THE YALE LAW JOURNAL ONLINE

GLENN R. SCHMITT, LOUIS REEDT & KEVIN BLACKWELL

Why Judges Matter at Sentencing: A Reply to Starr and Rehavi

In this Essay, researchers at the United States Sentencing Commission respond to criticisms by Sonja Starr and Marit Rehavi, published in the Yale Law Journal, *of the Commission's past analyses of demographic differences in federal sentences. The researchers explain the legal and practical foundation of their work and why these considerations support the Commission's methodological approach. The authors also question the representativeness of the data that Starr and Rehavi use in their alternative analyses and the assumptions they make about how the federal criminal justice system operates.*

Among the many issues that arise in any consideration of federal sentencing policy, one of the most important is whether current law or practice produces unwarranted disparities in sentencing outcomes. Congress created the United States Sentencing Commission, in part, to establish sentencing policies and practices that would avoid unwarranted sentencing disparities among defendants with similar records who commit similar crimes.[1] The Commission works to discharge this duty in a variety of ways, among them by establishing an Office of Research and Data to collect, analyze, and conduct research on data related to federal sentencing policy and practices.

Recently, the Commission has used the analytical tool of multivariate regression analysis to examine demographic differences in sentencing. We are the Commission staff members responsible for that work. In a recent article published in the *Yale Law Journal*,[2] Sonja Starr and Marit Rehavi criticize that

[1] 28 U.S.C. § 991(b)(1)(B) (2006).

[2] Sonja B. Starr & M. Marit Rehavi, *Mandatory Sentencing and Racial Disparity: Assessing the Role of Prosecutors and the Effects of* Booker, 123 YALE L.J. 2 (2013).

work and offer alternative approaches to it. We disagree with their criticisms and have serious misgivings about their alternate approaches. In particular, we believe that the data they examine do not represent the federal system as a whole and that their analytical methods are based on incorrect assumptions about how the federal system operates. In this Essay we present our thoughts on these issues.

I. THE SENTENCING COMMISSION'S ANALYSES OF DEMOGRAPHIC DIFFERENCES IN SENTENCING

Although the Commission has used regression analyses in other contexts, significant interest in its use of this tool did not develop until 2010, when it published a report that examined demographic differences in sentencing across three separate periods.[3] These periods, which together spanned the time between May 1, 2003 and September 30, 2009, reflect three distinct eras in sentencing after the Sentencing Reform Act of 1984. The beginning and end of each period was marked by a change in law or Supreme Court jurisprudence that significantly affected the discretion given to judges in imposing sentences in federal cases.[4]

The Commission's analyses found statistically significant differences in sentencing associated with demographic factors such as race,[5] gender, citizenship, age, and education. The differences found were not present in all time periods studied and differed in magnitude when they were observed. The finding that received the most attention involved race—in particular, that in

[3] U.S. SENTENCING COMM'N, DEMOGRAPHIC DIFFERENCES IN FEDERAL SENTENCING PRACTICES: AN UPDATE OF THE *BOOKER* REPORT'S MULTIVARIATE REGRESSION ANALYSIS (2010) [hereinafter 2010 REPORT].

[4] The first period examined was the PROTECT Act period, during which time Congress had constrained courts to hew closely to the sentencing guidelines. In this period, the sentencing guidelines remained binding on the courts. The second period examined was the *Booker* period, the time from the date of Supreme Court's decision in *United States v. Booker*, 543 U.S. 220 (2005), which struck down the mandatory nature of the guidelines, to just before the date of its decision in *Gall v. United States*, 552 U.S. 38 (2007). The third period was the *Gall* period, beginning on the date of that decision in which the Court clarified the amount of discretion judges have to vary from the advisory sentencing guidelines. In the 2010 REPORT, *supra* note 3, the *Gall* period extended through the end of fiscal year 2009, the last year for which data was available at the time of that report.

[5] 2010 REPORT, *supra* note 3, at 22-23. The Commission classifies offenders by race as Black, White, Hispanic, and Other. *See* U.S. SENTENCING COMM'N, 2012 SOURCEBOOK OF FEDERAL SENTENCING STATISTICS 168 (2013) [hereinafter SOURCEBOOK], http://www.ussc.gov/Research_and_Statistics/Annual_Reports_and_Sourcebooks/2012/sbtoc12.htm.

the most recent period studied, Black male offenders received sentences that were 23.3% longer than those imposed on White male offenders, when all other factors were held constant.[6]

The Commission updated this analysis with data through fiscal year 2011 as part of a longer, more comprehensive report on federal sentencing released in December 2012.[7] The Commission found that these average sentence length differences continued to exist in the most recent period studied. The average sentence length for Black male offenders exceeded that for White male offenders by 19.5% in the *Gall* period, 15.2% in the *Booker* period, 5.5% in the PROTECT Act period, and 11.2% in the *Koon* period.[8] As in the *2010 Report*, the Commission reported differences in sentencing, if any, for all other racial groups, as well as differences based on gender, citizenship, age, and education.[9] The Commission also reported on demographic differences in sentence length for fraud, firearms, and drug trafficking offenses individually.

The analytical model used for both the 2010 and 2012 reports was developed specifically for them. To develop this model, the Commission convened a roundtable of eight academics—experts in criminology, statistical analysis, and sentencing law and policy—to advise it as to which factors and methodological approaches should be used. The Commission's first report using this new model, the *2010 Report*, was also reviewed by two additional outside experts—one a criminologist and the other an expert in regression analysis, both of whom are tenured professors at major research universities—prior to its release to ensure that our analysis was performed correctly and the results were accurately stated.

[6] 2010 REPORT, *supra* note 3, at 23.

[7] U.S. SENTENCING COMM'N, REPORT ON THE CONTINUING IMPACT OF *UNITED STATES V. BOOKER* ON FEDERAL SENTENCING pt. A, at 108-10 (2012) [hereinafter 2012 REPORT]. The fiscal year for the federal government ends on September 30.

