Racial disparity in the wake of Booker/Fanfan
Making sense of “messy” results and other challenges for sentencing research

Rodney Engen
University of Arkansas

In the three decades since the first comprehensive sentencing guidelines were adopted in MN, PA, and WA scholars have produced a substantial body of research examining if, when, and how race, ethnicity, or gender still “matter” in the sentencing process (in state and federal systems). Yet it is difficult to comment on the impact of sentencing guidelines on sentencing disparity because there simply is little empirically rigorous research examining the effects of actual policy changes (i.e., the introduction, modification, repeal, etc., of sentencing laws) for sentencing practices (Engen, 2009). In this context, the study by Ulmer, Light, and Kramer (2011, this issue), and the U.S. Sentencing Commission report (USSC, 2010) to which it responds, constitute important contributions to the sentencing literature.1

Like the USSC study, Ulmer et al. (2011) examine racial/ethnic and gender disparity in sentencing under the U.S. guidelines prior to, and subsequent to, several key pieces of legislation and U.S. Supreme Court decisions affecting judicial discretion, most notably the Booker and Gall decisions. However, Ulmer et al. also employ numerous methodological “choices” (a wonderfully diplomatic phrase) different from those of the Commission’s study. Consequently, Ulmer et al.’s findings also differ in some important ways from those of the USSC study. They find that the increase in disparity between White and Black males is primarily related to imprisonment decisions, as opposed to sentence length, and is largely due to sentencing disparity in immigration cases. Ulmer et al. also find that controlling for criminal history explains a substantial proportion of the disparity found by the Commission.

Direct correspondence to Rodney Engen, Department of Sociology & Criminal Justice, University of Arkansas, 211 Old Main, Fayetteville, AR 72701 (e-mail: rengen@uark.edu).

1. See also Wooldredge, Griffin, and Rauschenberg (2005) and Wooldredge (2009) on the impact of reforms in Ohio. While conceptually similar, the guidelines in Ohio have little in common with the federal guidelines.

DOI:10.1111/j.1745-9133.2011.00773.x © 2011 American Society of Criminology

Criminology & Public Policy • Volume 10 • Issue 4

1139
and that departure sentences continue to contribute to sentencing disparity. Perhaps most importantly, they challenge the Commission’s (USSC, 2010) main conclusion that disparity in sentence length associated with race and gender increased significantly post-Booker and post-Gall, finding that disparity post-Booker/Gall is not necessarily greater than it was in some previous guideline periods. If anything, any recent changes may reflect a return to previous patterns of sentencing behavior under the guidelines.

Overall, Ulmer et al. (2011) conclude that little compelling evidence exists of the large increases in sentencing disparity that some feared might follow in the wake of these rulings. I concur. At best, the evidence is inconsistent regarding whether disparity worsened post-Booker and post-Gall, but there clearly is no evidence of an urgent need for legislation to counteract the supposedly deleterious effects of increased judicial discretion. However, just as Ulmer et al.’s reanalysis of the USSC data demonstrates the value of independent analyses employing different strategies and methodological “choices,” previous policy essays in CPP confirm the value of having more than one “set of eyes” examine and contemplate empirical findings. In this spirit, I offer a few observations regarding some of Ulmer et al.’s own methodological choices and their findings, some of which differ from the authors’ interpretations. I then turn to some larger issues beyond the scope of their study, which may prove challenging for research assessing the impact of changes in sentencing policy. In short, like most good research, the study prompts as many questions as it provides answers.

Some Additional Thoughts re Ulmer et al.

Among Ulmer et al.’s (2011) methodological “choices,” perhaps the most consequential one, with respect to policy implications, is the decision to include cases sentenced prior to the PROTECT Act. Including the previous time periods revealed that recent changes in sentencing practices—assuming they are in any way consequences of policy changes—may have more to do with undoing the PROTECT Act than with Booker/Gall. Regardless of the interpretation, however, the findings remind one of how sensitive interrupted time-series designs can be to the duration of the time series selected and the specific comparisons made. If Ulmer et al. had examined sentencing practices only in the same time periods included in the USSC’s study they might have concluded, like the USSC, that disparity increased post-Gall. Of course, strictly speaking, this might be true, but the meaning of any such increase changes when viewed in this broader historical context.

