MEASURING INTER-JUDGE SENTENCING DISPARITY

BEFORE AND AFTER THE FEDERAL SENTENCING GUIDELINES

James M. Anderson
Defender Association of Philadelphia

Jeffrey R. Kling
Princeton University and NBER

Kate Stith
Yale Law School

Discussion Paper #207

December 1998

Discussion Papers in Economics
Woodrow Wilson School
Princeton University
Princeton, NJ 08544

MEASURING INTER-JUDGE SENTENCING DISPARITY BEFORE AND AFTER THE FEDERAL SENTENCING GUIDELINES*

James M. Anderson
Defender Association of Philadelphia

Jeffrey R. Kling
Princeton University and NBER

Kate Stith
Yale Law School

December 1998

ABSTRACT

This paper evaluates the impact of the Federal Sentencing Guidelines on inter-judge sentencing disparity, which is defined as the differences in average nominal prison sentence lengths for comparable caseloads assigned to different judges. This disparity is measured as the dispersion of a random effect in a zero-inflated negative binomial model. The results show that the expected difference between two typical judges in the average sentence length was about 17 percent (or 4.9 months) in 1986-87 prior to the Guidelines, and fell to about 11 percent (or 3.9 months) from 1988-93 during the early years of the Guidelines. We have not sought to measure the effect of parole in the pre-Guidelines period, other sources of disparity such as prosecutorial discretion, or the proportionality of punishment under the Guidelines as compared with the pre-Guidelines era.

JEL Classification: K4, C5

Keywords: Interjudge sentencing disparity, Federal Sentencing Guidelines
Zero-inflated negative binomial random effect model

"Simply stated, unwarranted disparity caused by broad judicial discretion is the ill that the Sentencing Reform Act seeks to cure."[1]

I. INTRODUCTION

One of the chief objectives of the Sentencing Reform Act of 1984[2] was to reduce sentencing disparity among similar offenders. The Act described this purpose as "avoiding unwarranted disparities among defendants with similar records who have been found guilty of similar criminal conduct."[3] Both the Act and its legislative history demonstrate that Congress' overriding concern was to reduce disparity thought to result from the exercise of judicial discretion in sentencing. The Act was sponsored and shepherded through Congress by an unusual coalition of liberals and conservatives. Liberals expressed particular concern that permitting the exercise of discretion compromised the ideal of equal treatment under the law, while conservatives were concerned as well with perceived undue leniency in sentencing.[4]

To accomplish these ends, the Act created the United States Sentencing Commission to develop federal Sentencing Guidelines. These Guidelines, which became effective on November 1, 1987, restrict the exercise of judicial discretion to a narrow sentencing range in each case and limit judicial departures from that range. The sentencing ranges of the Guidelines are a small percentage of the statutory ranges[5] that were available to judges in the pre-Guidelines era; it was presumably hoped that sentencing disparity would be reduced commensurately.

This paper examines the impact of the Guidelines on one particular type of disparity: inter-judge disparity in the average length of prison sentences of criminal defendants in federal district courts. After a brief review of the Guidelines and the mechanisms through which they may affect disparity, we explore methods of measuring inter-judge disparity.

Using the fact that cases are randomly assigned to judges in many districts, we define inter-judge disparity as relative differences in the average sentence length for otherwise comparable caseloads of defendants assigned to different judges in the same district. We compare estimates of inter-judge disparity before and after implementation of the Guidelines.[6] In our preferred specification, the dispersion in the judge's effect on sentence length is represented by the variance of a judge-specific random variable in a statistical model of prison sentence length. Existing econometric random effects models for count data (such as the number of months of prison sentence length) are extended in several ways. A zero-inflated negative binomial model is developed in order to account explicitly for the fact that many cases end in dismissal or acquittal or have a sentence that involves no prison time at all. A parametric model of the random effect allows the judge effects to be correlated over time and allows a convenient interpretation of the dispersion in terms of a Gini coefficient—or twice the expected absolute difference in sentence length between two judges in the same district office relative to the office mean.

While noting that other kinds of disparity may have been exacerbated by the Guidelines and there may be unwarranted uniformity in sentencing under the Guidelines, we conclude that *inter-judge* disparity in nominal sentencing is less pronounced in the Guidelines era than it was in the era of discretionary sentencing.

II. Defining Inter-judge Disparity

The enabling legislation and legislative history of the Sentencing Reform Act refer to reducing unwarranted disparity as "the major premise of the sentencing guidelines."[7] Neither in

statute nor legislative history did Congress define or explain what constituted "unwarranted disparities among defendants with similar records, who have been found guilty of similar criminal conduct."[8] As the Senate Report accompanying the Sentencing Reform Act noted, "The key word in discussing unwarranted sentence disparities is 'unwarranted.'"[9]

To avoid confusion, we distinguish between three distinct types of sentencing variation: *proportionality*, *disproportionality*, and *disparity*. Proportionality, under our definition, is sentencing variation among a set of decision-makers in the criminal justice system that is justified by relevant differences among offenders and their crimes. Conventionally, these differences include various characteristics of the criminal offense (such as the amount of harm caused) and of the criminal offender (such as prior criminal record). Its converse, disproportionality, is any variation in sentencing outcomes for a given set of decision-makers that is not attributable to relevant sentencing factors; in this sense, disproportional variation is akin to what many previous commentators have referred to as disparity. In contrast, we define disparity as variation in sentencing between the sets of hypothetical decision-makers that could potentially be involved in the disposition of an offender's case. Under this approach, disparity can be thought of as the variation in sentence that would result if a single offender were processed through the criminal justice system by every possible combination of sentencing decision-makers.

Our definition of disparity is centered on the sentencing decision-maker, rather than on characteristics of the offense or the offender. Thus, for example, the fact that sentences for equal amounts of crack-cocaine and powdered-cocaine are dissimilar is not disparity under our definition. The crack/powder difference is either proportional variation or disproportional

variation, depending on one's judgment as to whether the difference in the type of cocaine is relevant to the proper amount of criminal punishment. Similarly, some have argued that the sentencing Guidelines may have increased overall penalty variation among offenders because, before the passage of the Guidelines, judges could consider the reputational consequences and the loss of potential earnings suffered by white collar offenders.[10] Under our definition, this type of variation is not "disparity". It is, rather, a form of disproportionality (or proportionality, depending upon one's views about whether reputational and earnings consequences should be considered by sentencing judges). We have adopted a definition of "disparity" that considers solely that variation caused by the identity of the decision-maker; the concern about "white-collar inequity," on the other hand, concerns variation in total punishment among different classes of offenders. The definitions of disparity, proportionality, and disproportionality that we have adopted permit us to distinguish variation that is attributable to the identity of the decision-makers in the criminal justice system from variation in sentences based on differences between criminal cases (that may or may not be justified depending on one's theory of the purposes of sentencing).

We measure only inter-judge disparity and do not attempt to gauge potential disparity at other stages in the sentencing process or potential disproportionality. We limit our inquiry for several reasons. First, the difficulties are formidable in rigorously measuring disparity in other stages in the process, despite its likely presence.[11] Second, in focusing on the variation attributable only to disparity between judges rather than on disproportionality we avoid the inevitably contentious issue of the purpose of sentencing itself. Finally, inter-judge sentencing disparity has generated more concern than disparity at any other stage of the sentencing process.