[8] In the *2012 Report* the Commission also analyzed cases from an earlier timeframe, the *Koon* period, which spanned October 1, 1998 to the date of enactment of the PROTECT Act on April 30, 2003. *Id.* pt. E, at 1.

[9] We remind readers of an important disclaimer that the Commission made in the *2012 Report*:

> Because judges make sentencing decisions based on many legal considerations, such as violence in an offender's past, or an offender's employment history, which are not controlled for in the Commission's multivariate regression analysis, these results should be interpreted with caution and should not be taken to suggest race or gender bias on the part of judges.

Id. pt. A, at 9.

In the new model, an offender's sentence length (the length of imprisonment as well as the length of any alternative confinement imposed) was the dependent variable and the independent variables included both case characteristic variables and demographic variables.[10] The Commission did not report any finding unless it was statistically significant at the ninety-nine percent confidence level (i.e., a finding was only reported if the likelihood that the result was due to chance was less than one in a hundred).[11]

Each of the factors in the model was selected based on what we have learned about federal sentencing practices from our combined fifty years of experience studying federal sentencing, as well as on the recommendations of the ten experts who advised us in developing it. The model was designed to measure the way in which sentences are actually imposed in the federal courts in felony and Class A misdemeanor cases,[12] both as a matter of law and in practice. Our goal was to examine the federal system as a whole and to report our findings. By contrast, Starr and Rehavi study less than twenty percent of the offenders convicted in the period they examine, excluding the two largest groups of federal offenders from a key part of their analysis, and use analytical methods based on less reliable data and incorrect assumptions about how the federal system operates.[13]

II. STARR AND REHAVI'S CRITICISMS OF THE COMMISSION'S WORK

We understand Starr and Rehavi's central thesis to be that the role of prosecutors in sentencing has been overlooked by researchers who use quantitative analytical tools to examine sentencing data. They assert that tremendous power is concentrated in the hands of prosecutors, and that the choices prosecutors make concerning what charges to bring determine the sentences imposed today to a greater extent than they did before the sentencing guidelines were established. They believe that prosecutors gain this power through the findings of fact often found in plea agreements, as most federal cases involve a guilty plea, and judges "typically lack the incentive, . . . and may

[10] For a description of these variables, see *id.* pt. E, at 7-8, 31-32.

[11] In both reports we presented the computer output of all of our data variables and analyses in a detailed appendix so that other researchers could recreate our variables and replicate our work, a standard practice in the social sciences.

[12] The Commission does not collect information on petty offenses (Class B and Class C misdemeanors or infractions). The sentencing guidelines do not apply in those cases. U.S. SENTENCING GUIDELINES MANUAL ch. 1, § 1B1.9 (2012) [hereinafter USSG].

[13] *See infra* notes 43-55 and accompanying text.

lack the information, to diverge from what the parties have agreed upon."[14] Moreover, Starr and Rehavi believe that prosecutors have continued to wield this power even after the guidelines became advisory—both because the prosecutor decides which charges to file, thus setting statutory maximums and minimums that limit judges' sentencing decisions, and because prosecutors can present facts to judges through plea agreements and at sentencing hearings.[15]

Starr and Rehavi attempt to study the impact of prosecutorial decisionmaking on sentencing outcomes by looking to data about offenders collected at points in time before the date of sentencing. They presume that these data are largely unaffected by the actions of prosecutors and, therefore, paint a more accurate picture of the true criminal conduct of offenders.

Many actors in the criminal justice system may have influence on the sentences judges ultimately impose, but we reject the notion that prosecutors alone principally determine the sentences imposed in the federal system. We think Starr and Rehavi wrongly disregard the influence of the defense community, and especially the federal public defenders, who we believe quite effectively represent the interests of their clients. We also do not accept the notion that judging in the federal system is limited largely to picking a sentence from a table after adding up a numerical score. In our experience, judges are very concerned with determining the fair sentence and assess all of the evidence presented to them with a discerning eye. We think this is especially true in the federal system, which uses a modified "real offense" system based on charged conduct as well as other "relevant conduct,"[16] a critical point that we discuss further below.

We would be among the first to congratulate researchers who develop datasets and methodologies that can assess the influence of each person or group in the criminal justice system. But we don't think that has yet happened. And while we find Starr and Rehavi's work interesting, we believe they are unduly harsh in their criticism of the Commission's past work, and of other researchers who have found similar results using methods similar to that of the Commission. We also have serious doubts that the alternative approaches Starr and Rehavi use do what they assert. In fact, we believe their new approaches

[14] Starr & Rehavi, *supra* note 2, at 12.

[15] *Id.* at 10-16.

[16] *See* USSG, *supra* note 12, ch. 1, subpt. A.4(a). *See generally* William W. Wilkins, Jr. & John R. Steer, *Relevant Conduct: The Cornerstone of the Federal Sentencing Guidelines*, 41 S.C. L. REV. 495 (1990) (discussing the importance of relevant conduct in assessing offender culpability).

are so flawed that conclusions based on them are misleading to policymakers and others who may rely on them.

A. The Presumptive Sentence

The sentences imposed on any two offenders are likely to differ for a number of important reasons, the most obvious of which is the severity of the crime each offender has committed. Researchers seeking to examine sentencing trends across differing crime types must develop a means to control for differences in the severity of the crime in each case. Starr and Rehavi criticize our measure of offense severity and propose their own, calling this the "key difference" between their approach and ours.[17]

In our work, we use an offense severity factor called the "presumptive sentence," which is the low end of the sentencing range called for under the sentencing guideline that the court applied in the case, and which reflects any minimum or maximum statutory limits on the court's discretion.[18] This number results from a consideration of two key factors: the severity of the offense committed and the offender's criminal history. These two broad considerations are incorporated into the sentencing guidelines and the advisory sentencing ranges that they provide.

