Journal of Criminal Law and Criminology

Volume 90
Issue 1 *Fall*

Article 6

Fall 1999

Technical Appendix

Recommended Citation

Technical Appendix, 90 J. Crim. L. & Criminology 307 (1999-2000)

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

TECHNICAL APPENDIX

The basic logic and methodology used in this study is described in Part III.B. This appendix is for technically-oriented readers who are interested in more detail concerning the databases, research design, statistical models, and computer packages used in the study, and in how various indices were calculated.

I. SENTENCING AS A UNITARY OR BIFURCATED DECISION

Researchers must decide whether to treat sentencing as a unitary or bifurcated decision. Those taking the bifurcated approach treat it as two separate decisions. First, the judge decides whether to incarcerate a defendant or use probation or an alternative sanction. Then, for those defendants who are incarcerated, the judge decides how long a term of imprisonment is appropriate. The two stages are modeled separately. Logistic regression is used to analyze the imprisonment decision for all offenders. Least squares multiple regression is used to model the decision regarding sentence length only for those offenders who receive a sentence of imprisonment.[158]

The bifurcated approach has an important advantage: it does not assume that the same set of factors affect both decisions equally. Empirical analysis suggests that many factors, such as employment, affect one decision but not the other.[159] For example, judges may be influenced by a defendant's potential job loss in deciding whether to incarcerate, so the employment factor would receive some weight in that model. But if the decision is made that incarceration is necessary, employment may have little influence on the decision concerning how long to imprison. Separate models for the in/out and imprisonment length decisions permit more precise weighting of the factors that affect each particular decision.

[158] *See, e.g.*, MCDONALD & CARLSON, *supra* note 18, at 59.

[159] RICH ET AL., *supra* note 16, at 82.

Several considerations prevented us from using this standard bifurcated approach, however. First, our model—with judges nested within cities and interactions among judges and offense types—has a very large number of parameters. The iterative calculations needed for the maximum likelihood estimates in logistic regression exceeded the capacity of the computer resources available at the Commission and at a large research university.[160] Second, excluding offenders who received no prison time from the second analyses concerning imprisonment length would destroy the comparability of caseloads needed for our analysis. Because judges differ in the proportion and types of cases wherein the defendant is sent to prison, studying sentence lengths only among the imprisoned caseloads would make the caseloads among different judges incomparable.

Treating the decision as a unitary decision permits us to include all offenders in the analysis, but it also raises new issues. Researchers taking a unitary approach have viewed the sentence imposed as a reflection of the judge's ranking of the case on an unobservable unitary dimension—the severity of punishment deserved by the offender. Imprisonment, as opposed to a non-incarcerative alternative, results when a case falls above a certain threshold level on this unobserved severity scale. This view has led researchers to adopt one of several research strategies for modeling the sentencing decision as a single decision.

Some studies attempt to translate the variety of observable sentences, such as fines, probation, or various months of imprisonment, into values on a single severity scale. Statistical analysis then identifies factors that predict where on this scale particular cases fall. Severity scales have been constructed for use with federal cases.[161] But several problems with this approach are apparent. Validation of the scales has been difficult, and severity ratings are not as meaningful to policymakers as are

[160] Discussion with representatives of the SAS Institute confirmed that the complexity of the model made the calculation very resource-intensive.

[161] *See* MICHAEL HINDELANG ET AL., SOURCEBOOK OF CRIMINAL JUSTICE STATISTICS—1974, at 401 (1975); L. PAUL SUTTON, FEDERAL CRIMINAL SENTENCING: PERSPECTIVES OF ANALYSIS AND A DESIGN FOR RESEARCH 23 (1978).

observable and concrete decisions, such as whether to imprison and for how long. For these reasons, we chose to use months in prison as our unitary outcome measure.

