Journal of Criminal Law and Criminology

Volume 90
Issue 1 *Fall*

Article 5

Fall 1999

Effect of the Federal Sentencing Guidelines on Interjudge Sentencing Disparity

Paul J. Hofer

Kevin R. Blackwell

R. Barry Ruback

0091-4169/99/9001-0239
THE JOURNAL OF CRIMINAL LAW & CRIMINOLOGY
Vol. 90, No. 1

Printed in U.S.A.

CRIMINOLOGY

THE EFFECT OF THE FEDERAL SENTENCING GUIDELINES ON INTER-JUDGE SENTENCING DISPARITY

PAUL J. HOFER, KEVIN R. BLACKWELL, & R. BARRY RUBACK*

I. INTRODUCTION

A. THE GOALS OF SENTENCING REFORM

The sentencing reforms of the past twenty-five years have had several goals, including "truth in sentencing," control of prison populations, and reduction of unwarranted disparity. The first goal was easily achieved when parole was abolished and the sentence imposed became the sentence served. Control of prison populations proved more difficult, and careful evaluation is needed to determine whether sentencing reform helped to check, or may have accelerated, the steady rise in prison in-

* Paul J. Hofer, Senior Research Associate, U.S. Sentencing Commission; Kevin Blackwell, Research Associate, U.S. Sentencing Commission; and Barry Ruback, Professor of Crime, Law, & Justice and Sociology, Pennsylvania State University. The views in this article are those of the authors and do not necessarily reflect an opinion or policy of the U.S. Sentencing Commission. We would like to thank Patricia Valentino for extensive copy editing and Yosefi Seltzer for his conscientious help in calling court clerks around the country and coding survey results. In addition, we acknowledge the assistance of staff of the Office of Policy Analysis of the U.S. Sentencing Commission and the Research Division of the Federal Judicial Center for providing access to the results of the FJC survey. Responsibility for the use made of these results or for errors of analysis rests solely with the authors.

mates and crowding of prisons.[1] Among the most important goals motivating reform at the federal level was reduction of unwarranted sentencing disparity.[2] Yet more than ten years after enactment of the Sentencing Reform Act of 1984 and implementation of the U.S. Sentencing Commission's guidelines, there is no consensus on whether they have reduced unwarranted disparity in federal sentences.

As reviewed below, the evidence is persuasive that, in the pre-guideline era, differences among judges in sentencing philosophies were the primary sources of unwarranted disparity. There are reasons to hope that the federal guidelines have been effective at controlling these differences: they are the most detailed guidelines ever developed, and they are mandatory. Judges can depart from the guidelines only for limited reasons that must be stated on the record, and that are then subject to appellate review. But these departures from the guidelines, as well as plea agreements that ignore the guidelines and inconsistencies in application of the guidelines, might reintroduce disparity into the system.

In this paper, after reviewing problems that have plagued earlier research, we conclude that the "natural experiment" created by the random assignment of cases to judges in many courthouses around the country provides the best opportunity to evaluate the success of the guidelines at reducing disparity. Although the natural experiment approach does not permit evaluation of disparity in individual cases, and does not permit direct examination of prosecutor-created disparity, it does allow us to draw some conclusions about the effect of the guidelines on inter-judge disparity. In particular, it permits precise measurement of changes in the "primary judge effect"—the overall tendency of some judges to be more lenient or severe than others, as measured by differences in average sentences among

---

[1] *See generally* MICHAEL TONRY, SENTENCING MATTERS (1996). *See, e.g.*, Thomas B. Marvell, *Sentencing Guidelines and Prison Population Growth*, 85 J. CRIM. L. & CRIMINOLOGY 696 (1995).

[2] *See* Kate Stith & Steve Y. Koh, *The Politics of Sentencing Reform: The Legislative History of the Federal Sentencing Guidelines*, 28 WAKE FOREST L. REV. 223 (1993); William W. Wilkins, Jr. et al., *The Sentencing Reform Act of 1984: A Bold Approach to Unwarranted Sentencing Disparity Problem*, 2 CRIM. L.F. 355 (1991).

judges with comparable caseloads. We believe this is an important barometer of the success of the guidelines because it represents the "tip of the iceberg" of disparity that would be observed—and is observed in simulation studies—when identical cases are sentenced by different judges.

What we find in the results of this natural experiment is that the guidelines have significantly reduced overall inter-judge disparity in sentences imposed, much as the parole guidelines reduced disparity in time actually served prior to implementation of the sentencing guidelines. Together with the other research reviewed below, these findings suggest that the sentencing guidelines have had modest but meaningful success at reducing unwarranted disparity among judges in the sentences imposed on similar crimes and offenders.

The success, however, is uneven. Some types of cases show no improvement, or show improvement in some cities but not in others. Further, there is evidence that some regional sentencing differences have increased under the guidelines, particularly in drug trafficking cases. These results demonstrate the need for further exploration of the sources of these remaining disparities, and of what additional changes are needed to make fair and uniform sentencing attainable for all types of cases in all places.

B. DEFINING AND MEASURING UNWARRANTED DISPARITY

To evaluate the guideline's success at reducing disparity we first must define the problem. A truism of sentencing research is that sentences should vary according to the seriousness of the crime and the dangerousness of the offender, but that "unwarranted disparity" is undesirable and unfair. Deciding what types of variations count as unwarranted disparity has been controversial, however. Most people readily agree that dissimilar treatment of similar offenders would be unwarranted, just as similar treatment of dissimilar offenders would improperly fail to reflect relevant distinctions among them. But this definition is purely formal, and thus incomplete, because it does not tell us

the proper criteria for classifying offenders as similar or dissimilar.[3]

Treating offenders differently based on a legally impermissible ground, such as race, results in the clearest case of unwarranted disparity—discrimination.[4] Discrimination may reflect intentional or conscious bias toward a group, or be a result of the distorting effects on rational judgment of unconscious stereotypes or fears. Either way, it is the most onerous type of unwarranted disparity and sentencing reform was clearly designed to eliminate it.[5] But discrimination is not the only type of unwarranted disparity.

To most people, different treatment based on *any* factor that is irrelevant to the rules and purposes of sentencing—so-called "extralegal" factors, such as an offender's physical appearance—is unwarranted. The most common statistical approaches used to study disparity, multiple regression and related analyses, are compatible with this definition. In this research, as many of the legally relevant factors as is practically possible are measured for a large sample of cases, along with additional factors that are not legally relevant but that may help explain the sentences imposed.[6] These explanatory factors are then statistically correlated with one or more outcome measures, typically whether the offender received probation or imprisonment and the length of any prison term imposed. The amount of variation in sentences that is correlated with the legally relevant factors is statistically removed, or "controlled for," and the remaining variation is examined.

For example, it may be that, on average, offenders receive six months more prison time for each prior conviction. The

---

[3] *See* Kevin Cole, *The Empty Idea of Sentencing Disparity*, 91 NW. U. L. REV. 1336, 1340-41 (1997).

[4] *See* 1 ALFRED BLUMSTEIN ET AL., RESEARCH ON SENTENCING: THE SEARCH FOR REFORM 8 (1983).

[5] *See* 28 U.S.C. § 994(d) (1999) ("The Commission shall assure that the guidelines and policy statements are entirely neutral as to the race, sex, national origin, creed, and socioeconomic status of offenders.").

[6] *See* John Hagan & Kristin Bumiller, *Making Sense of Sentencing: A Review and Critique of Sentencing Research*, *in* 2 RESEARCH ON SENTENCING: THE SEARCH FOR REFORM 1-54 (Alfred Blumstein et al. eds., 1983) (describing the development of increasingly sophisticated statistical techniques for studying the factors influencing sentencing).

variation in sentences correlated with prior convictions is subtracted from the outcome measure, and the *remaining* variation in prison lengths is examined to determine what *it* correlates with. This process is repeated for all of the legally relevant factors. If the remaining variation correlates with an impermissible factor, such as race, the data are consistent with discrimination in sentencing. Other types of unwarranted disparity are suggested if factors such as the region in which the case was prosecuted are correlated with sentences. Variation that is not accounted for by *any* of the factors in the model may also be taken as a measure of unwarranted disparity, although other problems affect this measure.[7]

A long line of regression studies of this kind, even before implementation of sentencing guidelines, has found that the most important influences on sentences are not impermissible factors such as race, or extra-legal factors such as region, but are the legally relevant factors that should determine sentences, particularly offense seriousness and an offender's criminal history.[8] Statistical analyses conducted at the U.S. Sentencing Commission subsequent to implementation of the guidelines reaffirmed that legally relevant factors remain the most important determinants of sentences today, explaining as much as 70% of the variation in imprisonment lengths. While such findings are reassuring, they do not obviate the need for further research and vigilance against unwarranted disparity, because most people think that any significant influence of legally ir-

---

[7] William Rhodes, *Federal Criminal Sentencing: Some Measurement Issues with Application to Pre-Guideline Sentencing Disparity*, 81 J. CRIM. L. & CRIMINOLOGY 1002, 1028 (1991). Variance that is unaccounted for by the model may reflect unwarranted disparities, but it is difficult to know whether and to what extent it represents a real problem. Such variance may reflect idiosyncratic inclusion and weighting of factors by individual judges from case-to-case, or legal or extralegal factors that were not included in the model, or measurement error, or some type of model misspecification, such as failure to account for the complex ways that factors interact. *But see* Joel Waldfogel, *Does Inter-Judge Disparity Justify Empirically Based Sentencing Guidelines?*, 18 INT'L REV. L. & ECON. 293 (1998) (exploring implications of assuming that variance unaccounted for by a particular model is nonetheless appropriate variation and contributes to proportionate sentencing).

[8] Gary Kleck, *Racial Discrimination in Criminal Sentencing: A Critical Evaluation of the Evidence with Additional Evidence on the Death Penalty*, 45 AM. SOC. REV. 783, 798-99 (1981).

relevant factors would be undesirable, even if these factors are not the primary determinants of sentences.[9]

In fact, sentencing reform seems to have been a boon to research on disparity using regression approaches, with over ten such studies appearing in recent years on just the federal guidelines alone.[10] This type of research is popular because the guidelines explicitly identify many, if not most, of the factors that are legally relevant to the sentencing decision, and because sentencing commissions have collected huge amounts of data on these factors. But for several reasons, we believe that these studies can add relatively little, if anything, to our understanding of disparity under the guidelines, and are even less helpful in evaluating whether the amount of disparity has increased or decreased.

First, there are significant methodological obstacles to using regression studies for measuring unwarranted sentencing disparity under the guidelines, and especially for comparing the amount of disparity before and after guideline implementation. Many of these obstacles are discussed in a later section. But perhaps the most significant for present purposes is the difficulty of obtaining comparable data for both time periods.

Second, theoretical disagreements may arise over whose views about a factor's legal relevance should prevail.[11] For example, whether different treatment is warranted based on of-

---

[9] We reserve for another day the issue of whether some different treatment is too minimal to be of concern. See interchange between the U.S. Sentencing Commission and the General Accounting Office in GENERAL ACCOUNTING OFFICE, SENTENCING GUIDELINES: CENTRAL QUESTIONS REMAIN UNANSWERED 13, 26, 178 (1992) [hereinafter GAO REPORT]. We also do not address whether it might be possible to equate different *forms* of punishment, such as home confinement and imprisonment, so that similar punishment might be administered in different forms.

[10] For a review of these studies and a critique of the use of regression analysis in research on discrimination, see Paul J. Hofer & Kevin Blackwell, *Searching for Discrimination in Federal Sentencing* (submitted for publication; on file with the authors). This article is an expansion and revision of a previous paper, Paul J. Hofer & Kevin R. Blackwell, *Identifying Sources of Unfairness in Federal Sentencing.* (Paper presented at the annual meeting of the American Society of Criminology, Nov. 21, 1997) (on file with authors).

[11] *See* TONRY, *supra* note 1, at 48; *see also* Anthony Doob, *The United States Sentencing Guidelines, If you Don't Know Where You Are Going, You May Not Get There, in* THE POLITICS OF SENTENCING REFORM (Chris Clarkson & Rod Morgan eds. 1995).

fenders' potential for rehabilitation depends on whether one believes rehabilitation is a legitimate purpose of sentencing. Views of rehabilitation have clearly changed over time, and disagreement over its role in sentencing still exists. We believe that because the Sentencing Commission is now legally charged with deciding what criteria should be legally relevant, it is appropriate to evaluate current and past practice in light of the Commission's policy decisions. At the same time, we recognize that what is considered disparity today has not always been considered disparity, and that some ways of evaluating past practice in light of current policies may "stack the deck" in favor of finding a reduction of disparity.[12]

In addition, even if we agree that the Sentencing Commission's views of what factors are legally relevant should prevail, no one, including the Commission, claims that it has identified every possible factor.[13] Judges are permitted to take into account factors not specified in the guidelines when deciding whether to depart, or where within the guideline range to sentence an offender. Disagreements continue to arise over what factors may be considered. For example, an offender's responsibility for the care of young children strikes many as legally

---

[12] Relying on very detailed guideline definitions of the precise weight that should be assigned to each factor may arguably "stack the deck" in favor of finding that the guidelines have been successful. The General Accounting Office conducted a pre/post comparison as part of its evaluation of the guidelines. GAO REPORT, *supra* note 9, app. I, at 40-58. To form matched groups, it used a 1987 prison-impact study conducted by the Sentencing Commission. The Commission had determined the offense levels and criminal history scores that would have applied to these pre-guideline cases if they had been sentenced under the guidelines. Pre-guideline cases that would have fallen at a particular offense level and criminal history score were compared with guideline cases having those same scores. By using this matching period, each group of pre-guideline cases contained widely varying offense behaviors, but were compared with guideline cases that fell in a single cell of the sentencing table and were subject to narrow ranges of sentences. The GAO reported that variance was significantly reduced under the guidelines for 57 of the 58 groups studied. But it is no surprise that judges did not view the seriousness of the pre-guideline cases in *exactly* the same way as the detailed guidelines, which greatly altered the weight given to the amount of drugs involved in the offense, the defendant's role and characteristics, and other factors. For a study of how the guidelines changed the influence of these factors in federal sentencing, see Barbara S. Meierhoefer, *The Role of Offense and Offender Characteristics in Federal Sentencing*, 66 SO. CAL. L. REV. 367 (1992).

[13] UNITED STATES SENTENCING COMMISSION, GUIDELINES MANUAL, ch. 1, pt. A.4, at 6 [hereinafter U.S.S.G.].

relevant, warranting a non-prison sentencing for crimes that might otherwise warrant imprisonment.[14] However, this responsibility is not clearly related to any of the purposes of sentencing that federal judges are directed by statute to consider,[15] and lenient treatment for mothers might seem unfair to a childless offender guilty of the same crime.

Even if consensus is reached on which factors are properly legally relevant, and only those factors are used in imposing sentence, unwarranted disparity might still result if different judges *weigh* the factors in different ways. Regression studies of disparity do not generally assess these inter-judge differences in weighing factors.[16] Yet the result is different treatment of similar offenders by different judges. Few would argue that it is fair for similar offenders to receive disparate sentences solely because of the judge to whom they happened to be assigned.[17]

For these and other methodological reasons discussed below, in this study we do not use multiple regression to measure disparity, nor do we assess the effects of defendant characteristics, such as gender and race. Instead, we take advantage of the natural experiment created by the random assignment of cases to judges to focus on the one extra-legal factor that research suggests was the most important source of unwarranted disparity in the pre-guideline era—philosophical differences among judges.

---

[14] *See Special Issue: Gender and Sentencing*, 8 FED. SENTENCING REP. 130 (1995) (articles discussing impact of sentences on innocent third parties, such as dependent children, with arguments and cases supporting its relevance and irrelevance).

[15] *See* 18 U.S.C. § 3553(a)(2) (1999).

[16] WILLIAM D. RICH ET AL., NATIONAL CENTER FOR STATE COURTS, SENTENCING BY MATHEMATICS? AN EVALUATION OF THE EARLY ATTEMPTS TO DEVELOP AND IMPLEMENT SENTENCING GUIDELINES (1982) (discussing limitation of studies that do not separately model individual judges); *see* Rhodes, *supra* note 7 (proposing a solution to this problem in the context of a two-limit tobit model).

[17] *See* BLUMSTEIN, supra note 4, at 76.

C. IDENTIFYING THE GREATEST SOURCES OF UNFAIRNESS AND UNWARRANTED DISPARITY

1. Characteristics of the Defendant

Because concern about discrimination was central to the debate over federal sentencing guidelines, some discussion of the research on the role and of race and gender before and after guideline implementation is needed. Data have consistently shown that the average of sentences imposed on African-Americans and on males is longer than the average of sentences imposed on whites and on females. Surprisingly, the difference in average sentences between whites and African-Americans and Hispanics has grown *greater* in the federal system since the introduction of sentencing guidelines.[18] The crucial question is whether these sentencing differences are due to some form of racial bias, or whether it results from differences among the groups in factors that *are* legally relevant and *should be* legally relevant.