While Starr and Rehavi note that using the presumptive sentence variable is the traditional approach to studying sentencing outcomes under systems that use sentencing guidelines, they criticize methods that rely on sentencing guidelines determinations as "distant proxies for the seriousness of the underlying conduct" and "the end product" of decisions made by prosecutors about charging, plea-bargaining, and sentencing fact-finding.[19]

In contrast, we believe any analysis of sentencing outcomes must be based on the legal and practical realties of the sentencing process. In our view, then, Starr and Rehavi make a significant error when they ignore the process judges actually use when making sentencing decisions. Federal judges are required *by law* to consider the advisory guideline range in every case.[20] Indeed, the Supreme Court has held that the failure to properly calculate and consider the

[17] Starr & Rehavi, *supra* note 2, at 16.

[18] Numerically, the presumptive sentence is the number of months of confinement at the bottom of the sentencing range.

[19] Starr & Rehavi, *supra* note 2, at 17.

[20] *See* Gall v. United States, 552 U.S. 38, 49 (2007) (citing Rita v. United States, 551 U.S. 338, 347-48 (2007)).

advisory guideline sentence is reversible procedural error.[21] The decision by Starr and Rehavi to disregard the presumptive sentence as a factor in their analytical model ignores the reality that judges are required to consider it when they decide sentences.

Furthermore, disregarding the presumptive sentence ignores the fact that the guidelines actually do affect sentencing decisions. In its recent decision in *Peugh v. United States*, the Supreme Court characterized the guidelines as the "lodestone of sentencing"[22] and recognized that the guidelines range is "intended to, and usually does, exert controlling influence on the sentence that the court will impose."[23] The Commission has also reported to Congress—after extensive quantitative and qualitative analyses, numerous public hearings, and interviews across the country with prosecutors, defenders, and probation officers—that the guidelines continue to exert significant influence on judges' sentencing decisions for most offenses.[24] To omit consideration of the sentencing guidelines from a study of federal sentencing outcomes is simply incorrect as a matter of both law and actual courtroom practice.

Finally, Starr and Rehavi's study ignores the critical fact that under the modified "real offense" approach in the federal guidelines, a judge must calculate the offense level based not only on charged conduct but also on other "relevant conduct."[25] This includes uncharged conduct that can increase or decrease a defendant's culpability. The Commission adopted this approach precisely because it wanted judges to be able to account for prosecutorial charging decisions that failed to represent a defendant's actual conduct.[26] Indeed, in more than sixty percent of all cases in fiscal 2012 the sentencing guidelines range was modified (higher or lower) by the application of one or more specific offense characteristics (the guidelines provisions that capture relevant conduct).[27] Contrary to the assertion by Starr and Rehavi that the presumptive sentence is only a "distant proxy" for the seriousness of a

[21] *Id.* at 51.

[22] Peugh v. United States, No. 12-62, slip op. at 13 (U.S. June 10, 2013).

[23] *Id.* at 14.

[24] 2012 REPORT, *supra* note 7, pt. A, at 60.

[25] *See* USSG, *supra* note 12, ch. 1, subpt. A.4(a); *id.* ch. 1, § 1B1.3.

[26] *Id.* ch. 1, subpt. A.4(a).

[27] *See* U.S. SENTENCING COMM'N, USE OF GUIDELINES AND SPECIFIC OFFENSE CHARACTERISTICS, OFFENDER BASED, FISCAL YEAR 2012, at 2 (2013), http://www.ussc.gov/Data_and_Statistics/Federal_Sentencing_Statistics/Guideline_Application_Frequencies/2012/Use_of_Guidelines_and_Specific_Offense_Characteristics_Guideline_Calculation_Based.pdf.

defendant's conduct, we believe the presumptive sentence provides a nuanced estimate of that conduct, determined by the sentencing judge after considering the evidence presented by both parties.

B. Controlling for Departures from the Sentencing Guidelines

Starr and Rehavi next claim that our use of the presumptive sentence as a measure of offense severity "is even worse" because we separately control for cases where the judge departs or varies from the sentencing guidelines.[28] They assert that this converts our analysis into one that estimates "race gaps in the *size* of departures (and in sentence choices within the narrow Guidelines range), but filter[s] out *whether* there is a departure."[29]

Under current law, judges are required to first correctly determine and then consider the guidelines sentence. The guidelines themselves then allow judges to depart from the advisory sentencing range, often at the request of one of the parties, for any of several "departure" reasons described in the *Guidelines Manual*.[30] Finally, judges are required to consider a separate set of statutory factors[31] in determining the final sentence—in essence deciding whether to vary from the advisory sentencing range. We account for these separate steps in our analytical model by separately controlling for whether the judge chose to depart or vary from the presumptive sentence. Rather than "filtering out" a key part of the sentencing decision, as Starr and Rehavi assert, our model controls for the different steps in the process. To collapse these steps into one factor—as Starr and Rehavi apparently would have us do—would ignore the reality of the actual sentencing process in the federal courts.

Starr and Rehavi are wrong when they assert that our model analyzes only "sentence choices within the narrow Guidelines range."[32] Our model examines all aspects of the sentence, whether within the range or outside it. Although the departure variable statistically accounts for the judge's separate decision to sentence outside the guidelines range, that variable does not account for the entire sentence. Even when the sentence is a departure or variance from the guidelines range, the presumptive sentence variable continues to account for much of the sentence imposed due to the anchoring effect discussed above.

[28] Starr & Rehavi, *supra* note 2, at 19.

[29] *Id.* at 20.

[30] *See* USSG, *supra* note 12, ch. 5, pt. K.

[31] *See* 18 U.S.C. § 3553(a) (2006).

[32] Starr & Rehavi, *supra* note 2, at 20.