Concern about inter-judge sentencing disparity is not hard to understand. Attorney-General Robert H. Jackson, pithily expressed the intuitive unfairness of inter-judge disparity:

> It is obviously repugnant to one's sense of justice that the judgment meted out to an offender should depend in large part on a purely fortuitous circumstance; namely the personality of the particular judge before whom the case happens to come for disposition.[12]

Prior to the promulgation of the Sentencing Guidelines, a federal judge's sentencing discretion was enormous and virtually unreviewable. As the last actor in the determination of the offender's formal sentence, the judge was in a position either to remedy or to exacerbate any disparity in the earlier stages of criminal prosecution. In addition, the judge has by far the most visible role in the sentencing process, formally announcing the polity's exaction of punishment. Variation in law-enforcement, charging, and probation practices are far less visible.

By the early 1970s, considerable intellectual enthusiasm for the idea of reducing sentencing disparity developed as part of a wave of general efforts to attack indeterminate sentencing.[13] The theory that punishment was designed to rehabilitate the offender had become discredited, and both parole and inter-judge disparity came under attack. Perhaps the most influential critic of judicial sentencing discretion was Marvin E. Frankel, himself a distinguished judge in the Southern District of New York.[14] Frankel argued that the range of choice provided to the sentencing judge was "terrifying and intolerable for a society that professes devotion to the rule of law," and that it should be "unthinkable in a 'government of law, not of men.'"[15] He wrote, "[I]ndividualized justice is *prima facie* at war with such concepts, at least as fundamental, as equality, objectivity, and consistency in the law."[16]

Frankel's anecdotal arguments appeared to be confirmed by an experimental study that he

helped to organize in the district courts comprising the Second Circuit. The organizers of the study distributed identical pre-sentence reports to fifty district court judges; each judge was asked to impose sentences for each case. Sentencing ranges varied widely, in one case from 20 years in prison and a $65,000 fine from the most severe judge to three years in prison from the most lenient judge. Moreover, the disparity was not attributable to either a handful of judges or outliers in each case but appeared throughout the range of sentences.[17]

The Second Circuit study's finding of substantial sentencing disparity was repeated in other studies.[18] Commentators also cited prison unrest[19] and racial and class discrimination[20] as problems deriving from the exercise of judicial discretion in sentencing. Critics from the political right expressed dissatisfaction with the perceived leniency of sentencing judges and parole officials.[21] Both liberals and conservatives argued that sentencing disparity compromised the ideal of equal treatment under law. By the early 1980s, there had developed across the ideological spectrum a consensus that dramatic changes in the sentencing process were needed to reduce sentencing disparity stemming from the exercise of judicial discretion.

III. Guidelines as Mechanism for Reducing Inter-judge Disparity

The result of this political consensus was the enactment of the Sentencing Reform Act of 1984. The primary sponsors of the legislation were Senator Edward M. Kennedy, the liberal Massachusetts Democrat, and Senator Strom Thurmond, the conservative North Carolina Republican; President Reagan hailed the bill when he signed it in October of 1984.

The Act created a Sentencing Commission charged with developing and implementing a system of binding Sentencing Guidelines. At the same time, beginning in the mid-1980s,

Congress enacted a series of laws that mandated high minimum sentences for certain crimes—including for nearly all narcotics offenses, which now constitute some forty percent of all prosecutions in federal court. The Sentencing Reform Act itself also contained a variety of mandates, such as the requirement that repeat offenders receive sentences "at or near" the statutory maximum, which have also contributed to a substantial increase in the overall severity of federal criminal sentences.

The centerpiece of the Guidelines is a grid containing 258-boxes (termed the "Sentencing Table"). The grid's horizontal axis ("Criminal History Category") adjusts severity on the basis of the offender's past conviction record. The vertical axis ("Offense Level") reflects a base severity score for the crime committed, as further adjusted for those aspects of the crime that the Guidelines deem relevant to sentencing. The Guidelines, through a complex set of rules requiring significant expertise to apply, instruct the sentencing judge on how to calculate both "Criminal History Category" and "Offense Level." The box at which these two factors intersect then determines the range within which the judge may sentence the defendant. As required by the Sentencing Reform Act,[22] the sentencing range in each box is small, the highest point being twenty-five percent more than the bottom point. This twenty-five percent range represents one source of discretion retained by judges under the Guidelines.

The only other form of lawful sentencing discretion is authority to "depart" from the Guidelines. This authority is formally limited, however, to two circumstances. The first is where the defendant has provided substantial assistance in the prosecution of others,[23] in which event the judge may pronounce a sentence that departs downward from the Guideline range—with the important caveat that the prosecutor must first agree, in the words of the Supreme Court, to

"authoriz[e] the district court to depart."[24] If the prosecutor does make the appropriate motion for departure, the court may depart below not only the Guidelines range but also below any applicable statutory minimum sentence.[25] The second situation in which a judge may depart, up or down, from the Guideline range is where the judge is able to demonstrate on the record that there are factors or circumstances present in the case at hand that have not been "adequately" factored into the Guidelines' sentencing rules by the Commission and make the case "atypical." The Sentencing Commission has admonished that it expects the exercise of this departure power to be "rare."[26]

In a 1996 survey, however, 73 percent of district judges indicated that they felt mandatory guidelines were not necessary to direct the sentencing process, and "they strongly prefer a system in which judges are accorded more discretion than they are under the current guidelines."[27] In recent years, judges have departed downward on the basis of substantial assistance to authority in nearly twenty percent of all cases sentenced under the Guidelines, and have departed (upward or downward) due to "atypicality" in another ten to twelve percent of cases.[28] In many other cases, the defense and the prosecution stipulate to the "facts" that are relevant to sentencing under the Guidelines or stipulate even to a particular Guideline sentencing range; in these cases, there may be no departure as a formal matter, but it may be difficult to determine whether the Guidelines have been faithfully implemented.[29] For these various reasons, the ultimate ability of the Guidelines to control inter-judge disparity is an open question.

IV. Measuring Changes in Disparity

As we have noted, our definition of disparity is centered on the sentencing decision-maker, rather than on characteristics of the offense or the offender. In order to measure inter-judge disparity as we have defined it, one must observe the sentencing outcomes of similar cases assigned to different judges. Previous researchers have attempted to do this in three main ways.

First, simulated cases with a common set of facts have been distributed to judges who then provide a sentence. The influential Second Circuit sentencing study, and a similar study conducted nearly a decade later by the U.S. Department of Justice (1981),[30] used this approach. It is quite difficult, however, for a simulation to reconstruct the full depth of information available to a judge in a real case. Moreover, there is no assurance that judges approach simulation studies with the seriousness and deliberation that they would bring to a real case with a real defendant and real victims.[31]

Second, the variation among cases with common observable characteristics has been measured, with the residual variation among these observable similar cases attributed to the judges. A fundamental problem with this type of analysis is the difficulty in distinguishing between disparity, disproportionality, and proportionality.[32] A 1991 study by the Sentencing Commission, for instance, compared similar cases from a pre-Guidelines year (1985) and a post-Guidelines period of two years (1989-90).[33] The analysis conducted by the Commission compared the range of sentences and mean sentence for each of these categories under the Guidelines to corresponding measures from the pre-Guidelines period. The evaluation categorized cases according to factors deemed relevant by the Sentencing Guidelines, and hence concluded that "unwarranted disparity" existed if and only if there was deviation from the

Guidelines. The Commission's finding of less disparity post-Guidelines was simply a confirmation that post-Guidelines sentences are more likely to be in accordance with the Guidelines. This study illustrates the general problem with this measurement strategy—to compare "similar" cases, a study must rely upon an inevitably controversial theory of what constitutes proportional and disproportional variation. Moreover, variation due to unobserved differences in the cases (proportionality) cannot be readily distinguished from variation due to the decision-makers in the system (disparity).

