Mandatory Sentencing and Racial Disparity:
Assessing the Role of Prosecutors and the Effects of *Booker*

Sonja B. Starr* & M. Marit Rehavi**
Draft, February 7, 2013

*FORTHCOMING, YALE LAW JOURNAL, FALL 2013*

**NOTE TO PARTICIPANTS IN THE LAW AND ECONOMICS WORKSHOP:** This paper is pretty long, so when I present it at the workshop, I will assume only that you’ve read through page 38 (and looked at the accompanying graphs). In case you prefer to read the whole thing, the last ten pages address some counterarguments and potential limitations to the regression discontinuity design approach, and also provide some simpler linear estimates of trends in racial disparity over time.

Recently, the U.S. Sentencing Commission has prominently claimed that the loosening of the formerly mandatory U.S. Sentencing Guidelines in United States v. Booker caused racial disparities in federal sentencing to quadruple, and has invoked this finding to support calls for strengthening the Guidelines again. But the Commission’s methods suffer from two flaws that are, unfortunately, pervasive in the empirical literature on sentencing disparities (including other recent studies of Booker). The first problem is a focus on the sentencing stage in isolation. Studies control for the “presumptive sentence” or closely related measures that are themselves the product of discretionary charging, plea-bargaining, and fact-finding processes. Any disparities in these earlier processes are built into the control variable, which leads to misleading sentencing-disparity estimates. The second problem is specific to studies of sentencing reforms: they use loose methods of causal inference that do not disentangle the effects of reform from surrounding events and trends.

This Article explains these problems and presents analyses that correct them and reach very different results from the existing literature. We address the first problem by using a dataset that traces cases from arrest to sentencing to assess

---

*Assistant Professor, University of Michigan Law School.
** Assistant Professor, Department of Economics, University of British Columbia. The research underlying this Article has been presented at the Ninth Circuit Judicial Conference, the National Sentencing Policy Institute, the National Bureau of Economic Research Summer Institute, the American Law and Economics Association Annual Meeting, the American Society of Criminology Annual Meeting, the Duke Legal Theory Colloquium, the Loyola Law School Faculty Workshop, the UBC Empirical Lunch, the CIFAR-IOG Workshop, and the Fawley, Law and Economics, and Criminal Justice workshops at the University of Michigan Law School. Thanks to participants for their feedback, and thanks to Daron Acemoglu, David Abrams, Alberto Alesina, Joe Altonji, Alan Auerbach, Nick Bagley, John Bronsteen, Ing-Haw Cheng, Kristina Daugirdas, John DiNardo, Avlana Eisenberg, Leonid Feller, Nicole Fortin, Nancy Gallini, Nancy Gertner, Sam Gross, Don Herzog, Jim Hines, Jill Horwitz, Thomas Lemieux, Justin McCrary, Julian Mortenson, Brendan Nyhan, JJ Prescott, Eve Brensike Primus, Adam Pritchard, Craig Riddell, Jeff Smith, and Sara Sun-Beale for helpful comments and conversations. Sharon Brett, Michael Chi, Ryan Bersowitz, Michael Farrell, Seth Kingery, Matthew Lee, Mids Panikkar, Art Robiso, Sabrina Speianu, and Adam Teitelbaum provided able research assistance.
disparities across all post-arrest stages. We find that most of the otherwise-unexplained racial disparities in sentencing can be explained by prosecutors’ decisions to bring mandatory minimum charges. We next address the problem of disentangling background trends by using a rigorous method called regression discontinuity design, which we apply to assess Booker’s effects on sentencing as well as charging, plea-bargaining, and fact-finding. Contrary to the Commission’s conclusion, we find that Booker did not increase racial disparity, and may have reduced it.
CONTENTS

INTRODUCTION.................................................................................................................................1
I. Prosecutors, Sentencing, and the “Hydraulic Discretion” Theory.................................5
   A. Studies Estimating the Extent of Unwarranted Sentencing Disparities.............9
   B. Our Dataset ..........................................................................................................................15
   C. Our Estimates of Racial Disparities in Charging and Sentencing...............16
   D. Interpretations and Limitations ...................................................................................19
      1. Possible Unobserved Offense Differences .............................................................20
      2. Possible Differences in Offender Characteristics .................................................21
      3. Possible Sources of Disparity that Our Estimates Leave Out ............................21
      4. Race, Gender, and Their Interaction ........................................................................23
III. The Booker Question: Does Expanding Judicial Discretion Increase Racial
    Disparity? .................................................................................................................................24
    A. Studies Assessing the Effects of Sentencing Policy Changes .............................25
    B. Our Method ......................................................................................................................30
    C. Regression Discontinuity Estimates of Booker’s Effects ........................................33
       1. Changes to Charging .................................................................................................33
       2. Changes in Plea-Bargaining .....................................................................................36
    D. Limitations and Causal Inference Challenges ...........................................................39
    E. Longer-Term Trends ......................................................................................................42
CONCLUSION.................................................................................................................................46
Figures & Tables ........................................................................................................................49
INTRODUCTION

In the United States, one of every nine black men under the age of 35 is currently behind bars,¹ and of today’s young black men, one in three can expect to be incarcerated at some point in his life.² These rates far exceed those of any other demographic group—for instance, black males are incarcerated at nearly seven times the rate of white males.³ The impact of demographically concentrated incarceration rates on offenders, families, and communities is a critical social concern.⁴ But why do these gaps exist? Can they be explained by differences in criminal behavior, or by differences in how the criminal justice system treats offenders? If it’s the latter, can the process be improved by reforms, such as changes to sentencing law?

These questions are not new. For decades, racial and other “legally unwarranted” disparities in sentencing have been the subject of considerable empirical research, which has in turn helped to shape major policy changes. Most importantly, the U.S. Sentencing Guidelines and their state counterparts were adopted with the goal of reducing such disparities. In 2005, when the Supreme Court’s decision in United States v. Booker rendered the formerly mandatory Guidelines merely advisory, Justice Stevens’ dissent predicted that the “result is certain to be a return to the same type of sentencing disparities Congress sought to eliminate in 1984.”⁵ Whether this prediction was accurate is perhaps the foremost empirical question in sentencing policy today. The most prominent study to date, a 2010 report of the U.S. Sentencing Commission, gave an alarming answer: that Booker and its judicial progeny had quadrupled the black-white sentencing gap among otherwise-similar cases, from 5.5% to 23%.⁶ In January 2013, the Commission issued an update with similar figures, this time combined with explicit calls for legislation returning the Guidelines to something close to their prior binding status.⁷

This Article introduces a new empirical approach, and gives a very different answer. The Commission’s methods are hobbled by two serious flaws that, unfortunately, pervade the rest of the empirical literature on sentencing disparity as

---

⁵ 543 U.S. 220(Stevens, J., dissenting).
⁷ U.S. Sentencing Comm’n, Continuing Impact of Booker on Federal Sentencing (2013). The update extended the last period through 2011; the new reported increase was from 5.5% to 19.5%. Id. at 108.
well. First, these studies consider the judge’s final sentencing decision in isolation, ignoring crucial earlier stages of the justice process—and if there are disparities at those stages, the sentencing disparity estimates will be misleading. Second, studies of changes in disparity after legal changes (like *Booker*) have failed to disentangle the effects of the legal change from surrounding events and background trends.

This Article develops both these critiques and presents our own research on racial disparities among federal arrestees, using a method that avoids these problems. We find a substantial black-white gap that appears to be introduced during the criminal process—but it appears to stem largely from prosecutors’ charging choices, especially the use of “mandatory minimum” offenses. We then turn to *Booker*, estimating its effects not only on sentencing, but also on charging, plea-bargaining, and sentencing fact-finding, something no prior studies have assessed. We find no evidence that judges’ use of expanded discretion worsens disparity, and suggestive evidence cutting the opposite direction.

Our research seeks to close a surprisingly wide gap that separates two bodies of scholarship: the theoretical and qualitative literature on how the criminal justice system functions (which uniformly recognizes the critical role of prosecutors) and empirical research on sentencing disparities (which effectively ignores that role). The modern criminal justice process is prosecutor-dominated. Prosecutors have broad charging and plea-bargaining discretion, and their choices have a huge impact on sentences. A central claim made by critics of mandatory sentencing is that restricting judicial discretion further empowers prosecutors, who tend to exercise that power in ways that perpetuate or worsen disparity. This “hydraulic discretion” theory has been described as a near-consensus view of sentencing scholars.

Yet the empirical research on sentencing disparity has not tested these claims, and fails to account for the role of prosecutorial discretion. In federal courts and other jurisdictions with sentencing guidelines, researchers typically estimate sentencing disparities after controlling for (among other things) the recommended Guidelines sentence. But the Guidelines recommendation is itself the end product of charging, plea-bargaining, and sentencing fact-finding. Controlling for it filters

---

8 While we focus on race, these weaknesses are also common in research on other “unwarranted disparities” (e.g., inter-district) and much of the non-disparity-related research on sentencing policy.

9 The *Booker* results are new to this Article. We begin, however, by discussing the results of a study of racial disparities in charging and sentencing (which was not focused on *Booker*) that we prepared for an economics journal: M. Marit Rehavi & Sonja B. Starr, *Racial Disparity in Federal Criminal Charging and Its Sentencing Consequences* (Univ. of Mich. Law & Econ. Empirical Legal Studies Ctr., Paper No. 12-002, 2012) (under review), available at [http://ssrn.com/abstract=1985377](http://ssrn.com/abstract=1985377). Here, we explore that study’s implications for debates in sentencing policy and legal scholarship.

disparities in those processes out of the sentencing-disparity estimates, and gives an incomplete and misleading view of the scope and sources of sentencing disparity.\textsuperscript{11}

We take a different approach, constructing a dataset that links records from four different federal agencies, allowing us to trace criminal cases from arrest through sentencing. We focus on the gap between black men and white men. Instead of controlling for the Guidelines sentence, we control for the arrest offense and other characteristics that are fixed at the beginning of the justice process. This method allows us to assess aggregate disparities introduced throughout the post-arrest justice process, from charging through sentencing. Using decomposition methods developed by labor economists for the assessment of wage and income inequality, we can then assess the contribution of each procedural stage (and other underlying case differences) to the total black-white gap.

Our results suggest that the traditional method used by the existing literature risks leading policymakers seriously astray. We begin by analyzing post-Booker cases only. After controlling for the arrest offense, criminal history, and other prior characteristics, we find a black-white sentence-length gap of about 10\%.\textsuperscript{12} But judges' choices do not appear to be principally responsible. Instead, between half and the entire gap can be explained by the prosecutor's initial charging decision—specifically, the decision to bring a charge carrying a "mandatory minimum." After controlling for pre-charge case characteristics, prosecutors in our sample were nearly twice as likely to bring such a charge against black defendants.

We then turn to Booker, using a method that takes prosecutors into account and also corrects another serious flaw that pervades the empirical literature on sentencing-law changes. This second problem is exemplified by the Sentencing Commission's findings. It found that disparities after Booker (averaged over a period of years) were larger than disparities before it. Even assuming that is true, it is a huge logical leap to conclude that Booker caused this increase—a classic confusion of correlation and causation. Many things change over time, and the greater disparity in the post-period could easily have nothing to do with Booker. Indeed, even if Booker had slowed an underlying trend of increasing disparity, the Commission's methods would make it look like there was greater disparity after Booker.

A useful tool for isolating the effect of an event or law change and disentangling it from those underlying trends is called regression discontinuity design. We use it to assess whether, in the immediate aftermath of Booker, there is a sharp break in an otherwise-continuous trend, which would provide a much stronger basis for inferring causality. Our method focuses on Booker's immediate effects, not its long-term effects, which admittedly is both a strength and a weakness. The long-term effects are presumably what policymakers care most about, but there is no good way to identify them—the causal inference problem is too serious. The immediate

\textsuperscript{11}A few studies (most notably the Sentencing Commission's Booker study) compound this problem further by also controlling for differences in Guidelines departure rates, filtering out one of the key aspects of the final sentencing decision as well.

\textsuperscript{12}This is in non-drug cases; our main sample for this portion of our analysis excludes drug cases due to data limitations discussed below. When drug cases are added, the gap rises to about 14\%.
effects can be more rigorously assessed. Fortunately, there is good reason to believe that if Booker had substantially changed racial disparity patterns in judicial decision-making, we would have seen at least part of the effect right away. Booker's effects on Guidelines compliance were not slow or subtle—departure rates immediately and dramatically spiked. If judges were inclined to use discretion in ways that widen the black-white gap, one would expect to see disparity jump right after Booker.

We do not see such a jump. Right after Booker, sentencing disparity did not increase, and may have modestly dropped. If Booker did have any adverse effects on black defendants relative to whites, it was probably a second-order result of charging changes: the use of mandatory minimum charges increased for black defendants immediately after Booker, but this effect appears to have been quite short-term.

We are very cautious about these findings. Even with our approach, identifying Booker's effects is hard. While Booker has been described as a "natural experiment," as an experiment it leaves much to be desired—it changed the legal regime for every non-petty federal offense at once, leaving no plausible control group. Our method does not require a control group and filters out longer-term trends effectively, but could be tricked by month-to-month fluctuations. And Booker was not a clean break in settled law: it came on the heels of a period of serious lower-court confusion, further complicating causal inference. We conduct tests to evaluate these problems, but we cannot erase the noise in the data or the complexity of the history.

Still, what we can say is that nothing in these data suggests that judges' use of their post-Booker discretion exacerbated racial disparity. And indeed, even apart from the challenge of filtering out long-term trends, there does not appear to be any evidence that aggregate racial disparity in the post-arrest justice process is actually larger in the post-Booker period. Unexplained racial disparities in that process were actually slightly smaller by the end of the study period (which runs through 2009) than they were just before Booker. We do not seek to explain the causes of that change, which is minor in any event; it is unclear whether Booker had anything to do with it. But it does appear that the Sentencing Commission's more alarming estimates were an artifact of its methods and its focus on sentencing in isolation.

Understanding the relative role of prosecutors and judges in producing disparities is important. The specter of increased disparity after Booker has been prominently cited to support new constraints on judicial discretion. For instance, the Department of Justice in the Bush Administration advocated mandatory "topless guidelines"—effectively, mandatory minimums but no maximums. The Sentencing Commission has recently advanced a multi-prong proposal to strengthen legislative and appellate-court constraints on judicial sentencing discretion—a proposal that in effect would restore the Guidelines very nearly to the legal status they enjoyed before

---

Such “solutions” could be counterproductive. Constraints on judges generally empower prosecutors by making their choices more conclusive determinants of the sentence. Our research suggests that mandatory minimums—prosecutors’ most powerful tools for constraining judges—are particularly important sources of disparity. Note that we do not claim our findings prove “discrimination” by prosecutors or anyone else. We are limited to what our data can capture, and unobserved differences between cases could justify different outcomes. Still, we have rich controls, including detailed arrest offense information, criminal history, and other demographic, geographic, and socioeconomic fields, yet substantial unexplained racial differences remain.

In Part I, we briefly introduce the federal sentencing framework and review the legal scholarship on prosecutorial and judicial discretion. In Part II, we present our critique of the “sentencing only” approach and the results from our alternative method. We discuss our method’s limitations, including sources of disparity that even our broader approach does not measure, and discuss the way race and gender disparities combine to affect black men, drawing on Starr’s recent findings in a related gender study. In Part III, we turn to Booker’s effects on racial disparity, beginning with our critique of the causal inference methods used by existing sentencing-reform research and proceeding to our regression discontinuity analyses. We conclude with possible policy implications.

I. Prosecutors, Sentencing, and the “Hydraulic Discretion” Theory

Federal prosecutors, like their counterparts in the states, have always possessed very broad discretion. Prosecutors choose what charges to bring, and the complex criminal code often provides a wide range of choices. Over 95% of cases then result in guilty pleas, and prosecutors control the terms of the deals they offer defendants. These can include the charges of conviction (“charge bargaining”), sentence recommendations and requests for “departures” from the usual range, and stipulations about sentencing-relevant facts (“fact bargaining”).

Traditionally, prosecutors’ discretion was matched by vast judicial discretion in choosing sentences, which was constrained only by broad statutory ranges—for instance, zero to twenty years. Statutory minimum sentences were not widespread before the 1980s, and still apply in only a minority of cases. Within the statutory ranges, judges were free to tailor sentences to the facts and the offenders’ circumstances. The disadvantage was that there was no good way to ensure that similar cases resulted in similar sentences.