Using months of imprisonment as the outcome raises an additional issue. Since cases receiving probation or other alternatives receive zero months of imprisonment, this may represent a truncation of the true underlying severity scale. Some probation cases deserve more severe punishment than others, but these differences are not reflected in the observable measure. Treating all non-prison sentences as zero months leads to a form of selection bias that can distort the weights assigned to the predictive factors in the model.[162]

Several methods have been proposed to correct for this distortion, but all have problems of their own. TOBIT analysis is commonly used to study sentencing as a unitary decision, but it requires that identical factors affect both the imprisonment and sentence length decisions.[163] Furthermore, examination of data on sentences imposed for federal offenses does not support the use of a TOBIT distribution.[164] Other methods have been developed in the econometric literature,[165] but these require extensive *a priori* specification of the factors likely to affect each decision, which often is impossible.[166]

Given these countervailing considerations, we chose to conduct a unitary analysis on the entire population of offenders treating sentences of probation or any form of non-prison alternative, including home or community confinement, as zero months imprisonment. Given that our concern is not to establish the weights of an exhaustive list of factors that explain sentence length, but only to examine changes in the influence of

[162] *See generally* Jeffrey A. Dubin & Douglas Rivers, *Selection Bias in Linear Regression, Logit and Probit Models*, *in* MODERN METHODS OF DATA ANALYSIS 410 (John Fox & J. Scott Long eds., 1990).

[163] *See generally* Steven Klepper et al., *Discrimination in the Criminal Justice System: A Critical Appraisal of the Literature*, *in* 2 RESEARCH ON SENTENCING: THE SEARCH FOR REFORM 55 (Alfred Blumstein et al. eds., 1983).

[164] *See, e.g.*, MCDONALD & CARLSON, *supra* note 18, at 59.

[165] *See generally* G.S. MADDALA, LIMITED-DEPENDENT AND QUALITATIVE VARIABLES IN ECONOMETRICS (1983) (switching regression); James J. Heckman, *Sample Selection Bias as a Specification Error*, 47 ECONOMETRICA 153 (1979) (sequential equation model).

[166] *See* Rhodes, *supra* note 7, at 1031 n. 47.

one such factor—the identity of the sentencing judge—over time, we believe this approach is adequate and valid for our limited purposes.

II. DEFINITION OF THE VARIABLES

A. DATABASES

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.[167]

The FPSSIS file contains information obtained by probation officers and organized into computer data files by the Administrative Office of the U.S. Courts. This file includes a description of the offense, information about the defendant's background and criminal history, the disposition of the case, and the sentence imposed. The FPSSIS database was maintained for the years 1984-1990. The Sentencing Commission performed additional reliability checks on the data obtained from the Administrative Office.

The Monitoring data file is based on data collected to meet the Sentencing Commission's statutory obligation to disseminate information regarding federal sentencing practices.[168] The Commission asks probation officers to submit five documents for each defendant sentenced under the guidelines: the indictment, the presentence report, the written plea agreement, the statement of reasons for imposing the sentence, and the judg-

[167] 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, *supra* note 121.

[168] 28 U.S.C. § 995(a)(12), -(15), -(16) (1994).

ment of conviction order. From these documents staff at the Commission code demographic information about the defendant, the offense of conviction, the applicable guidelines and final sentence, and the reasons for any departure from the guidelines.

B. EXPLANATORY VARIABLES

The same explanatory control variables were used in both analyses: offense type, criminal history, city, and judge. Because we wanted the results from the two time periods to be strictly comparable, we used only variables that could be measured precisely the same way in the two datasets. Offenses were categorized into twenty-five general types. The offenses are: Murder, Kidnaping/Hostage, Sexual Abuse, Assault, Robbery (including Bank), Drug Trafficking, Firearms, Burglary/Breaking and Entering, Auto Theft, Larceny, Fraud, Embezzlement, Forgery/Counterfeiting, Bribery, Tax Offenses, Racketeering, Gambling/Lottery, Civil Rights, Immigration, Pornography/Prostitution, Prison Offenses, Administration of Justice, Environmental Offenses, National Defense, and "Other offenses." There are, of course, variations in the seriousness of offenses within each of these types, which, if measured, would undoubtedly increase the amount of sentence variation accounted for by offense type.

For some analyses, only selected subsets of these offenses were used in order to reduce the number of cells in the three-way (offense type x criminal history x judge (city)) factorial design that had zero or small numbers of cases. Offenders were categorized into two criminal history groups—those with and those without any previous convictions. Due to the limited information available in the 1984/1985 dataset on the nature or the prior record, no more precise measurement of prior record was possible. In 1984-1985, 51% of the defendants had no prior criminal record; in 1994-1995, 48% had no prior criminal record.