In the third approach, caseloads randomly assigned to judges have been deemed to be comparable, and the average sentencing outcomes for these caseloads compared, with differences attributed to the judges. We adopt this third approach, and examine the average sentences of cases to which judges were randomly assigned within a particular federal district office to assess whether there was more disparity in these averages before or after implementation of the federal Sentencing Guidelines.[34]

To make the following discussion of measurement methodology more concrete, consider a simple example of a district in which cases are randomly assigned to two judges, Judge Harsh and Judge Lenient. In measurement of inter-judge disparity, we focus on the difference between judges within a time period (D_t) in the mean of prison sentences for each judge (θ) relative to the mean level of prison sentence length in the district as shown in (1).

$$D_t = \frac{\theta_h - \theta_l}{E[\theta]} \qquad \textbf{(1)}$$

In evaluating the effect of the Guidelines on inter-judge disparity, we examine the magnitude and statistical precision of the change in this disparity measure before and after the Guidelines (time

$$\Delta D = D_2 - D_1 \qquad (2)$$

periods 1 and 2), denoted as ΔD in (2). This empirical strategy is the most straightforward, but it has several important implications.

First, the null hypothesis that inter-judge disparity is the same in both periods ($\Delta D = 0$) can be tested directly.[35] This is conceptually distinct from statistical tests which rely on a null hypothesis of no disparity. Assume, for example, that the judge means were the same in both periods and the hypothesis of no disparity was rejected in period 1. Suppose that the null hypothesis of no disparity were accepted in period 2, because of a smaller sample size or increased variance in the distribution of sentence length. Under these assumptions, a false inference that there was a change in inter-judge disparity may be drawn when there was in fact no change.[36]

Second, the disparity measured by ΔD is variation relative to the mean of sentences in all cases. This has the desirable property of being an inequality measure that is scale invariant. For example, simply multiplying all sentences by a constant factor (say, to make sentences stiffer in the later period) will not affect the magnitude of ΔD. Defendants who are acquitted or whose cases are dismissed are assigned a sentence length of zero, because cases (and not convictions) are randomly assigned to judges and because only complete caseloads are comparable between judges rather than just convictions.

Third, the judges compared in the two periods are the same. This allows isolation of

changes in behavior of the same individuals, and avoids convolution with the potentially different sentencing patterns of other judges who heard cases in only one of the periods.[37]

To move beyond the illustrative statistical model in (1) and (2), we now discuss features of models suitable for estimation with data on multiple judges. We begin with the model in (3) based on the Gini coefficient, where g is the expected difference of J judge means relative to twice the overall mean.

$$g \equiv \frac{1}{J(J-1)} \sum_{j=1}^{J} \sum_{k=1}^{J} \frac{|\theta_j - \theta_k|}{2E[\theta]} \qquad \textbf{(3)}$$

If the true judge means θ were known, this would be an attractive measure. Unfortunately, the fact that θ is measured with sampling error results in an upward bias in the sample estimates of g, substantially complicating matters. One way to see this immediately is to examine the special case where the true judge means are all the same. In a finite sample of cases, the estimated means will not be exactly the same, so g will always be estimated to be some positive number when the true value is zero. This bias from sampling error is an inherent feature of this and many other simple estimation strategies that summarize the dispersion of estimated parameters.[38] Another disadvantage is of these methods is that standard errors on changes in dispersion of estimated means are analytically intractable -- as with the Gini coefficient -- or have poor finite sample properties -- as with analysis of variance methods.[39]

These estimation concerns lead us to develop a parametric model to estimate inter-judge disparity. Instead of summarizing the distribution of imprecisely estimated judge means, we adopt a strategy that incorporates the estimation of the judge effects directly in a statistical model of the underlying distribution of sentence lengths. As a point of departure, we consider the

negative binomial model which has been used previously in econometrics and which is part of a larger class of Generalized Linear Models well-known in statistics.[40] This type of count data model is attractive for this application because it accounts explicitly for the fact that the data are non-negative integers. Introducing a random variable into this model that corresponds to the judge assigned to the case allows us to directly estimate parameters that capture the dispersion of this random variable, or the variance of the "random effect" due to the judge.

We extend the negative binomial random effects model in several ways. First, we separate the likelihood into two parts, allowing an additional "zero-inflation" parameter for the probability of receiving no prison sentence.[41] Second, the mean of the random effect is allowed to differ by covariates, so that we can measure dispersion relative to the each district mean. Third, a lognormal distribution is used for the random effect, which allows a convenient interpretation of the dispersion of the random effect as a Gini coefficient, the expected absolute difference in average sentence length between judges in the same district.[42] Finally, the random effects are allowed to be correlated over time, and the extent of the correlation can be directly estimated.

Let Y denote the number of months of a prison sentence. For judge j in period t, there are a total of N_{tj} realizations of y_{itj}. To model the distribution of sentence lengths in cases heard by a judge, we use a negative binomial distribution with parameters m and p, augmented with an extra parameter d that affects the probability that $y_{itj}=0$. The joint likelihood function L_{Tj} for cases

heard by a judge in T periods is given in equation (4), where $\Gamma(.)$ is the gamma function.

$$L_{Tj} \equiv Pr(y_{11j},\ldots,y_{N_{Tj}Tj} | J=j) =$$

$$\prod_{t=1}^{T}\prod_{i=1}^{N_{tj}} \frac{w_t \Gamma(m_t + y_{itj})}{\Gamma(m_t)\Gamma(y_{itj}+1))}\left[d_t + p_{jt}^{m_t}\right]^{1(y_{itj}=0)}\left[p_{jt}^{m_t}(1-p_{jt})^{y_{itj}}\right]^{1(y_{itj}>0)} \quad \textbf{(4)}$$

The density of the zero-inflated negative binomial is normalized to one by setting $w_t = (1+d_t)^{-1}$.