In 1984, citing studies finding widespread racial, gender, inter-judge, and inter-district disparities in sentencing, Congress adopted the Sentencing Reform Act, which created a Sentencing Commission to devise binding Sentencing Guidelines. Under the Guidelines, complex rules determine the “offense level,” which is based on the conviction offense plus additional aggravating or mitigating “sentencing facts,” such as drug quantity or the defendant’s role in a group offense. The offense level is one of two axes of a sentencing grid; the other is the defendant’s criminal history category. Within each grid cell is a narrow range: “8 to 14 months,” for instance. Departures from this range were permitted only for specified reasons.

By greatly reducing judges’ discretion, the Guidelines concentrated tremendous power in prosecutors’ hands. As Kate Stith explains, “when judges had discretion to impose any sentence [in the statutory range], prosecutorial power was potentially limited or counterbalanced by the possibility of judicial discretion.” But under the Guidelines, plea-bargaining much more tightly constrained the sentence. The one feature of the Guidelines that was intended to limit prosecutorial power was the judge’s sentencing fact-finding authority. This system (called “real offense” sentencing) allows the judge to base a sentence even on uncharged conduct, so long as it falls within the statutory range for the crime of conviction. In principle, this system should reduce prosecutors’ ability to offer to understate the defendant’s culpability in exchange for a guilty plea.

Still, studies suggest that real-offense sentencing has not constrained prosecutors very much, because in practice prosecutors very strongly influence judges’ findings of fact. Plea agreements usually include factual stipulations, and even though DOJ has long directed prosecutors not to bargain over these facts, many studies have documented the persistence of fact-bargaining. Judges are not bound by the factual stipulations, and the power to diverge from them (relying on sentencing-stage evidence or on a Probation Office report) is an important aspect of judicial discretion. Still, judges typically lack the incentive, and may lack the information, to diverge from what the parties have agreed. One 1996 survey found that only 8% of judges said they “go behind” plea agreements somewhat or very frequently; 25% said they never do, while the rest said they did so infrequently. As

18 Id.
Nancy King put it, “[e]stablishing facts in an adversarial system without the assistance of adversaries is an awkward business.”

To the Guidelines’ many critics, this empowerment of prosecutors was a disastrous flaw, leading to harsh results for defendants generally and undermining the Sentencing Reform Act’s disparity-reduction goals. As Albert Alschuler put it, “the price of whatever success the Guidelines have achieved in reducing judge-created sentencing disparities has been the burgeoning of prosecutor-created disparities.”

Scholars often refer to discretion as being “hydraulic,” such that attempts to constrain it in one place will merely shift it to another. Stephanos Bibas, for example, wrote:

The criminal justice system operates like a toothpaste tube, and departures that are squeezed out of the judge’s end of the tube will wind up in the prosecutor’s domain. This hydraulic pressure means that departures will still exist, but they will now occur more often on prosecutors’ terms.

This theory has long pervaded scholarship about the Guidelines. As Terence Miethe wrote in 1987, “this ‘hydraulic’ or ‘zero-sum’ effect is so firmly entrenched as a criticism of current reform efforts that most researchers begin with the assumption that the displacement of discretion exists.”

Note that although scholars’ language often refers to shifts in “discretion,” this is a slight misnomer; the Guidelines did not really increase prosecutors’ discretion, which was already almost boundless. Rather, they increased their power: the choices they made more conclusively determined the sentence. In a 1996 survey, 86% of judges said that the Guidelines gave prosecutors too much power, and 75% said that prosecutors were now the actor with the most influence on sentences—more than judges themselves. Prosecutors thereby obtained greater leverage in plea-bargaining—they could nearly promise that defendants would get more lenient sentences if they pled guilty and harsher ones if they refused. In 2004, Marc Miller wrote, “The overwhelming and dominant fact of the federal sentencing system...is the virtually absolute power the system has given prosecutors....There is a lot of evidence to support this claim, but it can be demonstrated with one simple and awesome fact: Everyone pleads guilty.” After adoption of the Guidelines, plea rates skyrocketed from 87% to 97% by the new millennium.

22 King, supra note 17, at 303.
25 Miethe, supra note 7, at 155-56.
27 JOHNSON & GILBERT, supra note 19, at 6-7.
29 Alschuler, supra note 21, at 112; accord Stith, supra note 15, at 1425.
Since then, however, federal sentencing law has undergone another major change. In January 2005, the Supreme Court decided *United States v. Booker*, which rendered the formerly-mandatory Guidelines merely advisory. The Court held that a mandatory sentencing scheme in which a defendant’s maximum sentence could be increased based on judicial fact-finding violated the Sixth Amendment right to a jury trial. The Court could have remedied that defect by requiring more jury fact-finding, but it chose an alternate remedy: maintaining “real offense” sentencing, but severing the provision of the Sentencing Reform Act that rendered the Guidelines mandatory. The Court’s remedial choice remains reversible by Congress, which has so far not taken action. District courts today may depart from the Guidelines so long as the ultimate sentence is “reasonable.” In December 2007, in *Gall v. United States* and *Kimbrough v. United States*, the Supreme Court further clarified that courts of appeal should not deem sentences unreasonable merely because they fall far outside the Guidelines, and that judges may depart from the Guidelines on the basis of policy disagreements.

*Booker* was widely seen as an earthquake in federal sentencing law. Still, rendering the Guidelines advisory is not the same as eliminating them. Federal judges are still required to calculate the Guidelines sentence, and although they are then free to depart from it, they usually do not. There are many possible reasons for this continued conformity: federal judges might believe in the goal of reducing disparity, wish to avoid open-ended, subjective sentencing assessments, seek insulation from criticism or reversal, or simply treat the Guidelines as an “anchor.”

To the extent judges keep following the Guidelines, the power the Guidelines conferred on prosecutors will presumably remain largely intact. In addition, even if judges felt totally unconstrained by the Guidelines, prosecutors would retain at least two powerful sources of sentencing influence. First, their charging and charge-bargaining choices shape the statutory minimum and maximum sentences, which remain mandatory. Second, because they negotiate the factual stipulations accompanying pleas and also may introduce evidence at sentencing hearings, prosecutors have enormous influence over the information that gets to judges, and what judges know presumably will influence sentencing regardless of whether they

---

30 Booker, 543 U.S. at 245-46.
31 Id. at 246.
32 Id. at 265 (noting that the “ball now lies in Congress’ court.”).
35 Most federal sentences are still in the Guidelines range. See, e.g., SENTENCING COMM’N, supra note 6.
36 See Alschuler, supra note 21, at 97 (discussing sentencing influence of courthouse norms).
37 The “anchoring” literature shows that when people have to translate subjective judgments onto a numeric scale, they are often highly influenced by hearing some number mentioned—even numbers that (unlike the Guidelines) are actually irrelevant to the question posed. See James J. Prescott & Sonja Starr, *Improving Criminal Jury Decision Making After the Blakely Revolution*, 2006 U. ILL. L. REV. 301, 326 (2006) (reviewing the literature).
follow the Guidelines. Thus, even in the post-
Booker era, prosecutors should be expected to play a crucial role in the processes that shape sentencing.

In short, then, legal scholars and justice system participants widely agree both that prosecutorial choices are key drivers of sentences, and that sentencing law reforms involve tradeoffs between judicial and prosecutorial power. One might expect that this broad consensus would shape empirical research on sentencing disparities and sentencing reforms, but, oddly, it has not, as we demonstrate below.


For decades, unwarranted disparities in sentencing have been a major focus of empirical research. Overwhelmingly, these studies focus exclusively on the judge’s final sentencing decision, ignoring the rest of the justice process. In Section A, we review those studies and explain why this problem is so serious. In Section B, we describe the dataset that we constructed to enable a broader approach, and in Section C we summarize the results of our alternative strategy. In Section D, we discuss some limitations and implications of our results. Note that in this Part, we focus on disentangling the reasons for the gap between black and white sentences, using recent (post-
Booker) data; in Part III we turn to 
Booker’s effects.

A. Studies Estimating the Extent of Unwarranted Sentencing Disparities

Sentencing disparity studies generally begin by pointing to a gap in observed sentence outcomes and asking what generated it. For instance, black male defendants receive much longer sentences on average than white males do—a major contributor to higher incarceration rates for blacks. But is the sentence gap because black defendants have committed more serious crimes, or have more extensive criminal histories? Or are they treated differently in the criminal justice process?

Mass incarceration of black males has serious social consequences regardless of the causes. But if different offending patterns are to blame, the problem might be better addressed with policies focused on addressing the causes of crime, like poverty. In contrast, if the criminal justice system is treating like cases differently, then policymakers should focus on fixing that problem. Researchers thus seek to isolate the component of the sentence gap arising in the criminal justice process by controlling for some measure of the underlying severity of the case. But what measure? The answer to that question is the key difference between our approach and those of prior sentencing studies.

When researchers focus on the federal courts or other guidelines-based systems, the typical approach is to control for the “presumptive” or recommended Guidelines sentence—generally, the bottom end of the Guidelines range. There are

---

38 E.g., U.S. Sentencing Comm’n, supra note 6; Jeffrey T. Ulmer et al., Racial Disparity in the Wake of the Booker/Fanfan Decision: An Alternative Analysis to the USSC’s 2010 Report, 10 Criminology & Pub. Poly 1077 (2011) (following this approach and also reviewing prior literature doing the same).
variations on this approach, but all of them estimate differences in the actual sentence relative to what the sentence “should have been” under the Guidelines. Many studies also include controls for the statutory mandatory minimum. Studies in systems without guidelines similarly control for conviction severity. The problem with these approaches is that the key control variables are only distant proxies for the seriousness of the underlying conduct. They are the end product of the discretionary processes described above: charging, plea-bargaining, and sentencing fact-finding. And those processes too might produce disparities.

Other scholars have noted this problem. This includes, to their credit, some of those who employ the presumptive sentence approach themselves, who note that their accounts of disparities are incomplete. But these caveats generally are not mentioned when the work gets cited, and their importance may well be overlooked by policymakers. This is a serious mistake. The problem is not just that these accounts of disparity are insufficiently comprehensive—they are also potentially misleading. Absent an account of disparity at the earlier stages of the process, it is difficult to interpret disparities found in the final stage.

For instance, consider the most prominent recent sentencing-disparity study, the Sentencing Commission’s post-Booker report. The report finds that in 2008 and 2009, blacks received 23% longer sentences than whites, controlling for the recommended Guidelines sentence. This was an explosive finding, and led to calls (including by the Commission itself) to reinstate stronger constraints on judicial discretion. But how should this result be interpreted? Consider just three of many possibilities concerning what might have happened earlier in the justice process:

40 E.g., Nowacki, supra note 39.
41 E.g., Darrell Steffensmeier et al., Gender and Imprisonment Decisions, 31 CRIMINOLOGY 411 (1993).
42 See Alschuler, supra note 21, at 86-87; Engen, supra note 24, at 324-28; Shermer & Johnson, supra note 7, at 395.
43 See, e.g., Ulmer et al., supra note 36, at 1107-08. The Sentencing Commission itself notes in the 2013 update to its analysis that prosecutorial practices could contribute to sentencing disparities, but it does not connect this insight to its race results, nor does it alter its methods of estimating racial disparity to correct any of the problems identified here with its earlier report. See SENTENCING COMM’N, supra note 7, at 7. We focus principally on the 2010 report, which explained its methods of assessing racial disparity in much more detail.
44 U.S. SENTENCING COMM’N, supra note 6, at 3.
45 For instance, the former Chair of the Sentencing Commission, Judge William K. Sessions, has warned, citing the Commission’s disparity report, that post-Booker disparities were likely to lead Congress to adopt more mandatory minimums; Sessions himself proposes a simplified mandatory Guidelines system instead. At the Crossroads of the Three Branches: The U.S. Sentencing Commission’s Attempts to Achieve Sentencing Reform in the Midst of Inter-Branch Power Struggles, 26 J.L. POL. 305 (2011).
A. Prosecutors charge whites more harshly and/or offer them worse plea deals, such that the resulting Guidelines recommendation averages 23% higher for whites than for blacks with similar offenses and criminal histories.

B. Prosecutors charge whites more harshly and/or offer them worse plea deals, such that the resulting Guidelines recommendation averages 30% higher for whites than for blacks with similar offenses and criminal histories.

C. Prosecutors charge blacks more harshly and/or offer them worse plea deals, such that the resulting Guidelines recommendation averages 30% higher for blacks than for whites with similar offenses and criminal histories.

Under Scenario A, what looked like a 23% sentencing disparity now looks like judges sentencing more or less “correctly,” relative to underlying criminal conduct—they are correcting the disparity introduced by prosecutors. Under Scenario B, it actually seems that judges are not “favoring” whites enough—to sentence based on true culpability, they would have to do more to compensate for prosecutors’ favoring blacks. In contrast, under Scenario C, judges are compounding the underlying charging and plea-bargaining disparities; the “true” sentencing disparity is actually much more than 23%. If you don’t know which of the scenarios (or others) is true, it is risky to use the 23% figure as a guide to policy.

While this serious problem pervades the sentencing-disparity literature, in the Sentencing Commission study specifically, the problem is even worse. In addition to controlling for the presumptive sentence, the Commission also controls for whether the judge departed upward or downward from the Guidelines range, using a set of dummy variables for departure status. In doing so, the Commission is not just considering the sentencing decision in isolation—it is filtering out a key part of the sentencing decision itself. In effect, the Commission is estimating race gaps in the size of departures (and in sentence choices within the narrow Guidelines range), but filtering out whether there is a departure and, if so, in what direction. This is, to say the least, a strange choice, and could easily produce misleading results even if one the judicial sentencing decision were the only relevant source of disparity. This same problem also appears in the most prominent recent study responding to the Sentencing Commission report, that of Ulmer, Light, and Kramer; the authors critique other aspects of the Commission’s methods, but their main analysis of sentencing disparities also controls for departure status as well as the presumptive sentence.46

Another recent study by Joshua Fischman and Max Schanzenbach recognizes the problem with the presumptive sentence approach (and also does not make the

\footnotesize{Commission’s recent hearings on strengthening the Guidelines centered on the disparity concerns raised by the Commission’s report. See supra note 15 and accompanying text.}

\footnotesize{46 See Ulmer et al, supra note 36. This study does consider downward departure rates separately as an outcome variable, and finds substantial racial disparity in those rates (favoring whites, especially in recent years after Booker). That disparity in departure rates is filtered out of the sentence length disparity estimates, however, which presumably biases them downward. Similarly, the Sentencing Commission’s recent update to its analysis, supra note 7, also assesses departure rates as an outcome and likewise finds disparity favoring whites, but continues to filter that disparity out of the sentence length estimates.}
Commission’s further mistake of controlling for departure status). Fischman and Schanzenbach instead control for the Guidelines “base offense level.” This is an improvement over the presumptive sentence approach; it provides a somewhat fuller measure of judicial sentencing disparity, and may be the best approach possible using only the sentencing-stage data from the Sentencing Commission. But it still means that the authors’ sentence disparity estimates do not incorporate components introduced by the various prosecutorial decisions and negotiations that determine the base offense level. The base offense level is affected not only by charging and charge-bargaining, but also by a large part of the fact-finding required by the Guidelines. It incorporates, for instance, drug quantity in a drug trafficking case, or, in an assault case, the degree of physical contact and injury, the defendant’s intent, and the use of weapons. Disparities in any of those factual determinations, or in the prior charging or plea-bargaining processes, could bias the resulting sentencing-disparity estimates for essentially the same reasons explained above. To really avoid the problem with the presumptive sentence approach, one needs a measure of case severity that precedes all those discretionary processes.

The problem with the presumptive sentence control is compounded by a distinct source of potential bias that the existing literature has overwhelmingly failed to acknowledge: sample selection shaping the pool of sentenced cases. Nearly every study of sentencing disparity is confined to a sample consisting of sentenced defendants only—in federal court studies, typically only those sentenced for felonies or Class A misdemeanors (“non-petty offenses”), which the Sentencing Commission collects data on. To make it into the sample, cases must get through the criminal justice “funnel”: they must be arrested, charged, and convicted of a non-petty offense.

If these earlier processes are subject to demographic disparities, it could introduce sample selection bias into the estimates of sentencing-stage disparity. Suppose that all else equal, blacks are more likely to be convicted of a non-petty offense. If such a disparity exists, the base offense level measure could overstate the racial disparity in sentencing decisions if blacks are more likely to have a higher base offense level than whites.