Further, we think it is important to point out that departure decisions are not wholly judicial. More than sixty percent of all sentences below the guideline range result from a request by the government for the lower sentence,[33] often due to the offender's "substantial assistance" to the government "in the investigation or prosecution of another" offender[34] or the offender's willingness to participate in one of the government's early disposition programs (which leads to an expedited conclusion to the case).[35] Generally, courts may not reduce sentences based on these factors absent a motion by the prosecutor.[36] By separately controlling for departures, including these, we essentially control for this prosecutorial influence.[37]

In the *2012 Report*, we performed a separate analysis on the size of demographic differences in departure cases—that is, where the sentence was a departure below the guidelines range at the request of the government or a departure or variance that the judge made on his or her own. In those cases, we found no difference between the sentences for Black and White males in the *Gall* period (the most recent) after controlling for all other factors. If our overall analysis that found a 19.5% racial gap were actually estimating "race gaps in the size of departures (and in sentence choices within the narrow Guidelines range)," as Starr and Rehavi assert,[38] we would not have found these very different results in our departure case analyses.

As a final observation, it is particularly puzzling that Starr and Rehavi would criticize us for separately controlling for departures and variances given the outcome if we had not done so. Without these controls, the racial differences between Black males and White males increase significantly. We believe this is because White males are far more likely to obtain a court-

[33] This figure was sixty-one percent in fiscal year 2012. *See* SOURCEBOOK, *supra* note 5, at tbl.N.

[34] 18 U.S.C. § 3553(e) (2006); *see* USSG, *supra* note 12, ch. 5, § 5K1.1.

[35] *See id.* § 5K3.1. For further background on early disposition programs, see Memorandum from Deputy Att'y Gen. James M. Cole for All U.S. Att'ys, Department Policy on Early Disposition or "Fast-Track" Programs 1-2 (Jan. 31, 2012), http://www.justice.gov/dag/fast-track-program.pdf.

[36] 18 U.S.C. § 3553(e); USSG, *supra* note 12, ch. 5, §§ 5K1.1, 5K3.1.

[37] In the Commission's model, the departure controls do not differentiate between departures that the prosecution requests and departures and variances that the judge gives on his or her own initiative. However, when we reanalyzed the Commission's model with separate departure controls for government-sponsored departures and judge-initiated departures and variances, there was virtually no change in the size of the Black male vs. White male differences.

[38] Starr & Rehavi, *supra* note 2, at 20 (emphasis omitted).

imposed departure or variance than Black males.[39] Without accounting for this, the demographic differences in sentencing would seem even larger.

C. Controlling for the Type of Crime Involved

Starr and Rehavi further criticize the use of the presumptive sentence factor in traditional sentencing research because "it controls only for differences in crime severity . . . not for differences in crime *type*."[40] While they do note that many studies separately control for the type of crime involved, they assert that the controls used often involve only broad categories.[41]

The authors are correct that the presumptive sentence variable itself does not control for crime type, but we find it misleading that they imply that using this factor means that the resulting analysis has no crime-type control. Most researchers include such controls in their analytical model, as do we. The Commission's analytical model uses seven separate controls for crime type, albeit ones involving broad categories.[42]

Starr and Rehavi claim that their narrower crime-type controls are more precise. Unfortunately, they do not provide any way for other researchers to assess that claim. Most researchers using regression analysis report a statistical measure, commonly called the R-squared (R^2), that explains how much of the variation found in the data is accounted for by the variables in the empirical model used. While not the only test of the "fit" of a regression model, a low R^2 indicates that there is more to be explained in a particular set of data than the model used accounts for. The Commission's model explains over sixty percent of the variation in sentences, a relatively high level of explained variation. Starr and Rehavi do not report this statistic for their work overall, or whether the additional crime-type controls they use increased the R^2 of their model above what it would have been using broader categories.

[39] *See* 2012 REPORT, *supra* note 7, pt. E, at 22. White males were 25.2% more likely to receive a court-imposed departure or variance than Black males in the *Gall* period. The extent of the departure or variance did not differ based on race.

[40] Starr & Rehavi, *supra* note 2, at 23.

[41] *Id.*

[42] The crime type controls are: violent crime, drug trafficking, sexual abuse, pornography, immigration, white-collar crime, and "other" crime. 2010 REPORT, *supra* note 3, at C1-C2.

III. MAJOR LIMITATIONS OF THE STARR-REHAVI MODEL

Starr and Rehavi claim to have developed a more precise approach to studying sentencing outcomes, one that specifically accounts for the impact of prosecutors. We believe the data they choose to study, and the model they employ for that analysis, have a number of serious flaws.[43] Because of this, the conclusions they draw are not only unrepresentative of federal sentencing practices as a whole but are misleading to policymakers and others who might rely on them.

A. The Starr-Rehavi Alternative Measures Are Unreliable

Starr and Rehavi make much of the fact that their work includes a separate assessment of charging decisions in addition to sentencing outcomes. For their charging decision analysis, they devise several new variables, which they use in place of the "presumptive sentence" and crime-type variables that we use. They assert that with these variables they are able to compare "defendants who look similar near the *beginning* of the justice process" and assess "the aggregate sentencing disparity introduced by decisions throughout the post-arrest justice process."[44] We believe that their alternative measures of offense severity proceed from several false assumptions, fail to achieve their intended goal, and lead to unreliable results.