The judge variable was treated as nested within cities, since judges sit in only one city (i.e., the judge factor does not cross the city factor). The judge variable was defined as a categorical variable with each judge given a separate value. This is equivalent to creating a dummy variable for each judge, and this is in fact what the GLM procedure that was used for data analysis does automatically.[169]

C. OUTCOME VARIABLES

The unitary outcome variable, for reasons described above, was months of imprisonment imposed. In cases of split sentences, the length of imprisonment was the portion of the sentence that was spent behind bars. In accordance with procedures used in other Commission research, life sentences were treated as 470 months, which is 65 months longer than the highest within-guideline imprisonment sentence and approximates the average life expectancy of federal defendants receiving sentences of life.

Because a sizeable portion of defendants receive zero months of imprisonment and some receive life in prison, the outcome variable distribution is W-shaped rather than normal and shows some right "skewness." Analysis of variance assumes a normal distribution, but tests of normality performed on our data suggest that the effects of these violations are likely to be slight.

We conducted some analyses[170] using as outcome the expected time to be served. This was estimated with algorithms tailored to the conditions of each time period. For the pre-guidelines cases, the expected time to be served was based on a simulation of the parole guidelines, which were the mechanism for calculating the time to be served for pre-guideline defendants. For the post guideline defendants, sentences of greater than one year were reduced by the fifty-four days of good time that can be earned for such sentences in a year. The calculations for both time periods were based on the assumption that the defendant would receive all good time available.

[169] *See* SAS Institute, Inc., SAS/STAT USER'S GUIDE, Release 6.03, at 7.

[170] *See supra* note 125.

III. STATISTICAL MODELING AND COMPUTER ANALYSIS

A. ANALYTICAL PROCEDURE FOR THE SENTENCE LENGTH DECISION

Analyses of the sentence length decision were conducted using the General Linear Model (GLM) procedures available in SAS. GLM is a versatile package that can be used for analysis of variance (ANOVA), multiple regression, and other analyses using the method of least squares to fit linear models. GLM allows users to specify a variety of models and provides sums of squares and mean squares for each source of variation specified in the model. Other statistics of interest, such as the proportion of the total variance accounted for and the effect size associated with a factor, can be calculated from the sums of squares. In addition, GLM performs an F-test of the statistical significance of each effect in the model.

Because our design is a "natural experiment" with random assignment of cases to judges in each city, we find it easiest to conceive of the statistical analysis as a fixed-effects ANOVA with judges nested within cities and offense type and criminal history treated as categorical "covariates" (used as control variables). GLM is the preferred SAS procedure for ANOVA with unbalanced designs, that is, designs with unequal numbers of cases in the cells, which is unavoidable in a natural experiment. Through the concept of estimability, GLM can test various hypotheses regardless of the amount of confounding or missing data.[171] One can specify which variables are nested and can include interaction terms for any factors that are crossed.

The Type I (or "sequential") sums of squares output provided by GLM is the incremental improvement in the error SS as each factor is added to the model, in the order specified in the MODEL statement.[172] The analysis is identical to a multiple regression using a hierarchical analytical strategy, with control variables entered first, city entered next, the main effect for

[171] SAS Institute, *supra* note 169, at 584.

[172] *Id.* at 586.

judges entered next, and the interaction terms entered last.[173] The random assignment ensures there is no correlation among control variables and judges within each city, although there is some intercorrelation among control variables and cities.

Lindman gives the hierarchy of effects that is appropriate for an ANOVA with both nested and crossed factors.[174] The offense type by city and the judge(city) effects are at the same level in the hierarchy. We placed all main effects, including the judge(city) effect, before any interaction. No interaction terms were included for the criminal history factor because it was associated with such a small effect. The test of the null hypothesis for each factor is uncontaminated by effects preceding the factor being tested. The exact model was specified as follows: Months-in-prison = crimhist + offtype + city + offtype*city + judge(city) + offtype*judge(city)

B. RELATIONSHIPS AMONG THE VARIABLES AND THE ROLE OF UNMEASURED VARIABLES

Intercorrelation of explanatory variables must be carefully considered when structuring a multivariate statistical model and interpreting the results. For example, because defendants convicted of burglary offenses are more likely to have a criminal record than those convicted of tax offenses, determining whether offense type or criminal history accounts for longer average sentences for burglars is difficult. As shown in the tables in the text, our explanatory factors together account for between 34-39% of the variation in sentences, depending on the time period. About 4% of that is a shared variance that cannot be unambiguously assigned to a single explanatory factor.