The general finding from multiple regression studies of the effect of race in the pre-guideline era is that differences in sentences imposed on African-Americans and whites, once the effects of legally relevant criteria are taken into account, either disappear or are relatively small.[19] Studies of federal sentences under the guidelines similarly find little or no effect for race af-

---

[18] *See* DOUGLAS C. MCDONALD & KENNETH E. CARLSON, BUREAU OF JUST. STATS., SENTENCING IN THE FEDERAL COURTS: DOES RACE MATTER? 181 (1993).

[19] *Id.* at 21-35 (reviewing studies in federal courts prior to 1993); *but see Sentencing Guidelines: Hearings Before the Subcomm. on Criminal Justice of the Committee on the Judiciary, House of Representatives,* 100th Cong., 1st Sess. 39, 687-694, tbls. 1-8, (1987); *Id.* at 685 (testimony of Ilene Nagel, Commissioner, U.S. Sentencing Commission) (presenting preliminary analyses said to provide "strong empirical support for the Congressional proclamation that *widespread, unwarranted* sentence disparity is characteristic of the federal system") (emphasis in original). The Commission reported to Congress that sex, race, region where prosecuted, marital status, and other impermissible or irrelevant factors affected federal sentences prior to implementation of the guidelines. However, the Commission reported as statistically significant factors that did not rise to the traditional .05 confidence level. In addition, it is questionable whether the matching of offenders used in this preliminary study adequately captured offenders who were sufficiently similar in legally relevant ways. *See generally* MICHAEL TONRY, MALIGN NEGLECT: RACE, CRIME, AND PUNISHMENT IN AMERICA 69 (1995).

ter controlling for legally relevant factors. For example, an analysis published in the popular press concluded that, after controlling for offense seriousness and criminal history, the average difference between African-American and white sentences was about 10%, or about two months.[20] Closer analysis suggests that even this relatively modest finding is overstated, since part of the racial difference can be explained by legally relevant factors not included in the regression equation.[21] Gender has more consistently been shown to influence sentencing, with women generally receiving relatively lenient treatment, even after controlling for their less serious crimes and less extensive criminal histories. Under the guidelines, women are somewhat more likely than men to be sentenced in the lower part of the guideline range, but gender has much less of an effect on sentences than do the legally-relevant factors.[22]

Given the importance of eliminating any trace of race or gender discrimination in sentencing, research to measure its impact will continue. But two lessons should be learned from previous studies. First, different treatment of similar offenders by different judges—not different treatment based on race or gender—appears to have been the greatest source of unwarranted sentencing disparity in the pre-guideline era. Inter-judge disparity remains the greatest threat to the success of the guidelines today. The next section reviews in detail the re-

---

[20] *See* Laura Frank, *Tennessee's East District Among Worst*, THE TENNESSEAN, Sept. 24, 1995, at 17A. *See also* Mary Pat Flaherty & Joan Biskupic, *Rules Often Impose Toughest Penalties on Poor, Minorities*, WASH. POST, Oct. 9, 1996, at A26 (using different methodology, but also finding small effects for race of the defendant, ranging from a few extra days for short sentences to a few extra weeks for multi-year sentences).

[21] A reanalysis by research staff of the Sentencing Commission found that the race effect disappeared entirely once the effects of departures and mandatory minimum statutes were taken into account. (Commission analysis on file with the authors.) U.S. Sentencing Commission Internal Memorandum from Kevin Blackwell to Paul Martin and Commissioner Michael Galacek (Sept. 29, 1995) (on file with the author). For a general discussion of methodological problems with studies of the effect of race under the guidelines, *see* Hofer & Blackwell, *Searching for Discrimination in Federal Sentencing*, *supra* note 10.

[22] *See* Ilene H. Nagel & Barry L. Johnson, *The Role of Gender in a Structured Sentencing System: Equal Treatment, Policy Choices, and the Sentencing of Female Offenders Under the United States Sentencing Guidelines*, 85 J. CRIM. L. & CRIMINOLOGY 181, 221 (1994); Phyllis J. Newton et al., *Gender, Individuality and the Federal Sentencing Guidelines*, 8 FED. SENTENCING REP. 148, 152-53 (1995).

search demonstrating how differing judicial philosophies affect sentences.

Second, the gap between the average sentences of African-Americans and whites is *not* due to extra-legal factors, but is a result of the *legally relevant factors* themselves. The gap has grown wider in the past decade because Congress increased the punishment for the types of crimes disproportionately committed by African-Americans, Hispanics, and men, particularly drug trafficking and firearm offenses. Multiple regression studies that "control for" the type of crime committed thus miss an important point. The crucial questions about the fairness of sentencing today concern whether factors that are legally relevant, and that have a disproportionate adverse impact on African-Americans, Hispanics, and men, *should be* legally relevant. Are there sufficient policy reasons, grounded in the purposes of sentencing, to treat some types of offenses much more harshly than others, given that doing so will exacerbate sentencing differences among racial and gender groups?

Douglas McDonald and Kenneth Carlson, in the most sophisticated study to date of the effects of race on federal sentencing,[23] report that the increase in the gap between average sentences for whites and African-Americans is explained largely by the larger portions of African-American offenders sentenced for drug trafficking crimes, especially crack cocaine offenses.[24] Their study did not attempt to assess whether the 100-to-1 quantity ratio between crack and powder cocaine found in current law was justified on policy grounds.[25] The Sentencing Commission's own subsequent study determined that there was no sufficient policy basis for this ratio.[26] As of this writing, efforts to change these penalty statutes have been unsuccessful. But in any event, it is clear that the most important research and policy

---

[23] *See* MCDONALD & CARLSON, *supra* note 18.

[24] *Id.* at 13-14.

[25] *See* 21 U.S.C. § 841(b)(1)(A) (1999).

[26] U.S. SENTENCING COMMISSION, COCAINE AND FEDERAL SENTENCING POLICY 196-97 (1993). The Commission proposed to equalize treatment for similar amounts of crack and powder. *See* Amendments to the Sentencing Guidelines, 60 FED. REG. 25 (1995) (proposed May 12, 1995). However, Congress disapproved this amendment. Pub. L. No. 104-138, 109 Stat. 334 (1995).

debates about the racial and gender fairness of sentences will not concern the effects of discrimination but will be about what factors should be legally relevant to the purposes of sentencing.

2. *Characteristics of the Judge and the Court*

The popular conception of a range of judicial temperaments—from "hanging" judges to "soft" ones—has a basis in truth. Judges have different sentencing philosophies.[27] John Carroll et al. have demonstrated that a constellation of individual differences, including one's background and personality, relates to one's ideology, which in turn is reflected in how one thinks about the causes of crime and the goals of sentencing.[28] In the simplest terms, "liberals" tend to believe that factors external to the offender are responsible for criminal behavior. Rehabilitation is more of a sentencing goal for these judges, leading to greater reliance on probation and less concern with retribution. "Conservatives" believe that offenders choose to commit crimes. They are more punishment-oriented and tend to impose longer prison terms.

Several lines of evidence suggest that philosophical differences among judges accounted for a significant portion of the differences in sentences imposed on offenders in the federal courts prior to enactment of the sentencing guidelines. Shari Diamond and Hans Zeisel, discussing results obtained from a study of federal "sentencing councils" (at which panels of judges would independently review cases and set sentences, and then meet with the judge responsible for imposing the binding sentence), concluded that "it is reasonable to infer that the judges' differing sentencing philosophies are a primary cause of the disparity."[29]

---

[27] JOHN HOGARTH, SENTENCING AS A HUMAN PROCESS 22 (1971) ("Indeed, it would now be considered naive to assume that judges and magistrates can be expected to process information impartially and apply mechanically the appropriate legal principles to sentencing problems. Few sentencing judges themselves would claim that sentencing is a completely rational and mechanical process.").

[28] John S. Carroll et al., *Sentencing Goals, Casual Attributions, Ideology, and Personality*, 52 J. PERSONALITY & SOC. PSYCHOL. 107 (1987).

[29] Shari S. Diamond & Hans Zeisel, *Sentencing Councils: A Study of Sentence Disparity and Its Reduction*, 43 U. CHI. L. REV. 109, 114 (1975).

In another well-known early study—the Federal Judicial Center's Second Circuit Study—Anthony Partridge and William Eldridge found dramatic differences among judges in the sentences they imposed hypothetically on identical offenders.[30] Judges were sent actual presentence reports for twenty defendants representing a range of typical offenses, and asked what sentence they would impose. Differences of several years were common; in one case more than seventeen years separated the most severe from the least severe sentence. The data showed that a handful of judges were consistently much more severe or more lenient than their colleagues. However, in a finding with implications for the present study, the researchers noted that the majority of judges, on average, gave sentences of comparable severity to their fellow judges. But the judges had very different opinions about *which* particular cases deserved more severe or more lenient punishment. In the long run, these differences canceled out and average sentences were fairly similar. But the relative similarity in the overall averages masked a greater disparity at the individual case level.

Researchers at the Institute for Law and Social Research (INSLAW) conducted an expansion of the Federal Judicial Center (FJC) study that measured the precise contribution of philosophical differences to sentencing disparity. Brian Forst and Charles Wellford analyzed the role of sentencing goals in a sample of 264 federal judges sentencing a series of hypothetical cases.[31] They found that the judges who were oriented towards utilitarian goals (incapacitation and deterrence) gave sentences that were at least ten months longer than judges with other goals in mind. Their analysis divided inter-judge disparity into two types. The "*primary judge effect*" was defined as the general tendency for "toughness or leniency among the various judges;" as measured by differences in their overall average sentence lengths. The "*interaction effect*" was defined as disagreement

---

[30] ANTHONY PARTRIDGE & WILLIAM B. ELDRIDGE, FEDERAL JUDICIAL CENTER, THE SECOND CIRCUIT STUDY: A REPORT TO THE JUDGES OF THE SECOND CIRCUIT 36 (1974).

[31] Brian Forst & Charles Wellford, *Punishment and Sentencing: Developing Sentencing Guidelines Empirically From Principles of Punishment*, 33 RUTGERS L. REV. 799, 813 (1981).

among judges about the seriousness of particular types of cases, regardless of any overall tendency to be harsh or lenient.

In a multiple regression analysis of these data, Forst and Wellford found that 21% of the variance in prison terms was explained by the primary judge effect.[32] Like Partridge and Eldridge, however, they found that even more variation was explained by the interaction of judges with the particular characteristics of different offenses and offenders. Significant disparity arises from the general tendency of some judges to be relatively severe or lenient, but even more disparity arises from differences in opinions about the seriousness of particular offenses and the purposes of sentencing in specific types of cases.[33] In other words, the primary judge effect measured only the "tip of the iceberg" of the underlying disparity that arose in specific case types.

Other researchers have noted that judicial philosophy is likely to evolve over time and with experience.[34] Martha Myers found that prior experience as a prosecutor, as well as the judge's religion, were significantly related to the use of incarceration as opposed to probation and to sentence length.[35] Results are inconsistent on whether other general demographic judicial characteristics, such as race or social background, might also be associated with disparate sentences imposed on similar offenders. Malcolm Holmes et al. reported that the ethnicity of

---

[32] *Id.* The "variance" is the most common statistical measure of the distribution of scores on a measure, such as years of imprisonment. It is calculated by determining the mean, or average, of all the scores in a population and then calculating the distance of each individual score from the mean. The sum of these distances divided by the number of scores is the total variance. (For statistical reasons, the distances are squared before summing.) Depending on the particular research design that is employed, multiple regression or other statistical techniques discussed below may allow researchers to apportion the percentage of the total variance that is accounted for by various factors measured in the research.

[33] *See also* Kevin Clancy et al., *Sentence Decision Making: The Logic of Sentence Decisions and the Extent and Sources of Sentence Disparity*, 72 J. CRIM. L. & CRIMINOLOGY 524 (1981) (offering additional analysis of these same data confirming that judges' perceptions and philosophy explain a large portion of variance in sentences imposed).

[34] Rod A. Bond & Nigel F. Lemon, *Training, Experience, and Magistrates' Sentencing Philosophies*, 5 L. & HUM. BEHAV. 123 (1981).

[35] Martha A. Myers, *Social Background and the Sentencing Behavior of Judges*, 26 CRIMINOLOGY 649 (1988) (finding that older judges gave slightly longer sentences).

the judge was related to favoritism.[36] Cassia Spohn reported that African-American defendants are sentenced more harshly than whites, but that this was true for sentences imposed by both white and African-American judges.[37]

In addition to differences in philosophy among individual judges, several studies have found geographical differences in sentencing patterns by federal judges, suggesting that different political climates or court cultures can affect sentences. Research sponsored by the Department of Justice in the 1970s used data from six federal districts representing four geographical areas.[38] Sentences for eight offenses—bank robbery, bank embezzlement, counterfeiting, larceny from interstate commerce, auto theft, narcotics offenses, and Selective Service Act violations—were selected for analysis. The results showed that judges differed in the importance they placed on various factors depending on the region in which they sat.[39]

Studies in both federal and state systems have repeatedly found that offenders sentenced in rural areas tend to receive harsher sentences than those sentenced in urban courts.[40] Beverly Blair Cook, studying Selective Service cases in the federal courts, found that "[t]he severity . . . of the judges . . . differed according to the population of the city in which they decided the case, with the judges in the smallest cities giving the most severe sentences."[41] John Kramer and Jeffery Ulmer found that

---

[36] Malcolm D. Holmes, *Judges' Ethnicity and Minority Sentencing, Evidence Concerning Hispanics*, 74 SOC. SCI. Q. 496, 502 (1993).

[37] Cassia Spohn, *The Sentencing Decisions of Black and White Judges: Expected and Unexpected Similarities*, 24 L. & SOC. REV. 1198 (1990).

[38] L. Paul Sutton, *Federal Sentencing Patterns: A Study of Geographical Variations*, 18 U.S. DEP'T. OF JUST. 7 (1978).

[39] *See* Glen T. Broach et al., *State Political Culture and Sentence Severity in Federal District Courts*, 16 CRIMINOLOGY 373, 379-81 (1978) (finding influence of judicial philosophy and local political environment on sentencing in Selective Service cases).

[40] Thomas L. Austin, *The Influence of Court Location on Type of Criminal Sentence: The Rural-Urban Factor*, 9 J. CRIM. JUST. 305, 314 (1981). *See* John Hagan, *Criminal Justice in Rural and Urban Communities: A Study of the Bureaucratization of Justice*, 55 SOC. FORCES 597, 608-10 (1977); Carl E. Pope, *The Influence of Social and Legal Factors on Sentencing Dispositions: A Preliminary Analysis of Offender Based Transaction Statistics*, 4 J. CRIM. JUST. 203, 217 (1976).

[41] Beverly Blair Cook, *Sentencing Behavior of Federal Judges: Draft Cases—1972*, 42 U. CIN. L. REV. 597, 612 (1973).

even the promulgation of formal sentencing guidelines had not completely eliminated regional disparities in Pennsylvania.[42]

Clearly, structured sentencing systems have a major challenge in controlling disparities among judges and among regions. The U.S. Sentencing Commission sought to meet this challenge with extraordinarily detailed and mandatory sentencing rules.

D. HOW THE FEDERAL GUIDELINES SEEK TO REDUCE UNWARRANTED DISPARITY

Sentencing reform aimed to replace the discretion of individual judges with centralized decision-making by a Sentencing Commission, which would settle questions of policy in a uniform way through research, deliberation, and rule-making. In short, sentencing guidelines aimed to replace discretion with the rule of law.[43]

The federal guidelines represent perhaps the most ambitious attempt to structure sentencing decisions. Prior to November 1, 1987, federal judges could base their decisions on virtually any information they considered important. They did not have to explain what information they took into account, what weight they attached to the information, or how they combined the information to reach a decision. Moreover, sentences were not reviewable as long as they were within the broad statutory limits.[44]

In contrast, under the guidelines, the information relevant to sentencing is dictated by detailed rules. These rules concern

---

[42] John Kramer & Jeffery Ulmer, *Court Communities Under Sentencing Guidelines: Dilemmas of Formal Rationality and Sentencing Disparity*, 34 CRIMINOLOGY 383 (1996).

[43] M.E. FRANKEL, CRIMINAL SENTENCES: LAW WITHOUT ORDER (1972). Others, of course, have a less charitable way of defining this mission. *See* KATE STITH & JOSE A. CABRANES, FEAR OF JUDGING: SENTENCING GUIDELINES IN THE FEDERAL COURT 7 (1998) (describing the guidelines' "neoclassical preoccupation with artificial order [which] may seem anachronistic in what many intellectuals insist is our 'post-modern' or 'post-enlightenment' age. . . . [T]he Guidelines represent the continuing triumph of the administrative state.").