When p is defined as a particular function of m, d, and a judge-specific parameter θ, the mean sentence length for judge j in period t depends only on the judge effect θ as in (5).

$$p_{tj} \equiv \left(1 + \theta_{tj}\frac{(1+d_t)}{m_t}\right)^{-1} \Rightarrow \mathrm{E}(Y_{itj} | T=t, J=j) = \theta_{tj} \quad \textbf{(5)}$$

Since we are interested in the distribution of the judge effects, we directly model the distribution of θ. For the complete data in any one period (T=1), the likelihood L_1 is obtained by integrating out θ (the judge "random effect") using a lognormal density function with mean μ_t and standard deviation σ_t, denoted as f_1 in equation (6).

$$L_1 \equiv \prod_{j=1}^{J}\int_0^{\infty} L_{1j}(\theta_{1j}) f_1(\theta_{1j} | \mu_1, \sigma_1) d\theta_{1j} \quad \textbf{(6)}$$

For any single district office, the number of judges is relatively small (2-21) and there are a reasonably large number of cases per judge ($N_{tj} \geq 30$). Thus, estimates of σ correspond to the dispersion about the overall office mean in consistent estimates of average sentencing for an observed small sample of judges.[43] Data from multiple district offices are pooled for efficiency

of estimation. We allow the mean of the judge effects to differ by office, since cases are assigned randomly within district offices but not between offices. Denoting X as a set of indicators for each office, we let $\mu_t = X\beta_t$.

In order to account for correlation of judge effects across two periods, we also formulate a model in which the likelihood L_2 for two periods uses joint lognormal density function, denoted as f_2 in equation (7).

$$L_2 \equiv \prod_{j=1}^{J} \int_0^{\infty}\int L_{2j}(\theta_{1j},\theta_{2j}) f_2(\theta_{1j},\theta_{2j}|\mu_1,\sigma_1,\mu_2,\sigma_2,\rho)\, d\theta_{2j}\, d\theta_{1j} \qquad \textbf{(7)}$$

A primary interest in this study are measures of inter-judge disparity in sentencing. Denote γ as the Gini coefficient of concentration of the judge means derived from (6) or (7), as opposed to g in (3) which is computed using the estimated judge means. Using the properties of a lognormal parametric form for θ, the Gini coefficient measuring relative disparity in average sentence length between judges in two periods depends only on the two variance parameters of the random effect. A change in γ can be computed as in equation (8), where Φ is the cumulative normal distribution function.

$$\gamma_2 - \gamma_1 = 2\Phi\left(\frac{\sigma_2}{\sqrt{2}}\right) - 2\Phi\left(\frac{\sigma_1}{\sqrt{2}}\right) \qquad \textbf{(8)}$$

When data from multiple offices are pooled and μ includes indicators for each office, then γ measures the overall dispersion in judge means relative to their own district office mean. Furthermore, since γ is the expected absolute difference between two judges relative to twice the overall mean, it has the straightforward interpretation of inter-judge disparity as a percentage

difference relative to the overall level of sentence length.

One simple hypothesis to explain changes in inter-judge disparity over time is that the types of offenses within a judge's caseload are changing over time. We would like to distinguish between changes in the behavior of judges and changes in the types of cases to which they are assigned. What would trends in inter-judge disparity look like if judges had been assigned caseloads with the same shares of offense types every year? One way to answer this question is to statistically adjust by reweighting the caseloads. For example, the adjusted average sentence length for a judge in 1982-83 and 1992-93 might take the average in each period for drug cases and for non-drug cases, and compute a weighted average using the same weights in both periods. In this paper, we use a set of weights for each district office based on the shares of offense types within that office in 1986-87. Denote $N_{86\text{-}87}$ as the total number of cases in a district office during 1986-87. Let superscript z refer to a type of offense, so that N_{jt}^{z} is the number of cases assigned to judge j in time period t for offense type Z. Weights w_{ijt} are then defined in equation (9).

$$w_{ijt} \equiv \left(\frac{1}{1+d_t}\right)\left(\frac{N^{z}_{86\text{-}87}}{N_{86\text{-}87}}\right)\left(\frac{N_{jt}}{N^{z}_{jt}}\right) \qquad \textbf{(9)}$$

For results in the next section that use weighting for offense type comparability over time, we change the weights w in equation (4), using w_{ijt} from (9) instead of $w_t = (1+d_t)^{-1}$.

In this paper we focus on disparity in the overall average prison sentence length, so we note that disparity in the overall average is heavily influenced by cases with long sentence lengths. If there are differences in disparity by type of offense, the offenses with longer sentences will have a large effect on disparity in the overall average. To see this, say that Judge Harsh

sentences a fraud offender to 11 months and a violent offender to 63 months, while Judge Lenient sentences a fraud offender to 9 months and a violent offender to 57 months. The Gini coefficient between these two judges for the fraud cases is 0.1, and for the violent cases is 0.05. The Gini coefficient for the overall average of 33 and 37 is 0.057, which is indicative of the weight given to the offense type with the larger average sentence length.[44]

V. Data Description

In order to implement the measurement strategies outlined in the previous section, we minimally required data on the universe of cases filed within various districts, and the judge, disposition, and prison sentence length in the case. A special extract was prepared for this research by the Statistics Division of the Administrative Office of the U.S. Courts that included a previously unavailable non-identifying code that was used to group together cases heard by the same judge.[45]

In order to create a dataset of cases randomly assigned to judges, we excluded judges who did not hear a full caseload (and therefore were unlikely to have fully participated in the randomization). This selection rule was based on the number of cases heard.[46] Under random assignment, the caseload should be approximately balanced across judges. Based on this logic, we constructed a sample in which judges were deemed to be "active" in a particular year.[47] In order to have a sufficient number of cases to consistently assess the sentencing patterns over time, cases were dropped from the sample if the assigned judge had less than 30 cases within a two year period. Since random assignment is usually done within each of several offices in any district, we restricted our data to offices that had at least two judges. When judges were assigned cases in more than one office, cases were only included for the office from which the judge was

assigned the largest number of cases. Since judges are randomly assigned to cases, but the cases may have more than one defendant, we randomly selected one defendant from each case.

A central premise of this analysis is that cases are randomly assigned to judges. We marshal two types of evidence in support of our claim that random assignment was used in the districts included in our analysis. A primary source of evidence regarding randomization is the distribution of offense types among the caseloads of each judge. For example, the proportion of drug cases, embezzlement and fraud cases, violent and firearms cases, and other crimes should be the same for each judge in a district office except for sampling error. Differences in these proportions form the basis for the chi-square test of independence of the offense types and the judges. We performed these chi-square tests for each district office for judges with at least eight cases during 25 six month periods from July 1981 to December 1993. Under random assignment, we would not only expect that 95 percent of these tests would have chi-square test statistics less than the .95 critical value, but that the test statistics for the various time periods would be uniformly distributed over the (0,1) interval with an average p-value of 0.5. A statistical exclusion rule based on this logic was used to identify districts unlikely to have used random assignment throughout the 25 periods, where districts with a mean p-value below a threshold (the 5th percentile of the mean of 25 uniform random variables, .405) were excluded. Districts were also excluded if there were not at least two active judges in eight or more six month periods both before and after implementation of the Guidelines.

As a second source of information on random assignment, we drew upon interviews with the court clerks in the districts. We conducted 40 interviews ourselves, and also utilized a similar investigation by researchers at the U.S. Sentencing Commission.[48] The statistical exclusion rule

based on the mean p-value of chi-square test for independence of offense types identified all four offices that had consistently had at least two active judges but were reported to have used non-random assignment of cases according to the qualitative research, as well as nineteen other offices. For the 26 offices included in the analysis, we present quantiles of the p-values from the chi-square tests of independence of judges and offense types in Table II. The mean of the 606 test statistics is 0.49 and they are distributed fairly uniformly from zero to one.