The amount of shared variance was determined by comparing the sum of square attributable to the entire model with the sum of the Type II sums of squares for the individual factors, obtained from the SAS GLM output. The Type II sum of squares is the reduction in error attributable to each factor after all other factors have been taken into account. It is equivalent

[173] *See* JACOB COHEN & PATRICIA COHEN, APPLIED MULTIPLE REGRESSION/CORRELATION ANALYSIS FOR THE BEHAVIOR SCIENCES § 3.8.1 (2d ed. 1983).

[174] *See* LINDMAN, *supra* note 128.

to the square of the semi-partial correlation coefficient for the factor. The difference between the sum of these for all factors combined and the model sum of squares is the portion of the total explained variance that cannot be unambiguously attributable to a particular factor.

We structured the hierarchical model so that the legally-relevant factors were "credited" with explaining as much variation in sentences as possible. The effects of city and judge were calculated after adjusting for the part of the variation in sentences that could be accounted for by offense type and criminal history. 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.

The variance accounted for by our explanatory factors should be considered a conservative estimate of the influence of these factors in sentencing, because the limitations of measurement and statistical modeling inevitably introduce error. A factor with an effect size of four percent, though accounting for only 4% of the total variance in our outcome, represents over 10% of the variation that we can explain with our model. Some of the 60+% of variation in sentences that remains unaccounted for—an amount that decreases only marginally under the guidelines—may represent arbitrariness and disparity. But most of this amount probably reflects factors that we did not measure, such as other elements of offense seriousness or offender culpability. Other multivariate studies by Commission researchers that used a more complete set of explanatory factors have accounted for up to 70% of the variation in sentences under the guidelines.

Since judges are nested within cities in our model, there is a confound among judges and cities that must be kept in mind when interpreting results. Some differences in sentences that arise from diverse judicial philosophies will appear in our model as differences among cities. For example, if the judges in a particular city share a more lenient philosophy than in other cities, this city-wide philosophy is reflected in the city average and measured by the city effect, not by the judge effect (even though the individual judges might carry this leniency with them if they were assigned to a different city). The judge effect reflects differences among judges relative only to the other judges in the same city, and is thus a conservative estimate of the influence of judicial philosophy on sentencing nationwide. The city effect reflects differences among cities in shared judicial philosophies and also reflects local norms, practices, etc., that would impinge on the decision-making of any judge assigned to the city. It also, of course, reflects differences in the caseloads of various cities, for which we partially control with our offense type and criminal history independent variables.

In early analyses we included as control variables the race, gender, and age of offenders. These factors accounted for statistically significant but small amounts of variation. Interpretation of their effects is very difficult, however, because the factors are intercorrelated with each other and with offense type and criminal history. Most important, demographic characteristics are correlated with unmeasured factors that legally influence sentences. For example, race of the defendant is correlated with type of cocaine. (Our data for 1984-1985 do not distinguish powder from crack cocaine). Since larger penalties are imposed on crack cocaine, this difference in the proportion of African-American offenders involved with crack appears in our data as a race effect. A richer range of control variables is needed to adequately assess the independent roles of race, gender, or age in sentencing, which is not possible in a pre/post study in which variables must be measured identically at both time periods. Research on discrimination today, using a rich source of control variables, is presently underway at the U.S. Sentencing Commission.

C. MEASURES OF THE JUDGE EFFECT AND CHANGES BETWEEN THE TIME PERIODS

The comparison of greatest interest for this study is the effect associated with judges at the two time periods, after controlling for differences in case characteristics and other inter-city variation. To assess the size and importance of the judge effect in the analysis of sentence length, we calculated the unbiased estimate of the proportion of the total variance in sentence lengths associated with each factor, according to the procedure described in Lindman.[175] This is an appropriate measure of the importance of a factor in ANOVA. (The square root of this statistic is analogous to the semi-partial correlation coefficient associated with the factor in a hierarchical regression.) In discussing this proportion, we converted it to a percentage—the percentage of the total variance in sentences that can be attributed to judges. Only percentages greater than 1% were considered meaningful.[176] The effectiveness of the guidelines was evaluated by subtracting the percentage obtained in 1994-1995 from the percentage in 1984-1985.