48 Id. at tbl. 5 (listing controls).
50 In one portion of their analysis—the assessment of the effects of the Supreme Court’s decisions in Gall and Kimbrough on sentencing disparities—Fischman and Schanzenbach also present alternative results without the base offense level control. Fischman & Schanzenbach, supra, at Table 6. But without the base offense level control, there are no controls for the severity of the underlying cases. That approach is plausibly valid for the purposes of their assessment of these Supreme Court decisions’ effects if one makes the strong assumption that the underlying case mix did not change in racially disparate ways during their three-and-a-half-year study period. But the approach cannot be used for our purpose here: to disentangle the share of the observed racial disparity that appears to be explained by differences in underlying criminal conduct from the share that appears to be introduced in the criminal process. It also does not identify the amount of disparity or of the overall increase that is attributable to judges, the key figure for recent policy debates. For that purpose, it is not sufficient simply to drop the problematic controls—one must replace them with a better measure of the severity of the underlying case. As we explain below, we believe the arrest offense is the best available measure. Note that Fischman and Schanzenbach explicitly explain that they do not seek to assess the amount of disparity that is unwarranted by legitimate case differences; they focus only on change over time. Id. at 11. Some further differences with their approach are discussed in Part III.
offense, such that it takes a less serious case to get a black defendant sentenced. If so, we would expect blacks and whites who get sentenced to be unobservably different: blacks’ cases are less serious in a way that controlling for observable variables cannot capture. Sentencing disparity estimates within that sample will be biased because they cannot account for this unobserved difference. Again, without assessing the “funnel,” one cannot know whether to expect such a bias to exist, or if so, which direction it should cut.

Unfortunately, the empirical research on demographic disparities earlier in the justice process is relatively limited. It focuses almost entirely on certain measures of charge-bargaining, such as the rate of dropping charges; studies typically do not assess severity reductions. More importantly, few studies (and no federal studies) have assessed disparities in initial charging, even though without doing so, it is difficult to interpret charge-bargaining results. A few state-level studies have found racial disparities in the use of certain particularly harsh mandatory minimums, including a study of “habitual offender” charges in Florida and Pennsylvania and a Maryland study of add-on mandatory minimums for firearms.

At the federal level, many observers, including the U.S. Sentencing Commission, have pointed to racial gaps in the rate of mandatory minimum convictions. Fischman and Schanzenbach’s study provides useful new evidence that mandatory minimums may be an important contributor to sentencing disparities. But these studies raise important further questions. Because they do not control for underlying pre-charge case features affecting a defendant’s eligibility for mandatory minimums (such as the arrest offense), they do not examine the reasons for the mandatory minimum gap. They do not tell us whether black defendants have simply committed more crimes to which mandatory minimums apply, or whether there are racial disparities in prosecutors’ exercise of charging or charge-bargaining discretion.

---

51 See Shermer & Johnson, supra note 7, at 399 (reviewing the charge-bargaining literature, which mostly finds disparity favoring whites, and presenting their own findings showing no such disparity).
52 One early study by Spohn et al. found disparities favoring whites in the rate of filing felony charges in Los Angeles County, but did not analyze charge severity within felony charges. The Impact of the Ethnicity and Gender of Defendants on the Decision to Reject or Dismiss Felony Charges, 25 CRIMINOLOGY 175 (1987). See also sources cited note 7 and accompanying text (discussing another charging study).
56 U.S. SENTENCING COMM’N, supra note 6.
57 See supra note 47.
58 One study in Illinois courts evaluates whether judges differ from one another in their racial disparity patterns, finding that they do. See David Abrams et al., Do Judges Vary in their Treatment of Race?, J. Legal Studies (forthcoming). This study does not need to control for “presumptive sentence” because it can take advantage of the random assignment of cases to judges. The result interestingly shows that judicial discretion matters to racial disparity patterns. However, it does not answer the more basic question whether judges are actually treating similar defendants differently based on race, as opposed
The specification of an empirical model of disparity may seem like a purely scientific decision. But as Albert Alschuler has observed, it is bound up with normative questions: what kinds of disparity do we think are important? The choice of control variables determines what kind of disparity one is measuring, and so it should be shaped by a sense of the types of disparity policymakers care about. There are many reasons one might worry about demographic disparities in the justice process. For instance, such disparities might violate the Equal Protection Clause, exacerbate the social consequences of mass incarceration within particular communities, interfere with retributive or utilitarian punishment objectives, or undermine the justice system’s credibility.

We do not intend in this Article to resolve what policymakers’ objectives should be. But none of the reasons we can think of for caring about demographic disparities suggest that policymakers should confine their interest to equalizing sentences for cases in the same Guidelines cell. Rather, all imply that the key question is whether people with the same underlying criminal conduct (arguably including prior criminal history) receive the same sentence. Between the underlying criminal conduct and the sentence, there are many places in the process where disparities could be introduced. Policymakers should care about all of them.

A final disadvantage to the “presumptive sentence” approach is simpler: it does not control for differences in crime type, only for differences in crime severity according to the Guidelines. Judges might be more likely to depart from the Guidelines for some crimes than others, for reasons that have nothing to do with race. Such tendencies might well have racially disparate impacts, but they are not necessarily “unwarranted”—the nature of the offense is certainly a relevant sentencing consideration. Sentencing studies often do include controls for case type in addition to the presumptive sentence, but only for broad categories such as “drugs” or “violent crime,” which do not capture much nuance.

More precise crime-type controls, which we provide, can enable us to better distinguish the “disparate impact” component of racial disparity (the component that can be explained by non-racial factors like case type) from the component that we cannot explain with the variables we can measure, which could represent “disparate treatment” on the basis of race. The disparate impact versus disparate treatment distinction is crucial as a matter of constitutional law, although the extent to which

---

59 Alschuler, supra note 21, at 85-88.
60 E.g., U.S. SENTENCING COMM’N, supra note 6, at B1-B2 (using seven categories); Ulmer et al., supra note 36, at 1100.
it is normatively important is open to debate. We think all factors contributing to racial disparity in sentencing—whether legally “warranted” or not—are important for policymakers to understand, a point we return to below. But we believe that disentangling the reasons can help policymakers figure out what to do about them. In any event, studies like the Sentencing Commission’s generally purport to estimate “legally unwarranted” disparities, and if that is one’s objective, one should filter out legally relevant factors like case type.

B. Our Dataset

Our broader approach to the estimation of racial disparities requires something most researchers have not had: a dataset that traces federal cases from arrest through sentencing. We constructed it by linking files from four federal agencies: the U.S. Marshals’ Service (USMS, data from arrest and/or booking), the Executive Office of the U.S. Attorneys (EOUSA), the Administrative Office of the U.S. Courts (AOUSC), and the U.S. Sentencing Commission (USSC). It covers two stages of the process that the Sentencing Commission data alone (the sole source for most federal studies) do not include.

First, it includes the arrest offense, coded with 430 codes, and a string field describing the offense based on the arresting officer’s notes. This information allows us to substitute the arrest offense, instead of the presumptive sentence, as the key case-severity control. This substitution means that we are estimating sentencing gaps between black and white defendants who look similar near the beginning of the justice process, rather than between those whose cases have come to look similar near the end of it. We thus estimate the aggregate sentencing disparity introduced by decisions throughout the post-arrest justice process. In addition, the arrest offense codes provide far more detail on crime type than sentencing studies typically control for. The arrest offense is not a perfect proxy for underlying criminal activity, to be sure; we discuss its limitations below.

Second, our dataset includes rich information on initial charges, in addition to final charges. Specifically, we know the statutory section(s) under which the defendant was charged and convicted—for instance, 18 U.S.C. § 924(c). To assess charges quantitatively, we had to translate each combination of statutory sections into a numeric measure of total charge severity. This is not a simple task, which may be an additional reason prosecutorial decision-making is under-researched. Based on comprehensive research on every federal crime charged during the study period, we

---

62 Many critics have argued that the Supreme Court’s focus on discriminatory purpose is overly formalistic, instead advocating a focus on harms imposed on subordinated groups. See Jack M. Balkin & Reva B. Siegel, The American Civil Rights Tradition: Anticlassification or Antisubordination?, 58 U. MIAMI L. REV. 8, 8-10 (2004) (reviewing this literature). We are sympathetic to this view, but this longstanding debate need not be resolved for our purposes; empirical differentiation of the reasons for disparities has practical uses regardless.

63 These data are collected by the Bureau of Justice Statistics and made available to researchers under security conditions by the National Archive of Criminal Justice Data. We provide much greater detail on the construction and coding of the dataset in Rehavi & Starr, supra note 6, Data Appendix.
developed four different charge severity measures. The first three were grounded in sentencing law: the statutory maximum and minimum and a Guidelines-based measure. The fourth measure was based on sentencing practice: the mean sentence given in a baseline period before the study period. We then calculated the combined severity of all charges on all these measures.

Sometimes, the statutory provisions in the data contained multiple sentencing schemes depending on the facts of the case. Where possible, we resolved such ambiguities based on the other charges in the case, or otherwise used reasonable assumptions about which subparagraphs were likely to apply in most cases. However, in drug cases, the ambiguities were extreme—most cases were charged under omnibus provisions (such as 21 U.S.C. § 841(B)) encompassing all drug types and quantities. We could not meaningfully code the severity of such provisions, and thus could not assess initial charging disparities in drug cases. We subjected drug cases to more limited analysis focusing on disparities in the final mandatory minimum recorded at sentencing, a separate data field. We excluded child pornography cases because of a similar ambiguity. We also excluded immigration cases for different reasons: their stakes typically turn on deportation, and they involve different “fast-track” procedural environments. What remained were property and fraud offenses, regulatory offenses, violent crimes, and weapons cases.

We focus on the race gap between black and white U.S. citizen males. Starr’s separate gender study, discussed below, also assesses the race gap among women. Other racial groups’ outcomes were not analyzed because their numbers were very small. Hispanic defendants are included among the black and white defendants.

C. Our Estimates of Racial Disparities in Charging and Sentencing

We assess disparities introduced throughout the post-arrest justice process and break those aggregate disparities down by procedural source. We intend in future research to assess the specific contribution of every major stage of the justice

64 This measure is the Guidelines sentence that would apply if all of the statutory elements of all charged offenses were proven, but no other aggravating or mitigating facts were proven at sentencing. It is thus a “charge only” measure, allowing the effects of subsequent Guidelines sentencing fact-finding to be separated out.

65 Per the Guidelines, we assumed concurrent sentencing unless a statute specified otherwise.

66 For instance, if Charge 1 contained a heightened penalty if a gun was involved and Charge 2 was a gun charge, we applied the heightened penalty for Charge 1. Implementing this system required extensive legal research. In addition to coding all the triggering conditions for statutory or Guidelines sentencing enhancements for all federal crimes, we also had to code every crime’s elements to identify possible factors that could raise the penalty for any other charge in the case.

67 The Sentencing Commission’s Booker study includes immigration, but we agree with other scholars who have argued that it should be considered separately. See Ulmer et al., supra note 36, at 1085.

68 The USMS data do not identify Hispanic ethnicity. Among sentenced defendants, the Sentencing Commission’s data show that almost all persons of Hispanic ethnicity identify as white. If outcomes for Hispanics fall somewhere between those of blacks and non-Hispanic whites (as the Sentencing Commission, supra note 6, suggests), then our disparity estimates will be somewhat smaller than one would find if one looked only at blacks versus non-Hispanic whites.
process, but we began by focusing on initial charging and its role in explaining sentencing disparities. This stage has been almost entirely ignored by existing research, and it is especially important. In most federal cases, the initial charge is the final charge; charge-bargaining is the exception, not the rule. During our study period, dropping charges required a supervisor’s special approval. In initial charging, however, the line prosecutor has considerable discretion. In addition, before one can even begin to make sense of plea-bargaining disparities, one has to first know whether the baseline charges already reflect disparities.

The statistical analysis and the resulting estimates are described in detail in Rehavi and Starr (2012), an economics paper; accordingly, we do not re-present the tables or figures here. Instead, we summarize some key findings and explore their legal and policy implications. We had three main research questions:

(1) Do prosecutors charge otherwise-similar black and white arrestees differently?
(2) Do otherwise-similar black and white arrestees ultimately receive different sentences?
(3) How much of the sentencing disparity can be explained by the charging disparity?

By “otherwise-similar,” we mean similar in terms of the pre-charge case and defendant characteristics that we can observe. In the charging analysis (Question 1), we controlled for arrest offense, district, age, whether there are multiple defendants in the case, and county-level poverty, unemployment, income, and crime statistics. In the sentencing analysis (Question 2 and 3), we added additional controls that are recorded only for sentenced defendants: criminal history category and education level. Other variables were available only for subsets of the sample, but we checked to make sure that within those subsets, the results did not change when we added them. These included defense counsel type, marital status, and Hispanic ethnicity, as well as flags for whether certain facts were recorded in the written arrest offense description: guns, other weapons, drugs, conspiracy, racketeering, child victims, and official victims.

On question (1), while we did not find any racial disparity in the rate of facing felony charges, we did find significant racial disparities in charge severity, no matter which of our four charging measures we used. Using the statutory maximum,
guidelines, and past-sentence measures, the race gaps were moderate, ranging from 6 to 9%. The disparities in mandatory minimums were much more dramatic. After controlling for the variables above, black men were still nearly twice as likely to be charged with a mandatory minimum offense: 8% of white males faced such a charge, compared to nearly 16% of otherwise-similar black males.

Question (2) focuses on the aggregate sentencing disparity introduced by the entire post-arrest justice process. Black and white defendants were equally likely to be convicted of non-petty offenses (and thus to make it into the sentencing sample), but among those convicted there were significant unexplained sentencing disparities favoring whites. Most of the large raw sentencing gap (which was 50%) could be explained by the controls, and we used decomposition methods to identify which controls were the most important. The factors that could explain by far the largest components of the black-white gap were arrest offense and criminal history. But even after controlling for these and other variables, a gap of about 10% remained unexplained in non-drug cases; when drug cases were added, the unexplained gap rose to 14%. In some subsets of the sample the gap was larger, especially among more serious cases.

Initial charging is an important driver of these sentencing disparities—especially the decision to bring mandatory minimum charges. Out of the 10% otherwise-unexplained sentence gap in non-drug cases, half disappeared when we controlled for mandatory minimum charges. However, that estimate almost certainly understates the impact of mandatory minimum charges, because of the very conservative coding method we used—when our charge information was ambiguous, we assumed there was no mandatory minimum, which means we missed a substantial number of them. When we instead controlled for the final mandatory minimum (which is unaffected by the coding ambiguities), all the otherwise-unexplained racial disparity in the average sentence disappeared. This latter analysis could be performed for drug cases too. When drug cases were added to the sample, no significant disparity remained after controlling for the final mandatory minimum.

We subjected all these findings to a battery of robustness checks to assess whether varying the control variables, the sample definition, or the estimation method changed the results. Similar disparity patterns appeared in all specifications and subsamples. For example, the results did not vary substantially by region of the country or based on the arresting agency. Mandatory minimum charging disparities were similar across offense types, but the most common non-drug mandatory minimum in our data, and the one most responsible for driving sentencing disparities, was 18 U.S.C. § 924(C), the enhancement for crimes involving firearms. This statute has particularly harsh penalties: at least five years, running consecutively to other charges. There are higher minimums if the firearm is brandished or discharged and astonishing minimums (at least 30 years) if there is more than one § 924(c) count, 73

This difference reflects the coding issue, not new disparities introduced by charge-bargaining. Our analyses (using our coding for both the initial and final charges) do not show racial disparities in the rate at which mandatory minimums are dropped during plea-bargaining.

73
which could just mean there were two guns found in a defendant’s car. Prosecutors have considerable discretion in applying this statute, especially when the facts make the relationship of a gun to an offense ambiguous (for instance, the gun is simply found in the defendant’s car), and a lenient prosecutor may “swallow the gun” entirely.\textsuperscript{74} Michelle Alexander, in her recent book about race and incarceration, quotes a former U.S. Attorney describing one such incident:

I had an [assistant U.S. attorney who] wanted to drop the gun charge against the defendant [in a case in which] there were no extenuating circumstances. I asked, “Why do you want to drop the gun offense?” And he said, “He’s a rural guy and grew up on a farm. The gun he had with him was a rifle. He’s a good ol’ boy, and all good ol’ boys have rifles, and it’s not like he was a gun-toting drug dealer.” But he was a gun-toting drug dealer, exactly.\textsuperscript{75}

Our results suggest that this incident may not have been an anomaly.

