Linear Regression Model Results
Distance from the Guideline Range, All Sentences

	Mandatory Guidelines	PROTECT Act	Post-*Booker* I	Post-*Booker* II	*Kimbrough*/*Gall*
Constant	1.66 (1.76)	3.77 (2.77)	8.48* (3.14)	9.88* (3.78)	4.16 (5.41)
Judge4	2.39 (2.60)	–1.97 (4.06)	5.09 (4.14)	2.29 (5.26)	12.78 (8.66)
Judge5	0.88 (2.82)	–1.95 (4.09)	5.13 (4.56)	–1.34 (5.13)	—
Judge6	3.27 (2.57)	0.18 (3.89)	0.18 (4.24)	–1.90 (5.11)	6.67 (8.36)
Judge7	2.95 (2.84)	0.79 (4.12)	3.13 (4.74)	1.36 (5.29)	20.00* (7.22)
Judge9	3.98 (2.69)	6.20 (4.06)	–6.63 (4.14)	–5.04 (5.11)	10.80 (7.91)
Judge10	3.76 (2.74)	–1.83 (4.03)	–4.13 (4.45)	14.19* (5.60)	8.34 (7.73)
Judge11	4.57 (2.72)	3.65 (4.12)	–2.09 (4.70)	5.00 (5.56)	15.54* (7.38)
Judge13	3.26 (2.63)	-0.71 (4.09)	2.14 (5.08)	–1.29 (5.18)	1.14 (7.33)
Judge14	3.41 (2.46)	4.16 (3.80)	–0.65 (4.53)	–3.53 (5.38)	0.58 (8.24)
Number of cases	491	356	419	453	215
R^2	.010	.024	.036	.037	.066
Significance	.847	.482	.089	.048*	.073

*Note: * Significant at the .05 level*

Finally, Table A7 reports detailed regression results for distance from the guideline range, using only "discretionary" sentences:

TABLE A7
Linear Regression Model Results
Distance from the Guideline Range, Discretionary Sentences

	Mandatory Guidelines	PROTECT Act	Post-*Booker* I	Post-*Booker* II	*Kimbrough*/ *Gall*
Constant	2.24 (2.29)	4.90 (3.58)	9.53* (3.78)	11.81* (4.57)	5.20 (6.51)
Judge4	3.60 (3.43)	–2.36 (5.37)	9.33 (5.19)	4.66 (6.56)	19.47 (10.93)
Judge5	0.58 (3.46)	–2.59 (5.26)	5.68 (5.43)	–1.44 (6.23)	—
Judge6	4.41 (3.34)	0.20 (5.03)	2.41 (5.31)	–0.59 (6.42)	7.80 (9.94)
Judge7	3.40 (3.58)	1.44 (5.44)	5.08 (5.84)	3.07 (6.56)	23.43* (8.59)
Judge9	4.57 (3.38)	9.84 (5.44)	–7.09 (5.13)	–5.66 (6.26)	13.08 (9.46)
Judge10	4.82 (3.52)	–2.23 (5.26)	–3.23 (5.67)	18.08* (6.84)	10.59 (9.33)
Judge11	5.86 (3.46)	6.40 (5.58)	1.89 (6.44)	9.37 (7.06)	22.01* (9.10)
Judge13	4.53 (3.43)	–0.21 (5.26)	3.62 (6.23)	–1.28 (6.30)	0.92 (8.66)
Judge14	3.61 (3.09)	4.05 (4.77)	0.28 (5.48)	–2.91 (6.78)	0.80 (9.94)
Number of cases	384	269	319	349	172
R^2	.013	.036	.045	.051	.094
Significance	.835	.384	.105	.038*	.037*

*Note: * Significant at the .05 level*

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 A8
Linear Regression Models
Identity of Judge and Distance from Guideline Range, Three Periods

	% Variance Explained	Avg. Variance Explained	Model Significance
All Sentences			
Pre-*Booker*	1.2%	1.7 months	.338
Post-*Booker*	1.7%	3.0 months	.095
Kimbrough/*Gall*	6.7%	7.1 months	.073
Discretionary Sentences			
Pre-*Booker*	1.5%	2.1 months	.346
Post-*Booker*	2.7%	4.2 months	.031*
Kimbrough/*Gall*	9.4%	9.1 months	.037*

*Note: * Significant at the .05 level*

The results do not differ in any meaningful way from the models reported in the main text.[264] 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 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.[265]

264. *See supra* Table 3 and accompanying text.

265. *See supra* notes 235-36 and accompanying text.

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

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

	% Variance Explained	Avg. Variance Explained	Model Significance
All Sentences			
Mandatory Guidelines	4.5%	13.2 months	.001*
PROTECT Act	4.7%	13.3 months	.028*
Post-*Booker* I	3.8%	14.1 months	.028*
Post-*Booker* II	2.2%	9.7 months	.264
Kimbrough/*Gall*	6.1%	15.5 months	.044*
Excluding Mandatory Minimums			
Mandatory Guidelines	5.0%	9.7 months	.008*
PROTECT Act	4.4%	8.6 months	.195
Post-*Booker* I	6.1%	10.7 months	.028*
Post-*Booker* II	4.4%	10.0 months	.104
Kimbrough/*Gall*	8.0%	10.3 months	.180

*Note: * Significant at the .05 level*

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 (4.5%) and PROTECT Act periods (4.7%) individually than in the combined Pre-*Booker* period (3.1%).[266] They also show a substantial dip in the judge effect during the second eighteen-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 100% above 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 eighteen months after *Booker* than during any previous period, and grew even stronger (although not yet statistically significant) since *Kimbrough* and *Gall*. But again, the change appears more modest: 60-80% above pre-*Booker* levels rather than

266. *See supra* Table 1 and accompanying text.

several times pre-*Booker* levels.[267]

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.[268] The main text therefore relies, where possible and appropriate in light of the underlying legal changes, on longer period lengths.

267. *See supra* Table 2 and accompanying text.

268. *See* Anderson et al., *supra* note 23, at 290-91 (two years before the Guidelines, six years after); Hofer et al., *supra* note 47, at 284 (two years); Waldfogel, *Empirically Based Sentencing*, *supra* note 55, at 295 (four years).