The resulting sample used for estimation includes 77,201 cases, 27 percent of the total universe of over 285,000 cases from July 1981 to December 1993. About half the cases are excluded because an office does not have at least two judges who consistently had cases assigned to them throughout the period. Roughly another one quarter of cases are excluded because they were assigned in an office that did not appear to use consistently use random assignment of cases to judges throughout the time period under study. Both the regional composition and the offense types are similar in the estimation sample and in the universe of all cases.

In our analysis, we focus on comparison of inter-judge disparity for two year periods before and after promulgation of the Guidelines. Descriptive Statistics for the two-year periods from 1982-93 are given in Table I. The percentage with zero sentence length in our data include zeros for acquittals and dismissals, which have declined slightly as a share of all cases over time. The distribution of sentence length shifted toward higher sentences throughout the twelve and a half year period covered by these data as mandatory minimums and Guidelines took effect over time.

VI. RESULTS

Based on our interviews and statistical tests, we are fairly confident that the offices included in our sample used random assignment of cases to judges. Since the caseloads should therefore be comparable, differences in the average sentence length of these caseloads can be attributed to judges themselves. This section reports the results of the methods outlined in Section IV.

In Figure 1, we graph estimates of inter-judge disparity for two year time periods from 1982-1993 using the data described in Table I. The triangles in Figure 1 are Gini coefficients computed from the absolute difference of the judge means, based on g from equation (3) using the average of estimates for each district office weighted by the number of cases in that office. The circles in Figure 1 are from the dispersion in the random effect of the zero-inflated negative binomial model for each single period, based on estimates using equation (6) transformed into the Gini coefficient γ. The changes over time in g and γ are quite similar, with a peak in 1984-85, followed by a decline that accelerates from 1986-87 through 1988-89 and a leveling off thereafter. As discussed in Section IV, the estimates of g are biased upward by sampling error. Estimates of the magnitude of the bias depend on the assumptions used to model the judge means.[49] The estimates of γ account for sampling variability by explicitly modeling the underlying distribution of sentence lengths and the distribution of judge means, formalizing the intuition that larger deviations of a judge from the district office mean are increasingly likely to be due to sampling error. The point estimates of g are approximately 0.07 greater than γ for each time period. Based on the model of γ, the sampling error bias in g appears to be fairly constant over time.[50] Our main interpretation of these results is that the same temporal dynamics of inter-judge disparity are apparent in measurements of both g and γ. This increases our confidence that

our results about changes in disparity over time are not highly sensitive to the modeling strategy.

For the remainder of this section, we focus on estimates of γ from the random effects model so that we can account for sampling error bias, assess the statistical precision of changes between periods, and estimate the correlation of judicial sentencing patterns between time periods. To first verify that the model defined in equations (4) - (7) is an appropriate model for this data, we compare observed cell probabilities for the pooled data from 1986-87 with predicted values from a simple two-part negative binomial model, assuming that δ, γ, and θ are constant across all cases. The actual distribution and predicted distribution are presented in Table III. The model has been constructed to fit zero exactly, and does a reasonable job of representing the rest of the distribution -- even though the large amount of data results in a chi-square statistic (χ^2= 66) that rejects the hypothesis that the model exactly fits the data. Of course, the model does not account for the fact that within the cells of Table I the data are clustered at particular months (6, 12, 18, ...) rather than distributed smoothly across all months, but the model fits the basic features of the data quite well. Allowing θ to vary across the judges when maximizing the likelihood in (7) requires evaluating J double integrals for every function evaluation in nonlinear optimization, but these can be computed efficiently after an appropriate transformation of variables using Gaussian quadrature based on Hermite polynomials.[51] Standard errors for γ, the Gini coefficient in (8), are computed using a numerical approximation to the Hessian and the delta method.[52]

The estimates of γ for Figure 1 were estimated separately for each period based on equation (6) and data for all judges available in the period. The Gini coefficient estimates peak in 1984-85, fall by .019 in 1986-87 and again by .031 in 1988-89 before largely leveling off.

While not statistically significant, these results also suggest that inter-judge disparity may have been decreasing prior to 1986-87 -- an issue which we discuss further below. Taking the average over the three periods from 1988-93 as our post-Guidelines measure, inter-judge disparity fell from .085 to .054 from 1986-87 to 1988-93, a decrease of .031 with a standard error of .010. The changes are more pronounced than the mixed results of previous researchers.[53] Based on these results, the expected difference between two typical judges is twice the Gini coefficient -- about 17 percent in 1986-87 and about 11 percent in 1988-93.

Since overall sentence lengths are rising over time, a given percentage difference in inter-judge disparity implies a larger absolute difference in months of prison sentence length. Multiplying the percentage difference by the overall average sentence length in each period from Table I expresses our measure of inter-judge disparity in terms of months. For 1986-87 the mean sentence length was 29, the expected inter-judge difference was 4.9 months, which fell to 3.9 months in 1988-93 when the mean sentence length was 35.[54]

Between 75 and 86 percent of judges in any single two-year period were also in the sample during the following period. To ensure that results are not simply being driven by changes in the composition of judges, we also analyze changes based on the same judges in both periods. The lines in Figure 2 connect estimates of disparity between consecutive two-year periods, based on estimates of γ from equation (7). The changes over time are very similar to those reported in Figure 1. The decrease in inter-judge disparity before and after the promulgation of the Guidelines is sharper when comparing the same judges over time, as the γ falls by more than half between 1986-97 and 1989-90 from .090 to .039, a decrease of .051 with a standard error of .013. The estimates from 1988-93 range from .055 to .059 to .046, and these

differences are indistinguishable from sampling error. The most conservative estimate, not shown in Figure 2, compares the same judges in 1986-87 to those in 1989-90, omitting 1988 to allow a transition period for adjustment to the new regime and because legal challenges to the Guidelines were not resolved until January 1989. This estimate shows that γ falls from .083 to .067, and decrease of .016 with a standard error of .013. This estimate is conservative in the sense that it is a smaller change than that from 1986-87 to 1988-89 and that our other estimates of disparity for years after the Guidelines other than 1989-90 are all lower than .067. We conclude from measures using the same judges that the expected difference between two judges (twice the Gini coefficient) decreased from 17-18 percent in 1986-87 to 8-13 percent in 1988-90. This range using the same judges in both periods brackets decrease from 17 to 11 percent reported above for all judges based on Figure 1 for 1986-87 to 1988-93.

Another factor changing over time is the mix of offenses in the overall caseload. Table IV shows that the overall share of drug offenses increased from .21 to .33 from 1982-83 to 1992-93, while the share of "other" offenses (such as forgery) fell from .37 to .24. If the disparity in sentencing drug cases was always lower than disparity for other cases, then we might observe a decrease over time in measured overall inter-judge disparity that was due to a change in the caseload coming before judges. In an attempt to separate out changes in judicial behavior from changes in the types of cases, we compute weighted results, replacing w_t in (4) with w_{ijt} defined in (9). These weights statistically adjust so that the shares of offense types in the overall distribution for each judge in each time period are equal to the share for their district office in 1986-87. The four offense types used are violent & firearms, drug, embezzlement & fraud, and other cases. The choice of a base period does not affect trends over time, but does affect the

levels of the estimates; 1986-87 is chosen to address the counter-factual in which the Sentencing Guidelines were later adopted but the mix of offense types did not change.