[44] For general introductions to the federal sentencing guidelines by several original members of the Commission and key staff, see Stephen Breyer, *The Federal Sentencing Guidelines and the Key Compromises Upon Which They Rest*, 17 HOFSTRA L. REV. 1 (1988); Ilene H. Nagel, *Structuring Sentencing Discretion: The New Federal Sentencing Guidelines* 80 J. CRIM. L. & CRIMINOLOGY 883 (1990); Wilkins et al., *supra* note 2.

the scope of conduct by defendants and their accomplices that is to be taken into account for sentencing purposes; the harms caused by the offense, particularly the amount of drugs or money involved and the degree of victim injury; the offender's post-indictment behavior; the offender's criminal history; and a host of other factors. Points are added and subtracted to obtain an offense level and a criminal history score, which locates the offender in one of 258 cells on a Sentencing Table. Each cell provides a presumptive imprisonment range, which by statute can be no wider than six months or 25% of the minimum of the range, whichever is larger.

Unlike "advisory" guideline systems that have been implemented in some jurisdictions, the federal guidelines are mandatory; judges are not free to ignore them. Departures from the guideline range are allowed in certain circumstances, however, which is why some commentators call federal-type guidelines "presumptive."[45] But to impose a sentence outside the guideline range, the judge must articulate reasons on the record identifying "an aggravating or mitigating circumstance of a kind, or to a degree, not adequately taken into consideration by the Sentencing Commission . . . that should result in a sentence different from that described."[46] All departures are then subject to appellate review. Upward departures may be appealed by the de-

[45] The sentencing guidelines differ in this way from mandatory minimum penalty statutes, which provide no general departure mechanism for cases that meet the statutory requirements (typically a certain amount of drugs trafficked) even if a case is in other respects unusual. For this reason, judicial opposition to mandatory minimum statutes is even more vehement than to presumptive guidelines. MOLLY T. JOHNSON & SCOTT A. GILBERT, FEDERAL JUDICIAL CENTER, THE U.S. SENTENCING GUIDELINES: RESULTS OF THE FEDERAL JUDICIAL CENTER'S 1996 SURVEY (1997). Certain waivers of the mandatory minimum statutory penalties are allowed in the federal system for persons who provide substantial assistance with the prosecution of other persons, *see infra* note 53 and accompanying text, and for certain low-level, non-violent, first-time offenders. *See* 18 U.S.C. § 3553(e)-(f) (1985 & Supp. 1999); *see also* U.S.S.G. §§ 5K1.1, at 323-35, §5C1.2, at 323-25, 356 (1998). Guideline departures or offense level reductions are permitted for these same groups.

[46] 18 U.S.C. § 3553(b) (1999). See *Koon v. United States*, 518 U.S. 81 (1996), for a recent description of how judges are to determine whether to depart from the guidelines, and how the appellate courts are to review such departures. *See also* Paul J. Hofer et al., *Departure Rates and Reasons after* Koon v. U.S., 9 FED. SENTENCING REP. 284 (1997); Paul J. Hofer, *Discretion to Depart after* Koon v. U.S., 9 FED. SENTENCING REP. 8 (1996).

fense; downward departures may be appealed by the prosecution. Any party can appeal a sentence if it involves an incorrect application of the sentencing guidelines.[47]

1. Potential Problems

With detailed rules such as these, one might expect that unwarranted disparity would be dramatically reduced, if not eliminated altogether. But potential problems with guideline systems have been recognized from the beginning. Some of these might be found in any guideline system and some are unique to the federal system.

Judges could simply fail to comply with the guidelines. Although judges often resist the introduction of mandatory guidelines, data from several states and the federal system show that guidelines have changed sentencing practices. With important caveats discussed below, judges generally comply with them.[48] The Sentencing Reform Act and the Sentencing Commission anticipated that the bulk of cases would be sentenced within the applicable guideline range.[49] The Commission recognized that the guidelines could not anticipate every circumstance and encouraged judges to depart in some situations. But the Commission did not expect departures to be very frequent, since the guidelines were based largely on an empirical analysis of the factors that judges had typically held to be relevant to sentencing.

Data confirm that departures based on aggravating or mitigating circumstances have been just a small portion of sentences under the guidelines. Although there has been slow growth in the rate of downward departures, this has been offset to some extent by a reduction in the rate of upward departures. Departures based on circumstances not taken into account by the guidelines have generally represented less than 10% of all sen-

---

[47] 18 U.S.C. § 3742(a)(2) (1999).

[48] *See* TONRY, *supra* note 1, at 32-39; U.S. SENTENCING COMMISSION, Vols. 1 & 2, THE FEDERAL SENTENCING GUIDELINES: A REPORT ON THE OPERATION OF THE GUIDELINES SYSTEM AND SHORT-TERM IMPACTS ON DISPARITY IN SENTENCING, USE OF INCARCERATION, AND PROSECUTORIAL DISCRETION AND PLEA BARGAINING (1991) [hereinafter FOUR-YEAR EVALUATION].

[49] U.S.S.G., ch.1, pt. A.4(b), at 6, § 5K2.0, at 357-59.

tences in any given year.[50] Even the 1995 Supreme Court decision in *Koon v. United States*, which was interpreted by some as a signal of a more permissive departure standard, has not significantly increased the departure rate.[51] This level of compliance, along with judicial review of departures to ensure that they truly represent unusual cases, should be sufficient to "cure wide disparity."[52]

However, two other kinds of "departures"—one legally recognized and one not—could reintroduce disparity. First, the federal guidelines authorize departure for defendants who offer "substantial assistance to authorities" in the prosecution of other persons.[53] This departure is meant to provide an incentive for persons to provide evidence against other offenders or to help law enforcement in various ways. Motions for departure on this basis must be made by the prosecution. If the motion is granted, the judge has discretion to sentence outside the guidelines and in some cases, depending on the precise motion, outside the otherwise applicable statutory mandatory minimum.[54]

The portion of cases receiving these departures has increased over 400% in recent years.[55] Because Congress and the Commission decided that this assistance is a legally relevant factor, sentence reductions based on it are not unwarranted under our definition and do not necessarily create disparity. However, some research suggests that these departures are used as a means for avoiding the guideline sentence in cases where the judge and the prosecutor believe the applicable sentence is in-

---

[50] U.S. SENTENCING COMMISSION, SOURCEBOOK OF FEDERAL SENTENCING STATISTICS 39, fig. G (1996) (showing percentage of sentences within guideline range 1989 to 1996).

[51] *See* Hofer et al., *Departure Rates*, *supra* note 46, at 284-91. *See also* Hofer, *Discretion to Depart*, *supra* note 46, at 8-13 (arguing that the "heartland" analysis by which judges are to decide whether departure is appropriate is subject to varying interpretation and inconsistent application and could lead to disparity if used extensively).

[52] U.S.S.G., ch.1, pt. A.5, at 11.

[53] U.S.S.G., § 5K1.1, at 356.

[54] *See* United States v. Melendez, 518 U.S. 120, 128-30 (1996).

[55] *See* U.S. SENTENCING COMMISSION, *supra* note 50, at 39 (showing increase from 3.5% in 1989 to 19.7% in 1995).

appropriate, even if no real assistance is provided.[56] Research has also shown that prosecutors in different cities have different policies about when a motion for departure is justified, and that the policies that exist are not followed consistently.[57] Further, even though the guideline offers some guidance on the factors to be considered by judges when determining the extent of the reduction,[58] different judges may weigh these factors differently, creating a form of inter-judge disparity.

A second type of departure that could reintroduce unwarranted disparity is covert guideline circumvention, or what some observers have called "hidden departures."[59] Some judges may relax strict adherence to the actual facts of the case, and instead adopt a result-oriented approach that begins with the sentence they wish to impose and works backward to identify the facts leading to that result. It may not even be necessary to ignore facts. Many findings required by the guidelines are sufficiently subjective to afford significant discretion to those who wish to use it. The resulting sentence is not a true departure because there is nothing unusual about the case that merits a sentence outside the guideline range. Instead, the judge simply disagrees with the choices of Congress and the Commission and imposes

---

[56] Ilene H. Nagel & Stephen J. Schulhofer, *A Tale of Three Cities: An Empirical Study of Charging and Bargaining Practices Under the Federal Sentencing Guidelines*, 66 S. CAL. L. REV. 501, 550 (1992) ("The section 5K1.1 motion is also used to avoid guideline ranges or mandatory minimum sentences for sympathetic defendants—even when there has been no genuine substantial assistance.").

[57] Linda Drazga Maxfield & John H. Kramer, Substantial Assistance: An Empirical Yardstick Gauging Equity in Current Federal Policy and Practice (1997) (unpublished manuscript, U.S. Sentencing Commission, available on the Commission's website <http://www.ussc.gov/research.htm> (visited Feb. 24, 2000) and on file with authors).

[58] U.S.S.G., §5K1.1(a), at 356. Some circuits have rules for determining the amount of reduction that is appropriate, but the relevant statutes and guidelines prescribe no set amount of reduction for various forms of assistance.

[59] *See* HON. JOSEPH F. WEIS ET AL., REPORT OF THE FEDERAL COURTS STUDY COMMITTEE (1990) ("[T]he rigidity of the guidelines is causing a massive, though unintended, transfer of discretion and authority from the court to the prosecutor . . . . Some prosecutors (and some defense counsel) have evaded and manipulated the guidelines . . . . Some district judges report feeling enormous pressure to accept pleas even though they clearly do not comport with the guidelines."); *see also* Daniel J. Freed, *Federal Sentencing in the Wake of Guidelines: Unacceptable Limits on the Discretion of Sentencers*, 101 YALE L.J. 1681 (1992); Stephen J. Schulhofer, *Assessing the Federal Sentencing Process: The Problem is Uniformity, Not Disparity*, 29 AM. CRIM. L. REV. 833 (1992).

a different sentence than the one called for by the statutes and guidelines.

Another problem that could reintroduce disparity is ambiguity and complexity in the guidelines. Judges wishing to apply the guidelines literally may nonetheless be frustrated by the complexity of some of the guideline provisions. The relevant conduct guideline, on which all other guideline calculations rest, is notoriously complicated and subject to differing interpretations.[60] Application of the role in the offense guideline depends on drawing distinctions among highly ambiguous phrases such as "minor" and "minimal" roles.

Finally, plea bargaining has long been recognized as a threat to the effectiveness of the guidelines at reducing unwarranted disparity.[61] Because sentences in a guideline system are tied directly to the charges of conviction and the facts of the case, they are both more predictable and potentially more subject to control by prosecutors. Prosecutors and defense attorneys may engage in "fact bargaining"—stipulating to facts that do not accurately reflect the findings that could be established at the sentencing hearing.[62] Federal prosecutors can often cap a defendant's exposure to punishment by "charge bargaining."

---

[60] Pamela B. Lawrence & Paul J. Hofer, *An Empirical Study of the Application of the Relevant Conduct Guideline § 1B1.3*, 4 FED. SENTENCING REP. 330 (1992). A description of a four-defendant drug conspiracy was sent to a random sample of 46 probation officers around the country. Different defendants were involved with different amounts of drugs, and the offense level applicable to each defendant depended on how the probation officer interpreted the relevant conduct guideline. Considerable variation was found, particularly in the sentencing of the lowest-level defendant, with sentences ranging from one to five years imprisonment. The relevant conduct guideline was amended to clarify its application later in 1992, so it is unclear to what extent the disparities found in this study would continue today. *See* Paul J. Hofer, *Implications of the Relevant Conduct Study For the Revised Guideline*, 4 FED. SENTENCING REP. 334 (1992).

[61] STEPHEN J. SCHULHOFER, PROSECUTORIAL DISCRETION AND FEDERAL SENTENCING REFORM: REPORT TO THE FEDERAL JUDICIAL CENTER (1979); *but see* Terance D. Miethe, *Charging and Plea Bargaining Practices Under Determinate Sentencing: An Investigation of the Hydraulic Displacement of Discretion*, 78 J. CRIM. L. & CRIMINOLOGY 155 (1987) (arguing that displacement of discretion from judges to prosecutors is not inevitable); *see also* Albert W. Alschuler, *Sentencing Reform and Prosecutorial Power: A Critique of Recent Proposals for "Fixed" and "Presumptive" Sentencing*, 126 U. PA. L. REV. 550 (1978).

[62] *Special Issue: Assessing the Probation Officers' Survey: Does Fact Bargaining Undermine the Sentencing Guidelines?*, 8 FED. SENTENCING REP. 299 (1996) (reporting survey results that plea agreements often do not contain the full offense conduct).

In some cases the relevant conduct guideline will bring the uncharged conduct into consideration at sentencing. But in no case can a sentence exceed the statutory maximum for the offense of conviction. Prosecutors also control whether certain statutory sentencing enhancements will be applied, for example, for a defendant's use of a firearm or for a prior record of similar crimes.

Unless the government exercises its discretion to bring and press charges and prove facts in a similar way in similar cases, unwarranted sentencing disparity can easily result. Indeed, in a detailed presumptive guideline system, the role of the judge and the parole board are diminished or eliminated altogether, so disparity arising from prosecutorial discretion will go largely unchecked by later decisions. While some have argued that reducing disparity due to plea bargaining is not a concern of *sentencing* guidelines, which were intended to curtail only judicial discretion, it is hard to see the advantages of a system that replaces inter-judge with inter-prosecutor disparity. Prosecutors generally lack the experience of judges and have many considerations acting upon their decisions other than achieving the goal of uniform sentences.[63]

Due to these concerns, Congress, the Commission, and the Department of Justice took steps to ensure that plea bargaining would not undermine the guidelines. Congress authorized the Commission to promulgate policy statements concerning judicial review of plea agreements.[64] The Commission did so, directing judges to review plea agreements before accepting them to ensure that they would not undermine the sentencing guidelines.[65] Further, the Commission made the reduction given defendants as a reward for pleading guilty a standardized and explicit part of the guideline structure in the "acceptance of responsibility" guideline.[66] A partly "real offense" system was cre-

---

[63] *See* Jeffrey Standen, *Plea Bargaining in the Shadow of the Guidelines*, 81 CAL. L. REV. 1471 (1993) (reviewing the dangers of prosecutors' "monopsonist" control of discretion under the federal guidelines).

[64] 28 U.S.C. § 994(a)(2)(E) (1988).

[65] U.S.S.G., SENTENCING GUIDELINES AND PLEA AGREEMENTS, ch. 6, at 365-72.

[66] U.S.S.G., ch. 3, part E.1.1, at 285-87.

ated through the relevant conduct guideline so that sentences would depend less on the charges of conviction and more on the actual facts of the case as found by the judge.[67] Finally, the Department of Justice issued policies to discourage undercharging and other practices that could result in some offenders receiving sentences shorter than the guidelines applicable to their full offense conduct.[68]

The success of these efforts, however, is unclear. The most comprehensive empirical study of plea bargaining under the guidelines concluded that, while adherence to the guidelines is the predominant pattern, circumvention of the guidelines occurs in 20-35% of cases, especially in drug and weapon cases in which the guideline sentence is tied to mandatory minimum statutes instead of to data on past sentencing practices. The authors concluded that prosecutorial discretion "if unchecked, has the potential to recreate the very disparities that the Sentencing Reform Act was intended to alleviate."[69] The Commission's own *Four-Year Evaluation* found that 17% of all guilty plea cases indicate some form of plea impact, such as oral or written plea agreements that dismiss charges or stipulations that underrepresent the seriousness of the offense.[70] A recent survey of district court judges revealed that 75% believe that prosecutors have the greatest influence on the final guideline sentence relative to the judge, defense attorney, and probation officer, and 73% believe that plea bargains are a source of hidden dispar-

---

[67] For a discussion of the guidelines' principle of "Relevant Conduct," see William W. Wilkins & John Steer, *Relevant Conduct: The Cornerstone of the Federal Sentencing Guidelines*, 41 S.C. L. REV. 495 (1990).

[68] *Special Issue: Justice Department Guidance for Prosecutors: Fifteen Years of Charging and Plea Policies*, 6 FED. SENTENCING REP. 298 (1994).

[69] *See* Stephen J. Schulhofer & Ilene H. Nagel, *Plea Negotiations Under the Federal Sentencing Guidelines: Guideline Circumvention and Its Dynamics in the Post-*Mistretta *Era*, 91 NW. L. REV. 1284, 1284 (1997); *see also* Ilene H. Nagel & Stephen J. Schulhofer, *Negotiated Pleas Under the Federal Sentencing Guidelines: The First Fifteen Months*, 27 AM. CRIM. L. R. 231, 232-88 (1989); Nagel & Schulhofer, *supra* note 56.

[70] FOUR-YEAR EVALUATION, *supra* note 48, at 412. In 14% of all guilty plea cases, the plea agreement resulted in a sentence below the minimum of the original guideline range, resulting in an average reduction of 40 months.

ity.[71] Probation officers report that "fact bargaining" is undermining the sentencing guidelines.[72]

While the evidence is insufficient to reach firm conclusions about how frequently the guidelines fail to structure discretion due to misuse of departures, inconsistent interpretation of complex or ambiguous provisions, or plea bargains that underrepresent the seriousness of the offender's conduct, there is reason to believe that it may be enough to introduce significant disparity into the system. Clearly, the hypothesis that disparity remains, or has even increased, under the guidelines cannot be dismissed out of hand. Empirical study is needed to measure the effectiveness of the guidelines.