This observation also leads me to wonder what we might find if we could extend the timeline back even further, to the pre-guideline era. How would post-Gall sentencing compare with sentencing prior to the sentencing guidelines? Then again, I am not even sure this is a meaningful question. In other words, what is the appropriate comparison or benchmark for assessing the impact of recent court decisions, sentencing practices as they were twenty five years ago, ten years ago, or practices just prior to these court decisions? Neither seems an ideal reference. It may be more meaningful to ask how the exercise of discretion is related to the structure of sentencing laws in place at any given time.
Ulmer et al. (2011) also emphasize that a “substantial amount of racial–gender disparity can be attributed to immigration offenses” (p. 1098) and they suggest future research examining “the greater racial disparity among immigration cases” explicitly. I would qualify and expand on this observation in several ways. First, if sentence length disparity is greater in immigration cases, their results suggest it does not pertain uniquely to Black males, as implied in their conclusions. With the exception of “other male” and “other female,” all the race-gender effects on sentence length seem to be smaller when immigration cases are removed (see Figure 3 and Table 1 of Ulmer et al.). Second, the effect of being a Black male on the likelihood of incarceration is nominally larger when they remove immigration cases. Among the full sample, “Black male” is not significant until post-Gall, whereas among nonimmigration cases they find significant Black male effects pre-PROTECT and post-Booker as well. It seems, then, that including immigration cases suppresses a significant Black male effect in most eras and that Black males may be relatively less likely than other groups to receive a prison sentence in immigration cases. If so, then is it really Hispanic/non-Hispanic imprisonment disparity that is concentrated in immigration cases? What about drug offenses? Should we expect the effects of increased discretion to vary by offense type, generally? Future research probably should consider whether the impact of Booker and Gall differs by offense.

Making Sense of Messy Results

The consequences of policy changes are seldom, if ever, predicted accurately either by scholars, by practitioners, or by the policy makers themselves. This result is well known, at least among social scientists, since Merton described “The unanticipated consequences of purposive social action” in 1936. Ulmer et al.’s findings are no exception (see also Wooldredge, Griffin and Rauschenberg, 2005). Although they find limited evidence of the “anticipated” increase in racial disparity (namely, between Black and White men) the study is also replete with “unanticipated” results. Would anyone have anticipated an increase in disparity only between Black and White males? Why did disparity between Hispanic and non-Hispanic White males not increase (it actually decreased, in the sense that Hispanic men received slightly shorter sentences pre-Gall)? And why did disparity decrease post-Gall for Black and Hispanic women versus White men? Equally interesting to me, no significant difference was found in the likelihood of incarceration between White and Black men in any of the earlier periods; yet in these same periods, judges were less likely to sentence women of all races to prison and were more likely to sentence Hispanic men to prison. Why are they discriminating against Black men only now, if indeed that is what these findings represent? In short, no consistent evidence exists of an increase in disparity and, taken at face value, the findings suggest that disparity overall declined. What is more, this decline suggests that three of four minority groups received preferential (or more preferential) treatment before Gall.
How are we to make sense of such “messy” and unanticipated results? In short, I am not sure we can make sense of them, or that we should even accept them at face value. Rather, in the discussion that follows I suggest that some of these confusing findings (as with many results of sentencing research) may be a consequence of other methodological choices made by Ulmer et al. (2011) and by most sentencing researchers, and of data limitations plaguing most sentencing research.

The Problem with Statistical Significance

Ulmer et al. test whether unwarranted disparity increased or decreased in the wake of Booker and Gall by estimating race-gender coefficients before and after these rulings and by testing whether they differ across time periods by statistically significant margins. Unfortunately, statistical significance has never been a good measure of the substantive importance or “size” of effects in multivariate analyses, and it can be especially misleading when analyzing large data sets like those commonly used in sentencing research. The problem is not, as I have heard it described, that very large “Ns” somehow produce significant differences or effects where none exist but that large Ns allow us to measure associations with great precision, thus increasing our confidence that the slope coefficients in our models were not obtained by chance (which is not a bad problem to have). The consequence, however, is that even small associations often achieve “significance,” and we often will reject the null hypothesis of no difference in slope coefficients obtained in two or more time periods, or across groups, even when those differences are small.