\textbf{D. Interpretations and Limitations}

These results suggest the post-arrest justice process—especially mandatory minimum charging—appears to introduce sizable racial disparities. But are these gaps \textit{really} the result of racially disparate treatment? Or do they stem from unobserved differences that might be appropriate bases for different treatment? As Judge Nancy Gertner has warned, the quest to eliminate improper disparities should not lead us to seek “false uniformity” among cases that are actually dissimilar despite superficial similarities.\textsuperscript{76}

No observational study can fully tease out the causes of demographic disparities, because no dataset can ever capture all the subtle ways in which cases can differ.\textsuperscript{77} So one must tread cautiously when discussing causation—we speak in terms of “unexplained disparity,” rather than claiming to have proven “discrimination.” Still, our data are rich enough to shed light on some plausible causal theories, and we use them in this Section to explore whether the “unexplained” disparity is truly “unwarranted.” In addition, we point to some ways in which our disparity estimates may be under-inclusive—they do not encompass every discretionary choice shaping

\begin{footnotesize}
\begin{itemize}
\item \textsuperscript{75} \textsc{Alexander}, \textit{supra} note 4, at 118 (alterations in original).
\item \textsuperscript{77} In other settings involving potential unobserved variables, economists have developed a variety of useful quasi-experimental approaches, but these are of little help here. Such methods can help to analyze \textit{differences} in disparities (for instance, before and after policies or among different decision-makers), and we use one such approach below to assess \textit{Booker’s} effects. But they are not of much use in determining whether an apparent racial disparity is “real.” Race is inextricable from the rest of the person—there are no clever econometric tricks for isolating the effect of race from the effects of unobserved characteristics that might be correlated with race.
\end{itemize}
\end{footnotesize}
the black-white gap. Finally, we discuss the way these racial disparities appear to interact with gender disparities to produce particularly bad outcomes for black males.

1. Possible Unobserved Offense Differences

A first potential concern is unobserved differences in the underlying crime. This concern is less severe than it might have been: the detailed USMS offense codes, together with the written offense description field, capture considerable nuance in offense facts. In particular, they seem to effectively capture whether a gun was involved with the offense, which is important because of the substantial contribution of 18 U.S.C. § 924(c) charges to racial disparities.78 The multi-defendant case variable also captures an important offense characteristic, because multi-defendant cases often involve more serious crimes and often trigger conspiracy charges.

In drug cases (in addition to the limitations to the charge data), the arrest offense codes do contain an important ambiguity: they do not specify drug quantity. The EOUSA suspect investigation files record the drug quantity seized at arrest, but, when we examined patterns in the quantity distribution over time, we discovered a serious problem with this field beginning in 2004, when EOUSA adopted a new data entry system.79 We therefore reran our drug-case analysis on the last three years of data before this change (2001-03), and found that adding quantity controls did not substantially reduce the observed racial disparities. We also found racial disparities in the drug quantities found at sentencing fact-finding, after controlling for the seizure quantity and drug type recorded at arrest.80 This suggests that white defendants may be negotiating more favorable plea stipulations on quantity.

Similarly, the arrest data do not record the dollar value of losses in economic crimes. In some cases, the arrest codes suggest the scale (for instance, pickpocketing or vehicle theft), but in others (for instance, wire fraud) they do not. It is unlikely that differences in loss quantity could explain the racial disparities we found, however—in fact, they probably cut in the opposite direction. At least as recorded at sentencing fact-finding, whites tend to be involved in significantly higher-value property crime cases, after controlling for the other covariates.

Another important factor not captured by the arrest data is the defendant’s relative role in group offenses. We do not know of any anecdotal reason to believe that such differences could explain the racial disparities, i.e., that white defendants tend to be minor players in conspiracies while blacks tend to be leaders. If this were

78 Use of guns is usually clear from the arrest codes, and our description flags also included guns, drugs, and the combination thereof. Some cases might have been missed, but we seriously doubt that the number is large enough to explain the large racial disparity in 924(c) charges.

79 After extensive examination, we suspect the problem relates to the addition of a decimal point to the field—perhaps some (but not all) prosecutors are not noticing it. Unfortunately, it was impossible to determine which individual cases were affected. Comparisons to the Sentencing Commission’s quantity data do, however, make it apparent that the problem is with the new system, not the old one. It would be a service to future researchers and the public if EOUSA investigated this problem.

80 Quantities were converted into implied offense levels according to the Guidelines tables to allow comparisons across drug types.
the basis for the ultimate gaps, one would expect to see a noticeable difference in role adjustments at the sentencing fact-finding stage. But blacks get only very slightly worse role adjustments on average: a difference of 0.04 offense levels on the 43-level Guidelines scale, after controlling for the observed variables.\textsuperscript{81} This difference is statistically significant, but it is very small, and suggests that role differences are unlikely to explain much of the black-white sentencing gap.

2. Possible Differences in Offender Characteristics

Beyond the offense characteristics, there might be relevant offender characteristics that contribute to the race gap. We control for criminal history, the main offender characteristic built into sentencing law.\textsuperscript{82} The most obvious other possibility is socioeconomic differences, which are highly correlated with race. While poverty would not be a “warranted” reason for worse case outcomes, it would be a non-racial one and might suggest different policy approaches. However, the unexplained disparities we identify exist even after controlling for education level, defense counsel type (an excellent proxy for poverty), marital status, and county-level crime and economic statistics. Perhaps more remarkably, these socioeconomic factors taken together do not contribute significantly to the “explained” share of the racial disparity. This appears to be because poverty itself (as reflected by these indicia) is not an important predictor of higher sentences.\textsuperscript{83} Representation by a public defender is associated with slightly lower sentences, all else equal. This absence of socioeconomic disparity is good news, and we return to it below.

3. Possible Sources of Disparity that Our Estimates Leave Out

Although it is possible that our estimates of “unexplained” racial disparities include components that in fact have legitimate but unobserved explanations, in another sense these estimates are arguably under-inclusive. We estimate disparities across a much broader swath of the criminal justice process than existing studies do, but even our method does not encompass all of the key decision points. In addition to prosecutors and judges, other decision-makers shape criminal case outcomes—most notably, law enforcement agents and policymakers.

Any disparities produced by those actors’ choices will be found in the “explained” portions of the race gap—that is, the portions that come from the control variables. It is important not to overlook those portions when thinking about what should be done about racial disparity. Rather than simply using regression methods to filter them out, as most studies do, we therefore used decomposition methods that allow us to estimate the relative contribution of each control variable to the total observed black-white gap. These methods showed that the variables with by far the

\textsuperscript{81} The range of possible role adjustments is from -4 to +4.

\textsuperscript{82} Criminal history was not included in our main charging analysis because it is only recorded for the subset of charged defendants who were eventually sentenced. But within that subset, the charging disparities persisted after controlling for criminal history.

\textsuperscript{83} Some of these variables have significant effects on some outcome variables, but these effects are small and inconsistent in sign. There is no overall pattern suggesting that poverty worsens outcomes.
most explanatory value are arrest offense and criminal history. These variables may capture important differences that we want sentencing law to reflect, but they also reflect discretionary choices.

First, the recorded arrest offenses will be affected by law enforcement choices. This is a key limitation to our strategy of controlling for the arrest offense. We stated earlier that policymakers should ideally ask whether those who committed the same crime end up with the same sentence, but this is a very hard question to answer empirically. Researchers cannot observe what the defendants actually did. The arrest offense is a much better proxy for actual conduct than the presumptive Guidelines sentence, but it is not a perfect proxy. If it diverges from actual conduct in a racially disparate way, our “unexplained” disparity estimates will not capture that divergence. Nor do our estimates capture sample selection introduced by police decisions that determine who lands in the federal criminal justice system at all.85

In theory, these limitations could bias our results in either direction, but we think they probably mean we are understating the total disparities in the justice system. For arrest-stage disparities instead to explain our results, even partially, one would have to believe that federal law enforcement favors blacks. We think this is unlikely. Many criminal justice scholars have argued that black males are disproportionately targeted by law enforcement, while virtually nobody claims the opposite.86 Blacks are arrested for drug crimes much more often than whites are, even though blacks self-report both drug use and dealing at equivalent or lower rates.87 Beyond comparing arrest rates to reported crime rates, policing disparities are hard to study empirically because the underlying criminal behavior usually cannot be observed by researchers. But the existing quantitative evidence either supports the conventional wisdom or at least does not cut in the opposite direction.88

---

84 If the prosecutor’s pre-arrest involvement in the case influenced the arrest offense, this omission may leave out an aspect of prosecutorial discretion. When we drop cases with pre-arrest indictments (the cases with the most extensive pre-arrest involvement), disparity estimates increase.

85 Blacks are 45% of the sample, a rate far exceeding their general population share; the question is the extent to which this overrepresentation reflects actual crime rates versus policing patterns. This gap is included neither in the “explained” part nor the “unexplained” part of our disparity estimates; we can only decompose disparities within the set of cases we have data on.

86 See, e.g. ALEXANDER, supra note 4. In addition, surveys consistently find white Americans believe the police are fair while black Americans do not; these perceptions may reflect real differences in experience. See Jon Hurwitz & Mark Peffley, Explaining the Great Racial Divide: Perceptions of Fairness in the U.S. Criminal Justice System, 67 J. POLITICS 762 (2005).

87 See Alexander, supra note 4, at 99-100 (reviewing these studies); see also William J. Stuntz, Race, Class, and Drugs, 98 COLUM. L. REV. 1795 (1998) (observing that drug enforcement targets open-air markets, which are dominated by black men).

88 See, e.g., Andrew Gelman et al., An Analysis of the NYPD’s Stop-and-Frisk Policy in the Context of Claims of Racial Bias, 102 J. AM. STAT. ASSOC. 813 (2007) (finding evidence of disparities against blacks in stop-and-frisk procedures, and reviewing the policing-disparity literature). Much of the existing research focuses on traffic stops, and reaches mixed results. Blacks and males are likelier to be stopped and searched. Some researchers have found a lack of disparity in the “hit rate” of stops and searches (e.g., the rate of finding drugs), which they argue makes the policing pattern “rational.” See, e.g., Nicola Persico & Petra Todd, Generalising the Hit Rates Test For Racial Bias in Police Enforcement, With an Application to Vehicle Searches in Wichita, 116 ECON. J. F351 (2006). Others find lower hit rates for
To be sure, federal law enforcement could be different, but we are likewise unaware of any anecdotal suggestions that federal agents favor black suspects.

In addition, both the arrest offense and criminal history components of the “explained” disparity reflect subjective policy choices: important sources of disparity may simply be built into the law.\(^9^9\) In the Fair Sentencing Act of 2010, Congress responded to such a concern, partially mitigating the sentencing framework’s notoriously harsh treatment of crack cocaine cases.\(^9^0\) But the crack laws are not the only example of particularly heavy punishments being given to crimes disproportionately involving blacks. The harsh gun enhancements under 18 U.S.C. § 924(C) are another example—because black men are more frequently arrested with guns, these would disparately impact black men even if they were neutrally applied. Similarly, black males are also more frequently arrested for violent crimes, and sentencing law is often harsher on these crimes than on nonviolent crimes that might reasonably be considered more serious.\(^9^1\) These sentencing-law features are built into the “arrest offense” component of the measured disparities.

The criminal history component likewise reflects a subjective policy judgment to assign heavy weight to past crimes, even though those crimes have already been separately punished. There are many competing considerations surrounding that judgment. But it has a racially disparate impact. Moreover, this choice magnifies whatever racially disparate treatment exists in the criminal justice system, by carrying its impact from one case to the next: the criminal history score may be influenced by disparate treatment in past cases.\(^9^2\) That past disparity will appear as part of the “explained” disparity, so it is easy to lose sight of it—it will be filtered away by controlling for criminal history.\(^9^3\) Underlying “unwarranted” disparity can thus come to appear “legally warranted.”

4. Race, Gender, and Their Interaction

Finally, another limitation is that we only include men. Starr’s related study examines gender disparities and race-gender interactions.\(^9^4\) She finds unexplained gender disparities that dwarf the racial disparities our joint study found: men receive sentences that are over 60% longer than those of comparable women, conditional on

---

\(^{9^9}\) See Alschuler, supra note 21, at 87-88.


\(^{9^1}\) For instance, suppose X, who is unarmed, obtains $20 from Y by threatening to hit Y, and runs off with it. With no aggravating factors, his Guidelines offense level for robbery would be 20—the same offense level that would have applied had he defrauded Y out of nearly $1 million.

\(^{9^2}\) See Stith, supra note 15, at 1432-33.

\(^{9^3}\) Through its “career offender” and “armed career criminal” provisions, federal sentencing law is particularly harsh on cases that combine violent or (especially) gun cases with extensive criminal history—another structural feature with particularly harsh effects on black men. See 18 U.S.C. § 924(e) (2006); U.S. SENTENCING GUIDELINES MANUAL §§ 4B1.1, 4B1.4 (2011).

\(^{9^4}\) Starr, supra note 13.
the arrest offense, criminal history, and other pre-charge observable characteristics. These gaps are much larger than most other studies have estimated because—as with race—they mostly appear to arise prior to the final sentencing decision.\textsuperscript{95} The data suggest that differences in offender characteristics not captured by the main control variables may explain substantial shares of this gap, particularly differences in childcare responsibilities and perceived role in group offenses. But Starr finds large unexplained disparities (over 50\%) even among non-parents and in one-defendant cases, so these explanations do not appear to come close to explaining the whole gender gap, nor do any of the other theories Starr is able to test.

Notably, the gender gap was substantially larger (about 75\%) among black defendants. Among females, the racial disparities we found for men do not recur—there is no significant unexplained black-white gap in female sentences. The black female/white female gap appears to be explained entirely by differences in arrest offense and criminal history—although, again, it is possible that these factors build in structural, arrest-stage, or other hidden sources of disparity.

As noted above, black males are incarcerated at extremely high rates in the United States, and in assessing this problem, policymakers should consider both the race and gender dimensions and their interactions. Black male defendants appear not only to face the harsher side of both the racial and gender disparities, but also an additional interaction effect—an extra apparent “penalty” for being \textit{both} black and male. Gender disparity need not be seen as being about “special” treatment of women—rather, one could ask why the criminal justice system appears to treat males so much more harshly. If it did not, Starr’s data suggest that many fewer black men would be in prison.

\section*{III. The \textit{Booker} Question: Does Expanding Judicial Discretion Increase Racial Disparity?}

The results above suggest that while the role of prosecutors in producing the black-white sentencing gap has been overlooked by the empirical literature, the role of judges may have been overstated. Still, we do not claim judges’ contribution is necessarily trivial. Depending on coding choices, we found up to several percentage points, on average, of racial disparity that neither the charges nor underlying case features could explain, mostly affecting more serious cases. Given this fact, one might wonder whether the Supreme Court’s recent expansion of judges’ discretion was a mistake, and whether disparity could be reduced by restoring tighter constraints, as some have proposed. We thus now turn to the issue that so worried Justice Stevens in his \textit{Booker} dissent: has freeing judges to sentence outside the Guidelines led to an increase in unwarranted disparities?

\textsuperscript{95} In the gender context, an especially large share of the disparity appears to arise in sentencing fact-finding. \textit{Id.} at 4. Mandatory minimums are also important, but only in drug cases, presumably because women are rarely arrested for the kinds of non-drug crimes to which mandatory minimums apply.
The Sentencing Commission has given the most prominent answer to this question so far, and its answer is a resounding yes. Its race findings have garnered understandable attention, because they are shocking: Booker and its progeny appear to have led to a more-than-fourfold increase in racial disparity in sentencing, from 5% to 23%. We disagree. In Section A, we explain our methodological objections, which extend beyond the Sentencing Commission’s study to the rest of the literature examining the disparity consequences of sentencing law reforms. In Section B, we present the results of an alternative analysis reaching tentative conclusions that contradict the Commission’s. Unlike the results above, these findings are completely new to this Article and are therefore presented in more detail. In Section C, we discuss the limitations on our analysis, and why researchers will probably never be able to give an entirely definitive answer to the question of Booker’s effects.

A. Studies Assessing the Effects of Sentencing Policy Changes

A subset of the sentencing disparity literature focuses on measuring \textit{changes} in disparity resulting from changes to sentencing law, such as Booker. These studies have two pervasive methodological problems. First, they typically control for the presumptive sentence, ignoring any effects of sentencing reform on pre-sentencing decision-making. Second, they fail to disentangle the causal effects of policy changes from underlying trends and surrounding events. We explain both problems here.

The problem with the presumptive-sentence approach is largely explained above, but it impacts sentencing-reform studies in a slightly different way. In principle, studies focusing on \textit{changes} in disparities have an advantage over those that estimate the \textit{extent} of “unwarranted” disparity: the ability to ignore the possibility of differences between groups that the observed variables do not capture.\footnote{See Fischman & Schanzenbach, supra note 47, at 11.} Suppose the control variables amount to only a “broken yardstick” for measuring the defendant’s underlying criminal behavior—for instance, suppose the “presumptive sentence” variable diverges from true case severity in racially disparate ways. In a policy-change study, so long as the same broken yardstick is used before and after the policy change, one can validly estimate the policy’s relative effects on different groups. This “advantage” is a mixed blessing: estimates of changes in disparity are less policy-relevant if we do not know whether the disparity in either the pre- or the post-period is “real.” Still, not every study needs to answer every question, and research that brackets the “is this real?” question can be useful.