II. PREVIOUS EVALUATIONS OF UNWARRANTED DISPARITY IN THE FEDERAL SYSTEM

Students of sentencing reform have recognized the need for more and better research to evaluate how well these reforms have reduced unwarranted disparity. As noted by a recent panel of experts, "[t]he past 20 years have produced many accusations but few studies documenting the misuse of discretion by judges, parole boards, and corrections officials, resulting in unwarranted disparity."[73] The panel concluded that "more research is needed to assess whether guidelines and other forms of structured sentencing are reducing sentencing disparity."[74]

In a recent review of research on the success of sentencing guidelines,[75] Michael Tonry noted that every commission has claimed success, but conceptual and methodological difficulties with the evaluation studies make it difficult to say with certainty whether and how much disparity has been reduced. Evaluations of state systems have been few and independent evaluations al-

---

[71] JOHNSON & GILBERT, *supra* note 45, at 6-11.

[72] Douglas A. Berman, *Editor's Observation: Is Fact Bargaining Undermining the Sentencing Guidelines? Probative Officer's Survey*, 8 FED. SENTENCING REP. 300 (1996) (reporting results of a probation officer survey and commenting on the findings that fact bargaining is reintroducing sentencing disparity).

[73] BUREAU OF JUSTICE ASSISTANCE, NATIONAL ASSESSMENT OF STRUCTURED SENTENCING 5 (1996).

[74] *Id.* at xvii.

[75] *See* TONRY, *supra* note 1, at 40.

most nonexistent.[76] The available data suggest that disparity has likely been reduced in these jurisdictions from what it would have been without the guidelines, but the results are far from definitive.

Because of the greater visibility of the federal system and the completeness of the data base, the federal system has been evaluated more thoroughly than any other, both by the U.S. Sentencing Commission itself and by outside researchers. However, this greater scrutiny has not increased the consensus about whether disparity has been reduced. Indeed, opinions vary from those who believe disparity has been reduced,[77] to those who cannot tell whether there has been significant change,[78] to those who think disparity has actually gotten worse under the guidelines.[79] Divergent points of view are common in the arena of sentencing policy. But such a range of opinion about an important *empirical* matter indicates a failure of research to provide objective, quantified answers to these essentially factual questions. This failure begs for explanation.

A. METHODOLOGICAL PROBLEMS IN DISPARITY RESEARCH

Disparity research has involved a remarkable variety of methods, each of which has strengths and weaknesses. Because no one method can provide definitive answers, conclusions about the existence of disparity and the effectiveness of sentencing guidelines will necessarily involve "triangulation" between

---

[76] The exception is Minnesota, where a thoughtful independent evaluation showed that the guidelines did have some success at reducing sentencing disparity, though some of the initial success eroded with time. In addition, disparity arising from plea bargaining remained. *See* DALE G. PARENT, STRUCTURING CRIMINAL SENTENCES: THE EVOLUTION OF MINNESOTA'S SENTENCING GUIDELINE (1988); Terance D. Miethe & Charles A. Moore, *Socioeconomic Disparities Under Determinate Sentencing Systems: A Comparison of Pre- and Post-Guideline Practices in Minnesota*, 23 CRIMINOLOGY 337 (1985); Lisa Stolzenberg & Steward J. D'Alessio, *Sentencing and Unwarranted Disparity: An Empirical Assessment of the Long-Term Impact of Sentencing Guidelines in Minnesota*, 32 CRIMINOLOGY 301 (1994).

[77] FOUR-YEAR EVALUATION, *supra* note 48, at 13; Theresa Walker Karle & Thomas Sager, *Are the Federal Sentencing Guidelines Meeting Congressional Goals?: An Empirical and Case Law Analysis*, 40 EMORY L. J. 393 (1991).

[78] GAO REPORT, *supra* note 9.

[79] Gerald W. Heaney, *The Reality of Guidelines Sentencing: No End to Disparity*, 28 AM. CRIM. L. REV. 161 (1991).

different methods and "generalization" from the limited samples in the studies to the larger question of disparity in the federal system. The first step toward reaching reasonable conclusions about the success of the guidelines is to understand the different methods and the questions that can be answered by each.[80]

1. Simulations Versus Actual Cases

A great deal of important research on sentencing disparity has asked judges to impose sentences in simulated cases. The Federal Judicial Center's Second Circuit study and the INSLAW study, both discussed earlier, were perhaps the most important evidence establishing the existence of disparity in the pre-guidelines era.[81] The primary advantage of this approach is clear: because *different* decision-makers are asked to sentence *identical* cases, any differences observed in their sentences can readily be attributed to differences among the judges. The disadvantage is that hypothetical situations may be so different from actual sentencing that the results cannot be generalized to the real world. In a real sentencing, the judge receives a lengthy presentence report, conducts a hearing, and observes a live defendant. Hypothetical cases may contain relatively impoverished information. Simulation studies use a limited sample of cases, which may be selected to illustrate differences among judges instead of accurately reflecting the overall caseload. Thus, simulations can exaggerate the amount of inter-judge disparity among routine cases.

On balance, however, simulation studies can provide powerful evidence of the existence of inter-judge differences and the

---

[80] We do not include in our review interview results or opinion surveys asking judges and others whether disparity has increased or decreased under the guidelines. *See, e.g.*, FOUR-YEAR EVALUATION, *supra* note 48, ch. 3, at 31-268 (reporting results of interviews and surveys of key participants in the sentencing process). While such studies are appropriate to gauge participant perceptions, they are an unsound basis for drawing objective conclusions because they may be influenced by the respondent's limited sample of experience or by their expectations.

[81] For a short review of the studies relied upon in the legislative history of the Sentencing Reform Act see *Exhibits on the 25 percent Rule: Exhibit A, Summary of Pre-Guidelines Sentencing Disparity Discussed in the Legislative History of the Sentencing Reform Act*, 8 FED. SENTENCING REP. 189 (1995).

personality and philosophy factors that contribute to disparity. But they do not provide good estimates of the amount of such disparity in the real world. Nor do they seem likely to permit clear evaluation of success of the guidelines at reducing such disparity.[82] For this we need data collected on actual sentences.

2. *Controlled and Quasi-Experiments*

When using real-world data, the task of isolating the effect of judges and the effect of the guidelines becomes much more complicated. A controlled experiment is an ideal method to isolate the effect of one factor. If the guidelines had been implemented in a randomly-selected half of the states, the other half could have served as a "control" group. Both groups would have been subject to similar historical trends occurring at the same time as guideline implementation, such as new mandatory minimums, "get tough" prosecution policies, changes in law enforcement budgets, and changes in the ideological composition of the bench. Comparing the guideline states with the control states would have permited us to focus on the effects due solely to the guidelines. But such an experimental design was not used and ethically could not be used, so the problem of identifying which changes were caused specifically by the guidelines is more difficult.

Social scientists have developed other methods—called quasi-experimental designs—for isolating the effects of reform measures when controlled experiments are not possible.[83] Time

---

[82] *See* Milton Heumann, *Empirical Questions and Data Sources: Guideline and Sentencing Research in the Federal System*, 6 FED. SENTENCING REP. 15 (1993) (calling for use of simulations to study disparity in the guideline system). Replication of the FJC's Second Circuit Study is the obvious choice for using simulation to evaluate whether the guidelines have reduced disparity, but we are skeptical that meaningful pre/post-implementation comparisons are possible. The cases used in the FJC cases may not contain sufficient information for guideline application. Changing the cases could render the pre/post-comparison problematic. Even if the cases include sufficient information for guideline application, they might lack information that could justify a departure. The role of plea bargaining in creating disparity would also go unmeasured in such a study.

[83] *See* THOMAS D. COOK & DONALD T. CAMPBELL, QUASI-EXPERIMENTATION (1979) (describing non-experimental approaches to inferring causality when strictly experimental research designs are not possible).

series analysis is one such method that involves collecting data over a long period and projecting what changes would be expected in some outcome measure, for example, the rate of guilty pleas, if the same linear, cyclical, or other trends continue after an intervention as were observed before. Actual changes are compared to the projections, and any differences found are attributed to the intervention. Time series analysis is not feasible, however, if insufficient data are available to establish clear trends or if it cannot be assumed that the trends would continue absent the intervention. Further, while time series analysis may be appropriate for studying the rate of guilty pleas or changes in average sentence length, it is not easily applied to studying changes in the amount of *disparity* in sentence lengths among different judges.[84]

A simpler quasi-experimental design than a time series analysis is a pre/post comparison. The amount of disparity at one time before guideline implementation is compared to the amount at a time afterward. This method cannot detect trends that were occurring before implementation, and the results may mistakenly be taken to suggest that all differences between the two times are due to the guidelines. But pre/post comparisons can demonstrate whether any change has occurred, and can be supplemented with other methods to help identify the event between the two points that might have caused the change. The analysis presented in later sections is a pre/post comparison design.

Strict isolation of the specific effects of the guidelines is not always necessary. Some changes, though technically not part of the guidelines, are intrinsically bound up with guideline implementation. For example, the guidelines affected plea bargaining by creating new forms of bargains, by altering the incentives bearing on the parties, and by making the plea

[84] For an example of a time series analysis of the effect of the guidelines, see FOUR-YEAR EVALUATION, *supra* note 48, at 394-410, which found that neither the Anti-Drug Abuse Act of 1988, the implementation of the guidelines, the *Mistretta* decision, nor the "Thornburgh Memo" had any apparent effect on major indices of case processing in the federal courts. The Anti-Drug Abuse Act of 1986 appeared to be associated with increased case filings, resolutions, and guilty pleas, but the effect was confounded with data changes occurring at the same time.

agreement more determinative of the final sentence. These changes naturally accompany implementation of a detailed presumptive guideline system, and these effects can fairly be considered part of the change to a guideline system of the type found in the federal courts.

Other changes occurred at the same time as the guidelines but were not a necessary part of the switch to a guideline system. Foremost among these was the rise of retribution—and the decline of rehabilitation—as the primary goal of sentencing. This trend was reflected in the "war on drugs," in mandatory minimum statutes and other legislation, and to some extent in changes in the composition of the federal bench. Some simple strategies, for example studying only judges who sentenced in both time periods, allow for partial isolation of these effects. But disentangling implementation of the guidelines from all other contemporaneous developments is impossible.[85]

---

[85] William W. Schwarzer, *Sentencing Guidelines and Mandatory Minimums: Mixing Apples and Oranges*, 66 SO. CAL. L. REV. 405 (1992) (arguing that the effects cannot be disentangled). *But see*, A. Abigail Payne, *Does Inter Judge Disparity Really Matter? An Analysis of the Effects of Sentencing Reforms in Three Federal District Courts*, 17 INT. REV. L. & ECON. 337 (1997). Payne attempted to isolate the effects of the guidelines from the effects of the mandatory minimum statutes by separately analyzing embezzlement, fraud, and theft cases (EFT cases) and drug trafficking cases, on the assumption that the former were controlled entirely by the guidelines while the latter were partly controlled by the statutes. *Id.* at 343. In some drug cases, a mandatory minimum statute overrides, or "trumps," the otherwise applicable guideline and directly controls the minimum prison term. But even in cases where the guideline range controls, the statutes "affect" the sentence, because the guidelines were established based on an extrapolation of the quantity thresholds and ratios among drugs found in the statutes. In some limited contexts it may be sensible to separate the effects of the guidelines from those of the statutes. *See, e.g.*, U.S. SENTENCING COMMISSION, SUPPLEMENTARY REPORT ON THE INITIAL SENTENCING GUIDELINES AND POLICY STATEMENTS 70 fig. 2, 72 fig. 3 (1987) (showing increase in prison time served and growth in prison populations for 15-year period, and separating the portion of growth attributable to various statutes and to the guidelines); MCDONALD & CARLSON, *supra* note 18, at tbl. 11.8 (analyzing average sentences for crack and powder cocaine under alternative legal rules, including a rule in which sentences are based only on the drug quantity levels mandated by the statutes without additional increases required by the guidelines). In the context of disparity research, however, isolating the effects of the guidelines from the statutes may not be empirically possible or conceptually sensible. Further, because under our definition the guidelines determine the full list of factors that are legally relevant, the biggest problem with sentences controlled by the statutes is that they involve unwarranted *uniformity*, not disparity.

3. The Problem of Case Comparability Between the Time Periods

If we measured disparity one way before guideline implementation and another way afterward, any differences between the two time periods could arise from our use of a different "ruler"—not from the guidelines changing the amount of disparity. One of the advantages of the "natural experiment" methodology used in our analysis is that it consistently measures the contribution of inter-judge differences to sentence variability in the total caseload in the pre- and post-implementation periods.

Matched-group comparisons have often been used to address this comparability problem. Similar cases in the two time periods are identified, using the best data available. The amount of variation in sentences among the matched groups is calculated, and statistically significant differences are taken to reflect the effect of the guidelines, since the cases are assumed to be essentially identical. But this approach has several limitations. Matching procedures require difficult decisions about how many factors, and which ones, should be used for the matching. Using too few factors results in "matched" cases that are not actually comparable. Using too many greatly reduces the number of cases that are available for the analysis, since only a few cases match on all the factors. Results are then based only on a subset of cases that may not be representative of the entire population.

Finally, the data available to create the matched groups may be inadequate to form truly comparable groups. In the federal courts, the switch to the guidelines was accompanied by the creation of a new data collection system. These data files vary somewhat in the variables collected and in their definitions, and attempting to identify similar cases using these different datasets is extremely difficult.[86] Even if the data collection system remains constant, evaluators must be sensitive to possible changes

[86] *See infra* Part III.A (discussing the Federal Probation Sentencing and Supervision Information System (FPSSIS) and the Monitoring datasets). The FPSSIS data are plagued by reliability problems. For the disparity study in its FOUR-YEAR EVALUATION, the Sentencing Commission needed to recode case files and develop a specialized dataset.

in the meaning of the data. For example, if plea bargaining is affected by the guidelines, offenders convicted of drug *possession* may actually contain larger numbers of drug *traffickers* after guideline implementation than before. Greater sentencing disparity among "possessors" *after* the guidelines are implemented would not be due to the guidelines, but to a wider range of offense types resulting in conviction for "possession."[87]

Three early evaluations of disparity reduction under the federal guidelines illustrate the problems involved in creating comparable groups. Theresa Walker Karle and Thomas Sager compared cases sentenced between November 1, 1985, and October 31, 1987, to cases sentenced after November 1, 1987, in three districts in the Fifth Circuit.[88] They found a statistically significant reduction in variation in sentences imposed for ten out of thirteen offense types, but they matched cases only on offense type, so other differences in the seriousness of the offenses or offenders between the two time periods were uncontrolled. For example, the amount of drugs or criminal history of defendants at the two time periods may have differed. In addition, the "guidelines" group included nonguidelines cases, since the cases were divided based on sentencing date, not offense date.

Judge Gerald Heaney studied guideline and non-guideline cases sentenced in 1989 in four districts in the Eighth Circuit. He carefully reviewed the way that police, prosecutors, probation officers, and judges may introduce disparities in sentencing.[89] The gap between the average sentences for African-Americans and whites was larger for the guideline than for the non-guideline cases.[90] However, these cases are not comparable

---

[87] For additional discussion and illustrations, see Michael Tonry, *Sentencing Commission and Their Guidelines, in* 17 CRIME AND JUSTICE: A REVIEW OF THE RESEARCH 137-195 (Michael Tonry ed. 1993).

[88] *See* Karle & Sager, *supra* note 77.

[89] *See* Heaney, *supra* note 79.

[90] Judge Heaney did not attempt to assess whether this gap was due to legally relevant or irrelevant factors. For other critiques of Judge Heaney's study, see Joe B. Brown, *The Sentencing Guidelines are Reducing Disparity*, 29 AM. CRIM. L. REV. 875 (1992); Schulhofer, *supra* note 59; William W. Wilkins, *Response to Judge Heaney*, 29 AM. CRIM. L. REV. 795 (1992).

because the non-guideline cases involved offenses that occurred more than a year before sentencing (before November 1, 1987, the effective date of the guidelines) while the guideline cases involved offenses that had occurred since the guidelines became applicable. Thus, the guideline cases contained more offenses of the type that reach sentencing quickly, such as drug trafficking, which are sentenced harshly and in which African-Americans are more highly represented.