For example, I refer readers to Ulmer et al.’s (2011) Appendix B, where their full models of sentence length are presented. Most coefficients in these models are statistically significant, and a majority of them differ significantly over time in at least one comparison. This includes the effect of the presumptive sentence on sentence length, which was larger post-Booker and post-Gall. Taken at face value, this suggests that judges sentenced more closely to the guidelines when they ostensibly were free to deviate (we might have predicted the opposite). I would find this unanticipated pattern interesting, but the differences are so small I am more inclined to view them as random fluctuations having little substantive importance. Whereas a 1% increase in the presumptive sentence length produced a 0.669 (or 67%) increase in sentence length pre-PROTECT, it produced a 0.689 (or 69%) increase in sentence length post-Gall \(z = -13.217; p < .001\). For an offender with a 5-year presumptive sentence, this difference amounts to a little more than 1 month. I cannot help wondering, also, how large really are the differences over time in the size of the race/ethnicity/gender coefficients? One challenge for policy analyses such as this, and for disparity research generally, is in determining when the observed disparities, or changes in disparity, are substantively meaningful (see Langan, 2001). Calculating predicted sentence lengths and probabilities of incarceration for each offender group, under different contingencies, in each time period, might give us a better sense of which of the observed changes are both “real” and substantively important.
The Problem of Departures

A final methodological choice made by Ulmer et al. (2011), and by many sentencing researchers, that may have substantive import is the decision to control for sentence departures in the models predicting sentence length and incarceration. Although this is common practice in the literature, controlling for what may be the main source of disparity under the guidelines—departures—means that, in essence, we are focusing our attention on disparity that cannot be attributed to departures. This is akin to pointing out that Black and Hispanic defendants are less likely to receive the benefit of downward departures but then limiting our analyses of disparity to cases that were denied the best breaks of all (departures). It should perhaps come as no surprise, then, that the findings are inconsistent. Although an interesting question, analyses that only examine disparity that is net of the disparity in departure do not reveal disparity in the sentencing process overall.

Making sense of any changes in sentencing or in disparity over time (and subsequent to major policy changes) also is difficult when we control for departures, especially in light of the evidence that disparity in the likelihood of receiving downward departure also has changed (see Appendix C in Ulmer et al., 2011). If the concern is that judges or prosecutors will be even more discriminatory in their use of departures post Booker/Gall, I believe we could learn more by first estimating changes in disparity in incarceration and sentence length prior to controlling for departures (i.e., changes in the “total” effects of race, ethnicity, and gender) and then testing whether controlling for departures explains the changes in disparity. Without first knowing the main or total effects of race-gender, we cannot say with much confidence whether disparity in the sentencing process has changed, remained constant, or simply moved.

Judging Judicial Discretion and Other Challenges for Sentencing Research

At this point, I would like to call attention to some larger challenges and limitations of sentencing research, generally, for assessing the effects of court rulings like Blakely, Booker, and Gall, or any other policy changes. I find it difficult to say much on the topic that is new, especially in light of the thoughtful essays in the 2007 special section of Criminology & Public Policy addressing this question (see Bushway and Piehl, 2007; Frase, 2007; Hofer, 2007; Wellford, 2007), but some ideas may bear repeating and further development. I encourage interested readers to consider those essays carefully. In my subsequent comments, I attempt to build on some of those arguments as well as two points raised by Ulmer et al. (2011), as they pertain to research testing the impact of policy changes on sentencing and disparity.

The Problem of Prosecutorial Discretion in the Sentencing Process

The general unavailability of data on charging and plea bargaining remains, in my opinion, the greatest challenge to the validity of sentencing research and perhaps especially research assessing the effects of policy changes on sentencing (Bushway and Piehl, 2007; Frase, 2007). It is well known that prosecutors and defense attorneys routinely negotiate over
charges and other “facts” with an eye toward the sentence that is likely to result (Ulmer, 2005; Ulmer et al., this issue). Acknowledging this reality, sentencing researchers, including Ulmer et al., are careful to limit their inferences to “judicial” discretion and disparity at the formal sentencing stage, and to note that important sources of disparity might be overlooked. However, the lack of data on charging and plea bargaining also threatens the validity of inferences about decision making at the sentencing stage. Blumstein et al. (1983) described this as:

[T]he problem of classifying “like cases”. . . . Cases that appear alike initially may, on closer scrutiny, differ in subtle ways . . . or in not-so-subtle ways (e.g., two cases in which the conviction offenses are the same as a result of plea negotiations may differ substantially in the actual underlying offense behavior). (p. 267)

If judges take it into account these unmeasured but “real” differences, it might result in very different sentences for offenders who appear to be “similarly situated” based on their conviction offenses and criminal histories (Wilmot and Spohn, 2004). If these unmeasured differences are related to defendants’ race, ethnicity, and sex, then estimates of sentencing disparity may be biased.