However, a serious problem arises if one cannot be confident that the yardstick itself has not been affected by the policy change. Consider again the Sentencing Commission 2010 study, discussed above. It found that the black-white gap went from 5% before Booker to 15% after, and finally to 23% after Booker’s successor cases Kimbrough and Gall. Below, we discuss other problems that make it very problematic to infer that these changes were \textit{caused} by either Booker or Kimbrough/Gall. But let’s start with a more basic question: Do these numbers actually tell us that racial disparity in sentences has grown?
In each period, the Sentencing Commission estimates sentencing disparities conditional on the presumptive sentence (likely a “broken yardstick” for the reasons discussed above), and then compares the disparities across time periods. If one were certain that racial disparities in the processes determining the presumptive sentence remained constant pre- and post-Booker, then this would be a “same broken yardstick” comparison. Whatever biases were hidden in the presumptive sentence variable would affect the estimates for both time periods similarly, so the comparison would be apples-to-apples. The problem is that Booker may have replaced one broken yardstick with a different one, if it affected charging, plea-bargaining, or sentencing fact-finding in racially disparate ways. Cases with the same presumptive sentences may represent different actual conduct pre- and post-Booker in ways that vary by race. Sample selection bias is also a potential problem: Booker may have changed which cases are winnowed out by the “funnel.”

There is good reason to worry about these potential biases. One clear lesson from the legal scholarship and practitioner surveys reviewed in Part I is that the stages in the criminal justice process are interrelated. Charging, plea-bargaining, and fact-finding all occur in anticipation of and in an attempt to influence the sentencing consequences. It is not even remotely safe to assume that changes in sentencing law do not affect decision-making at those earlier stages. After all, consider what happened after the Guidelines were adopted: a drastic increase in guilty pleas, which law scholars (very plausibly) attribute to prosecutors’ sharp increase in leverage.

There are many theoretically plausible ways decision-making prior to sentencing could change after Booker. For example:

• Prosecutors might have to offer more favorable plea deals to induce guilty pleas, potentially resulting in more favorable findings of fact, reduced charges and presumptive sentences, and perhaps more trials.\(^9^7\)
• Prosecutors could respond to the reduction in their power over Guidelines sentences by expanding use of their other tool for constraining judges: statutory mandatory minimums.
• Judges might become less willing to make findings of facts that diverge from the plea stipulations, because doing so is no longer necessary to achieve what they perceive as a just sentencing result—they can depart instead.

These changes would only bias estimates of post-Booker changes to racial disparity if they had a racially disparate impact on the presumptive sentence or on the composition of the sentenced sample.\(^9^8\) It is, of course, possible that this is not so, but one cannot simply assume it is not so—it must be tested. However, all of the existing studies of Booker (and prior studies of the initial shift to mandatory

---

\(^9^7\) If cases thereby became more resource-intensive, one might expect prosecutors to bring fewer cases or fewer charges per case. See James J. Prescott, Empirical Evidence of Prosecutorial Charging Manipulation (2006) (unpublished manuscript) (on file with author).

\(^9^8\) This could be the case even if the changes looked superficially equivalent by race. For instance, if prosecutors doubled their use of mandatory minimums for both blacks and whites in response to Booker, but their underlying use of mandatory minimums was twice as common for blacks, the doubling would look racially neutral, but would have twice the impact on blacks’ sentences.
sentencing) do assume exactly that, usually implicitly. Other studies have criticized various other aspects of the Sentencing Commission’s Booker study and have reached different conclusions. But these studies too have taken the sentencing-stage-only approach, controlling either for the presumptive sentence or something closely related (the Guidelines “base offense level”), and are subject to the same concern.99

These studies, in short, ignore the “hydraulic discretion” theory that has dominated theoretical scholarship about sentencing reform. Conversely, key aspects of the hydraulic discretion theory remain almost completely untested empirically.100 No empirical studies have yet used case data to assess changes in disparities in charging, plea-bargaining, or sentencing fact-finding in the wake of Booker. One study judges and public defenders about their perceptions of whether aspects of plea-bargaining had changed.101 However, the researchers did not evaluate these perceptions’ accuracy, and the perceptions of judges and defense counsel varied quite substantially.102

Just a few studies have looked at changes in charging and plea-bargaining disparities in response to earlier changes to sentencing law and policy. Wooldredge et al. found that Ohio’s shift to mandatory sentencing reduced race-related disparities in charge-bargaining, yet increased racial disparities in sentencing (a surprising result), but the authors did not evaluate changes in initial charging, without which the results are harder to interpret.103 In a 1987 study of Minnesota’s adoption of mandatory sentencing guidelines, Miethe did evaluate initial charging and found a small but significant increase in gender disparity and no significant change in racial disparity; plea-bargaining disparities were unchanged.104 No studies have evaluated changes in disparities in sentencing fact-finding.

Beyond the failure to account for pre-sentencing stages of the process, recall that the Sentencing Commission’s study of Booker has an additional problem: it also controls for departure status, thereby also filtering out some of the potential disparities in the sentencing decision as well. This is an especially surprising choice for a study of Booker’s effects, because—as we will see below—Booker dramatically changed the probability of a departure from the Guidelines by authorizing departures that were previously forbidden. So it is odd to compare racial disparities in

---

99 E.g., Ulmer et al., supra note 36 passim (finding a post-Booker increase in racial disparity in incarceration rates but not length); Fischman & Schanzenbach, supra note 38, at 2-3 (finding mixed results in an analysis of multiple doctrinal changes affecting judicial discretion, but concluding that expanded discretion does not increase racial disparity); Nowacki, supra note 39.
100 Engen, supra note 24, at 324-25.
101 Jeffrey T. Ulmer & Michael T. Light, The Stability of Case Processing and Sentencing Post-Booker, 14 J. GENDER, RACE, & JUST. 143 (2010) (finding perceptions of increased detail in factual stipulations and appeal waivers in plea agreements, but also increased entry of “open pleas” with no agreement).
102 Id.
104 Miethe, supra note 7, at 167-71 (1987). The Miethe and Wooldredge et al. studies are rare examples of studies that focus on “hydraulic” effects of sentencing reform; however, both are subject to the other critique raised below concerning causal inference from changes over time.
sentencing before and after Booker only after filtering out those mediated by racial differences in departure rates.

In addition to the use of inappropriate control variables, there is another serious methodological concern with these studies: causal inferences from changes over time are always risky, because many things change over time. Comparisons of averages between periods before and after a policy change, while appealingly simple, can be misleading.

These studies generally compare the average disparity before and after a policy change. In most, disparities are estimated separately for each period using a regression model that controls for the presumptive sentence and other observed variables.\(^{105}\) The recent federal studies have focused not just on Booker, but also on other recent policy changes affecting judges’ sentencing discretion. One such change was the PROTECT Act of 2003, which imposed rules intended to discourage downward departures from the Guidelines. It required courts to report to Congress on departure rates, required written justifications for departures, provided for de novo appellate review of departures in some cases, restricted the Sentencing Commission from creating new grounds for downward departures, limited judicially initiated downward adjustments for “acceptance of responsibility,” and directed DOJ to adopt an action plan for reducing departures.\(^{106}\) Another important change was the Supreme Court’s December 2007 decisions in Kimbrough and Gall (discussed above), which reinforced the Booker holding.

The Sentencing Commission focused on three primary time periods, with cases classified by sentencing date: (1) PROTECT-to-Booker (nearly two years), (2) Booker-to-Kimbrough/Gall (nearly three years), and (3) post-Kimbrough/Gall (nearly two years). It found the lowest black-white disparities in period 1, when judicial discretion was the most limited, and the greatest in period 3, when discretion was broadest.\(^{107}\) A competing study by Jeffrey Ulmer, Michael Light, and John Kramer criticized aspects of the Commission’s method, but it too compared averages across these time periods (as well as earlier periods).\(^{108}\) It similarly found increases in racial disparity in the post-Booker and post-Kimbrough/Gall periods, although these were concentrated in the decision to incarcerate defendants rather than in sentence length among those incarcerated.\(^{109}\) A recent study by Jeffrey Nowacki similarly compares

\(^{105}\) The use of separate regressions means that the case mix is not controlled for across time periods, another problem. Some studies report single regressions with race-time period interactions, which is preferable (but does not solve the other problems raised here). See Ulmer et al., supra note 36 (reporting both methods).


\(^{107}\) Ulmer et al., supra note 36, at 1091-94 also include the pre-PROTECT period.

\(^{108}\) Ulmer et al., supra note 38.

\(^{109}\) Id. Much of the authors’ criticism focused on the Commission’s failure to separate the incarceration decision from the length decision. In our view there is no theoretical reason the two decisions necessarily need to be separated; either approach is acceptable. The main problem with the Commission’s approach is that even though it kept the zeros in the main sample, it log-transformed the outcome variable; since the log of zero is undefined, the Commission assigned the value of 0.01 months to the non-incarceration sentences. The problem is that the choice of 0.01 is arbitrary, and the resulting effect estimates are sensitive to these arbitrary choices.
the cases from 2002-2004 (pre-
Booker) to those from 2005-2008 (post-
Booker), and finds a fairly small but significant increase in disparity in the latter period, controlling for final Guidelines offense level, criminal history, and other variables.110

But comparison of averages across such broad periods is at best suggestive. This kind of comparison is too blunt a tool for causal inference. Differences in the averages between periods might merely reflect longer-term trends or other intervening events. If racial disparity were rising steadily throughout the period, for instance, the average disparity after 
Booker would necessarily be larger even if 
Booker had no effect on racial disparity. In fact, this would be true even if 
Booker actually slowed the rate of increase in disparity.

Sentencing disparity might well be affected by numerous other developments over periods of that length. These could include, for instance, the policies and law enforcement priorities of a Presidential administration changing or taking time to trickle down to line prosecutors; changes in the composition of the judiciary and the U.S. attorneys’ offices; or something as simple as administrative changes in supervision of prosecutors. Even if these developments had no racial purpose, they might have had racially disparate impacts. Causal inferences would be more credible if effects were visible in a much shorter time window, such that one could more confidently assume that 
Booker is the only important change that could have driven the outcome. One can also filter the surrounding trends out of the estimates of the policy’s effects by including them in the regression.

Among the recent 
Booker studies, Fischman and Schanzenbach’s offers an improvement on the standard approach. Their model filters out year-to-year variation in sentencing patterns for different categories of crimes and judicial districts, which captures an important subset of the things that might vary over time. They focus on changes in appellate review of sentencing, and find that in general, looser review has not been associated with increased racial disparity, although (like the Sentencing Commission) they do find a recent increase in disparity after 
Kimbrough and 
Gall.111 However, their approach only filters out trends in racial

---

110 See Nowacki, supra note 39.
111 The authors further examine 
Kimbrough/Gall and their predecessor case 
United States v. Rita, 551 U.S. 338 (2007) with an “event study” approach that effectively averages disparities over six-month periods, rather than the Commission’s longer periods. Id. Because they leave out the five months between 
Rita and 
Kimbrough/Gall, the last pre-period and first post-period are actually nearly a year
disparity if they are mediated by the crime category or district; any trends driven by other factors are left in. Below, we set forth an approach that filters out continuous trends in racial disparity itself (rather than trends in particular factors that contribute to it), and that uses monthly data to capture within-year variation as well.

B. Our Method

Our alternative method is called regression discontinuity design ("RD").\textsuperscript{112} Rather than comparing racial disparities averaged over periods of years, we create flexible regression models that filter out month-to-month trends (including non-linear ones) in sentences and other relevant outcomes. We then look for sharp breaks in these trends—discontinuities—immediately after \textit{Booker}. Like other studies, we base our causal inferences on changes over time, and any unmeasured changes that coincide with \textit{Booker} could trick us. But because we are looking for \textit{immediate sharp} changes, this concern is less grave. While a lot can change in a couple of years, usually a lot less changes suddenly in a couple of months. In addition, even if continuous background trends did have a noticeable effect on disparities in those couple of months, our method filters the trends out. We are looking only for sharp breaks that coincide with \textit{Booker}. If the surrounding trends are fairly smooth and there is a sudden break at \textit{Booker}, the inference that \textit{Booker} caused the change depends only on the assumption that no other unobserved factor affecting sentencing disparity \textit{suddenly} changed at the time of \textit{Booker}.

We use the same dataset described in Part III, but draw a broader sample from it. The sample runs from fiscal years 2001 to 2009 and includes women and non-citizens. It includes all non-immigration cases except identity theft, which was subject to other major sentencing-law changes near \textit{Booker}.\textsuperscript{113}

Our overall research interest is measuring the effect of changes to judicial sentencing discretion on sentencing and case processing disparities. We begin by looking at Guidelines departure rates, not because that is the ultimate outcome of interest, but because departure rates help us determine which legal reforms amounted to important changes to judges’ discretion in practice. They directly measure Guidelines compliance, and thus are the most logical measure of the extent to which the Guidelines actually constrained judicial behavior at any given time. We focus our attention on \textit{Booker} itself, not on its progeny \textit{Kimbrough} and \textit{Gall} or the PROTECT Act’s tightening of the Guidelines. The reason can be seen plainly in Figure apart. Although this analysis improves on the Sentencing Commission’s, we think an even finer-grained approach to time trends yields greater payoffs for causal inference, and we also prefer to focus on \textit{Booker}, the bigger legal change.

\textsuperscript{112} RD estimators are widely used in the education, public finance, political economy and labor economics literatures to recover causal estimates when randomized experiments are not possible. \textit{See} David Lee & Thomas Lemieux, \textit{Regression Discontinuity Designs in Economics,} 48 J. ECON. LIT. 281 (2010); Guido W. Imbens & Thomas Lemieux, \textit{Regression Discontinuity Designs: A Guide to Practice,} 142 J. ECONOMETRICS 615 (2008).

\textsuperscript{113} Including women and non-citizens improves the precision of the estimates by increasing the sample size within each month. However, the results are substantively similar if these groups are excluded. We include controls for gender and citizenship.
1, which plots departure rates by sentencing month. Note that 96% of these departures are downward.

The vertical lines in Figure 1 mark four key events: the PROTECT Act, Blakely (Booker’s immediate predecessor), Booker, and Kimbrough/Gall (which clarified and strengthened Booker’s holding). As this graph makes clear, Booker was a major shock to the sentencing discretion afforded to judges. Departures increased immediately and substantially, from about 30% to about 40%. Although there are other month-to-month fluctuations, Booker marks by far the most dramatic break. The sharpness of this change helps to alleviate one substantial concern about RD—its inability to capture effects that occur slowly. It is certainly possible that the full effects of Booker took a while to take hold, and the inability to test that possibility is a disadvantage to our method. But if judges were inclined to use broader sentencing discretion in ways that disadvantage blacks, one would expect to see at least some of that effect in the immediate vicinity of Booker, because Booker immediately changed judges’ willingness to depart from the Guidelines.

In contrast, PROTECT and Kimbrough/Gall were not nearly as dramatic a change to the sentencing regime in practice. PROTECT appears to have caused no sudden change at all in departures. Kimbrough and Gall may have been more important—departure rates did rise afterwards—but the rise continued a trend that began three months before the decisions, and there was no sudden break in the trend (nor was there a sudden break at the time of United States v. Rita, five months earlier). Even if Rita, Kimbrough, and Gall collectively led to an increase in departures, the fact that the decisions were separated by five months makes it too diffuse a change to judges’ sentencing discretion to assess with this method. And even combined, the change over that whole period is still much smaller than the change at Booker. One should not expect small changes to have big effects, and if they appear to, one has to suspect some confounding factor. Booker, as the bigger change, is the more logical place to test the effects of changing judicial discretion.

We now turn to the effects of increasing judges’ departure discretion on other stages and outcomes in the justice process. Because criminal cases have several key dates, the RD method can be used to isolate Booker’s effect on each key stage in the process. However, it cannot be used to directly estimate the aggregate effect of Booker on all stages. The Sentencing Commission and other Booker researchers have always divided cases by sentencing date, but many cases’ processing dates straddle

---

114 For reasons explained below, this graph and all others are limited to district courts in the Second, Fourth, Fifth, Sixth, and Eleventh Circuits. The nationwide departure pattern looks similar.

115 The graph includes all departures. In 46% of departure cases, the departure was requested by the government as a reward for “substantial assistance” in another case; if these were excluded, the pattern would look similar, but the rise at Booker would be even steeper.