The unweighted results using w_t from Figure 2 are reproduced in the first three columns of Table V and weighted results using w_{ijt} are shown in columns four through six, and the number of judges active in the consecutive two-year periods is shown in column seven. In addition to making the shares of offense types comparable over time, weighting equalizes the shares of offense types within a period. The magnitudes of the weighted point estimates of inter-judge disparity are slightly lower than the unweighted estimates, because the correction for variability due to the fact that the shares are similar but not exactly equal when cases are assigned randomly.[55] The point estimates differ slightly, but the overall pattern of results is quite similar for the unweighted results and those weighted for comparability over time. For example, the weighted estimate of γ in 1986-87 is .079, and falls to 0.042 in 1988-89, implying that the expected difference between two judges fell from 16 percent to 8 percent.

Our main interpretation of these results is that changes in inter-judge disparity are not due to changes in the types of offense in the overall caseload. The aspect of the results that appears to be most sensitive to changes in specification is the change in disparity from 1984-85 to 1986-87. In all specifications, disparity appeared to be stable or increasing through 1986. For example, the weighted estimates for using the same judges in 1983-84 and 1985-86 (not shown in Table V), were .088 and .092. Decreases in disparity appear to be concentrated during 1987-89, but there are not enough cases per judge to more precisely identify the timing of the changes.

Our preliminary research on disparity for particular offense types suggests that the decrease in inter-judge disparity is concentrated within the violent, weapons, and drug crimes.

Estimation for particular offense types substantially reduces the number of cases per judge in each period, however, and Monte Carlo simulations suggest that the methods used in this paper are substantially less reliable when there are less than 30 cases per judge used in the estimation. In future research we intend to pool additional years of data and model the dispersion in judge means as a parametric function that is changing over time, in order to obtain reliable estimates for various offense types.

Regarding the correlation of the judge effects, we find that the behavior of judges appears to be fairly consistent over time prior to the Guidelines. Table VI reports the correlation of judge effects between time periods. Prior to 1986-87, this measure is greater than .70.[56] There is some evidence on the consistency of judicial sentencing patterns declined thereafter, but the results are mixed. It is increasingly difficult to reliably estimate the correlation between two periods when the variance of the random effect is small in both periods. In comparisons for two-year periods subsequent to those reported in Table VI, the correlation varies from .44 to -.24. However, the standard errors are at least .21, and we cannot draw any credible conclusions from these very imprecise estimates in the later periods.

Finally, we return to the question of causality. Were these changes over time in inter-judge sentencing disparity caused by the Federal Sentencing Guidelines? Clearly, the largest change in disparity occurs between 1986-87 and 1988-89, which corresponds to the effective date of the Guidelines in November, 1987. Disparity from 1988-93 has remained at levels lower than observed from 1982-1987. While this timing is suggestive, the Guidelines only applied to offenses committed after November 1, 1987. Because of the lag from commission of offense to case filing and because of constitutional challenges to the Guidelines, about half of the

cases filed in 1988 and 1989 were not sentenced under the Guidelines.[57] We suspect that the 1986 enactment of mandatory minimum sentences for drug offenders (which applied only to crimes committed after October 1, 1986) may have substantially contributed to the decrease in disparity after 1986.[58]

VII. Discussion

By focusing our analysis on the *nominal* length of prison sentences, we have not considered inter-judge disparity in the length of time *actually served* in prison. It is possible that parole policies in the pre-Guidelines period reduced inter-judge sentencing disparity in the time actually served by an offender. The Sentencing Reform Act eliminated parole, so parole cannot affect time-served in cases sentenced under the Guidelines.

For defendants who were sentenced to terms of imprisonment prior to the Sentencing Guidelines, parole authorities actually determined the date of release from prison, and thus the time actually served by the offender. These determinations were based upon the Parole Guidelines, which were similar in function to the Sentencing Guidelines. Like the Sentencing Guidelines that were modeled after them, the heart of the Parole Guidelines were a grid of boxes that indicated actual sentence length based on the offender's prior record and the seriousness of the offense. Since this determination was independent of, and subsequent to, the offender's sentencing, the Parole Guidelines may have substantially mitigated inter-judge disparity in nominal sentences. None of the influential studies that indicated widespread disparity in the pre-Guidelines era considered the effect of the Parole Guidelines in reducing inter-judge time-served disparity. In future research, we intend to examine inter-judge disparity in time-served.

Despite this caveat, we believe our focus on disparity in nominal sentences is appropriate. First, as we have related above, inter-judge sentencing disparity was a critical impetus to the passage of the Guidelines and reducing it was a central goal of the still-controversial Sentencing Guidelines. Second, inter-judge sentencing disparity is an interesting phenomena apart from its ultimate outcome on an offender's sentence. The actual ceremony of sentencing has an expressive function which is important independent of the actual time served. The sentencing is the moment at which the community publicly expresses its disapproval of the offender's action. The prosecutor and defense counsel offer a few words, offered more for the victim, the defendant, their friends and family and any press than for the judge. The defendant is asked to speak, if she wishes, to accept responsibility, to ask for forgiveness, or to say nothing. Finally, the judge formally articulates a measure of the defendant's offense against the community.

That different judges publicly express different measures of justice for the same offenses is therefore both interesting and troublesome, even if the offenders ultimately serve the exact same sentence. Frankel cited the fact that some judges sentenced draft-evaders to the maximum sentence while other judges imposed almost no prison time for defendants who broke the law in adherence to principle.[59] This disparity is notable because it shows that two judges, each purportedly expressing the will of the community, differ dramatically in their representation of the proper sentence for a particular crime. As Frankel noted, "It is not directly pertinent here whether either category of judge is right,"[60] but the fact of their disagreement is pertinent. The disagreement undermines the expressive function of sentencing by suggesting that a sentence is not so much a measure of the offense to the community, but simply the personal judgment of the judge.

Despite the importance that the progenitors of the Guidelines placed on inter-judge sentencing disparity and our focus on it in this paper, it would be a mistake to equate inter-judge sentencing disparity with "unwarranted sentencing disparity," and consider our findings a simple vindication of the Sentencing Guidelines. We note, first, that we have sought to measure only disparity among judicial participants in sentencing. There are, of course, other sources of sentencing disparity in the federal criminal justice system. As many commentators have noted, considerable disparity exists in charging policies at various U.S. Attorney's offices,[61] in the policies of law-enforcement personnel, and the manner in which the probation officer conducts an independent investigation of the offense.[62] The Guidelines did nothing to address these sources of disparity.