D. LIMITATIONS OF OUR TEST OF CHANGES IN THE AMOUNT OF INTER-JUDGE DISPARITY

Because our method is not a true experiment (permitting control over the number of cases in each experimental condition), no test of the statistical significance of the difference in the R-squared between the two time periods was available. A model that would provide significance tests—a within-judge repeated-measures design—requires "orthogonality" of the experimental factors, i.e., that the number of cases for each judge at each time period be equal, or at least that the number of cases be proportional.[177] This condition is not met with our data. Some progress has been made in developing significance tests with unbalanced, nested data,[178] but these were not available for use as part of this project. Thus, no test of the statistical significance of changes in the judge effect at the two time periods was possible.

Finally, we are aware that more structured statistical models are possible, some of which would permit a test of the significance of changes in inter-judge disparity and overcome a bias in our estimates of the judge effect.[179] The model used for our

[175] *Id.* at 38-41.

[176] *See* Jacob Cohen, *A Power Primer*, 112 PSYCH. BULL. 155, 155-56 (1992).

[177] *See* WILLIAM L. HAYS, STATISTICS (3rd ed. 1981).

[178] *See generally* ANTHONY S. BRYK & STEPHEN W. RAUDENBUSH, HIERARCHICAL LINEAR MODELS: APPLICATIONS AND DATA ANALYSIS METHODS (1992).

[179] Letter from Jeffrey R. Kling, National Bureau of Economic Research, Cambridge, Mass. (Dec. 9, 1997). Interview with Jeffrey R. Kling (Jan. 29, 1998).

analysis may underestimate the size of the primary judge effect if sampling bias plays a role in our estimate of the judge means. The result of this bias, however, is that we may underestimate the amount of inter-judge disparity and the degree of improvement brought about by the guidelines. We believe that the general finding of our study, however, still holds. We hope that refinements of statistical modeling will continue and a more precise picture of the effects of the guidelines will emerge.

IV. TESTING THE RANDOMNESS OF CASE ASSIGNMENT

Random assignment has been considered an important mechanism that prevents attorneys from "shopping" for judges most likely to be favorable to their case. For this reason, courts strive for truly random assignment, although they may use other blind and case-independent procedures that, although technically not random, approximate it for practical purposes. A computer program is even available to the federal courts for unbiased case assignment. But ethical and logistical requirements sometimes limit the principle of randomness. For example, judges cannot hear cases in which they have some personal or family affiliation with one of the attorneys. Cases that are related (e.g., all defendants in a drug conspiracy) may be assigned to a single judge rather than randomly distributed.

In general, however, a reasonable approximation of randomness is attempted in most cases and in most districts. At an April 1996 Federal Judicial Center seminar attended by the chief judges of eighty-four of the ninety-four federal judicial districts, sixty-nine (82%) said case assignment is always random. Ten (12%) said it is random most of the time, with rare exceptions, one (1%) said it is usually random, but with frequent exceptions, and four (5%) said it is never random.

Because the assumption of caseload comparability is fundamental to the logic of our study, we identified and excluded from subsequent analyses those districts that did not have random assignment at either of the two time periods of our study. We began with interviews. In 1989, researchers from the Federal Judicial Center visited all district courts and interviewed the court clerks.[180] Clerks were asked how cases were assigned to judges. Most reported that cases were assigned randomly or "rotationally," meaning that cases were assigned to the next judge on a list in the order that they were filed. For the court's purposes, and for ours, the rotational method seems a reasonable approximation of random assignment. In April and May 1996, we called court clerks in cities that met our initial criteria for inclusion and determined that most continued to use random or quasi-random assignment. While most districts appeared to use assignment procedures that appeared random, in some cases it was difficult to tell from the information reported. In addition, it was possible that some districts that reported random assignment might in fact have violated their procedures. For these reasons we determined that the proper criteria should be a statistical test.