4. The Commission's Four-Year Evaluation

The disparity study included in the U.S. Sentencing Commission's *Four-Year Evaluation* deserves more extended discussion because it represents the best pre/post matched group comparison yet undertaken. Congress instructed the U.S. Sentencing Commission to conduct an evaluation of the Guidelines four years after they were implemented. For this 1991 report,[91] the Commission conducted several disparity-related studies but featured a "Distributional Analysis," which compared bank robbery, cocaine distribution, heroin distribution, and bank embezzlement cases sentenced during fiscal year 1985 to cases sentenced between January 19, 1989, and September 30, 1990.[92]

---

[91] FOUR-YEAR EVALUATION, *supra* note 48. Because implementation of the guidelines was delayed in many jurisdictions due to constitutional challenges to the Sentencing Reform Act, the guidelines had been fully implemented only two-and-a-half years at the time of the Commission's evaluation.

[92] The Commission reported preliminary results from two other studies in the "Disparity in Sentencing" chapter of its report. "Judicial Sentencing Patterns under the Guidelines" reported the location of sentences relative to the guideline range (e.g., downward departure, bottom of the range, middle of the range, etc.) for four groups of matched offenses and offenders. *See id.* at 300. Locations were compared for offenders of various races, genders, and other demographic characteristics to assess whether unwarranted disparity might affect the choice of location. Inadequate numbers, inconsistent findings, and statistically insignificant results led the Commission to conclude that "interpretation of any findings is difficult." *See id.* at 324. A study of "Judicial Discretion and Its Relationship to Disparity" reported on how the guideline range was used for various offenses and offenders. Violent offenses were sentenced at the top of the range more often than other offenses, economic offenses were more often sentenced near the bottom of the range, and drug offenses were more often below the range. Women were more frequently sentenced near the bottom of the range. No consistent pattern was discernable for race. Because these analyses do not bear directly on inter-judge disparity, we discuss these studies no further in this report.

Thus, only cases sentenced after the Supreme Court's decision in *Mistretta*, which upheld the constitutionality of the guidelines and led to their full, nationwide implementation, are included in the post implementation sample.

Using FPSSIS data and original data collection from paper case files,[93] offenders were matched on factors that the guidelines deem relevant to sentencing, including the approximate amount of drugs, injury caused to any victims, the defendant's role in the offense, criminal record, and whether the defendant pleaded guilty or went to trial. Offenders who cooperated with the government in the prosecution of other persons were excluded from the analysis, a decision that has been criticized by some reviewers.[94] Because of the restricted number of crimes and the requirement of matching, the number of defendants in each comparison group ranged from small (e.g., thirteen pre-guideline bank robbers with weapons, seventeen without weapons) to moderate (eighty-one guideline cocaine traffickers).

For each defendant at each time period, the Commission determined: (1) the sentence imposed by the judge at the sentencing hearing, and (2) the expected time that the defendant

---

[93] TONRY, *supra* note 1, argued that data limitations and comparability problems make the Commission's findings suspect. While it is true that the switch from FPSSIS to the Commission's own monitoring data changed the definitions of some of the variables needed for pre/post comparisons, Commission researchers went back to the original paper records and recoded some data to ensure that data used to form the matched groups were valid and reliable. While better data would always be desirable, no fatal flaw in the data used to form the matched groups is apparent and the analyses conducted were performed only with data meeting acceptable standards of validity and reliability.

[94] The Commission removed from pre-guidelines cases those in which the presentence report indicated that the defendant had cooperated with the government, and from guideline cases those that involved departures based on "substantial assistance to the government." *See* U.S.S.G. § 5K1.1. Matching defendants on whether they cooperated is appropriate and necessary to create comparable groups. However, as Michael Tonry has noted, substantial assistance motions are sometimes used to avoid guideline sentences. The available data are inadequate to identify which defendants inappropriately received a substantial assistance departure. Since judges have complete discretion within statutory limits once such a motion is granted, eliminating these cases may have artificially reduced the range of disparity in the guideline sample. A possible alternative analysis—*including* all cooperating defendants in the comparison groups—may be a better approach (although the portion of cooperating defendants is probably higher in the guidelines group since the longer guideline sentences serve as an incentive for cooperation).

would actually spend in prison, that is, the "expected time to be served." For pre-guideline defendants, the expected time to be served was calculated using an algorithm developed by the Commission, which incorporated the presumptive parole release date, expected good-time deductions, and other factors known to influence time served in the pre-guidelines era. This resulted in the best available estimate of the date a defendant could expect to be released provided that he or she maintained good behavior while in prison. The expected time to be served for guidelines defendants was the sentence imposed, less the maximum 13% reduction for good behavior in prison that is permitted by the Sentencing Reform Act. Thus, both expected time-to-be-served measures were calculated based on a presumption of good behavior while in prison.

Comparisons of the distribution of sentences for each of the matched groups demonstrated that the guidelines had reduced disparity in sentences imposed. In every offense group, there were large variations in sentences imposed in the pre-guidelines period. And for every group, variation in sentences imposed was reduced under the sentencing guidelines. Less variation was found in expected time to be served in the pre-guidelines era, and under the sentencing guidelines variation in time to be served was similar to the pre-guideline period. In only three of the eight matched groups was the reduction in variation of time-to-be-served statistically significant, although for all groups the change was in the direction of less variation under the guidelines.

Several reviewers have cited only the results for expected time-to-be-served and concluded that the evidence for the effectiveness of the guidelines is weak.[95] Apparently, these reviewers

[95] BUREAU of JUSTICE ASSISTANCE, *supra* note 73, at 85 ("The fact that only three of the offense categories showed statistically significant reductions in disparity suggests that pre-guidelines cases were already exhibiting a relatively high degree of uniformity in court disposition. Further declines in sentencing disparity might have occurred independently of the introduction of guidelines and may be the product of chance."); MCDONALD & CARLSON, *supra* note 18, at 32; Rhodes, *supra* note 7, at 155 (criticizing use of an expected time to be served measure that incorporated the presumptive parole date because it "may not correspond closely with the prison time actually served prior to the guidelines. This may render the comparisons of time served inaccurate").

believe the only meaningful comparison is between time served under the guidelines and time served in the pre-guidelines era. However, we think it counts as success if the guidelines achieved truth-in-sentencing while maintaining at least the same uniformity in time served that had previously been achieved through parole guidelines. The goal of the parole guidelines was clearly the same as the sentencing guidelines—reducing unwarranted disparity—and by most accounts, they were a success.[96] Indeed, the parole guidelines provided the model for the sentencing guidelines. We believe that the findings from the *Four-Year Evaluation* suggest that the guidelines achieved truth-in-sentencing *while* decreasing disparity in sentences imposed *and* maintaining or increasing the uniformity in time to be served that had previously been obtained under the parole guidelines.

The Commission concluded, and perhaps overstated, its belief that the *Four-Year Evaluation* demonstrated that unwarranted disparity was reduced by the guidelines.[97] In any event, the Commission's self-evaluation was met with considerable skepticism. At the request of Congress, the GAO reanalyzed the data on cocaine and heroin offenses, using somewhat different techniques, and also conducted additional tests.[98] Their analyses replicated and confirmed the Commission's basic findings, yet they decided that there was insufficient evidence to conclude

---

[96] Michael R. Gottfredson, *Parole Guidelines and the Reduction of Sentencing Disparity*, 16 J. RES. CRIME & DELINQ. 49 (1979) (reporting that data strongly suggest a considerable reduction in disparity when examining time served compared to sentence imposed and, when read in light of other research, imply that this reduction is a result of the federal parole guidelines).

[97] David Weisburd points out that the Commission overstated its results in its summary of findings in the report. *See* David Weisburd, 5 FED. SENTENCING REP. 150 (1992). The text of the Commission's report says that disparity had been reduced for cocaine and heroin offenses both in terms of sentence imposed *and* time to be served. *See* FOUR-YEAR EVALUATION, *supra* note 48, at 299. But since the findings for time-to-be-served were not statistically significant, the possibility that they were due to chance cannot be eliminated, even though the variances were reduced under the guidelines. Weisburd, *supra*.

[98] GAO REPORT, *supra* note 9, at 58-59. The GAO's analysis involved all heroin and cocaine distributors in Criminal History Category I. The GAO also conducted additional analyses aimed at determining whether legally irrelevant characteristics such as race and gender were affecting sentences under the guidelines. *See id.* apps. II & III. Since these studies do not bear on whether inter-judge disparity has been reduced from the pre-guidelines era, we will not discuss them here.

that the guidelines had reduced disparity.[99] Subsequent reviewers have reached the same judgment. Several criticisms that seem to us incorrect have been discussed in previous footnotes and in the discussion that follows. However, several other methodological problems qualify our confidence in the *Four-Year Evaluation* findings and suggest the need for additional research to complement the matched-group approach.

First, several reviewers have argued that small sample sizes make the Commission's findings suspect.[100] In its review, the GAO noted that because only cases that were part of narrowly-defined matched groups were studied, only 479 of the 25,940 cases sentenced in FY1990 were analyzed. Six of the matched groups involved fewer than 30 cases and none was larger than 81. But in fact, small sample sizes make it more difficult to get statistically significant results. The reductions in variances for sentences imposed were statistically significant in the Commission's study because the *effect sizes*—the amount of reductions—were very large. For example, for weaponless bank robbers who had prior convictions, the range of months between sentences at the 10th and 90th percentiles fell from 138 months in the pre-guidelines era to 26 months under the guidelines. For heroin distributors without a criminal history, the range fell from 78 months to 28 months. Precisely because the reduction in variance was so large, the Commission was able to achieve statistically significant results with relatively small sample sizes.

The problem with the Commission's analysis, it seems to us, is not the *size* of the samples but their *representativeness.* What can we learn about the overall effectiveness of the guidelines in reducing disparity by examining eight groups of offenders guilty of four types of crimes? The Commission viewed the four crimes as broadly representative of federal offenses, since they

---

[99] *Id.* at 10. ("[L]imitations and inconsistencies in the data available for pre-guidelines and guidelines offenders made it impossible to determine how effective the sentencing guidelines have been in reducing overall sentencing disparity."). The Commission disagreed with the GAO's conclusion, which was reflected in the subtitle of the GAO report "Central Questions Remain Unanswered." Instead, the Commission argued that the research justified a different conclusion as noted in "Disparity Reduced, But Some Questions Remain." *See id.* app. VI, at 182.

[100] *See* TONRY, *supra* note 1, at 47.

included one "street" crime, one "white collar" crime, and two drug crimes. But we and others have been less convinced.[101]

One more problem mars the *Four-Year Evaluation.* In the pre-guideline era, judges anticipated that the time imposed would be reduced by one- to two-thirds through parole release. Accordingly, they imposed longer sentences for some offenses than are required by the guidelines. Several of the groups in the Commission's study showed significant reductions in the lengths of sentences imposed. Variances in sentence imposed would be smaller for these crimes under the guidelines simply because the sentences are shorter and the range of sentences narrower, even without any reduction in unwarranted disparity. The GAO appears to have recognized this problem and addressed it through use of a "coefficient of variation." Unfortunately, this part of the GAO study is flawed on other grounds.[102] To make meaningful comparisons, an adjustment in the measure of variation is needed to allow for changes in the lengths of sentences imposed in the guideline era.

Several of the limitations of the Commission's matched group comparisons could be overcome.[103] However, given the trade-off between creating precisely defined groups and creating groups of sufficient size for meaningful statistical comparison, some loss of representativeness is inevitable. Problems inherent with matched-group research designs convince us that we cannot rely on their findings alone to conclude that the

---

[101] Weisburd, *supra* note 97, at 151 ("[T]he restricted samples employed by the Commission make it very difficult to generalize broadly from their findings.").

[102] The "coefficient of variation" adjusts the measure of dispersion to account for differences in the average sentences at the two time periods. It appears to have been used in the analysis, reported in Appendix I of the GAO Report. GAO REPORT, *supra* note 9. However, this analysis matched offenders on criteria that were too general. *See supra* note 12 and accompanying text. Simple comparisons of standard deviations appear to have been used in the other replication of the Commission's analysis.

[103] Using more recent guidelines data, a reanalysis could: (1) create as many matched groups as possible; (2) compare groups both including and excluding cooperating defendants; (3) exclude cases where mandatory minimum statutes truncate the range (or use the guideline sentence in such cases) to ensure that any reduction in variation is not due to these statutes; and (4) compare coefficients of variation or some other measure of variation that takes account of differing lengths of sentences at the two time periods.

guidelines have been either a failure or a success.[104] A different method that does not require elimination of so many types of offenses is needed.

Statistical control has been proposed as one alternative to matched-group comparisons.[105] It requires researchers to measure characteristics at the two time periods, but instead of using these data to match cases, it uses them to estimate the contribution of each characteristic to the final sentences. These contributions are then removed from the sentences and the "residual" variation that remains is examined to see if it has changed. In this way, differences between the types of cases at the two times are "controlled for" statistically. This method also has problems, however, that we believe limit its usefulness for studying the effect of the guidelines on inter-judge disparity.[106]

B. USING RANDOMIZATION TO CREATE COMPARABILITY

The comparability problem has been called "insurmountable."[107] The GAO ultimately concluded that because of limitations in the data, such research would never yield definitive answers as to whether the guidelines had been effective at reducing unwarranted disparity. It recommended that attempts at pre/post comparisons in the federal system be ended.[108] But given the importance of the policy issue involved, we believe it is premature to abandon these efforts, especially since matching is not the only means of achieving the comparability needed for pre/post evaluations.

---

[104] For other reviews of these studies, see BUREAU OF JUSTICE ASSISTANCE, *supra* note 73, at 84-89; MCDONALD & CARLSON, *supra* note 18, at 26-35; Michael Tonry, *The Failure of the U.S. Sentencing Guidelines*, 39 CRIME & DELINQ. 131 (1993).

[105] William Rhodes, *Sentence Disparity, Use of Incarceration, and Plea Bargaining: The Post-Guideline View from the Commission*, 5 FED. SENTENCING REP. 153 (1992).

[106] These problems include uncertainty about how variance shared by several characteristics should be apportioned, which can confound comparisons if the prevalence of these interrelated factors differs between the two time periods. In addition, removing the contribution of various factors introduces the possibility of systematic error into the residual variance if the form of the relationship between the factors and the outcome is misspecified.

[107] *See* TONRY, *supra* note 1, at 47.

[108] *Id.* at 24 ("As a practical matter, it is not possible to rectify the shortcomings of the pre-guidelines data to develop a more meaningful baseline for comparing sentencing outcomes before and after the guidelines . . . .").

In much scientific research, comparisons are made not between individuals, but between averages of comparable *groups* using random assignment to ensure comparability between the groups. For example, researchers assessing the effect of a treatment, such as a new drug for high cholesterol, administer the drug to an experimental group and to a control group. Subjects are randomly assigned to the two groups. There is no need that the *individuals* in the two groups be similar; we know that they will vary in pre-treatment cholesterol. By using random assignment, in the long run these differences are evenly distributed across the groups and the *average* pre-treatment average cholesterol levels will be fairly similar between the groups.

The laws of probability guarantee certain things. First, the larger the size of the groups, the more similar their average cholesterol levels will be. We recognize intuitively that in larger groups, the effects of extremely high or low levels on the group average will be attenuated. Second, based on the size of the groups and the amount of variation among individuals within them, we can determine how much difference in group averages is expected by chance, that is, *due to the random assignment.* With twenty persons in each group, it may be quite common for the average of two randomly created groups to differ by ten points. It would be extremely unlikely, however, for their difference to be 80 points. If we find an 80-point difference after the experimental group has undergone treatment, then we are safe to conclude that it was the treatment, and not any pre-existing differences among the individuals in each group, that accounts for at least some of the difference.

Statistical "hypothesis testing"—the bread and butter of scientific research—applies these laws of probability to real-world data. Statisticians calculate the probability that any observed difference might be due to the random assignment instead of some other difference between the groups. If the experimental group's average score is so much different that it would arise from random assignment less than five times in 100, the difference is said to be "statistically significant" at the .05 probability level.

Random assignment plays a role in legal procedures as well as in experimental research. In most federal courts, judges in the same location are assigned cases randomly to prevent "judge shopping" and to help ensure fair procedures. This creates a "natural experiment" that permits us to assess how judges influence sentences. Judges are like treatments. We know from the laws of probability the differences in average sentences that we can expect by chance among judges who are part of the same random assignment pool. If the observed differences are larger than this, we can infer that it is something about the judges—most likely differences in sentencing philosophies—that explains some of the differences. It is *not* just differences in the types of cases sentenced.

Capitalizing on this natural experiment avoids perplexing questions about how to match offenders and permits us to study a much larger number of sentences, not only the small number that meet a matching criteria. We pay a price, however. We study differences in *averages* among judges, not differences in sentences imposed on identical or similar cases. Our focus becomes the "primary judge effect"—overall differences among judges in leniency or severity—not the amount of variation imposed on a particular type of case. How this approach complements simulation studies and matched-group comparisons will be discussed further after presentation of our results.