Moreover, the reduction in the exercise of judicial discretion resulting from the Sentencing Guidelines and the imposition of mandatory minimum sentences increased the impact of any disparity at these earlier stages in the criminal process. Reduced discretion for judges at the end of the process magnifies the importance of decisions made by the prosecutor, probation office, and law enforcement officials. Since the sentence will be determined by what is proven by a preponderance of the evidence under the Guidelines, the prosecutor exerts far more influence over the sentence than she did pre-Guidelines. Similarly, the offender's sentence will directly reflect any disparity between probation officers, because they are the "Guidelines experts" who initially advise the judge of the facts of the case and how the Guidelines apply to these facts. Under the Guidelines, law enforcement officials also wield more influence over final sentences. Many Guidelines sentences (including narcotics sentences and sentences for all crimes with monetary losses) depend upon the measurable quantities involved in the offense. In

many investigations of such crimes, governmental authorities exert substantial control over the quantity that will be used to calculate the defendant's sentence under the Guidelines. For instance, in narcotics investigations, the undercover agent often determines the amount of drugs either purchased from or sold to a putative defendant. Her decision as to the quantity to attempt to buy or sell from the target offender will play a large role in determining the ultimate sentence under the Guidelines.[63]

By giving prior actors (law-enforcement officials, probation officers, and prosecutors) more influence over the ultimate sentence, the Guidelines provide opportunities for these earlier actors to pursue their own agendas that did not exist pre-Guidelines. If these prior actors vary in their willingness to engage in manipulative tactics aimed at achieving a higher or lower Guidelines sentence for particular defendants, prosecutorial sentencing disparity will actually have been increased by the Guidelines. Pre-Guidelines, any disparity resulting from these practices was comparatively less because the judge and the parole board exercised overwhelming control over the sentence.

In addition, mean sentence length has substantially increased under the Guidelines. Critics of this increase may argue that any reduction in judicial sentencing disparity was achieved primarily by reducing the frequency of sentences that, pre-Guidelines, were relatively lenient,[64] and that the result is a plethora of long sentences disproportionate to the crime and the offender. In this respect the Guidelines may be contrasted with a regime that successfully reduced disparity without dramatically increasing sentence length.

Lastly, we note that there are other consequences of the Guidelines besides their effect on sentencing disparity and sentence length. Judges have complained that the Guidelines rules

themselves are as arbitrary as any exercise of judicial discretion may have been prior to the Guidelines, and that their arbitrariness and their extraordinary complexity have made the sentencing process incomprehensible and inaccessible to victims, defendants, and the general public.[65]

VIII. CONCLUSION

Our study indicates that the Guidelines (and concomitant statutory minimum sentences) have been successful in reducing inter-judge nominal sentencing disparity. To the extent that this was the central goal of the Sentencing Reform Act of 1984, Congress successfully achieved this goal. The Guidelines have reduced the net variation in sentence attributable to the happenstance of the identity of the sentencing judge. The expected difference in the sentence lengths of two judges receiving comparable caseloads was 16 to 18 percent in the pre-Guidelines period of 1986-87. Comparing inter-judge disparity before and after the Guidelines, we find this measure declined substantially, with estimates of the expected difference ranging between 8 and 13 percent during 1988-93.

Unfortunately, the very success of the Guidelines in reducing inter-judge disparity by constraining judicial discretion may have exacerbated the impact and the degree of disparity at earlier stages of the criminal justice process, through the elimination of parole and the severe reduction in the judiciary's ability to compensate for inter-actor disparity earlier in the criminal justice process. Also, although our empirical results show large changes in inter-judge nominal sentencing disparity, we have not measured disparity in time-served. Disparity in time-served and disparity among decision-makers earlier in the criminal justice process are both subjects in

need of further research.

We conclude by noting that elimination of disparity is only one objective of a just sentencing system. Other commentators have argued that the present regime has purchased a reduction in inter-judge sentencing disparity at the price of undue severity in sentences, undue uniformity of those sentenced,[66] and unwarranted complexity.[67] Even if disparity from all sources (judges, prosecutors, law-enforcement agents, probation officers, and so on) could be eliminated, the result would not necessarily be a just or fair sentencing system. In particular, implementation of the Sentencing Guidelines and statutory minimum sentences have led to complaints of undue *uniformity* in sentencing. If a sentencing regime fails to take account of characteristics of the offense or of the offender that are believed to be relevant to a just sentence, then its results may be unwarranted even if there is no measurable disparity due to the identity of particular decision-makers.

BIBLIOGRAPHY

Alschuler, Albert W. "The Failure of Sentencing Guidelines: A Plea for Less Aggregation." *University of Chicago Law Review* 58 (1991): 901.

Austin, William & Thomas A. Williams III. "A Survey of Judges' Responses to Simulated Legal Cases: Research Note on Sentencing Disparity." *Journal of Criminal Law & Criminology* 68 (1977): 306.

Bartolomeo, John, *et al. Sentence Decision Making: The Logic of Sentencing Decisions and the Extent and Sources of Sentence Disparity.* Washington, D.C.: U.S. Department of Justice, 1981.

Bowman, Francesca D. "Probation Officers Advisory Group Survey." *Federal Sentencing Reporter* 8 (1996): 303.

Brown, Joe. "Quo Vadis? What Congress and the Department of Justice should do in Response to the Justice Department's Analysis of Non-Violent Drug Offenders with Minimal Criminal Histories." *Federal Sentencing Reporter* 7 (1994): 25.

Cameron, A. Colin & Pravin Trivedi. *Regression Analysis of Count Data.* New York: Cambridge University Press (1998).

Casper, Jonathan D. "Determinant Sentencing and Prison Crowding in Illinois." *University of Illinois Law Review* (1984): 231.

Chib, Siddhartha *et al.* "Posterior Simulation and Bayes Factors in Panel Count Data Models." *Journal of Econometrics.* 86 (1995): 33.

Cook, Beverly Blair. "Sentencing Behavior of Federal Judges: Draft Cases." *University of Cincinnati Law Review* 42 (1973): 597.

Davis, Kenneth Culp. *Discretionary Justice.* Baton Rouge: Louisiana State University Press, 1969.

Diamond, Shari S. & Zeisel, Hans. "Sentencing Councils: A Study of Sentencing Disparity and its Reduction." *University of Chicago Law Review* 43 (1975):109.

Edmunds, Robert H. "Guidelines Sentencing and Department of Justice Policies Under the Reagan-Bush Administrations." *Federal Sentencing Reporter* 6 (1994): 306.

Evans, Gwynne. *Practical Numerical Integration.* New York: John Wiley & Sons, 1993.

Frankel, Marvin E. "Lawlessness in Sentencing." *University of Cincinnati Law Review* 41 (1972): 1.

—. *Criminal Sentences: Law Without Order.* New York: Hill & Wang, 1973.

Freed, Daniel J. "Federal Sentencing in the Wake of the Guidelines: Unacceptable Limits on the Discretion on Sentencers." *Yale Law Journal* 101 (1992): 1681.

Gaudet, Frederick J., *et al.* "Individual Differences in the Sentencing Tendencies in the Sentencing Tendencies of Judges." *Journal of Criminal Law & Criminology* 23 (1933): 811.

Gaudet, Frederick J. "The Differences Between Judges in the Granting of Sentences of Probation." *Temple Law Quarterly* 19 (1933): 471.

Gurmu, Shiferaw, *et al.* "Semiparametric Estimation of Count Regression Models." *Journal of Econometrics* 88 (1999): 123.

Hausman, Jerry, *et al.* "Econometric Models for Count Data with an Application to the Patents-R&D Relationship." *Econometrica* 52 (1984): 909.

Heaney, Gerald H. "The Reality of Guidelines Sentencing: No End to Disparity," *American Criminal Law Review* 28 (1991): 161.