1. Data Lost in the Linking Process

The data Starr and Rehavi use were gathered from four federal agencies, including the Commission. Linking datasets from different federal agencies can be difficult. Cases that cannot be matched usually must be dropped from any analysis. Starr and Rehavi were able to match only sixty-eight percent of the cases in the four datasets.[45] While this problem is somewhat unavoidable, there

[43] The authors have not made their data available for further analysis, and so we are unable to independently verify the results they report. However, they do point to another article they wrote that presents the "full results" of their analysis including "a full technical explanation of [their] methods." Starr & Rehavi, *supra* note 2, at 5-6 n.9 (citing M. Marit Rehavi & Sonja B. Starr, *Racial Disparity in Federal Criminal Charging and Its Sentencing Consequences* (Univ. of Mich. Program in Law & Econ., Working Paper No. 12-002, 2012) (under review), http://ssrn.com/abstract=1985377). This article too is incomplete in its description of variables and so does not allow for replication of their work.

[44] *Id.* at 24 (emphasis omitted).

[45] Rehavi & Starr, *supra* note 43, Data App. at 3.

is no reason to believe that the data lost is random, and so the resulting dataset they use is unlikely to be representative of the federal system as a whole from the outset.

2. *Variables Based on Arrest Codes*

In an effort to assess whether offenders committing similar crimes are charged in similar ways, Starr and Rehavi use "arrest codes" found in data collected by the U.S. Marshals Service to develop separate crime-type variables. The authors claim that these variables provide more information not only on the type of crime involved in the case, but also on the severity of the crime. Indeed, they state that it is a "much better proxy for actual conduct than the presumptive Guidelines sentence."[46]

We accept that using 430 different codes grouped into 107 categories may measure crime type to a high degree.[47] We wish that the authors had reported data to support their claim that these additional variables added precision to their model over the more common approach of using fewer (and broader) crime-type controls. But we disagree with their assertion that using arrest data as a measure of offense severity enables them to compare offenders who "look similar near the *beginning* of the justice process."[48]

In particular, we doubt that offenders with similar arrest codes are as similar as Starr and Rehavi assert. In fact, using arrest codes could cause dissimilar offenders to be treated identically in their analysis. For example, a fraud case with a small loss will have the same arrest code as one involving a large loss. Using arrest codes also does not account for any criminal conduct that is not specifically charged—either because it was not a necessary element of the offense that was charged or because it was not the primary reason for the arrest—or conduct by other defendants in a jointly undertaken enterprise. For example, two offenders who committed the same crime will have the same arrest code, but the one who possessed a gun in connection with the offense is certainly more culpable, yet might not be arrested for a weapon offense. Most people would accept that the leader of a gang is more culpable than a junior member, but arrest charges will not capture this difference in roles. Arrest codes also do not capture additional criminal conduct that is discovered after

[46] *Id.* at 35.

[47] Rehavi & Starr, *supra* note 43, Data App. at 6.

[48] Starr & Rehavi, *supra* note 2, at 24.

the initial arrest. The presumptive sentence variable that the Commission uses does capture this conduct.

Finally, we doubt that the arrest decision is as disconnected from the influence of prosecutors as Starr and Rehavi believe. In our experience in the federal system, law enforcement officials usually work very closely with prosecutors during the investigation stage of a case. Often, arrests are not made until prosecutors are satisfied that sufficient evidence has been gathered to obtain an indictment from the grand jury, as is required in most federal cases. This is especially true in fraud cases, which often involve long investigations. Fraud offenses are a major component of Starr and Rehavi's limited dataset. Starr and Rehavi's presumption—that arrest decisions are independent from prosecutorial decisionmaking and, therefore, that arrest data measure offender conduct prior to the involvement of a prosecutor—simply doesn't hold in the federal system.

3. Variables Based on Statutory Maximums and Minimums

In conducting their analysis of charging decisions, Starr and Rehavi consider four separate charging measures and attempt to examine whether racial differences exist in charging decisions in light of these measures. While Starr and Rehavi are not specific about how they developed these measures, they appear to be based on the statutory maximum penalty of the crime charged, the statutory minimum penalty (if any), an (incomplete) sentencing guidelines calculation, and the average sentence imposed on White males for a period prior to the data examined.

We believe their reliance on statutory maximum and minimum punishments as variables in their model is misplaced. Congress does not make fine distinctions among offense severity when setting statutory maximums and minimums in proposed legislation. Using these statutory ceilings and floors, which are often expressed only in multiples of five (e.g., five- and ten-year mandatory minimum penalties,[49] or ten- and twenty-year maximum penalties[50]) is a rough cut, at best. The Commission devotes significant time

[49] *See, e.g.*, 21 U.S.C. §§ 841, 960(b) (2006) (establishing five-, ten-, and twenty-year minimum penalties in certain drug cases); 18 U.S.C. §924(c) (2006) (establishing five-, seven-, and ten-year minimum penalties in certain firearms cases); 18 U.S.C. § 2252(b)(1) (2006) (establishing five- and fifteen-year minimum penalties in child pornography cases).

[50] *See, e.g.*, 8 U.S.C. § 1326 (2006) (establishing two-, ten-, and twenty-year statutory maximum penalties in immigration cases); 21 U.S.C. §§ 844, 960(d) (2006) (establishing ten and twenty-year statutory maximum penalties in drug cases); 18 U.S.C. §§ 1029(c),

and analytical resources, both legal and statistical, to determine the relative severity of various crimes with similar statutory maximums and minimums set by different groups of members of Congress in different years. Starr and Rehavi seem to assume that if the maximums (or the minimums, when they do exist) are the same, then the crimes must be similarly severe. This has not been our experience.

B. The Starr-Rehavi Dataset Does Not Reflect the Real World of Federal Sentencing

1. Principal Analyses Limited to a Non-Random Fraction of Federal Cases

While the Commission's analyses include virtually all offenders for whom complete sentencing information was reported to the Commission by the courts,[51] the offenders Starr and Rehavi study for most of their analyses account for less than twenty percent of the offenders convicted during the period they study.[52] Their principal sample contains only offenders convicted of "property and fraud crimes, weapons offenses, regulatory offenses, and violent crimes."[53] This group is far from representative of the federal system, where crimes involving immigration and drugs alone account for more than sixty percent of all cases.[54] Even more problematic, they include data on only about half of the White males in the population and just under half of the Black males, the very groups on which they purport to focus. This non-random exclusion of data on the two groups they study further undermines the reliability of their results.