Controlling for prosecutorial decisions might be even more important in research assessing the impact of policy changes on sentencing. If we wish to make strong inferences about the consequences of policy changes we ideally must either include measures of offending behavior and/or plea bargaining that are independent of the conviction offense (Frase, 2007) or we must assume that prosecutorial practices remained relatively stable. If prosecutorial behavior did not remain stable—and Ulmer et al. report that, at least with regard to departures, it did not—then we have the same problem of being unable to compare “like with like” over time. Unfortunately, when sentencing policies change, court actors—including prosecutors—often adapt in unanticipated ways (Engen and Steen, 2000).2 Referring to Blakely and related decisions, Richard Frase (2007: 404) remarked, “the cases [may] have their most important lasting effects not on sentencing practice but on charging decisions and on the design of sentencing laws and guidelines.”

The Problem of Disparity in the Guidelines
A potentially more difficult issue raised by both Hofer (2007) and Bushway and Piehl (2007), and undoubtedly familiar to many readers, is that the sentencing guidelines themselves are

---

2. Hofer (2007), in his analysis of aggregate sentencing trends pre- and post-Booker, found both an increase in departures post-Booker and an increase in average presumptive sentences. Among the possible explanations he offered for this somewhat paradoxical pattern is that federal prosecutors may have increased the severity of charges and aggravating facts presented in response to the increase in judicial discretion. If this is correct, then cases sentenced for identical crimes pre- and post-Booker are almost certainly not alike.
not neutral. Racial disparity may be built into the guidelines, especially in the realm of drug offenses (Tonry, 1995). The crack versus powder cocaine differential is a textbook example (literally) of institutionalized bias. Similarly, because of the counting of drug offenses, Black offenders disproportionately “qualify” for the career criminal enhancement that, according to Hofer, contributes significantly to the disparity in average sentence lengths between Black and White offenders (see also Bushway and Piehl, 2007). Undoubtedly, we could find other examples where one group or another is inherently disadvantaged, and I second Bushway and Piehl’s call for research examining the structure of the guidelines in this way.

This poses several difficult challenges for sentencing research: First, how are we to evaluate the exercise of judicial discretion, or changes in sentencing disparity, relative to guidelines that many observers, including federal judges, believe are unjust? Should we be relieved or concerned if several studies find that, for the most part, judges are continuing to sentence in accordance with the guidelines even post-Booker/Gall? If the laws are indeed unfair, then the uniform application of such laws would only perpetuate injustice. Although I expect many researchers would agree with this last statement, we implicitly assume guidelines neutrality when, in Bushway and Piehl’s (2007: 479) words, “the sentencing grid defines the starting point of analysis.” The problem, then, is if we cannot use the guidelines as the benchmark against which to judge judicial behavior, what do we use? Ultimately, sentencing research will need to develop some criteria for assessing punishment disparity, or measures capturing the seriousness of offending and other “legitimate” concerns that are independent of the guidelines.

Second, equal treatment (i.e., the absence of disparity) is not the only criterion relevant to judging the impact of these recent decisions or of sentencing guidelines generally. Substantive “justice” also is important. Although this concept is much harder to define than “disparity,” evidence of widespread dissatisfaction with the federal sentencing guidelines (Frase, 2007) suggests the laws do not always reflect normative expectations for appropriate punishment. The important point for the current discussion is that the exercise of judicial discretion under the guidelines—and changes to the guidelines themselves—may increase one type of justice while diminishing the other. For instance, a number of studies suggest that the frequent use of departures in the sentencing of drug offenders may contribute to unwarranted disparity, undermining justice defined as equal treatment. At the same time, the use of departures in these cases may actually enhance substantive justice by mitigating the impact of sentencing laws that many view as excessively harsh. In principle, Booker and Gall could have similar effects. If judges use their increased discretion to counteract some of the more extreme and controversial aspects of the guidelines (e.g., mandatory minimums or the career offender enhancement), they might hand down sentences that most observers would agree are more “appropriate,” on average, than if they had followed the guidelines closely. If so, it is most likely that all races, ethnicities, and genders will benefit from this discretion. At the same time, experience shows that these groups might not benefit equally. In this hypothetical, but highly plausible, scenario wherein all groups benefit, but
some benefit more than others, would we conclude that the *Booker/Gall* rulings improved or undermined the quality of justice in U.S. courts? The answer might depend on how we define justice.