A. STATISTICAL TESTS OF RANDOMNESS

In his pilot study, Waldfogel used standard statistical techniques to test whether the proportion of male and female offenders assigned to each judge were within the range that is expected if cases are in fact randomly assigned. He found that in each of the courts, judges received "approximately the same fraction of cases with female defendants." Waldfogel used gender of the defendants for these tests because he reasoned that characteristics such as offense type can be changed (through superseding indictments and charge dismissals) after the case is assigned. Because the judge might influence these processes, he thought it was better to use an immutable defendant characteristic like gender to test whether the proportion of cases was consistent with random assignment.

We also used statistics to test the assumption of random assignment, but we used race of offender rather than gender. The overall percentage of females in the federal system is small, about 15% in 1995, resulting in little chance for wide variation in the percentage of female defendants across judges. We constructed two-dimensional tables, with race on one dimension

[180] Interview protocol, FJC Time Study (on file at the Federal Judicial Center).

and judges on the other. If case assignment were random, the proportions of the racial groups would be about the same for each judge. In statistical terms, the two dimensions would be "independent."

Several statistical tests were used to assess the independence of the two dimensions. Chi-square analyses were conducted to compare the expected distribution of cases with the observed distribution and estimate the likelihood that the observed distribution could occur through random assignment. For small cities, with 300-500 cases per year, the chi squares were indeed nonsignificant. With large cities with more than 500 cases per year, however, the chi squares were significant at the .001 level. (Results of these analyses are available from the authors.)

A problem with the chi square is that it is very sensitive to sample size. Small effects become significant if the number of cases is very large. Thus, the chi-square analysis essentially identified and eliminated all large districts, even though the departures from randomness were minimal and similar to departures in smaller cities. To avoid this problem, we decided to measure the effect size using an R-squared analog. David Knoke and Peter Burke provide an R-squared analog for log-linear analyses, which we computed for each courthouse in our sample using the SAS CATMOD procedure.[181] We compared the baseline model (the geometric mean) with the alternative model (which included the main effects for offense type and judges). For both the baseline model and this alternative model, we computed the L^2 statistic using the formula $L^2 = 2\ f_{ij} \ln (f_{ij}/F_{ij})$, where f_{ij} is the observed cell frequency and F_{ij} is the expected cell frequency.[182] With these two L^2 values, we then computed the R^2 analog using the formula R^2 analog = [(L^2 baseline model) - (L^2 alternative model)]/(L^2 baseline model).[183] A model whose R^2 analog is greater than .90 may be presumed to be sufficiently close to a random distribution that systematic variation in case assignments to each judge may be ruled out. This was the criterion we used for including cases in the analy-

[181] DAVID KNOKE & PETER J. BURKE, LOG-LINEAR MODELS 40-42 (1980).

[182] *Id.* at 30.

[183] *Id.* at 41.

ses. Twenty-seven cities failed our test at one or both time periods and were eliminated from subsequent analyses.

B. MULTIPLE DEFENDANT CASES

Finally, we considered the effects of an additional limitation on random assignment. Our analysis concerns sentences imposed on *individual* defendants. But it is *cases*, not individuals, that are assigned randomly. Many cases contain multiple defendants and all defendants in a case are sentenced by the same judge. In the pre-guideline years covered by our analysis, 38.2% of all cases involved multiple defendants, and 55% of drug cases did. The percentages in the guideline years were similar, 35.6% of all cases and 53.4% of drug cases.

In cities where cases are randomly assigned, each judge will get a similar share of multiple defendant cases. The concern, however, is that our statistical test of differences between judges assumes random assignment of individuals.[184] We need to test the possibility that differences we uncover among judges might be due to some judges getting cases with multiple defendants that are much more or less serious than average.

To assess this possibility we performed several analyses using only single-defendant cases and compared these results with the analyses for all cases.[185] If multiple defendant cases are causing or exaggerating differences in average sentences among judges, we would expect the R-squared for judges to be smaller for single defendant cases than for all cases combined. Comparing the results of the two analyses shows no such pattern. In fact, the R-squared for judges are more often larger in single defendant cases than in multiple defendant cases. Thus, in our main analyses we combined single and multiple defendant cases.

[184] Technically, the ANOVA F-test used to assess the statistical significance of any differences found among judges assumes random assignment of individual cases. If assignment is not fully random, somewhat larger differences between judges might be expected to arise by chance. The p-value will then be underestimated and we could falsely fail to reject a true null hypothesis.

[185] Analyses on file with the authors.