The dataset Starr and Rehavi use also significantly underrepresents cases involving mandatory minimum penalties. For example, in the two-year time period they studied (a portion of the *Gall* period), 29.4% of all offenders were

1035, 1037(b), 1341, 1343 (2006) (establishing five-, ten-, fifteen-, twenty-, and thirty-year maximum penalties in various fraud offenses).

[51] The data used for the *2012 Report* included cases on 707,987 offenders.

[52] Starr and Rehavi do not provide statistics on the exact number of cases in any of their analyses, but the Commission attempted to replicate their population based on the parameters stated in their technical appendix. *See supra* note 43. The replicated population was 16.6% of the available data during the time period examined.

[53] Rehavi & Starr, *supra* note 43, at 7.

[54] *See* GLENN R. SCHMITT & JENNIFER DUKES, U.S. SENTENCING COMM'N, OVERVIEW OF FEDERAL CRIMINAL CASES: FISCAL YEAR 2012 (2013) [hereinafter 2012 OVERVIEW], http://www.ussc.gov/Research_and_Statistics/Research_Publications/2013/FY12_Overview_Federal_Criminal_Cases.pdf.

convicted of an offense carrying a mandatory minimum. In the data Starr and Rehavi use, however, only 13% of the offenders were convicted of an offense carrying a mandatory minimum penalty. Of the cases that Starr and Rehavi ignored, almost one in three (32.8%) involved a mandatory minimum penalty. Given the strong assertions they make about the role of mandatory minimum penalties, it is unfortunate that their principal sample analysis relies on data that lack many of the cases in which this type of penalty applied.[55]

Finally, the principal sample Starr and Rehavi use included data only from fiscal years 2007 to 2009, and then only cases in which both the charge and the sentence occurred in that period.[56] In comparison, the Commission's most recent report uses data on virtually all cases from 1998 through 2011. By focusing on such a short period of time, Starr and Rehavi exclude even more cases from their analysis, especially those involving complex criminal enterprises and multiple defendants. This only exacerbates the bias introduced by their decision to use an unrepresentative dataset of offenses. If nothing else, the implied claim in their article—that the results they report reflect the sentencing practices in the federal system as a whole—is unsupported by the data they have used.

2. Incorrect Mandatory Minimum Penalty Controls

Starr and Rehavi report a ten to fourteen percent difference in sentences between Black males and White males after controlling for various factors. They claim that these differences are cut by half when they control for "mandatory minimum charges" and completely eliminated when they control for the "final mandatory minimum sentence."[57] They point to this as proof that the prosecutor's decision to charge crimes carrying a statutory mandatory minimum penalty—and not any judicial decision-making—explains any racial differences in sentence lengths. The authors do not fully explain how they create these variables, but it appears the "mandatory minimum charge" variable applies in any case where the crime charged carries a mandatory

[55] Starr and Rehavi state that they bring drug cases into their analysis at a latter point, but they do not provide specifics on this step in their analysis. In particular, they do not state whether they are able to control for drug type or quantity. These are important considerations because there are racial and sentencing differences in federal drug cases based on the type of drug involved. For example, the rate at which drug cases involve a mandatory minimum charge varies by the type of drug involved. *See* SOURCEBOOK, *supra* note 5, tbl.43.

[56] They do use data from 2001 to 2009 in their separate discontinuity analysis. Starr & Rehavi, *supra* note 2, at 53.

[57] *Id.* at 29 (emphasis omitted).

minimum penalty and that the "final mandatory minimum sentence" variable includes cases in which the offender was convicted of an offense carrying a mandatory minimum penalty. If so, both significantly overrepresent the extent to which judges' sentencing discretion is limited by these provisions.[58]

In almost half of all cases in which an offender is charged with a crime carrying a mandatory minimum penalty, the offender is relieved of that penalty by the date of sentencing.[59] When that happens, the judge is no longer constrained from imposing a sentence lower than the one called for in the sentencing guidelines. Classifying these cases as involving a mandatory minimum penalty is legally incorrect, and would significantly overrepresent the impact of these penalties. Our model correctly controls for the presence of a mandatory minimum penalty by applying this variable only in cases where the offender did not receive relief from the penalty. When we control for the true effect of such a penalty, significant demographic differences in sentences remain.

3. Drug Offenders Are Excluded

Starr and Rehavi indicate that the major difference between their approach and the Commission's is the inclusion in their model of controls for "arrest offense and other characteristics that are fixed at the beginning of the justice process."[60] However, because information about drug type or drug quantity was not available in the data they used from the U.S. Marshals Service or the Executive Office for U.S. Attorneys,[61] they report that they "cannot assess initial charging disparities in drug cases."[62] Consequently, they exclude drug cases from all analyses in which they use their "principal sample." Because of

[58] Starr and Rehavi note that they are missing "nearly 40% of the mandatory minimums" and thus "understating mandatory minimums' role." *Id.* at n.90. This may not be correct, as the authors provide no analysis as to whether this forty percent error rate is random or, instead, concentrated in specific offense types or on specific type of offenders (e.g., those with more serious criminal histories). If this error is non-random, then it would necessarily affect the results of any analysis based on it.

[59] U.S. SENTENCING COMM'N, REPORT TO THE CONGRESS: MANDATORY MINIMUM PENALTIES IN THE FEDERAL CRIMINAL JUSTICE SYSTEM, at xxviii (2011) (documenting that "46.7% . . . of offenders convicted of an offense carrying a mandatory minimum penalty were relieved from the application of such a penalty at sentencing because they provided substantial assistance to the government or qualified for the safety valve provision, or both"); *id.* at 132.