116 This is not very surprising; PROTECT and Kimbrough/Gall were much subtler changes in the law. PROTECT and Gall did not directly speak to judges’ legal authority to depart; they affected appellate review standards and data collection procedures. Kimbrough did directly affect departure authority, but only in crack cases (where mandatory minimums applied regardless); it was uncontested that judges could depart on policy grounds in other cases. The crack holding could only have helped black defendants—a logical challenge for studies that point to Kimbrough as a source of racial disparity.
Booker, so one cannot simply deem cases "pre-Booker" or "post-Booker." We assess Booker's effects on charging, as well as the sentencing consequences of those charging changes, by assessing what happens when the charging date passes Booker. Cases charged shortly before Booker will overwhelmingly have been disposed of and sentenced after Booker, so focusing on the immediate effects as the charging date passes Booker means that the sentencing effects of changing charging practices can be separated from the sentencing effects of changes to other process stages.

Likewise, we assess plea-bargaining changes and their sentencing effects by assessing what happens when the conviction date passes Booker, and we assess changes in judicial behavior by assessing what happens when the sentencing date passes Booker. Note that the judicial behavior being measured is not just changes to the final sentencing decision, but also changes to sentencing fact-finding. Separately assessing the conviction date and the sentencing date helps to disentangle judges’ contributions to disparities in sentencing fact-finding from disparities in the negotiated plea stipulations.

The most serious complication in drawing causal inferences about Booker is that the decision was hardly a bolt from the blue. Rather, Booker followed six months after the Supreme Court's decision in Blakely (denoted by the second vertical line in Figure 1), applying the same Sixth Amendment holding to a state sentencing scheme. It was Blakely that was an unexpected "earthquake," rendering it fairly obvious that the federal Guidelines were in constitutional trouble. What was not clear was what the Supreme Court would do to remedy the constitutional defect. Instead of the advisory guidelines approach (which none of the circuits had adopted), the Court could have struck the Guidelines down entirely, left them mandatory but shifted fact-finding to the jury, or left the whole matter to Congress. The lower courts quickly began weighing in, and the Supreme Court quickly agreed to review Booker.

The Blakely decision raises a dilemma for causal inference for three reasons. First, it could mean that the effects we are looking for happened in a more diffuse manner starting before Booker, because courts or parties adjusted their behavior in anticipation of the mandatory Guidelines’ fall. In that case, estimating discontinuities at Booker alone might understate the effects of moving away from mandatory Guidelines. Second, the anticipation of Booker may have affected the mix of cases decided immediately before and after Booker, if district courts delayed sentencings while waiting for the Supreme Court’s opinion. If so, those changes in cases could confound estimates of Booker’s effects. Third, even assuming Booker did cause the measured changes, not all of Booker’s effects can necessarily be attributed to expanding judicial discretion. In addition to rendering the Guidelines advisory, Booker may have affected outcomes by ending the chaotic interregnum period and rejecting the alternative remedies that the Court could have chosen. These problems

---

117 The average time from charge to disposition in our sample is 5 months, and the average time from disposition to sentencing is a further 4 months.

are not unique to our method (they afflict all studies of Booker), but they cannot be ignored.

For this reason, we constrain our analysis to five federal judicial circuits: the Second, Fourth, Fifth, Sixth, and Eleventh. Within two to six weeks of Blakely, these five courts of appeals issued decisions holding that Blakely did not apply to the federal guidelines.\(^{119}\) In those circuits, Booker’s legal effects were simpler: it changed the governing law from the old regime (mandatory Guidelines) to the new one (advisory Guidelines). During the Blakely-to-Booker period, there was neither legal chaos nor a third legal regime. Figure 1, which is limited to these “business as usual” circuits, shows that nothing happened to departure rates at Blakely or during the interregnum—there was no trend break until Booker.

Our focus on these circuits is only a partial solution to the Blakely problem. While district courts were required to follow the “business as usual” approach, if the parties anticipated that the Supreme Court would change the law before sentencing, they were free to let that expectation affect their charging and plea-bargaining decisions.\(^{120}\) Therefore, as detailed below, we also analyze changes happening at Blakely to see whether there is evidence of such anticipation effects.

C. Regression Discontinuity Estimates of Booker’s Effects

Here we present our RD estimates for key charge severity, plea-bargaining, and sentencing measures. In addition to the results presented below, we also assessed changes in criminal justice “funnel,” which could have introduced sample selection bias into the RD estimates. However, we found no significant change in the rate of filing charges in district court as the charging date passed Booker, nor in the rate of non-petty convictions as the disposition date passed Booker.\(^{121}\)

1. Changes to Charging

The principal charging dynamic that we sought to analyze is whether Booker affected prosecutors’ use of mandatory minimums, which our (post-Booker) findings discussed in Part II show to be a key driver of the black-white gap. There is also a logical causal mechanism. Booker reduced prosecutors’ ability to use the Guidelines to control sentencing outcomes, an ability that confers massive leverage in plea-bargaining. Without being able to rely on the Guidelines, it is plausible that

\(^{119}\) See United States v. Mincey, 380 F.3d 102, 106 (2d Cir. 2004); United States v. Hammoud, 378 F.3d 426 (4th Cir. 2004); United States v. Pineiro, 377 F.3d 464, 473 (5th Cir. 2004); United States v. Koch 383 F.3d 436, 438 (6th Cir. 2004); United States v. Curtis, 380 F.3d 1308(11th Cir. 2004).

\(^{120}\) For instance, DOJ directed prosecutors to begin including aggravating factors in indictments, rather than waiting for the plea stipulation or sentencing hearing to allege them. See James Comey, Departmental Legal Positions and Policies in Light of Blakely v. Washington, Letter to All U.S. Prosecutors (July 2, 2004). This memo did not suggest changes in the actual charges, though, just in the alleged facts.

\(^{121}\) We treat January 2005 as the first month in the post-Booker period. There were six business days in January before Booker was decided, and the dataset gives dates only in months. Conflating the last week of the pre-period into the post-period is if anything likely to mean we slightly understate Booker's effects.
prosecutors might turn more often to their other tool for constraining judges: mandatory minimums.

Our findings above also clearly showed that it was the initial charging stage in which the mandatory minimum disparity emerged, so that is a key stage to analyze. As explained above, we could not code initial charges in drug or child pornography cases. We only know the final mandatory minimum. Fortunately, unlike in the analysis in Part II, in this part of our analysis there is a solution to this problem. RD allows us to assess changes to the final mandatory minimum when the charging date passes Booker. Even though the outcome variable is measured at the conviction stage, changes in it that are triggered by the timing of the charge are probably the result of charging changes. This approach allows us to assess all case types.

The results from the formal RD analysis are presented in Table 1, which shows the estimated discontinuous change in mandatory minimum convictions at Booker. Within each panel of the table, the first row ("Overall Discontinuity") estimates the change for the whole population at Booker, while the second ("Blk-Wht. Diff. Disc") estimates Booker-related change for black defendants relative to whites. That is, the second row measures the change in racial disparity at Booker. To see the estimated change for black defendants at Booker, one adds the estimates in the two rows. The estimated change for white defendants at Booker is simply the overall discontinuity. We estimate regressions that include separate non-linear time trends for blacks and whites, before and after Booker—that is, we filter out both the overall underlying trends and the underlying trends in the black-white disparity. The regressions also filter out the month-to-month variation in arrest offenses and other pre-charge features of the case.

The estimated discontinuities represent the break in the curve at Booker—that is, the difference between the intercepts of pre-Booker curve and the post-Booker curve.

---

122 The “charging date” is the date of the indictment when there is one. In cases with no formal indictment, we used either the arrest date or the date the prosecutor opened the file on the case, whichever was later (usually they are the same month)—that is, the date that the prosecutor had both the case and the defendant in hand, and declined to add to or change the charges from the complaint.

123 If prosecutors suddenly started charging mandatory minimum offenses more after Booker, that would presumably translate into more convictions of mandatory minimum offenses for cases charged after Booker, too. As noted above, initial charges usually are not dropped; doing so requires a supervisor’s special permission. See Ashcroft, supra note 69.

124 The controls include arrest offense, criminal history, gender, age, a multi-defendant case flag, U.S. citizenship, criminal history, and education. The results shown exclude district, which was not an important contributor to racial disparity in our initial study; including so many dummy variables was problematic given the sample size per month. District was added in robustness checks and the results were generally similar but often less precise. Note that controls serve a different function in RD than they do in other regressions—they are mainly there to absorb statistical noise. If there are underlying continuous trends in the effects of the control variables, those will be filtered out by the time-trend variables. Including the controls, however, protects against the possibility of sudden changes in underlying case features at Booker. In a perfect RD situation, if one could safely assume that other variables changed only in continuous ways, one would not need controls at all, but we do not rely on that assumption.
Within each panel of Table 1, the four columns show the results of multiple specifications that use different methods of fitting curves to the data—we vary the length of the time window used to estimate the curves on each side (12 months versus 18 months) as well as the degree of the polynomial function of time (quadratic versus cubic). There is no one “right” choice for the window or the polynomial. A result is more robust if it is consistent across specifications, such that it is not just an artifact of a subjective modeling choice.

We find that when the charging date passed Booker, there was a significant, discontinuous increase in the mandatory minimum rate—but only against black defendants (Panel 1A). The increase in the black-white disparity in mandatory minimums is significant and quite large in all specifications, ranging from six to thirteen percentage points. Most of the increase in disparity is because of an increase for black defendants, but there also appears to be a smaller reduction in mandatory minimums’ use against white defendants.

Figure 2a provides an approximate visual representation of this result. Although the RD is estimated based on a narrower window of time surrounding Booker, the graphs show longer surrounding trends to provide context. The hollow circles and dots represent the monthly averages in the residuals for whites and blacks, respectively, from a regression on all the variables from the RD. A residual is the difference between the actual outcome observed for an individual and the outcome predicted by a multivariate regression based on other observed characteristics (e.g. arrest offense). Figure 2a thus shows the trends in average black and white charges after controlling for the cases’ underlying characteristics other than race. Curves are then fitted to these monthly averages to approximate the month-to-month trends for blacks and whites, and the vertical distance between the black and white curves represents the unexplained racial disparity at any given time.

The figure shows that the estimated jump in disparity after Booker is heavily influenced by the charging patterns in the first three months after Booker, especially the first month. Although there is an unexplained race gap in mandatory minimums through most of the period (the black line is above the white line), the trends had converged in the period leading up to Booker. In the month of Booker, there was a huge spike in black mandatory minimums. After the first few months, however, things seem to have reverted more or less to the previous trends. The race gap fluctuated somewhat, but the dominant background trend is a steady rise in mandatory minimums for both blacks and whites, and that trend continued.

Overall, although there is a significant break, the patterns are much less dramatic than what we saw with departures (Figure 1), in which the changes were

---

125 To provide perspective, about 40% of defendants during 2004 faced a mandatory minimum.
126 The curves in the visual representations are fit slightly differently from the formal RD, so the correspondence between the figures and tables is only approximate. The figures contain the monthly average of the variable of interest along with curves fitted using kernel weighted local polynomial smoothing. The curves are fit separately on each side of Booker, and capture linear and non-linear trends over time. The vertical distance between the curves on either side at Booker (the difference in intercepts) is a visual approximation of the discontinuity estimated with RD.
much larger and stuck. When a trend break is driven largely by a one-month anomaly, one has to wonder if it is due to chance. Here, the divergence from the trend in that one month far exceeds the noise found in the rest of the data, so we suspect that it is connected to Booker, but nonetheless, it did not seem to last. Perhaps prosecutors responded to the immediate shock of Booker with some degree of panic, and hedged their bets against a possible coming wave of Guidelines departures by charging mandatory minimums (in a pattern disparately affecting blacks). If so, charging may have reverted to normal when prosecutors saw that Booker did not cause a major drop in sentences (which we shall see below). This, of course, is only speculation. What we do know is that, despite the significant discontinuity, Booker's longer-term effects on charging look fairly subtle.  

We next assess whether the ultimate sentence length was discontinuously affected by the charging date passing Booker—that is, did changes in charging translate into sentencing consequences? We find some evidence of this: racial disparity in the sentence jumps by between 4 and 13 months for cases charged immediately after Booker, depending on the specification (Panel 1C). Almost all of this consists of an increase for blacks. However, as Figure 2b shows, the noise in the sentence-length data, compared to which the break is not so remarkable. Unsurprisingly, then, the estimates are not very precise, and the statistical significance of the increase in black sentences varies across specifications.

2. Changes in Plea-Bargaining

We next assess Booker's effects on plea-bargaining by examining what happens when the disposition date passes Booker. Specifically, we assess three outcomes: the conviction mandatory minimum, the final Guidelines offense level, and sentence length. The mandatory minimum and the offense level represent two key subjects of plea negotiations: the charge of conviction and the stipulations of sentencing facts. By assessing the effects of the conviction date on the offense level, we can separate out Booker's effects on fact-bargaining from its effects on judicial fact-finding (which will be assessed below). We then turn to the ultimate sentencing consequences of any plea-bargaining changes.

These results can be quickly summarized: nothing dramatic happened, or at least, nothing that can be picked out from the noise of the surrounding data (Table 1, Column 2; Figs. 3a-3c). Mandatory minimum rates for whites are noticeably higher in the post-Booker period than before it (Figure 3a), but that increase actually occurred several months before Booker. Prosecutors, unlike judges, were free to adapt their behavior before the Court ruled, so these changes could have been in anticipation of Booker; if so, that would mean that that Booker could have increased white mandatory minimums, but too slowly for the RD analysis to detect. Booker does not

---

127 Panel 1B also shows a discontinuous increase in disparity in the final offense level for cases charged after Booker that is significant in two of the four specifications; the disparity comes mostly from an increase in black offense levels. This might simply reflect the apparent mandatory minimum shift, since the ultimate offense level is affected by any mandatory minimums that apply. In any event, this is further suggestive evidence that charging for blacks got somewhat harsher at Booker.
appear to have had any significant discontinuous effects on racial disparity in plea-bargaining or on plea-bargaining outcomes generally.


Finally, we assess changes in judicial decision-making by examining what happens when the sentencing date passes Booker. We consider three outcomes: departures, the final Guidelines offense level, and sentence length. Booker directly expanded judges’ legal authority to depart, and we showed in Figure 1 that this expansion had an immediate effect. In Figure 4a and Panel 3D, we break this effect down by race. The estimates all show a jump in white departure rates of about 8 percentage points, and the jump for blacks is similar, so there is no significant change in disparity in departure rates. Notice that both the black and white trends of declining departure rates after Booker are identical to the trends before it—but both curves are shifted upward. In other words, Booker’s boost to departures occurred immediately and clearly had a lasting effect.

Booker’s legal holding did not directly affect fact-finding, but it could have affected it indirectly (even setting aside any effects on the plea negotiations, which our focus on the sentencing date filters out). If a judge believes the sentencing range that follows from the plea agreement is inappropriate, she has two options for altering it: she can make findings of fact that “go behind the plea” or she can depart from the Guidelines. Expanded authority to do the latter might make it less necessary to do the former.

Panel 3B of the table gives some support for this theory, but only for whites. In white defendants’ cases, the increase in departures appears to be offset to some degree by an increase in the Guidelines sentence from which the judge departs. Estimates of the effect size vary between about 0.1 and 0.9 offense levels; each offense level is associated with a 10-15% increase in the Guidelines sentencing range. The change in racial disparity in offense levels is less precisely estimated and is thus not significant in every specification, but the sign is consistently negative, and (adding the two rows together) black offense levels do not seem to go up. Figure 4b shows that after Booker, the black and white trends, which had been diverging, converged again. While this may suggest that judicial discretion helps to close racial gaps, an alternative interpretation is less charitable to judges. Before Booker, there was a significant racial disparity favoring whites in findings of fact, after controlling for the observable variables. One possible interpretation of Booker’s effects is that there had been a thumb on the fact-finding scale in favor of whites, and after Booker, that thumb became no longer necessary.

---

128 We treat departures as a binary variable here, but one sees similar patterns in departure size.
130 As discussed above, survey data show that most judges do not diverge from the plea stipulations very often—but that does not mean they never do. The reasons they choose to do so in particular cases might be the same kinds of exceptional dissatisfaction with the sentence that the plea agreement would produce that might otherwise motivate departures.
Note that the upward shift for whites, although subtle, does not appear to be temporary. There is a long-term trend of gradual increase in the offense level. However, one cannot safely infer that the long-term trend is caused by *Booker*, because RD only estimates the local effect at the discontinuity. Nor can one assume that the trend is necessarily caused by judges—again, the offense level is jointly determined by the parties’ negotiations and sentencing-stage arguments and by the judge’s findings.\(^\text{131}\) As one moves farther away from the discontinuity, one quickly gets to a point in which most cases were both pled and sentenced after *Booker*, and it becomes impossible to isolate *Booker’s* effects on judicial fact-finding alone.