1. Waldfogel's Study of Three Cities

The natural experiment methodology was first used to study the effects of the sentencing guidelines in a pilot study by economist Joel Waldfogel published in December 1991.[109] He studied two time periods—pre-guideline sentencing from 1984 until 1987, and a transitional period from 1988 until 1990, which contains a mix of both guideline and non-guideline cases—in the Southern District of New York, the District of Connecticut, and the Northern District of California. First, he established through interviews and statistical tests that the cases were randomly assigned. Then he compared the average sen-

[109] Joel Waldfogel, *Aggregate Inter-Judge Disparity in Federal Sentencing: Evidence from Three Districts*, 4 FED. SENTENCING REP. 151 (1991).

tences among judges who had been at each location at both periods, using "mean absolute deviation" as the measure of inter-judge disparity. He found statistically significant differences among judges (which he called aggregate inter-judge disparity) at all locations prior to the guidelines. In the transitional period, Connecticut and New York continued to show significant differences among judges.[110]

Waldfogel's approach is promising, but his results are problematic for several reasons. First, the transitional period he studied was very early in guideline implementation, therefore the full effect of the new system was not realized. Second, any effect of the guidelines is "watered down" by the inclusion of a large proportion of non-guideline cases in the transitional group. Third, because he was forced to hand code data from court files in the various districts, he was able to include only three districts, which may not be representative of the nationwide effects of the guidelines.[111] Finally, Waldfogel's measure of inter-judge disparity—mean absolute deviation—suffers from the same problem as the Commission's use of variance in its study of the dispersion of sentences among matched groups: it does not take into account overall changes in the lengths of sentences for some types of crimes under the guidelines.[112]

---

[110] Waldfogel used a simple measure to compare the amount of inter-judge disparity before guideline implementation and during the transition. He calculated the number of months that the average sentence of each judge differed from the overall average sentence at each location. The average of these differences was the "mean absolute deviation" for a city. He found that the degree of inter-judge disparity in Connecticut and New York had actually increased, but it had remained the same in the Northern District of California. For the period 1984-1987, the mean deviations in months were 4.2, 5.8, and 4.2 for, respectively, New Haven, Manhattan, and San Francisco. For the period 1988-1990, the means were 9.9, 10.4, and 4.4. *Id.* at 152.

[111] Hand coding was necessary, despite the existence of computerized records, because of a long-standing policy of the Judicial Conference of the United States against releasing computer records with judge identifiers included.

[112] We replicated Waldfogel's work using our data but computing both his "mean absolute deviation" and R-squared, the measure of inter-judge disparity that we ultimately chose for our analysis. (Two of the cities used by Waldfogel did not meet our criteria for random assignment, and were excluded from our later analyses.) Our conclusions were the same as his when we calculated "mean absolute deviations." However, when we used R-squared, disparity decreased under the guidelines in all three cities. This contradiction arises from mathematical differences between the measures. Variances are sensitive to extreme differences because they are calculated

2. *Payne's Study of Three Cities*

In 1997, A. Abigail Payne published a replication and expansion of Waldfogel's approach.[113] Data were obtained for felony cases initiated by grand jury indictment between 1980 and 1991. This was supplemented with judge information obtained from the Southern and Eastern Districts of New York (SDNY and EDNY) and the Eastern District of Pennsylvania (EDPA). Only drug trafficking cases or cases involving embezzlement, fraud, or theft (the EFT cases) were studied.[114] Payne found that since implementation of the guidelines, sentences have grown somewhat longer for EFT cases and substantially longer for drug cases. She also reported that under the guidelines, more defendants are pleading guilty at their initial appearance, rather than at a subsequent hearing.[115]

To study inter-judge disparity, Payne conducted several statistical analyses for each city, separating the drug cases from the EFT cases. First, she examined differences among judges in the proportion of cases receiving a term of imprisonment. Results varied depending on the type of case and the district. Second, regression analyses were performed for each type of case in each city. These analyses measured the proportion of variance in sentences that could be attributed to (1) the specific type of offense, (2) the year the case was sentenced, and (3) differences among judges. (The latter is a measure of inter-judge disparity similar to that which we use for our analysis, and is discussed in

---

by squaring each judge's average sentence subtracted from the overall average for his or her city, thus giving more weight to judges who are farther from the norm. If the guidelines were effective at moving a few extreme judges closer to the norm, the R-squared could shrink even if more judges moved slightly farther from their city's average. Variance is the most widely accepted measure of dispersion when extreme outliers do not skew the distributions of interest. Given the results of our tests of normality, reported below, we chose to use variance rather than mean absolute difference.

[113] Payne, *supra* note 85.

[114] See discussion of the logic of using these case types to isolate the effects of the guidelines, *supra* note 85.

[115] This suggests that pre-indictment plea negotiations are more common under the guidelines. Payne also states that cases are not assigned to judges until after the defendant enters a plea. Payne, *supra* note 85, app. B at 359. This is not consistent with our understanding of the case assignment process in most districts, where cases are assigned to judges when they are filed.

detail in later sections.) Payne found an increase in disparity in all cities in EFT cases and inconsistent results among cities in drug cases. A measure of the statistical significance of the differences among judges yielded inconsistent results among cities in EFT cases but suggested uniform improvement in drug cases.[116] Finally, a non-parametric test of differences among judges in median sentence suggested a limited effect for the guidelines in EFT cases and modest or no effect in drug cases. Payne concluded from the inconsistent results of these analyses that the amount of inter-judge disparity was never very great, but that the guidelines appeared to have reduced it in some, but not all, of the district courts studied.[117]

The mixed results suggest that the effects of the guidelines on disparity are not the same in all places and with all types of cases—a finding that we replicate with our analyses. Several limitations in Payne's approach, however, suggest that a more powerful analysis may yield more robust and consistent findings. First, like Waldfogel, Payne used data from a transitional period, which included many non-guideline cases mixed with cases sentenced under the guidelines, thus diluting any effect of the guidelines. Second, because data were available only for three districts, four case types, and judges who sentenced both before and after the guidelines, the generalizability of the findings is limited.

Most important, the statistical analysis used by Payne to determine the proportion of variance in sentence lengths attributable to judges does not appear to be the most powerful model possible. A combined model treating judges as nested within cities is superior, and permits us to obtain an inter-city as well as inter-judge effect.[118] Further, Payne tested for differences in average sentences—the primary judge effect—but did not measure or discuss the interaction effect of judges with case type, which we know from previous research may be a more powerful influence on sentence disparity than the primary judge effect.

---

[116] Payne does not discuss the influence of variations in sample sizes on the results of the F test. *See infra* Part III.B (discussing our use of the F-test).

[117] Payne, *supra* note 85, at 357.

[118] *See infra* Part III.B (discussing our statistical model).

Given the promise of the natural experiment methodology but the limitations of Waldfogel's and Payne's studies, we set out to extend their work by (1) using more recent data that strictly separated guideline from pre-guideline cases in as many cities, and with as many judges, as possible, and (2) using a more complete statistical model.[119]

III. A NEW ANALYSIS USING THE NATURAL EXPERIMENT APPROACH

The basic logic of the analyses conducted for our study is to compare the amount of inter-judge disparity before and after guideline implementation. Inter-judge disparity is defined as differences in average sentences among judges who receive comparable caseloads. Thus, we are measuring the primary judge effect—the general tendency of judges to be more lenient or severe than their colleagues. We recognize that this measures only the "tip of the iceberg" of disparity created by judges, because it fails to account for interaction effects between judges and particular offense and offender characteristics.[120] We believe, however, that this measure provides a useful barometer of the success of the guidelines. The statistical model we developed allows us to aggregate the amount of inter-judge disparity across different cities while comparing each judge only to other judges in the same city. In addition, the statistical model will test whether the differences we find might be due to differences in caseloads arising from the random assignment and not due to differences in the judges' sentencing decisions.

---

[119] For a preliminary report of another attempt to extend Waldfogel and Payne's natural experiment methodology, *see* SMITH & CABRANES, *supra* note 43, at 122. These authors' initial findings appear to support Payne's conclusion that the guidelines have had some modest success at reducing inter-judge disparity, at least in some cities and for some types of cases.

[120] *See supra* Part I.C.2 (discussing the primary judge effect and interaction effects). A recent analysis by Joel Waldfogel further demonstrated the importance of the interaction effects. *See supra* Part II.B.1. Using real data from the pre-guideline era in the Northern District of California, he found that the primary judge effect accounted for only 2.3% of the variation in sentences, but that an additional 9% was due to these interactions, which is another aspect of inter-judge disparity. The significance of our findings concerning the primary judge effect is further discussed in the final section of this article. *See infra* Part IV.A.

The Technical Appendix provides greater detail on the databases, definition of variables, statistical model, and computer programs used for this study.

A. SAMPLE SELECTION

Our data came from two sources: (1) the 1984 and 1985 Federal Probation Sentencing and Supervision Information System (FPSSIS) and (2) the fiscal years 1994 and 1995 U.S. Sentencing Commission Monitoring data file. We used two-year time frames for both periods to reduce the possibility that unique case assignments would interfere with random assignment. Only felony convictions were included in the analyses.[121]

Federal district courts include active judges, who hear full caseloads and to whom cases are randomly assigned, and senior judges, who have smaller caseloads and who may be selective about the kinds of cases they hear. Because the caseloads of judges on senior status are systematically different from those of active judges, we checked the Federal Reporter[122] for each year to identify which judges were active and excluded senior judges from our analyses. Because no inter-judge comparisons are possible in jurisdictions where only one judge is present at either of our time periods, these locations were also eliminated from our analyses. Further, to reduce the likelihood of systematic case selection that might arise if only two judges are present in a courthouse, we used the nine cities that had at least three or more judges who were active at both time periods.

From these locations we further eliminated those where random case assignment could not be confirmed. Because ensuring random assignment is important to the logic of our analysis, we describe our test for randomness in detail in the

[121] Our datasets contain only offenders who were convicted and sentenced. We do not include defendants whose cases were dismissed before conviction or who were acquitted at trial. Since it is case *filings*, not convictions, that are randomly assigned to judges, some non-comparability in the sentenced caseloads could emerge if judges differed in their dismissal or acquittal rates. Analysis by other researchers of data on all filed cases suggest that differences in dismissal and acquittal rates do exist among judges, but that they are negligible and appear unlikely to explain differences in sentences. Interview with Jeffrey R. Kling, National Bureau of Economic Research (Dec. 9, 1997).

[122] FEDERAL REPORTER, Ser. 2, West Publishing.

Technical Appendix. Applying all these selection criteria left us with forty-one cities, containing 254 judges in 1984-1985 and 301 judges in 1994-1995. Nine cities had the same three or more judges active at both time periods, which were used for within-judge analyses.

B. DEVELOPMENT OF A STATISTICAL MODEL

To compare differences among judges before and after guideline implementation, we conducted a series of multivariate analyses. Sentence length imposed by the judge was considered the outcome, or dependent, variable in the statistical model. Several legal and extra-legal factors were included in the model as explanatory, or independent, variables. Most important was the sentencing judge. Because we wanted to compare each judge's sentences only with those of other judges in the same random assignment pool, each judge was treated as "nested" within his or her city and compared only to other judges with comparable caseloads. We also examined the effects of the city in which the case was decided. When interpreting results, remember that cases were not randomly assigned to cities. Any city effect found may reflect unmeasured differences in the types of cases prosecuted in different cities, as well as differences in regional philosophies and sentencing traditions.

In addition to judge and city, the influence of general offense type and criminal history was considered. Because we wanted the results from the two time periods to be strictly comparable, we used only variables that could be measured the same way in the two datasets. Offenses were categorized into seven general types. Offenders were categorized into two criminal history groups—those with and those without any previous convictions. In 1984-1985, 51% of the defendants had no prior criminal record; in 1994-1995, 48% had none.

We structured the model so that offense type and criminal history were "credited" with explaining as much variation in sentences as possible. Judges were treated as nested within their city and the judge effect was assessed after variation in sentences due to all other explanatory variables had been removed. Because cases were randomly assigned (thus assuring no correlation between case characteristics and judges within each city)

the judge effect can be interpreted unambiguously as the independent influence of judges on sentences.[123]

The statistical significance of each of the explanatory factors was tested with an F-test. Of greatest interest was the effect of judges. The F-test for this factor calculated the probability that any observed differences among judges might be due to differences in their caseload from random assignment rather than to differences in their sentencing decisions. We reported differences as statistically significant only if the result yields a p-level of .05 or smaller, indicating that the observed difference would be expected to arise from random assignment less than one time in twenty.

The magnitude of the influence of each of the explanatory factors was measured with the R-squared—the proportion of variance in sentences accounted for by the factor. R-squared measured the effect of a factor as a percentage of the total variance in sentences at a given time period. Unlike the measures used in the Commission's *Four-Year Evaluation* and by Waldfogel, the R-squared measured inter-judge disparity as a proportion of the total variation at a given time period, whatever the absolute total may be. In discussing R-squared, we converted it to a percentage—the percentage of the total variance in sentences that can be uniquely attributed to judges.

If the guidelines have been effective at reducing inter-judge disparity, the percentage of variation attributable to judges should be smaller after guideline implementation than before. We subtracted the percentage obtained in 1994-1995 from the percentage in 1984-1985. The size of this difference measures the magnitude of the change, the sign of the difference measures the direction: positive indicates an increase in, and negative indicates a decrease in disparity. Thus, we have two general approaches to examining differences among judges and between the two time periods. The F-test assesses the statistical *significance* of differences among judges in a city. The propor-

[123] See *infra* Technical Appendix for a discussion of how the nesting of judges within cities means that some of the effect of varying judicial philosophies may be confounded with the city effect.

tion of variance accounted for by judges assesses the *meaningfulness* of these differences.

C. RESULTS OF A NATIONWIDE ANALYSIS

1. Comparisons Among the Same Judges Before and After Guideline Implementation

The most straightforward way to assess the effect of the guidelines is to compare sentences of judges who sentenced before the guidelines and under them. Unfortunately, because ten years separates the two time periods—during which many judges retired and new judges came on the bench—only nine qualifying cities have at least three judges who were sentencing at both times. This makes forty-two judges available for comparison. Unlike the analysis in the next section involving different judges, the mix of Republicans and Democrats, liberals and conservatives, and women and men in this analysis are largely the same at the two time periods because the judges are the same judges. Differences in disparity at the two times are most likely due to policy changes and not to personnel changes.[124]

Table 1 displays results of this within-judge analysis for sentences imposed.[125] Only defendants convicted of offenses with at least 1,000 cases at both time periods are included—namely drugs, firearms, fraud, immigration, larceny, robbery, and

---

[124] Research has shown, however, that judicial attitudes do change with experience on the bench, generally in the direction of more severe sentences. In addition, experience on the bench exposes judges to a larger number and wider range of offenders and provides more opportunity to learn about the sentencing behavior of colleagues. Thus, some decrease in inter-judge disparity might be expected over the decade of our comparison, even without the introduction of sentencing guidelines. As shown by comparison with the later analyses of all judges, disparity among these judges decreased somewhat more than among judges in general. However, both groups show an aggregate decrease in disparity over the decade of our study.

[125] Analyses were also performed for estimated time-to-be-served, using the algorithms described in the Technical Appendix. As expected, before the guidelines judges had less influence over time served than over sentence imposed. This reflects the effect of the parole release decisions, which were made pursuant to parole guidelines that were designed in part to correct disparities in sentences imposed. Under the sentencing guidelines, time served is a direct reflection of sentence imposed and judges' influence on both is the same.

other.[126] Additional analyses were conducted using all offense types and the pattern of findings was essentially the same.

The top line gives the percentage of variation accounted for by the complete model, i.e., all of our factors combined: 34.68% in the pre-guideline era and 38.93% under the guidelines. In the middle of the table, we see that offense type explains by far the largest share of variation in sentences, both in the pre-guidelines era and under the guidelines. About a fifth of the variation in sentences is accounted for by the general type of offense being sentenced. Criminal history, somewhat surprisingly, explains much less. Recall that we categorized offenders into only two groups: first time offenders and those with any criminal history.

On the second line of the table we see the finding that directly concerns inter-judge disparity—the primary judge effect. In the pre-guidelines era, 2.32% of the variation in sentences was explained by differences among judges. Under the guidelines, this drops to 1.24%. As shown in the column on the right, this represents a reduction of 1.08%. The effect of judges is statistically significant at both time periods, but it is reduced almost by half under the guidelines. These findings indicate that the guidelines, while not eliminating all inter-judge disparity, have had a positive effect.

To get a better sense of what these effects mean, we translated the percentages into actual months of imprisonment, which are shown in separate columns on the table.[127] The 2.32% of variance in sentences imposed attributed to judges in 1994-1995 means that, on average across all cases and judges, defendants can expect their sentences to be about 7.87 months longer or shorter solely due to the particular judge to which they are assigned. This is, of course, an average across all judges and some judges may be more extreme than the average

[126] We limited analyses to these offense types because this number is needed to permit calculation of interaction terms, as reported at the bottom of Table 1 and discussed in the next section.

[127] We translated the effect size into actual months of imprisonment by (1) multiplying the total variance by the portion of the variance accounted for by judges, and (2) finding the square root of the result, thus translating the numbers back into absolute terms.

TABLE 1.