[60] Starr & Rehavi, *supra* note 2, at 7.

[61] *Id.* at 24, 32-33; Rehavi & Starr, *supra* note 43, at 2, 7-8.

[62] Starr & Rehavi, *supra* note 2, at 26.

this, each of their three major analyses is skewed in that it is limited largely to firearms and fraud cases and, therefore, is not representative of the federal population as a whole.

The lack of drug cases in Starr and Rehavi's major analyses undermines their conclusions about the impact of prosecutor decisions to charge offenses carrying a mandatory minimum penalty. Drug offenses are the second most common federal crime, accounting for over thirty percent of all cases in fiscal year 2012.[63] More importantly, drug trafficking cases are the most common offense type for Black male offenders and are among the offenses that produce the highest average sentences. Compounding this error is the fact that more than three-quarters of the federal cases involving a mandatory minimum penalty are drug trafficking cases. As Starr and Rehavi's major analyses lack these cases, their conclusions about the impact of mandatory minimum penalties in the federal system are unsupported.

Starr and Rehavi state that they add drug cases (as well as child pornography cases) for a "more limited" analysis,[64] presumably using an additional charge variable for these cases despite the fact that they had earlier explained that these cases involved "ambiguities" that were too extreme to be included in the principal sample analyses.[65] To perform this secondary analysis the authors use a variable for the effect of a mandatory minimum penalty based not on the offense as charged but on the offense at conviction. The authors do not explain why their analysis using sentencing data is not flawed, as the "end product of charging, plea-bargaining, and sentencing fact-finding,"[66] in the same way they claim the Commission's analyses are. Regardless, given that Starr and Rehavi find it acceptable to use sentencing data to make their ultimate claim—that there are no unexplained racial differences at the time of sentencing—we find it unfair that they so severely criticize the Commission's use of sentencing data in its analyses.

4. Immigration Cases Are Excluded

Starr and Rehavi state that they exclude immigration cases because "their stakes typically turn on deportation, making prison sentence length analysis a

[63] 2012 OVERVIEW, *supra* note 54, at 3 (reporting that drug cases accounted for 30.2% of all federal crimes in fiscal year 2012).

[64] Starr & Rehavi, *supra* note 2, at 7 n.13 and accompanying text.

[65] *Id.* at 26, 30.

[66] *Id.* at 6.

very incomplete picture of case outcomes."[67] We disagree. More than three-quarters of all noncitizen immigration offenders receive an incarceration sentence, and more than half are sentenced to a year in prison or longer. This is real punishment, and it is served before any deportation occurs. Also, in excluding all immigration cases Starr and Rehavi exclude a number of U.S.-citizen offenders, as over five percent of immigration cases involve citizen offenders.[68] Of course, citizen offenders are not subject to deportation. For Hispanic citizens, immigration crimes are the third most common offenses committed.

5. Cases with Non-U.S.-Citizen Offenders are Excluded

Starr and Rehavi exclude cases involving all non-U.S. citizens. This decision also limits their ability to generalize their finding to the federal system as a whole, as non-U.S.-citizen offenders are more than forty percent of the federal offender population.[69] Non-citizens also account for more than thirty percent of all drug offenders,[70] and so excluding them further undermines the authors' secondary analysis using drug offenses.[71]

6. Hispanic Offenders Are Combined with White and Black Offenders

Starr and Rehavi assert that they studied differences in the sentences between Black and White male offenders; however, they also have included Hispanic offenders in their analyses, as either White or Black offenders. Again, they explain that they took this approach because the charging data available to them did not include information about ethnicity. This limitation could bias their results because White, Black, and Hispanic offenders do not commit similar crimes,[72] have different offense severity levels,[73] and have different

[67] *Id.* at 26.

[68] *See* SOURCEBOOK, *supra* note 5, tbl.48.

[69] 2012 OVERVIEW, *supra* note 54, at 3 (46.1% in fiscal year 2012).

[70] SOURCEBOOK, *supra* note 5, tbl.36 (30.4% in fiscal year 2012).

[71] Starr & Rehavi, *supra* note 2, at 30.

[72] In the population that Starr and Rehavi studied, firearms offenses were committed by 38.1% of Hispanic offenders, compared to just 22.4% of White offenders but 44.5% of Black offenders.

[73] In the population that Starr and Rehavi studied, the average guideline minimum (including trumping mandatory minimums) for Hispanic offenders was fifty-six months, while for White offenders it was forty-seven months and for Black offenders it was seventy months.

criminal histories.[74] Including the dissimilar Hispanic offenders in the White or Black groups could make the data about those groups unrepresentative of the true population of White or Black offenders and confound any race effect. Further, this feature of their model makes any real comparison between our results and theirs impossible.

IV. THE IMPACT OF *BOOKER*

Starr and Rehavi further criticize the Commission's analyses for our decision to group cases into four time periods. They imply that these periods are too long to attribute any changes in sentencing outcomes to changes in federal sentencing law that mark the beginning of those periods. Instead, Starr and Rehavi posit that any measurable impact of a change in the law, such as from a Supreme Court decision, is limited to only a short period after that decision is announced. They suggest that beyond that point, other factors could be the cause of changes in sentencing practices—such as changes in the types of cases that are prosecuted, changes in the staffing at U.S. Attorney or defender offices, or changes in the government's enforcement priorities.[75]

To address this perceived shortcoming, Starr and Rehavi conduct an analysis in which they look for sharp breaks in sentencing immediately after the date of the *Booker* decision[76] as proof that the decision affected sentencing outcomes along racial lines. They conclude that because they did not find these sharp breaks, *Booker* had no effect on sentencing disparity. We believe the theoretical underpinnings of this approach are not well-founded.