Whatever its causes, the trend of increasing offense levels may help to explain what otherwise might have been a mystery: why (as Figure 4c shows) sentences did not go down in the long run after *Booker*, even though downward departures went way up and stayed up. As Figure 4c and Panel 3C of the table show, there appears to have been an immediate drop in black sentences at *Booker*, although there is considerable noise in the data. White sentences did not fall, however (even though white departures increased).\(^\text{132}\) Perhaps this is because the departures were offset by fact-finding changes. And perhaps this was also true in the long run for both blacks and whites: Figure 4b shows a sustained trend of increased offense levels.

The magnitude of that change looks fairly small, though—perhaps half of one offense level. One might wonder how such a subtle change in fact-finding could cancel out such a large increase in departures. The answer is that although the increase in departures at *Booker* was a very sharp break in the prior trend, it still only affected a small percentage of cases (about 8%, according to the RD). The average size of a departure from 2005 to 2009 was 29 months, so a back-of-the-envelope calculation suggests that *Booker* brought the average sentence down by only about 2.3 months. An increase of just one-half an offense level, applied to the average case in the sample, would raise the low end of the Guidelines range by 2 months, enough to cancel out most of that departure effect.

Thus, although *Booker* was the biggest sudden change to federal judges’ sentencing discretion since the Guidelines’ adoption, it nonetheless was perhaps less of a revolution than various observers either feared or hoped. *Booker* is only what federal judges make of it, and so far that appears to have been not much. This post-*Booker* stability should not be taken as especially good news for those concerned about incarceration rates for black men. If *Booker* does not change judicial behavior very much, then it cannot do what critics of the Guidelines hoped: substantially mitigate the Guidelines’ harshness. Sentences continue to increase, even after controlling for shifts in the pool of offenses and offenders. And with plea levels still over 96%, prosecutors’ tremendous leverage appears to remain intact.

\(^\text{131}\) It is possible that prosecutors could have litigated the sentencing stage more aggressively after *Booker*, especially in cases in which the plea did not include detailed factual stipulations.

\(^\text{132}\) Indeed, if anything there is a visible upward turn in the white trend at *Booker* (although it does not amount to a discontinuous break), while the black sentence trend stays flat.
D. Limitations and Causal Inference Challenges

Unlike the Sentencing Commission, we find no evidence that *Booker* increased racial disparity in the exercise of judicial discretion; if anything it may have reduced it. The only possibly adverse effects for blacks that we see arise from prosecutors’ shift to mandatory minimums, although that shift appears temporary. Like our results in Part II, these findings cut against the case for restoring constraints on judicial discretion. Still, there are some limitations to our method. As we have already discussed, it provides only local estimates of immediate effects. Beyond that, there are a few other things to keep in mind.

First, it is important to understand what our RD analysis does not assess. In the initial decompositions described in Part II, we sought to disentangle the share of the black-white gap that was explained by the disparate impact of factors such as criminal history from unexplained disparities that could represent racially disparate treatment. Here, in our *Booker* analysis, we only do that in a limited sense. We do control for the arrest offense and the other pre-charge covariates, so in that sense we are measuring changes in (apparently) “unwarranted” disparity. Controlling for those variables means that if the relative composition of the black and white defendant pools (in terms of the observable variables) changed suddenly right around *Booker*—either due to random or seasonal variation in crime or to reaction to *Booker* itself—it should not bias the results.

But the coefficients on those variables—the strength of the relationships between each of them and the outcome variable—are estimated only for the entire time period. While the trends will filter out any smooth (gradual) changes over time, they cannot filter out sharp sudden changes that coincide with *Booker*. We do not separately estimate, for instance, the relationship between criminal history and sentence length before and after *Booker*. If criminal history becomes a stronger predictor of sentence length gradually during the time period, the polynomial trends in our regression would filter that change out. But if the relationship between criminal history and sentence changes suddenly at *Booker*—if *Booker* changes it—our method will not filter out that change.

In effect, what that means is that we are focused on the question “Did *Booker* change racial disparity patterns in charging, plea-bargaining, and sentencing?” rather than “Why did *Booker* change those patterns?” If, for instance, prosecutors started using mandatory minimums more against black defendants, this need not have been motivated by race—it could have been motivated by wanting to crack down on gun crimes, for instance. In short, we are estimating *Booker’s* racially disparate impacts. We do not filter out the share of those impacts that are mediated by other variables, not just because doing so is impractical with our method but also because it is undesirable. If policymakers care about the effects of sentencing reform on black incarceration rates, filtering out everything that is not racially motivated would not convey those consequences fully. Together, our results in Part II and in this Part present a fairly rich picture of the static factors (case features) and dynamic factors (sentencing law reform) that contribute to outcomes at each procedural stage.
Second, while RD effectively filters out long-term trends, it is vulnerable to statistical noise that might generate false positives. If the graph is sufficiently noisy, one might be able to see discontinuities at lots of points. Of course, *Booker* need not have been the *only* shift over the course of the study period to be a real shift. But if there are frequent breaks, even at points where there is no other known triggering event, then not much can be made of finding a break at *Booker* as well.

We think that with appropriately cautious interpretation, this is not such a serious problem—far less serious than the causal inference problem that pervades other studies. This is why we fit the monthly trends with multiple kinds of functions, and do not put stock in an apparent discontinuity that appears only in one version. It is why we do not use even higher-order polynomials, which would likely over-fit the data. It is also why the graphs matter, perhaps more than the numbers. If a discontinuity cannot be picked out with the eye—or if it looks no different from many other unexplained breaks—then it is probably nothing to write home about.

As an additional precaution, we conducted “placebo tests” on every outcome variable, re-running all our analyses on twelve other arbitrary breaking points across the study period.\(^{133}\) These tests were reasonably successful. In the mandatory minimum variable, when the placebo tests were run by charging date, none showed a significant discontinuity in racial disparity in more than one specification (a “false positive”); when run by disposition date, one out of twelve did. This makes us more confident that the spike at *Booker*—even if brief—was likely something real. In the prison variable, which was visibly noisier in the graphs, there was one false positive when the placebos were run by charging month, four when run by disposition month, and two when run by sentence month. In the offense level variable, there were two false positives when the placebos were run by disposition month and one when they were run by sentence month. The departure variable had two false positives (run by sentencing month), but visual inspection makes clear that *Booker* was by far the cleanest break in the study period.

Finally, we return to the question of *Blakely* and anticipation of *Booker*. *Blakely* is marked with a dotted line in the figures, and we also repeated all the numeric analyses on it. There are no apparent breaks in departures, offense level, or sentence length when the sentencing date passes *Blakely*, nor are there strange patterns in the six months before *Booker* (Figs. 4a-4c). It appears that these courts really did follow the “business as usual rule.” But what about prosecutors?

It does not appear that initial charging changed much—both the mandatory minimum variable and sentence length continue on their existing trajectories after *Blakely* (Figs. 2a-2b). The sharpest effect of *Blakely* appears to have been on plea-bargaining. Just after *Blakely*, there are two anomalous months in which plea severity plunges, in particular for whites (Figs. 3a-3c). This effect is probably because one immediate reaction to *Blakely* in the world of federal criminal practice was anticipation of the possibility that sentencing facts might have to be proven to the

\(^{133}\) The tests ran every six months for three years in each direction before *Blakely* and after *Booker*.
jury or pled to by the defendant.\footnote{The Seventh Circuit had so held in its decision in \textit{Booker}, 375 F.3d 508 (2004), less than three weeks after \textit{Blakely}. \textit{Accord United States v. Ameline}, 376 F.3d 967 (9th Cir. 2004).} That expectation would give the defendant much more leverage in plea-bargaining, because if they refused to plead to adverse facts, it was not at all obvious that the prosecutor would nonetheless be able to argue them at sentencing. Nonetheless, by the third month after \textit{Blakely}—after all five of the circuits we focus on had established the “business as usual” rule—this anomalous period appears to have ended.

This anomaly at \textit{Blakely} could affect the interpretation of what happened at \textit{Booker}, either in plea-bargaining (because it affects the trend in the pre-period) or in sentencing (because cases sentenced around \textit{Booker} might have been plea-bargained around \textit{Blakely}). We do not find any significant effects on plea-bargaining at \textit{Booker} anyway, and based on the graphs, it does not appear that any discontinuities would have appeared if those anomalous months had not disturbed the trend. The implications for the sentencing results are harder to interpret. In the sentence-length and offense level graphs (Figures 4b and 4c), one can see a V-shaped pattern in the white trend—a dip right before \textit{Booker}. We suspect that one reason for that dip is the lingering effect of those lenient plea deals that got struck in the two months after \textit{Blakely}, most of which would have been sentenced before \textit{Booker}. The pattern does not exist for black cases, but black cases did not have an anomalous post-\textit{Blakely} period. Perhaps had it not been for those cases, the white sentence levels right before \textit{Booker} would have been higher, and one might have seen a drop in white sentences at \textit{Booker}, paralleling the drop in black sentences. Or one might not have seen the jump in white offense levels at \textit{Booker}.

While this interpretation is possible, there is also reason to believe that the \textit{Blakely} effect might not have actually changed the \textit{Booker} discontinuity estimates much. There is enough variation in the time between disposition and sentencing that it is not as though all of those anomalous-period cases were sentenced in the month or two right before \textit{Booker}. Rather, from one sentencing month to the next, there is a gradual increase and then a gradual decrease in the probability of the case having come from that period. In other words, whatever effect those plea-bargaining changes had on sentencing should be a roughly continuous trend, and to the extent it is continuous, the polynomials should filter out. To substantially affect the discontinuity estimates at \textit{Booker}, the probability of having been a post-\textit{Blakely} “sweetheart deal” case would have to have plunged suddenly in the month of \textit{Booker}. This is a substantial advantage of RD over other methods.\footnote{This is a principal reason we do not simply use a short-window differences-in-differences approach—for instance, comparing the three months before \textit{Booker} to the three months after. If we had, the \textit{Blakely} effects would have been very different in the pre- and post-periods.}

In addition, the absence of the plea-bargaining anomaly after \textit{Blakely} in black cases is itself substantively interesting—if anything it looks like there may be an \textit{increase} in severity there. That could be suggestive evidence that prosecutors respond to uncertainty and sudden change differently with black and white defendants (and more favorably for whites). This interpretation, while speculative, is
consistent with the fact that black mandatory minimum use spiked immediately (if briefly) at *Booker*, but white charge severity did not.

A final concern about the interregnum period is that cases could have been delayed until after *Booker*. Causal inference with an RD estimator requires that “individuals—even while having some influence—are unable to *precisely* manipulate the assignment variable,” and here, the assignment variable is the exact sentencing date relative to *Booker*. Note that this is a relatively modest requirement—it only requires that cases sentenced very near *Booker* were not subject to the court’s *precise* manipulation of which side of the line they fell on. If a court merely took steps to make it more likely that a case would be sentenced after *Booker*, such as scheduling the sentencing hearing for a faraway date, this would not be seriously problematic; the scheduling would get the case *near* *Booker*, but there would still be a chance element as to which side it landed on. The chance element is amplified by the fact that nobody knew when the Supreme Court would rule—legal observers did terribly at predicting *Booker*’s release, with many predicting a very fast decision after the October argument.\(^\text{137}\)

The case counts do not suggest problematic manipulation. The number of cases sentenced in December 2004 was 1703; the number in January 2005 was 1708. The mean elapsed time since the plea was identical: 4.08 months. The breakdowns by race and case type were almost identical. If anything, there may have been some delaying of cases in November 2004 (1576 cases, the lowest that year), when expectations of an early *Booker* decision were high, but that is a small dip, and when the Court did not release its decision quickly it appears the counts went back to normal. After *Booker*, the number of cases also stayed normal; they were slightly higher in March (1841 cases), but this was lower than four other months in 2004 and 2005. In short, there is very good reason to believe that the courts in these circuits really did do “business as usual,” or that any manipulation was too imprecise to have major effects right around the discontinuity.\(^\text{138}\)

**E. Longer-Term Trends**

Compared to the Sentencing Commission’s dramatic findings, our analysis tells a very different story about *Booker*. But our results are difficult to compare directly to the Commission’s given that the method, sample, and covariates are all different.

---

\(^{136}\) Lee & Lemieux, *supra* note 112, at 283.


\(^{138}\) In any event, manipulation would only bias our results if it occurred in a racially disparate way. And the manipulation concern applies only to our analysis of *sentencing* responses to *Booker*—there is no reason to believe that any prosecutor would wait to charge or plea-bargain a case until after *Booker*, nor would defendants likely take the large risk of stalling guilty pleas (risking their withdrawal) while waiting for a Supreme Court decision.
So one might wonder what our estimates suggest about the Sentencing Commission’s findings. Has the Commission identified a real increase in racial disparity over the time period, but simply misattributed it to *Booker*? Or is there no real increase to begin with, once one estimates disparity properly?

Our tentative answer is that there is no real increase in this period (at least within non-immigration cases, which are our focus). One very straightforward way to assess the average growth in racial disparity during a time period is to directly estimate it—for instance, with a simple linear trend model. To that end, we estimate the overall difference in black and white sentences as well as the average growth in that gap over time. We did this for cases sentenced between the PROTECT Act and the end of FY 2009, and focused on black and non-Hispanic white men: the time period and groups for which the Sentencing Commission found the purported quadrupling of disparity. The sample includes all non-immigration cases except those subject to major substantive sentencing-law changes during the study period: identity theft, obscenity/child sexual exploitation, and sex offender registration.

In Column 1 of Table 2, we show the estimated linear trend in average sentence when controlling for the arrest offense and other prior characteristics—that is, the trend over time in the aggregate black-white sentence disparity introduced during the post-arrest justice process. The estimated trend in racial disparity is insignificant, and its sign is actually negative: the model (noisily) estimates that the unexplained black-white sentence gap declined by 1.7 months, from about 12.5 months to about 10.8 months over the course of the period. The sign of this estimated change is consistent across a variety of estimation strategies and sample definitions. For instance, while Table 2 shows the results when sentence length is estimated in months (including non-incarceration sentences as zeros), we get similar results if we use a log-linear model excluding the zeros. We also get essentially identical results when we estimate yearly rather than monthly trends. Likewise, we also see no rise in disparity over time when, instead of estimating linear trends, we estimate the differences in the “black” effect among the three key time periods that the Sentencing Commission study identifies (PROTECT-to-*Booker*, *Booker*-to-*Gall*, and

---

139 All the other post-*Booker* racial disparity studies also found some increase in disparity, at least after *Kimbrough/Gall*. See Ulmer et al., *supra* note 38; Fischman & Schanzenbach, *supra* note 47; Nowacki, *supra* note 39.

140 Specifically, the regression includes an overall linear (monthly) time trend as well as an interaction between that trend and the “black” coefficient.

141 We use the time period covered in the main Sentencing Commission report on demographic disparities (from 2010). The Commission’s very recent update adds 2010 and 2011, but those years of data have not been made available to researchers.

142 Except for identity theft, these legal changes were sufficiently far from *Booker* that these categories did not have to be dropped from the narrower-window RD analysis.

143 For this assessment, we used the whole country, not just the “business as usual” circuits. As in Part II, we focused on black and white men (so as to be directly comparable to the Sentencing Commission’s most prominent result). We controlled for arrest offense, criminal history, district, education, age, and multi-defendant case structure.

144 This is obtained by multiplying the per-month linear trend estimate by the number of months.
And indeed, some reasonable variations on our approach produce significant and much larger estimated downward trends. For instance, had we included the offense categories that we dropped due to substantive changes in sentencing law, the estimated decline in disparity during the study period would have been significant and three times as large—about six months total.\footnote{We obtain the latter estimate with a single differences-in-differences regression containing time period dummies interacted with the “black” variable. This analysis shows, if anything, slightly smaller disparities in the later periods (with the smallest post-Gall), although the time period-race interaction terms are not significant. Standard errors are clustered on the month to account for possible events affecting many cases at once; all significant results remain so with alternative clustering (such as clustering on the district to account for correlated prosecutorial policies or courthouse practices).}

Why, then, does the Sentencing Commission find an increase? There may be a variety of reasons\footnote{This is likely because child pornography and child sexual exploitation arrestees are overwhelmingly white, so increasing sentences for those offenses tends to reduce black-white disparity. Although the Sentencing Commission included these categories, we exclude them because it keeps our estimates focused on changes in the administration of the law. Also, had we included Hispanics, the estimated decline in disparity would have been larger and statistically significant.} but a prime reason appears to be that racial disparity in the processes determining the presumptive sentence declined significantly over the same period. By controlling for the presumptive sentence, the Commission filtered out that reduction in disparity, and what is left is a misleading picture. The black-white gap in sentences relative to the presumptive sentence may have grown, but that is because the black-white gap in presumptive sentences shrank (after controlling for underlying case characteristics). When one controls for the presumptive sentence, the disparities look larger in the later period because the presumptive sentence control is filtering out less of the disparity. The presumptive sentence was not the “same broken yardstick” during this period—over time, the yardstick changed.