PERCENTAGE OF THE VARIANCE EXPLAINED (IN MONTHS) BY JUDGE AND OTHER FACTORS [A]

Combined Analyses in Cities with the Same Three (or More) Judges in Both Time Periods

PRE-GUIDELINE (1984/1985)		
Variable Group	Months	"R-Square"
Model	30.43	34.68
Judge	7.87	2.32
City	5.24	1.03
Offense Type	22.33	18.66
Criminal History	9.95	3.70
Offense Type * City	10.09	3.81
Offense Type * Judge (City)	11.73	5.15

POST-GUIDELINE (1994/1995)		
Variable Group	Months	"R-Square"
Model	42.65	38.93
Judge	7.61	1.24
City	12.94	3.59
Offense Type	30.69	20.16
Criminal History	10.50	2.36
Offense Type * City	14.74	4.65
Offense Type * Judge (City)	18.00	6.94

PRE/POST DIFFERENCE		
Variable Group	Months	"R-Square"
Model	12.22	4.25
Judge	-0.26	-1.08
City	7.70	2.56
Offense Type	8.36	1.50
Criminal History	0.55	1.34
Offense Type * City	4.65	0.84
Offense Type * Judge (City)	6.27	1.79

[A] Cases were assigned randomly among judges in each city, but not among cities.

suggests. Under the guidelines, the judge effect, as a percentage of variation, is shorter. But because sentences have gotten longer in the guideline era, judges' influence in absolute terms remains almost the same—7.61 months.

The findings on the effect of cities are much less encouraging, but also harder to interpret. Under the guidelines, the city in which one is sentenced appears to have a greater effect than it did in the pre-guideline era, rising from 1.03% to 3.59%. Because cases are not randomly assigned to cities, we cannot know whether this reflects a growth of inter-regional disparities or reflects some growing difference in the caseloads of various cities that is not captured by our general measures of offense type and criminal history. As discussed further below, there are reasons to believe it may be both.

2. *Comparisons Among All Judges*

The possibility that a changing mix of viewpoints on the bench accounts for changes in inter-judge disparity is limited by the same-judge analysis presented above. However, it also limits the number of judges and cases available for study. Two cities—New York and Los Angeles—together account for almost half the judges in the nine-cities analysis. To see if the results from the nine cities generalized to the system as a whole, we examined all cities that had three or more active judges during both of the two time periods, regardless of whether some or any of the judges were the same at both times. Forty-one cities met this criterion.

Table 2 displays the results of this analysis. The overall pattern is similar to the results using only nine cities. Offense type continues to be the greatest influence on sentences. Criminal history remains significant but small. But it appears that as we include more cities in the analysis, a greater range of differences among them emerges. The city effect grows from 1.81% in the pre-guidelines era to 5.64% under the guidelines.

The influence of judges on the sentence imposed is significant at both time periods—2.40% in the pre-guidelines era and 1.64% under the guidelines, a reduction of .76%. Because we are comparing different judges, we cannot interpret these

TABLE 2.

PERCENTAGE OF THE VARIANCE EXPLAINED (IN MONTHS) BY JUDGE AND OTHER FACTORS[A]

All Cities[B]

PRE-GUIDELINE (1984/1985)		
Variable Group	Months	"R-Square"
Model	33.26	33.76
Judge	8.89	2.40
City	7.70	1.81
Offense Type	23.21	16.44
Criminal History	10.05	3.08
Offense Type * City	11.41	3.97
Offense Type * Judge (City)	14.07	6.04

POST-GUIDELINE (1994/1995)		
Variable Group	Months	"R-Square"
Model	46.40	37.64
Judge	9.69	1.64
City	17.97	5.64
Offense Type	31.00	16.80
Criminal History	14.56	3.70
Offense Type * City	16.98	5.04
Offense Type * Judge (City)	16.58	4.80

PRE/POST DIFFERENCE		
Variable Group	Months	"R-Square"
Model	13.14	3.88
Judge	0.80	-0.76
City	10.27	3.83
Offense Type	7.79	0.36
Criminal History	4.51	0.62
Offense Type * City	5.57	1.07
Offense Type * Judge (City)	2.51	-1.24

[A] Cases were assigned randomly among judges in each city, but not among cities.

[B] The cities included in this analysis are those which pass the two statistical tests of randomness described in Appendix A.

results as an unambiguous effect of the guidelines. In general, however, the guidelines appear to have had modest success at reducing inter-judge disparity. But in this forty-four-cities analysis the effect of the guidelines appears more limited and differences among cities appear more important. The findings suggest the need for more fine-grained analysis of how sentencing is affected by the guidelines in different cities.

The last two lines on Tables 1 and 2 display interaction effects. The offense type by judge interaction effect measures whether the relative leniency or severity of a judge depends on the type of crime being sentenced. Because judges are nested within cities, the offense type by city interaction was also determined.[128] All the interactions were statistically significant and accounted for sizeable portions of variance. These effects suggest the need for more fine-grained analyses of how the guidelines have affected different judges' sentencing of different types of crimes.

Recall that the simulation studies reviewed in Part I found that sentences were influenced by a judge's overall tendency toward leniency or severity—the primary judge effect—but even more by disagreements among judges over the appropriate sentence for particular types of cases. For example, a judge may be on average four months more lenient than the average for his or her city. But if there is an interaction effect, the judge may be eight months more lenient for fraud cases, but only two months more lenient for drug trafficking. In fact, some judges may be more lenient than their colleagues in some types of cases but more harsh in others.

Examination of changes in the interaction effects between the two time periods reveals no consistent pattern between the nine-cities and forty-four-cities analyses. The offense type by city interaction went up in both analyses. The offense type by judge interaction went up in the nine-cities analyses and down in the

---

[128] Technically, the offense type by city interaction is above the offense type by judge (nested within city) interaction in the hierarchy of effects. The variance associated with offense type by city is partialled out of the dependent variable before the offense type by judge effect is determined. *See* HAROLD R. LINDMAN, ANALYSIS OF VARIANCE IN EXPERIMENTAL DESIGN 204 (1992).

forty-one-cities analysis. The presence of significant interactions and the lack of a consistent pattern suggest the need to examine different offense types separately.

D. RESULTS OF DISAGGREGATED ANALYSES

Researchers studying the guidelines have issued a warning: "the trees may be more significant than the forest."[129] Rand Corporation researcher Terence Dunworth and law professor Charles Weisselberg studied the impact of the guidelines on felony sentencing and reached the basic conclusion that the guidelines do not affect all case types and all districts equally.[130]

> It is extremely difficult, and perhaps, unhelpful, to draw general, system-wide, conclusions about the effect of the guidelines upon the district courts. A showing that different districts and cases are subject to different stresses is, in itself, significant, because it suggests that the guidelines mean different things in different contexts.[131]

We used several methods to test the hypothesis that the effects of the guidelines vary for different offense types and in different cities. Because the number of cases available for analysis drops dramatically when we disaggregate national caseloads into specific offenses in specific locations, it becomes difficult to achieve statistical significance and power with every desired comparison. But the results provide examples that support the hypothesis that the effect of the guidelines is not uniform across all offense types and places.

1. Findings for Different Offense Types

Table 3 shows the effect of judge and other factors on sentences imposed for various offenses. Because we are modeling the influences on each offense separately, offense type is not a factor and there are no interaction terms. The number of cases available for each offense is relatively small and some effects

---

[129] Charles D. Weisselberg & Terence Dunworth, *Inter-District Variation Under the Guidelines: The Trees May Be More Significant Than the Forest*, 6 FED. SENTENCING REP. 25 (1993).

[130] Terence Dunworth & Charles D. Weisselberg, *Felony Cases and the Federal Courts: The Guidelines Experience*, 66 SO. CAL. L. REV. 99 (1992).

[131] Weisselberg & Dunworth, *supra* note 129, at 27.

TABLE 3.

PERCENTAGE OF THE VARIANCE IN SENTENCE IMPOSED EXPLAINED BY JUDGES AND OTHER FACTORS

All Cities

PRE-GUIDELINE (1984/1985)

Variable Group	Drug	Fraud	Immig.	Robbery	Firearm	Larceny
Model	18.00	14.14	26.28	39.96	29.58	29.96
Criminal History	0.85	3.14	4.23	5.72	0.58	4.78
Race	0.45	0.91	2.20	0.85	1.15	1.40
Gender	1.44	0.84	0.74	1.10	0.25[1]	1.17
Age	1.60	0.06[1]	0.13[1]	1.11	0.14[1]	0.07[1]
Judge	7.47	6.90	5.89	15.80	18.08[1]	16.16
City	6.20	2.29	13.10	14.38	9.37	6.37

POST-GUIDELINE (1994/1995)

Variable Group	Drug	Fraud	Immig.	Robbery	Firearm	Larceny
Model	26.50	14.83	40.82	36.06	30.11	28.08
Criminal History	3.76	2.95	15.08	2.77	6.25	5.49
Race	3.60	0.62	1.16	2.28	0.42	1.04
Gender	1.35	1.10	0.31	0.65	0.40	2.10
Age	0.54	0.89	0.57	0.11[1]	0.81	0.40
Judge	4.55	6.34	10.36	18.88	14.00	13.61[1]
City	12.70	2.91	13.33	11.37	8.23	5.43

PRE/POST DIFFERENCE

Variable Group	Drug	Fraud	Immig.	Robbery	Firearm	Larceny
Model	8.50	0.69	14.54	-3.90	0.58	-1.88
Criminal History	2.91	-0.19	10.85	-2.95	5.67	0.71
Race	3.15	-0.29	-1.04	1.43	-0.73	-0.36
Gender	-0.09	0.26	-0.43	-0.45	0.15	0.93
Age	-1.06	0.83	0.44	-1.00	0.67	0.33
Judge	-2.92	-0.56	4.47	3.08	-4.08	-2.55
City	6.50	0.62	0.23	-3.01	-1.14	-0.94

[1] Variable not significant at a .05 level.

(indicated on the Table) do not reach statistical significance, but some general conclusions can be drawn.[132]

The size of the judge effect is substantially larger for separate offense types than it was for the total caseload, confirming that differences among judges are more apparent as the specificity of the case characteristics increases. In the analysis reported in Table 2, judges on average accounted for 2.40% of the variance in the pre-guideline era. Table 3 shows considerable variation among offense types but a uniformly larger effect for judges, ranging from 5.89% for immigration to more than 16% for larceny. (The 18.08% for firearms cases was not statistically significant.)

For most offenses, the judge effects decrease under the guidelines. The influence of judges on drug cases, for example, drops from 7.47% to 4.55%. Robbery and immigration are different. This result surprised us and we reanalyzed the data excluding "outliers"—cases in which a judge's average sentence was more than three standard deviations from the city mean for that type of case. The size of the judge effect decreased for postguideline cases about 1% for immigration and 4% for robbery. But they remained the largest significant judge effects of all offense types and the only two that increased under the guidelines. Some hypotheses about why disparity may have increased for these two offenses are discussed in Part IV.

For most offense types the city effect also decreases under the guidelines or increases less than one percent. But for drug cases the city effect increases 6.5%. Because such a large portion of the federal criminal docket is comprised of drug cases, the increase in the city effect found in the combined case analysis may be due largely to drug sentencing. The increase could reflect greater differences among the cities in the types of drug

---

[132] These results were obtained using a model that included gender, age, and race as explanatory variables. The effects of these characteristics were generally small at both time periods for every offense type. In several instances they do not reach statistical significance. Race is associated with a larger share of variation and its importance in drug cases increases in the guideline era, reflecting the impact of differential treatment of crack and powder cocaine under the mandatory penalty statutes, and the disproportionate number of African-American offenders sentenced under the harsher crack laws. *See supra* Part I.C.1; *see also infra* Technical Appendix.

cases sentenced now compared to 1984/1985, or the greater spread of sentences for drug offenses proscribed under the guidelines, ranging from probation to life in prison, which would magnify any caseload differences that existed among the cities.[133] Or it could reflect differences in the way that various cities are implementing the drug guidelines.

2. *Findings for Different Cities*

The findings for separate cities mirror the findings for offense types—while the guidelines have in most instances reduced inter-judge disparity, pockets of problems remain and in some cases disparity may have worsened. We determined the judge effect for each of the cities having at least three judges at both time periods.[134] In twelve cities the judge effect was statistically significant at both time periods. In seven of these inter-judge disparity was reduced by 1% to 5%. In two there was essentially no change. In three it increased by 1% to 3%.

The dramatic growth of the main city effect for drug cases, as shown in Table 3, suggests that different cities may be reacting differently to the drug guidelines and that these reactions are magnifying differences among them. An additional concern is that individual judges within each city are also reacting differently to the drug guidelines. To test this possibility we analyzed drug cases in the nine cities where the same judges sentenced during both time periods. Too few cases were available in five of the cities to achieve statistical significance at either time. However, in the four that achieved significance, two showed reductions of inter-judge disparity and two showed increases. This suggests that the drug guidelines are affecting different cities differently, both through development of distinct city-wide ad-

---

[133] For example, even if differences among cities in the portion of small-scale or large-scale drug offenders remained constant over the two time periods, the greater difference in the sentences imposed on these two types of offenders would appear in our data as an increase in the city effect for drug offenses.

[134] A simple one-way analysis of variance without any control variables was used. Since cases are randomly assigned in each city included in the analysis, each judge should get a comparable distribution of offense types and offender characteristics. The comparability of caseloads were checked in our tests of randomness. *See infra* Technical Appendix.

aptations and also in the degree to which the guidelines constrain individual judge discretion.

What should be made of these results? The simplest bottom line is that the guidelines appear to affect different crimes in different ways in different cities. While the guidelines, on balance, appear to be reducing aggregate inter-judge disparity and to be succeeding more often than they are failing, problem areas remain. Some evidence even suggests an increase in disparity in some places and with some types of offenses.

IV. EXPLAINING THE EFFECTIVENESS AND INEFFECTIVENESS OF THE GUIDELINES

The nationwide results, showing a decrease in overall inter-judge disparity, should hearten supporters of sentencing reform. But the unevenness of the results may cause some to question whether the guidelines have been worth the trouble. We believe there are reasons to judge the sentencing guidelines as a qualified success, and reasons to hope that additional improvements would reduce disparity further. But before discussing this conclusion, it is important to confront some of the questions that might be raised about our findings.

A. IS THE INTER-JUDGE DISPARITY WE FOUND TOO SMALL TO WORRY ABOUT?

The judge effects we found were statistically significant but generally small, particularly in light of the simulation studies discussed in Part I, which reported primary judge effects of up to 20% of the total variation in sentences. Our finding—of primary judge effects of less than 3% in the pre-guideline era for the combined caseload—suggests that the influence of judges was not as great in the real world as was indicated by the simulation studies. Perhaps the hypothetical cases used in the simulations accentuated differences of opinion to a greater extent than does the actual docket. Both types of studies demonstrate that the influence of judges on sentences is minor compared to the seriousness of the offense.

As described previously, however, the primary judge effect is like the tip of the disparity iceberg. Differences among judges

in their overall tendency to be lenient or severe are not as great as differences over the appropriate sentence in individual or similar types of cases. We know from the simulation and matched-group studies that the more specific the case, the more disparity is apparent. This disparity lies hidden below the caseload averages measured by the primary judge effect, which is only a rough barometer of the degree of individual case disparity. Two of our findings support this view. First, the judge effect was uniformly larger in our analyses of separate offense types than in our combined case analyses. Second, the interaction effect for offense type by judges was significant and generally larger than the primary judge effect, which indicates that knowing what type of case a judge is sentencing gives you more information than does knowing only his or her average leniency or severity.

Ultimately, whether a certain amount of disparity is enough to worry about is a value judgment best left to policymakers. Our findings suggest there was substantial disparity among similar cases in the pre-guidelines era, but less inter-judge disparity under the guidelines.

B. ARE THE DECREASES IN DISPARITY LARGE ENOUGH TO PROVE THE GUIDELINES' SUCCESS?

The success of the guidelines would be clear if the primary judge and city effects and the interaction effects fell to zero and were not statistically significant after implementation of the guidelines. The actual data are more complicated. The primary judge effect was cut almost by half in the nine-cities analysis but only by a third in the forty-one-cities analysis. The offense type by judge interaction fell in the forty-one-cities analysis, but actually increased in the nine-cities analysis. In the separate offense analyses, the judge effect fell for four offenses, but increased for two. Some may look at these results and conclude that the guidelines' success at reducing inter-judge disparity is a wash.

At this point we can only speculate about why inter-judge disparity in immigration and robbery cases may have become worse. Immigration has been the subject of numerous legislative initiatives in recent years that have generally increased pen-

alties. These often-controversial initiatives may be unevenly applied.[135] Other commentators have noted that charge bargaining is especially likely to involve guideline circumvention in bank robbery cases.[136] Charges for possession of a firearm during the commission of an offense under 18 U.S.C. § 924(c) are also unevenly prosecuted.[137] Since conviction for each additional bank robbery count and each firearm count under § 924(c) adds substantial time to the offender's sentence, any disparity in prosecution can have dramatic effects.