We disagree that the full impact of *Booker* would be seen in the weeks or months immediately after *Booker*. Although there was a sharp drop in the rate at which cases were sentenced within the guidelines range immediately after *Booker* was announced, we believe that it took considerable time (and numerous lower court rulings) for the full impact of *Booker* to be understood and reflected in sentencing data. In fact, we believe that it was not until the Supreme Court decisions in *Kimbrough*[77] and *Gall* that the full impact of the *Booker* decision was fully understood in some lower courts. This fact is

[74] In the population that Starr and Rehavi studied, 53.6% of Hispanic offenders are in Criminal History Category I, compared to 56.7% of White offenders but just 31.9% of Black offenders. Also, while 9.7% of Hispanic offenders are in Criminal History Category VI, 10.7% of White offenders and 18.4% of Black offenders fall into that category.

[75] Starr & Rehavi, *supra* note 2, at 8, 50-52.

[76] United States v. Booker, 543 U.S. 220 (2005).

[77] Kimbrough v. United States, 552 U.S. 85 (2007).

reflected in our own data, which show a downward trend in the rate of sentences imposed within the guidelines range that did not begin until the *Kimbrough* and *Gall* decisions.[78]

Starr and Rehavi do not fully explain how they conducted their analysis, and so we cannot replicate it. However, we did rerun the Commission's model using only data from six months before and after *Booker*, and limited that analysis to cases from the district courts in the same circuits that Starr and Rehavi examine.[79] We found statistically significant differences in the sentences imposed between Black male and White male offenders in the six months after *Booker*. We also found these sizeable differences to exist in the six months before *Booker* (which was roughly the time between the *Blakely* and *Booker* decisions),[80] although the differences were larger after *Booker*. When we ran our model for those same timeframes and circuits but limited the data further to only the offense types that Starr and Rehavi used, we continued to find differences, but the increase between the two timeframes was smaller, although still statistically significant.

We then separately examined data from the six months at the end of the PROTECT Act period (i.e., the six months just *before* the *Blakely* decision), again looking only at cases from the five circuits that Starr and Rehavi examined. Our model found no statistically significant difference between the Black male and White male sentences, whether using the full set of cases or the more limited set of offenses that Starr and Rehavi use. This suggests to us that the increases in racial disparity that we found after *Booker* may have begun to develop after the decision in *Blakely*, a case with a holding very similar to *Booker* but which applied only to state court cases.

If so, then the reason Starr and Rehavi failed to find any impact of *Booker* or the case that foreshadowed it, *Blakely*, was not because there wasn't one, but because they wrongly assumed that *Booker* would only have caused a sharp change in sentencing outcomes along racial lines if it had any effect. As our analysis shows, the impact of *Booker* appears to have been more gradual, supporting our proposition that the full impact of such a decision takes some

[78] *See, e.g.*, U.S. SENTENCING COMM'N, FINAL QUARTERLY DATA REPORT FISCAL YEAR 2010, at 10-12 (2011).

[79] Starr and Rehavi limit the data they use for this part of their analysis to cases from the district courts in the five circuits that held that the *Blakely* decision did not apply in the federal courts, presuming that sentencing practices in those courts would be more affected by the *Booker* decision than those in courts in the circuits that held *Blakely* did apply. Starr & Rehavi, *supra* note 2, at 57-58.

[80] The period between *Blakely* and *Booker* was not examined for the Commission's prior reports.

time before it can be measured, and may have begun even before the date of that decision.

The longer time periods used in the Commission's work reflect the eras in which very different legal constraints were imposed on the courts and prosecutors. While other factors could also have influenced sentencing decisions during those periods, some related to the legal constraints and some related to other changes in the criminal justice system, this possibility does not undercut the importance of the legal decisions that mark the periods we used in our work. We do not think it important to say (or to disprove) that *Booker*, or *Kimbrough* and *Gall*, alone caused the differences we find. Rather, we think it is more important to note that demographic differences in sentences existed to a much larger extent after those decisions than immediately before them and that policymakers, judges, and the Commission should consider this fact. The technique used by Starr and Rehavi hides these differences entirely.

CONCLUSION

We support any researcher who is able to shed more light on differences in sentencing that may be associated with demographic factors. But we do not believe Starr and Rehavi have done this. The measures of offense severity that they develop are imprecise and do not account for the actual practice in the federal courts. Their analysis of charging decisions is limited to selected crimes from a narrow period of time, and so cannot be generalized to the federal system as a whole—a key point that is largely ignored in their article. Finally, their presumption that any impact of the *Booker* decision would produce only sharp breaks in sentencing patterns ignores the reality that the impact of this decision has evolved over time.

We continue to believe that an offense severity measure based upon the application of the federal sentencing guidelines is the most complete and accurate available, because it reflects a judge's findings of the actual criminal conduct by an offender (whether charged or uncharged) after considering the evidence presented by both parties. Our analytic approach accounts for the fact that judges are required to accurately determine and consider these sentencing determinations, and are influenced by them. Our approach also accounts for the prosecutorial decision to request sentences below the guidelines range, and for cases in which mandatory minimum penalty provisions limit a judge's sentencing choices. Most importantly, our work reflects the federal system as a whole because we examine all offenders and all crime types that come before the federal courts.

To be sure, the demographic differences in sentencing that we have found in our work for the Commission warrant continued examination. But any such examination must be based on the reality of the sentencing process in the federal courts. Starr and Rehavi's work falls short of this standard.

Glenn R. Schmitt is the Director of the Office of Research and Data at the United States Sentencing Commission. Louis Reedt is the Deputy Director of the Office of Research and Data. Kevin Blackwell is a Senior Research Associate in the Office of Research and Data.

Preferred citation: Glenn R. Schmitt et al., *Why Judges Matter at Sentencing: A Reply to Starr and Rehavi*, 123 YALE L.J. ONLINE 251 (2013), http://yalelawjournal.org/2013/10/23/schmitt.html.