Columns 2 through 4 of Table 2 illustrate this point. In the regression shown in Column 2 of Table 2, rather than controlling for the arrest offense, we substituted the final offense level, the mandatory minimum indicator, and broad offense-type categories associated with the offense of conviction. This reflects a fairly typical version of the presumptive sentence approach; recall that the presumptive sentence is determined by the final offense level (and the criminal history category, which we control for in all regressions). The regression in Column 3 is identical except that we more closely approximate the Commission’s approach by also adding departure status controls.\footnote{Among other things, these may include the Commission’s arbitrary choice of the 0.01-month valuation of non-prison sentences, the lack of detailed offense-type controls, and the inclusion of immigration cases. These issues are discussed in Part II.} After these modifications, both of these regressions show a significant linear increase in racial disparity over time, albeit not as dramatic an increase as the Commission itself found. In the Column 2 version, the unexplained black-white gap increases from about 3.2 months at the beginning of the period to about 6.5 months at the end. When departure status is added as a control in Column

\footnote{As explained above, we believe that it is a mistake to control for departure status, even if one intends to focus on judicial behavior alone. We do so here purely for illustrative purposes.}
3, the black-white gap is estimated to rise from about 1 month at the beginning of the period to about 4.2 months at the end.\footnote{149}

The difference between the estimates in Columns 2 and 3 (using variations on the presumptive sentence approach) and those in Column 1 (which use our preferred approach) can be reconciled by looking at Column 4, which repeats the exercise but with the final offense level as the outcome of interest. After controlling for the arrest offense and other pre-charge characteristics, the unexplained black-white disparity in final offense levels declined by nearly one level during this period. For the average case in the sample, a change of one offense level is associated with a 5-month change in sentence length—almost exactly the difference between the disparity trend estimate in Column 1 (-1.7 months) and those in columns 2 and 3 (+3.3 months and +3.2 months, respectively).\footnote{150}

Thus, the overall unexplained racial disparity in the post-arrest justice process certainly does not appear to have increased over the 2003-09 period, and if anything it seems to have decreased. The linear trend results do suggest that the \textit{procedural sources} of disparity may have shifted over the course of the period, with the earlier stages in the process becoming a bit less important and the judicial sentencing decision becoming a bit more important. However, it bears noting that throughout the time period, the earlier process stages appear to be the dominant procedural sources of disparity.\footnote{151} This is consistent with the results we describe in Part II, which find that even in the most recent years, charging decisions appear to be the major driver of sentencing disparity.

Note that we make no claims as to the causes of these longer trends, and specifically, we do not claim to have established that \textit{Booker} caused them. The RD analysis suggests that \textit{Booker}'s effects on all parts of the justice process were fairly modest. Many factors might contribute to the longer-term trends, and further exploration is necessary to see what parts of the earlier process changed. For instance, change in sentencing fact-finding might be affected either by changes in plea-bargaining or by changes in judges’ willingness to “go behind the plea.” Exploring such possibilities further would be a fruitful area of future research.

\footnote{149}Note that even in the Columns 2 and 3 regressions, the overall estimated average sentences for black defendants are lower at the end of the period than at the beginning. That is, even with the presumptive sentence approach, the rise in black-white disparity is not large enough to offset the overall monthly trend—other factors equal, sentences for both black and white defendants declined in all of the models over time.

\footnote{150}When we repeat the Column 3 analysis (the one closest to the Commission’s) on a sample that includes the excluded obscenity, sex-offender, child sex crimes, and identity theft categories, we obtain a somewhat stronger estimated upward trend in disparity, totaling about four months. Likewise, when we repeat the Column 4 analysis for that sample, we find a larger decrease in offense-level disparity, totaling 1.5 levels. This supports the theory that inclusion of those categories contributed to the Commission’s findings of increased disparity in the later periods, largely because the underlying presumptive sentence measure was changing.

\footnote{151}Throughout the time period, the overall estimated racial disparities are much larger when one controls for the arrest offense—thereby incorporating disparities from those earlier procedural stages—than when one uses either version of the presumptive sentence model.
CONCLUSION

Determining the causes of racial disparities in criminal justice is not easy. We believe our approach improves substantially on existing research, but we do not offer definitive answers, and we doubt that anyone will soon. So what are policymakers to do? We do not seek to resolve that dilemma completely. Even if we did have crystalline empirical answers, criminal justice policy does not turn on demographic disparity alone—many competing objectives must be considered. That said, our results have implications for these dilemmas, and we fear that the contrary results of existing research may be distorted to support counterproductive “solutions” to disparity. We close with some brief thoughts on these points.

First, despite our concerns about the methods of the Sentencing Commission and others, we agree that the high rate of incarceration of black men is a serious social problem and that examining the possible contribution of criminal justice disparities is important. Our research suggests that in the federal system, disparities in the post-arrest justice process are contributors to this problem. After controlling for the arrest offense, criminal history, and other prior characteristics, sentences for black male arrestees diverge substantially from those of white males (by around 10% on average). While this disparity does not seem to be growing, it is persistent.

Second, the procedural source of this disparity matters, and it is myopic to focus on judicial sentencing alone. Our research suggests that racial disparities in recent years have been largely driven by the cases in which judges have the least sentencing discretion: those with mandatory minimums. Our assessment of Booker is more tentative, but we find no evidence that it increased racial disparity. The Sentencing Commission’s contrary conclusion is based on deeply flawed methods.

For these reasons, we are particularly concerned about proposals to respond to sentencing disparities by restoring tighter constraints on sentencing, especially those that entail expanding mandatory minimums. Our results suggest that this would not reduce disparities in the justice process. Quite the contrary: we find that prosecutors file mandatory minimums twice as often against black men than against comparable white men. Moreover, for those concerned about mass incarceration of black men, expanding mandatory minimums would be counterproductive. Sentencing law changes that increase severity have a particularly adverse impact on black men, who are disproportionately involved in the criminal justice system. Making sentencing law more rigid would likely exacerbate this problem even if it led to more equitable administration of the law—and our results suggest that it would likely lead to less equitable administration.

Third, we do not advocate attempting to reduce disparity by taking discretion away from prosecutors. Eliminating prosecutorial discretion is probably impossible. The Department of Justice has certainly tried. The disparities we found persisted despite the Ashcroft Memo ordering prosecutors to charge and pursue the “most serious readily provable offense,” as well as DOJ bans on fact-bargaining. If taken at their word, these policies would have stripped almost all discretion from line

152 See Gonzales, supra note 14.
prosecutors. But such policies are very difficult to enforce, because line prosecutors inevitably must subjectively evaluate the evidence. And even if constraining prosecutorial discretion did succeed, one might see another “hydraulic” effect. If prosecutors had to pursue to the fullest every case law enforcement brought them, their current power over case outcomes might shift another step back—to law enforcement, where it might be even harder to monitor. Prosecutors’ decision-making is notoriously difficult to observe—unlike judges, they do not publish written reasoning. But law enforcement is even more of a “black box.”

Even if all discretion could somehow be removed from the justice system, we doubt that would be a justice system anybody would want. Flexibility allows appropriate tailoring of both charges and sentences to the circumstances of individual cases, so as to avoid unduly harsh punishments when they are not justified. Efforts to eliminate unwarranted disparity are important, but they should not come at the cost of unwarranted uniformity. Instead, rather than looking for ways to curtail prosecutorial discretion, legislators could consider curtailing their power by dialing back existing mandatory minimums. If sentencing laws were less rigid, it would be less necessary for decision-makers to find ad hoc means of mitigating their impact. The Fair Sentencing Act of 2010, which reduced crack sentences, showed it is politically possible to reform excessive sentencing laws, and that empirical evidence of racial disparities can help to bring such changes about.

One potential next focus could be the severe gun enhancements in 18 U.S.C. § 924(c). These laws hit black men particularly hard both because they are more frequently arrested for gun crimes and because of large apparent disparities in prosecutors’ exercise of charging discretion. Certainly, policymakers must weigh this problem against gun violence concerns. Notwithstanding these serious concerns, we wonder whether the mandatory minimums in the statute are truly always necessary, such that judicial discretion should be precluded. For instance, is a five-year add-on sentence really necessary in every case in which a firearm has merely been “carried”—let alone a mandatory extra 25 years for a second gun, and yet another 25 for a third? And no prosecutor would need to “swallow a gun” if a gun did not automatically trigger a massive additional penalty.

Fourth, there was a piece of good news in our findings that also carries potential policy implications: case outcomes were not correlated with measures of socioeconomic status. This absence cuts against conventional wisdom. Can it really be that poor defendants do not fare worse? We do not conclude that the conventional wisdom is wrong, but it might not apply to the federal courts, where indigent defendants generally receive high-quality representation, especially from federal public defenders. We suspect that we would not have gotten the same

---

153 See Stith, supra note 15, at 1470; O’Sullivan, 1425-26; Miller, supra note 26, at 1257.
154 See supra note 84 and accompanying text.
155 The hurdles facing indigent defendants have long been the subject of extensive scholarship and policy attention. E.g., DOJ’s Access to Justice Initiative, at http://www.justice.gov/ati/ (describing the “access-to-justice crisis”).
156 Richard Posner & Albert Yoon, What Judges Think of the Quality of Legal Representation, 63
result had we studied states in which indigent representation is under-resourced and in disarray. The federal example offers a potential model for those states. When a justice system devotes sufficient resources to indigent defense to attract strong lawyers, train them well, and keep caseloads reasonable, poverty need not drive outcomes, and the race gap will likely be smaller than it might otherwise be.

Finally, while our approach is far more comprehensive than that of prior sentencing studies, there is enormous room for further exploration. For instance, we plan to explore further the possible role of sentencing fact-finding in producing racial disparities. More research is also necessary to see whether patterns like those we found are also found in state courts. More generally, we do not claim to have proven purposeful discrimination by prosecutors or anyone else—it would be difficult to do so with administrative data like ours. Other kinds of studies may be necessary to dig deeper into causal theories for racial disparities: perhaps experimental studies in which race is randomly assigned to otherwise identical prosecutor files, or qualitative studies involving reviews of case files and interviews. DOJ itself is well positioned to carry out such work. One easy step would be for DOJ to keep statistics on mandatory minimum charging decisions by race when it tracks prosecutors’ performance. Doing so would not only facilitate research, but could help prosecutors who do not want to contribute to disparities but might not be conscious of them. The government itself should take the elimination of disparities in criminal justice as seriously as other civil rights enforcement matters, and think creatively about solutions and strategies for answering the empirical questions that remain.

---


158 See Steven B. Bright, Legal Representation for the Poor, 75 Mo. L. Rev. 683 (2010) (noting the superior resources of federal public defenders).

159 In contexts such as employment and housing, disparity researchers can experimentally manipulate race while leaving other factors identical. See, e.g., Devah Pager, The Mark of a Criminal Record, 108 Am. J. Soc. 937 (2003). The federal government uses “testers” (fake applicants) to enforce its discrimination statutes. Similar field experiments in criminal justice would generally be illegal: a crime staged for research is still a crime, as is submitting fake information to authorities. But such studies could be legislatively authorized, under regulated conditions and perhaps carried out by DOJ itself.
| Table 1. Regression Discontinuity Estimates of Booker’s Effects on Charging, Plea-Bargaining, and Sentencing |
|---------------------------------------------------|---------------------------------------------------|---------------------------------------------------|
| **1: Charging Date** | **2: Disposition Date** | **3: Sentencing Date** |
| **A: Convicted of Mandatory Minimum Offense** | **A: Convicted of Mandatory Minimum Offense** | **A: Convicted of Mandatory Minimum Offense** |
| Overall Discontinuity | -0.023 | 0.037+ | 0.062* |
|                       | (0.014) | (0.022) | (0.027) |
| Blk-Wht Diff in Disc. | 0.062* | 0.048+ | 0.006 |
|                       | (0.027) | (0.026) | (0.021) |
| Blk-Wht Diff in Disc. | 0.062* | 0.048+ | 0.006 |
|                       | (0.027) | (0.026) | (0.021) |
| **B: Offense Level** | **B: Offense Level** | **B: Offense Level** |
| Overall Discontinuity | -0.126 | 0.179  | 0.377  |
|                       | (0.248) | (0.296) | (0.481) |
| Blk-Wht Diff in Disc. | 0.377  | 0.253  | 0.461  |
|                       | (0.481) | (0.378) | (0.473) |
| Blk-Wht Diff in Disc. | 0.377  | 0.253  | 0.461  |
|                       | (0.481) | (0.378) | (0.473) |
| **C: Prison Sentence (months)** | **C: Prison Sentence (months)** | **C: Prison Sentence (months)** |
| Overall Discontinuity | -1.43  | 2.81   | 4.55   |
|                       | (2.11) | (2.35) | (4.63) |
| Blk-Wht Diff in Disc. | 4.55   | -1.25  | 4.55   |
|                       | (4.63) | (4.08) | (4.63) |
| **D: Departures** | **D: Departures** | **D: Departures** |
| Overall Discontinuity | 0.075** | 0.076** | 0.075** |
|                       | (0.015) | (0.017) | (0.026) |
| Blk-Wht Diff in Disc. | 0.075** | 0.076** | 0.075** |
|                       | (0.015) | (0.017) | (0.026) |
| Degree of Polynomial | 2nd | 2nd | 2nd |
| Window Around Booker | ±18 | ±12 | ±18 |

Note: Regressions include polynomial time trends estimated separately before and after Booker, interacted with race, and controls for arrest offense, criminal history, education, multi-defendant case, gender, age, and citizenship. Coefficients reflect discontinuous changes in average white outcomes and the average black-white disparity as the date of charging, disposition, or sentencing passes Booker. Samples are cases with 18 months of Booker (N=57,000) or within 12 months of Booker (N=39,000); number of observations varies slightly depending on the outcome variable. The black discontinuity is the sum of the white discontinuity and the discontinuity in the black-white difference. Standard errors are clustered on the running variable (black*month). +p<0.1 *p<0.05 **p<0.01
Table 2. Average Monthly Changes in Sentence Length and Final Offense Level  
May 2003-Sept. 2009: Comparison of Linear Trend Models

<table>
<thead>
<tr>
<th>Outcome</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sentence Length (Months)</td>
<td>Final Offense Level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>12.49***</td>
<td>3.22***</td>
<td>1.01</td>
<td>1.15***</td>
</tr>
<tr>
<td>[Initial Black-White Disparity]</td>
<td>(1.17)</td>
<td>(0.79)</td>
<td>(0.78)</td>
<td>(0.11)</td>
</tr>
<tr>
<td>Month</td>
<td>-0.008</td>
<td>-0.092***</td>
<td>-0.107***</td>
<td>0.013***</td>
</tr>
<tr>
<td>[Overall Linear Time Trend]</td>
<td>(0.014)</td>
<td>(0.010)</td>
<td>(0.009)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>Black*Month</td>
<td>-0.023</td>
<td>0.043***</td>
<td>0.042***</td>
<td>-0.012***</td>
</tr>
<tr>
<td>[Trend in Black-White Disparity]</td>
<td>(0.023)</td>
<td>(0.015)</td>
<td>(0.015)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>Arrest Offense Controls?</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Final Offense Level/Category &amp; Mandatory Min Controls?</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Departure Status Control?</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Observations</td>
<td>122,140</td>
<td>121,950</td>
<td>121,950</td>
<td>121,129</td>
</tr>
<tr>
<td>Total Change in Black-White Disparity Implied by Estimates (Months)</td>
<td>-1.7</td>
<td>+3.3</td>
<td>+3.2</td>
<td>-0.9</td>
</tr>
</tbody>
</table>

Coefficients for sentencing month, black, and black*month from OLS regressions. Regressions also include controls for: criminal history, district, age, citizenship, education, and multi-defendant case status, in addition to those noted in the table. The implied overall change is calculated by multiplying the monthly disparity trend by 76 months (the length of the period). Standard errors clustered on race-month in parentheses. *p<0.10, **p<0.05, ***p<0.01.