On balance, we claim modest success for the guidelines at reducing inter-judge disparity for several reasons. First, we place greatest emphasis on the primary judge effect. The interaction effect is more difficult to interpret. Its increase in the nine-cities analysis may be an aberration affected by the relatively small number of cities available in that analysis. Similarly, we are heartened that the judge effect decreased in the two largest offense groups, drugs and fraud, which together account for over half of all defendants in the federal system.

Most important, the results from our natural experiment should be considered in conjunction with findings from previous research. Together, the simulation, matched group, and natural experiment methods provide converging evidence that the guidelines have had an overall positive effect. Our results are consistent with the basic finding from pre-guideline simulation studies: philosophical differences among judges were the chief source of unwarranted disparity in the pre-guideline era. The reduction in the primary judge effect we found is consistent with the Commission's *Four-Year Evaluation*,[138] which showed significant reductions in the dispersion of sentences imposed on

---

[135] Recent findings indicate that similar immigration offense conduct is prosecuted under different statutes and has different rates of departure from district to district, leading to substantial differences in final sentences. Linda Drazga Maxfield & Keri Burchfield, Immigration Offenses: Does Federal Guideline Practice Apply the Principles of Sentencing?, Paper Presented at the American Society of Criminology Annual Conference (Nov. 13, 1998).

[136] Nagel & Schulhofer, *supra* note 56, at 548.

[137] Paul J. Hofer, *Federal Sentencing for Violent and Drug Trafficking Crimes Involving Firearms*, 37 AM. CRIM. L. REV. 741 (2000) (showing that § 924(c) charges are brought in less than half of the cases in which they appear factually warranted).

[138] *See* FOUR-YEAR EVALUATION, *supra* note 48.

similar matched cases. In addition, the natural experiment provided a piece of the puzzle that no other method could provide. It enabled us to study the full felony caseloads of a large number of judges located in many cities throughout the country. We believe our findings are representative and can be generalized to the federal system as a whole.[139]

However, more information would always be useful. Replication of our findings would give us greater confidence in the results. Additional studies using matched group or simulation approaches are possible. In the methodological review of earlier studies in Part I we suggested ways that future studies using these methods might be improved. We urge continued research using a variety of methods.

C. HAVE OTHER SOURCES OF DISPARITY WORSENED?

1. Prosecutorial Discretion and Plea Bargaining

In Part I.D.1 we described several potential problems that might prevent the guidelines from reducing unwarranted disparity. Our findings suggest that whatever disparity may still be arising from judicial departure from the guidelines, or from lax or inconsistent application of vague or complex provisions, these problems do not offset the reduction of inter-judge disparity achieved by implementation of the guidelines. One of the most common criticisms of the new regime, however, is that the guidelines have transferred control of sentencing from the judge to the prosecutor.[140]

---

[139] Our requirement that cities have at least three active judges did eliminate single-judge courthouses and smaller cities from our analyses. We can think of no reason why the guidelines should restrict the discretion of judges in small cities any more or less than judges in larger cities; the same variation we found among larger cities is likely also present among smaller ones. There may, however, be differences between large and small cities that affect the sentencing system in unforeseen ways.

[140] WEIS, *supra* note 59; Jose A. Cabranes, *Sentencing Guidelines: A Dismal Failure*, 207 N.Y. L. J. 2 (1992) ("[T]he guidelines have sub silentio moved the locus of discretion from the judge to the prosecutor."); Bennett L. Gershmann, *The Most Fundamental Changes in the Criminal Justice System: The Role of the Prosecutor in Sentence Reduction*, 5 CRIM. JUST. 2, 4 (1990) ("Since the prosecutor is able to make a very precise selection of the ultimate sentence, the judge's role is simply to ratify the choice of sentence determined by the prosecutor."); Heaney, *supra* note 79, at 190-200; Standen, *supra* note

In addition to the discretion they have always enjoyed to bring and dismiss charges or to make sentence recommendations, the last ten years have witnessed the creation of new tools by which prosecutors can control sentencing.[141] These include mandatory minimum statutes that limit judges' discretion, motions for departure based on a defendant's substantial assistance in the prosecution of others, which are solely in the hands of prosecutors,[142] and factual stipulations accompanying plea agreements, which under the guidelines have a direct and predictable impact on the guideline range applicable to a case. The mechanisms designed to regulate this discretion—Department of Justice policies, "real offense" elements of the sentencing guidelines, and judicial review of plea agreements—may not be capable of preventing prosecutorial decisions from reintroducing unwarranted disparity.[143]

In this study we have not quantified the extent to which prosecutorial discretion may be introducing disparity back into the system. Some of the inter-judge disparity we found in our study could be inter-prosecutor disparity.[144] But if prosecutors are introducing disparities, and if those disparities are randomly

---

63, at 1473-74 (arguing that restrictions on judicial discretion give prosecutors almost limitless control over sentencing and thus plea bargaining).

[141] For a recent excellent discussion of the tools by which prosecutors can control sentences, see STITH & CABRANES, *supra* note 43.

[142] Kimberly S. Kelley, Comment, *Substantial Assistance Under the Guidelines: How Smitherman Transfers Discretion from Judges To Prosecutors*, 76 IOWA L. REV. 187 (1990).

[143] *See* Paul J. Hofer, *Plea Agreements, Judicial Discretion, and Sentencing Goals*, FJC DIRECTIONS, May 1992, at 3 (describing importance of judicial review in preventing bargains from undermining sentencing reform and citing obstacles to its exercise); Stephen J. Schulhofer, *Sentencing Issues Facing the New Department of Justice*, 5 FED. SENTENCING REP. 225 (1993) (discussing disparity caused by uneven respect for the guidelines among prosecutors making plea agreements and suggesting five areas where improved DOJ policies are needed).

[144] Inter-prosecutor disparity might appear as inter-judge disparity in our study to the extent that prosecutors are linked to judges in each city. In some locations, prosecutors are assigned to a particular judge. Disparity due to prosecutors would appear there as inter-judge disparity in our data, because the lenient or severe charging or bargaining decisions of a particular prosecutor would be reflected in the data for the judge to which he or she was assigned. (The judge, of course, might affect these charging and bargaining decisions in various ways.) In other locations, a given prosecutor may appear before many or all judges in the city. In these cities, differences among prosecutors would be distributed among the judges in our analysis and would be concealed from our analysis.

distributed among the judges in a city, our methodology would not detect it and could not measure whether prosecutor-based disparity had increased or decreased under the guidelines. Further, the city effect may reflect differences in charging and plea bargaining policies around the country, and the growth of the city effect may indicate that these differences are having a greater impact today than they did in the pre-guidelines era. In short, the guidelines may have constrained judicial discretion successfully but allowed prosecutorial discretion to replace it. Our methodology cannot disentangle these two sources.

Findings from other studies raise the real possibility that prosecutorial discretion is reintroducing disparity. Former Commissioner Ilene Nagel and Professor Stephen Schulhofer conducted the most thorough series of studies of plea bargaining under the guidelines.[145] They cataloged the ways that plea bargaining can undermine the goals of the Sentencing Reform Act and estimated, based on case reviews and interviews in ten districts, that in approximately 20-35% of cases resolved by guilty pleas the sentence imposed is not the one required by strict application of the guidelines.[146] It seems likely that in some of these cases similar defendants get dissimilar bargains.

Survey research has also suggested that plea agreements affect sentences. The vast majority of judges believe that plea bargains lead to disparity in at least some cases, and both prosecutors and judges report that plea agreements do not always reflect the total offense conduct.[147] Weakness in the evidence, the defendant's cooperation, the need to provide incentives for early pleas, and the need to avoid excessive sentences that

---

[145] *See* Nagel & Schulhofer, *supra* note 56. *See also* Milton Heumann, *The Federal Sentencing Guidelines and Negotiated Justice*, 3 FED. SENTENCING REP. 223, 223-24 (1991); David N. Yellen, *Two Cheers For A Tale of Three Cities*, 66 SO. CAL. L. REV. 567, 571 (1992) (arguing that the amount of manipulation will increase over time as the institutional commitment to the Sentencing Reform Act declines and caseload pressures mount).

[146] Shulhofer & Nagel, *Plea Negotiations*, *supra* note 69, at 1290.

[147] JOHNSON & GILBERT, *supra* note 45, at 9 (showing large majorities of judges and probation officers believe that plea bargains are a hidden source of disparity in the guideline system); FOUR-YEAR EVALUATION, *supra* note 48, at 161-98; *see generally* Berman, *supra* note 72, at 300 (reporting survey results that plea agreements often do not contain the full offense conduct).

would be unjust in particular cases are all cited as reasons for allowing defendants to reach plea agreements that do not reflect the full offense conduct.

These findings suggest a pressing need for quantitative empirical evaluation of how prosecutorial discretion is affecting the uniformity of sentencing. But the obstacles to such research are immense. There are no data on how much disparity was caused by prosecutorial decisions in the pre-guideline era, so pre/post comparisons are impossible. Even more limited questions about the effects of prosecutors today are not easily answered. If uniform sentencing is defined as treating offenders similarly based on what they *really did*, instead of what prosecutors are willing and able to prove they did, then we need an impartial assessment of the offender's real offense conduct.

The federal system has an advantage over virtually all state systems for the investigation of prosecutorial discretion, because under the real-offense guidelines the probation officer is instructed to act as the court's investigator. The presentence report is intended to reflect the offender's actual offense conduct.[148] But the probation office has limited resources and must often rely on the government's version of the offense. If researchers are to identify truly similar offenders, they must either accept the facts in the presentence report, win the cooperation of prosecutors, or go into the field and accompany law enforcement agents on their investigations.

Given the importance of establishing the impartiality of all aspects of the criminal justice system, cooperative efforts to evaluate the roles of prosecutors and the influence of offender characteristics would be in everyone's interest. Though perfect data may never be available, measures of the type of cooperation provided by a defendant and the quality of evidence supporting various charges would permit more careful evaluation than has so far been possible.

---

[148] PROBATION AND PRETRIAL SERV. DIV., ADMINISTRATIVE OFFICE OF THE U.S. COURTS, THE PRESENTENCE INVESTIGATION REPORT FOR DEFENDANTS' SENTENCES UNDER THE SENTENCING REFORM ACT OF 1984, Publication 107 at I-3 (Revised 1992) ("[I]t is crucial that a probation officer exercise independence as an agent of the court by developing factual and rule-based assertions.").

2. *Regional Differences*

In our view, the most troubling finding from this study is the growth of the city effect. In Part I we discussed how regional differences were a potential source of disparity in the pre-guideline era. Our findings show that the city in which the case is sentenced continues to matter for some types of cases. Although the findings are not as easily interpreted as the findings regarding judges (because cases are not randomly assigned to cities) the offense type and criminal history variables controlled to some extent for differences in caseloads. On balance, we are left with an uneasy suspicion that some types of inter-city disparities have increased under the guidelines.

Table 3 shows that the increase in inter-city disparity occurred almost entirely in drug cases. The drug guidelines are among the most-criticized in the manual,[149] and even the guidelines' strongest supporters recognize that changes in the drug guidelines may be needed.[150] Some persons believe that the severe punishments required by the mandatory minimum statutes and the guidelines are inappropriate for these offenses. Other critics accept that punishment is appropriate, but contend that the current rules do a poor job of apportioning punishment based on the seriousness of the crime.[151]

---

[149] GAO REPORT, *supra* note 9, at 17 (harshness and inflexibility of drug guideline most frequent problem cited by interviewees; examples of unwarranted disparity attributed to guideline); Peter Reuter & Jonathan P. Caulkins, *Redefining the Goals of National Drug Policy: Recommendations from a Working Group*, 85 AM. J. PUB. HEALTH 1059, 1062 (1995). Reuter and Caulkins reported the recommendations of a RAND corporation working group, which concluded:

> Federal sentences for drug offenders are often too severe: they offend justice, serve poorly as drug control measures, and are very expensive to carry out. . . . The U.S. Sentencing Commission should review its guidelines to allow more attention to the gravity of the offense and not simply to the quantity of the drug.

Id.

[150] Frank O. Bowman, III, *The Quality of Mercy Must Be Restrained, and Other Lessons in Learning to Love the Federal Sentencing Guidelines*, 1996 WISC. L. REV. 679, 747-49 (1996).

[151] Many have attributed these problems to the drug guidelines' emphasis on drug quantity. They argue that quantity often fails to discriminate among more and less culpable defendants and is based on manipulatable, arbitrary, or unreliable factors. *See* Judicial Conference of the United States (JCUS), *1995 Annual Report of the JCUS to the U. S. Sentencing Commission* (Mar. 1995), at 2 ("[T]he Judicial Conference . . . encourages the Commission to study the wisdom of drug sentencing guidelines which

Given this criticism, it is not surprising that pressures from the Department of Justice and from the Commission to apply the drug guidelines strictly and uniformly have met with resistance. It seems likely that this resistance has taken different forms in different places and that, as a consequence, regional disparity has increased. To compensate for the new controls from Washington, new types of discretion have been invented, in the form of charging and declination policies,[152] novel plea bargains such as fact bargains,[153] "hidden" departures,[154] and other arrangements. These bring back into the system some of the flexibility that judges, attorneys, and probation officers think is needed to keep the system working and to avoid unfairness. Ironically, the proliferation of new rules in the past ten years has created more ways that cities might differ from one another in how they deploy, and in some cases circumvent, the rules.

Based on anecdotal reports and some previous research, it seems likely that different adaptations in different cities accounts for the growth of inter-city disparity under the guidelines.[155] While inter-judge disparity for drug cases has decreased, differences among cities are creating more regional

---

are driven virtually exclusively by the quantity or weight of the drugs involved. The Conference commends the Commission's efforts to study the drug guidelines with new eyes, based on the experience of the past several years."). *See also* Hon. William W. Wilkins et al., *Competing Sentencing Policies in a "War on Drugs" Era*, 28 WAKE FOREST L. REV. 305, 320-24 (1993).

[152] U.S. Attorneys have discretion to decline to prosecute. In many cases concurrent federal and state jurisdiction provide prosecutors with a choice of forum, which often has profound consequences for the offender's exposure to punishment. *See* Richard Berk & Alec Campbell, *Preliminary Data on Race and Crack Charging Practices in Los Angeles*, 6 FED. SENTENCING REP. 36, 38 (1993) (showing disparity between state prosecutions and federal prosecutions for the same crime). Some U.S. Attorney's offices simply do not prosecute drug couriers or mules unless they are repeat offenders or involved with large amounts of drugs judged sufficient to make the case appropriate for federal court.

[153] *See generally* Nagel & Schulhofer, *The First Fifteen Months*, *supra* note 69, at 272-78; Berman, *supra* note 72.

[154] Freed, *supra* note 59, at 1723-24.

[155] *See generally* Nagel & Schulhofer, *supra* note 56; Heaney, *supra* note 79, at 556 (providing the best descriptions of the variety of procedures and local practices that have developed under the guidelines and examples of how this variety could result in disparity).

disparity than existed before the guidelines. And because sentences have generally gotten much longer in drug cases, variation in the use of the new types of discretion have dramatic effects on sentences. Especially given that the drug statutes and guidelines are also implicated in the growth of the gap between sentences for African-American, white, and Hispanic offenders, these variations in application raise special concerns.[156]

D. IMPROVING THE FEDERAL SENTENCING SYSTEM

The guidelines' mixed record helps account for why there is still no consensus as to whether they have been successful. We expect the mixed findings from this study to contribute, but not to end, the debate. Rather than argue about whether the glass is half empty or half full, we prefer to ask: What accounts for the failure of the guidelines to reduce disparity for some types of offenses and in some cities? What changes might make this goal attainable?

Research can help explore why the guidelines are not working everywhere.[157] But based on what we already know today, we do not believe uniform sentencing can be accomplished simply by insisting that the rules be followed more closely. In theory, disparity would be eliminated if prosecutors always charged fully, never bargained away counts or facts that could be proven, and if judges always strictly applied the guidelines to the facts and departed only in extraordinary cases. But in practice, major reforms can seldom be dictated from above if they result in outcomes that appear unworkable or unjust to the persons asked to implement them. For the goal of uniformity to be more fully achieved, judges should accept the policies of Congress and the Commission, even if they conflict with their own philosophy. But Congress and the Commission should also in-

---

[156] *See supra* Part I.C.1.

[157] The natural experiment methodology developed for this study may provide more answers. For example, future analyses could explore whether departure rates and the extent of departure are similar among judges in a given city. Comparative case studies of individual districts, showing connections between various prosecutorial and judicial practices, also appear promising. *See* Lisa M. Farabee, *Disparate Departures Under the Federal Sentencing Guidelines: A Tale of Two Districts*, 30 CONN. L. REV. 569 (1998).

form those policies with the experience of prosecutors and judges who know first-hand the varying circumstances of crimes and the individual characteristics of offenders.