Third, we should bear in mind that the consequences of any change to sentencing laws may depend on the level of analysis we choose to examine. Sentencing research characteristically focuses on the effects of individual offender and offense characteristics. However, because Black, White, Hispanic, male, and female defendants differ in the aggregate on such factors as average offense seriousness, criminal history, and conviction offense, any changes in the importance placed on these criteria—whether discretionary in origin, legislated, or the result of court rulings—will impact groups differently in the aggregate (Bushway and Piehl, 2007; Hofer, 2007). For example, steps taken in recent years to reduce the “crack versus powder cocaine” differential have probably done more to reduce racial disproportionality in incarceration than any other policy changes adopted since the guidelines were introduced. Importantly, this effect is likely to happen irrespective of disparity at the individual level. Similarly, recent court rulings increasing judicial discretion—or future rulings and legislation restricting it once again—could impact racial disproportionality in the aggregate differently than they impact disparity at the individual level. For instance, to continue my earlier example, if judges use their newfound discretion to reduce sentences for drug offenders substantially, but still give preferential treatment to Whites, we could even observe an increase in disparity at the individual level accompanied by a reduction in racial disproportionality in the aggregate. Again, would we conclude that *Booker/Gall* have increased or decreased racial justice?

**Conclusion: “There’s Nothing So Practical as a Good Theory” (Kurt Lewin)**

Perhaps the most difficult challenge for research hoping to make sense of major policy changes is that we have little pertinent theory to guide us. Why should we expect disparity to increase when guidelines are repealed or made advisory? This question assumes, first and foremost, that the guidelines worked to minimize disparity in the first place. Several previous essays appearing in this journal have questioned this assumption explicitly (Bushway and Piehl, 2007; Engen, 2009; Kramer, 2009). The prediction also rests on an assumption that, if left to their own devices, judges will discriminate on the basis of race and gender. And yet, this often seems not to be the case. John Kramer (2009) pointed out that local norms have always regulated sentencing, and this has continued even under the guidelines. Although they might disagree with some aspects of them, court communities long ago adapted to and incorporated the guideline model into their local legal culture (Kramer, 2009; Ulmer and Kramer, 1996). Were the guidelines to be repealed entirely it is likely that the philosophy, values, and even the content of those guidelines would be carried forward as a part of local legal culture.

Finally, we know that sentencing guidelines do more than simply provide a context in which social–psychological processes take place. Guidelines structure ways in which
substantive sentencing rationales and goals are achieved (e.g., justice, rehabilitation, and community protection), they structure the plea-negotiation process by providing specific rewards and penalties, they facilitate court actors’ abilities to manage caseloads, they provide “benchmarks” for determining appropriate punishment, they provide political “cover” for what might be unpopular decisions, they constrain the exercise of discretion, and in many instances, they probably determine the sentence despite judicial preferences (Bowen, 2009; Engen and Steen, 2000; Savelsberg, 1992; Ulmer, 1997, 2000; Ulmer and Kramer, 1996).

If we are to predict what is likely to happen in the wake of court rulings like Booker and Gall, or what might happen if legislation once again limits their discretion, then we need theory that goes beyond social–psychological models of judicial decision making. We need theory that explains how and toward what ends judges and other court actors use the laws in day-to-day decision making and how the structure of sentencing laws both facilitates and limits the ability of judges and prosecutors to achieve their objectives. Reviewing some of the earlier qualitative work by Ulmer, Kramer, and their colleagues might be a good place to start.

References


**Court Cases Cited**


**Statute Cited**

Rodney Engen is Associate Professor in Sociology and Criminal Justice at the University of Arkansas. His research examines racial and gender disparities in sentencing, the effects of prosecutorial discretion under sentencing guidelines, and the effects of state sentencing policies on imprisonment. His research has appeared in journals including the *American Journal of Sociology, Criminology, Justice Quarterly*, and *Social Problems*.