Heumann, Milton. "Empirical Questions and Date Sources: Guideline and Sentencing Research in the Federal System." *Federal Sentencing Reporter* 6 (1993): 15.

Hofer, Paul. Personal Communication. December 4, 1997.

Hofer, Paul *et al.* "The Effect of the Federal Sentencing Guidelines on Inter-judge Sentencing Disparity." Unpublished manuscript, U.S. Sentencing Commission, November 1998.

Hoover, J. Edgar. "The Dire Consequences of the Premature Release of Dangerous Criminals Through Probation and Parole." *F.B.I. Law Enforcement Bulletin* 27 (1958): 1.

Hopkins, Alec. "Is There a Class Bias in Criminal Sentencing?" *American Sociological Review* 42 (1977): 176.

Howard, Joseph C. "Racial Discrimination in Sentencing." *Judicature* 59 (1975-76): 121.

Johnson, Molly Treadway & Scott Gilbert. *The U.S. Sentencing Guidelines: Results of the Federal Judicial Center's 1996 Survey.* Washington, D.C.: U.S. Judicial Center, 1997.

Karpoff, Jonathan M. & John R. Lott, Jr., "Why the Commission's Corporate Guidelines May Create Disparity." *Federal Sentencing Reporter* 3 (1990): 140.

Lambert, Diane. "Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing." *Technometrics* 34 (1992): 1.

Lott, John & Stephen Bronars. "Time Series Evidence on Shirking in the U.S. House of Representatives." *Public Choice* 76 (1993): 125.

Lott, John R., Jr. "Do We Punish High Income Criminals Too Heavily?" *Economic Inquiry* 30 (1992): 583.

Morris, Norval. *The Future of Imprisonment.* Chicago: The University of Chicago Press, 1974.

Mullahy, John. "Specification and Testing of Some Modified Count Data Models", 33 *Journal of Econometrics* (1986): 341.

Nagel, Irene H. & Stephen J. Schulhofer. "A Tale of Three Cities: An Empirical Study of Charging and Bargaining Practices Under the Federal Sentencing Guidelines." *Southern California Law Review* 66 (1992): 501.

Newman, Donald J. "Court Intervention in the Parole Process." *Albany Law Review* 36 (1972): 257.

Newman, Jon O. Foreward, "Parole Decision-making and the Sentencing Process." *Yale Law Journal* 84 (1975): 810.

—. "A Better Way to Sentence Criminals." *American Bar Association Journal* 63 (1977): 1562.

Partridge, Anthony & William B. Eldridge. "The Second Circuit Sentencing Study." *A Report to the Judges,* 1974.

Payne, Abigail. "Does Inter-judge Disparity Really Matter? An Analysis of the Effects of Sentencing Reforms in Three Federal District Courts." *International Journal of Law and Economics* 17 (1997): 337.

Schulhofer, Stephen. "Assessing the Federal Sentencing Process: The Problem is Uniformity, Not Disparity." *American Criminal Law Review* 29 (1992): 833.

Searle, Shayle, *et al. Variance Components.* New York: John Wiley & Sons, 1992.

Seymour, Whitney North. "1972 Sentencing Study for the Southern District of New York." *New York State Bar Journal* 45 (1973): 163.

Stith, Kate & José A. Cabranes. *Fear of Judging: Sentencing Guidelines in the Federal Courts.* Chicago: University of Chicago Press, 1998.

Taha, Ahmed. "The Effect of the Federal Sentencing Guidelines on the Disposition of Criminal Cases." Unpublished manuscript, U.S. Department of Justice, August 1998.

Trott, Stephen S. Letter to Hon. William W. Wilkins, Jr. (April 7, 1987), reprinted in *Federal Sentencing Reporter* 8 (1995): 196.

—. Letter to Chairman of U.S. Sentencing Commission (November 9, 1994), reprinted in *Federal Sentencing Reporter* 8 (1995): 197.

U.S. Attorney General. *Annual Report,* 1940.

U.S. Board of Parole. *Biennial Report.* Washington, D.C.: U.S. Board of Parole, 1970-72.

U.S. House of Representatives. *Hearings on Corrections, Federal and State Parole System Before Subcommittee No. 3 of the House Committee on the Judiciary.* 92d Cong., 2d Sess., (1973).

U.S. Senate. *Of Prisons and Justice.* U.S. Senate Document No. 70, 88th Cong., 2d Sess. (1964).

—. *Report No. 225.* 98th Cong., 1st Sess. (1984), *reprinted in* 1984 U.S. Code Congressional & Administrative News: 56.

U.S. Senate. Committee on the Judiciary. *Reform of the Federal Criminal Laws: Hearings on Section 1347 Before the Subcommittee on Criminal Laws and Procedures.* 95th Cong., 1st Sess., (1977): 8580, 8995.

U. S. Sentencing Commission. *Annual Report.* Washington, D.C.: U.S. Sentencing Commission, 1990.

—. *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.* Washington, D.C.: U.S. Sentencing Commission, 1991.

—. *Special Report to Congress: Mandatory Minimum Penalties in the Federal Criminal Justice System.* Washington, D.C.: U.S. Sentencing Commission, 1991.

—. *Annual Report.* Washington, D.C.: U.S. Sentencing Commission, 1995.

—. *Sourcebook of Federal Sentencing Statistics.* Washington, D.C.: U.S. Sentencing Commission, 1997.

von Hirsch, Andrew. *Doing Justice: The Choice of Punishments.* New York: Hill & Wang, 1976.

Waldfogel, Joel. "Aggregate Inter-Judge Disparity in Sentencing: Evidence from Three Districts." *Federal Sentencing Reporter* 4 (1991): 151.

— . "Inter-judge Disparity in Federal Sentencing: Evidence from Three Federal Districts, 1984-90." New Haven: Yale University, 1992

Wheeler, Stanton, *et al. Sitting in Judgment: The Sentencing of White-Collar Criminals.* New Haven: Yale University Press, 1988.

Wicker, Tom. "Judging the Judges." *New York Times.* February 6, 1976: A29.

ENDNOTES

* Assistance in data production was generously provided by Ralph Mecham, Steve Schelsinger, and Cathy Whitaker at the Administrative Office of the U.S. Courts. Helpful comments were made by participants in the conference sponsored by the Journal of Law and Economics at which an earlier version of this article was presented, and at various stages of the project by Daron Acemoglu, Josh Angrist, Jushan Bai, Katherine Brownlee, José Cabranes, Gary Chamberlain, Ken Chay, Peter Diamond, Hugh Eastwood, Dan Freed, Jerry Hausman, Bo Honore, Larry Katz, Kara Kling, David Lee, Steve Levitt, Jeff Liebman, John Lott, Whitney Newey, Abigail Payne, Anne Peihl, Steve Pischke, Jack Porter, Jim Poterba. Kling acknowledges financial support from a National Science Foundation Graduate Fellowship and an Alfred P. Sloan Doctoral Dissertation Fellowship.

1. Stephen S. Trott, Letter to Hon. William W. Wilkins (April 7, 1987). The letter from Trott (who was Associate Attorney General writing on behalf of Department of Justice) to Wilkins (who was the first Chairman of the Sentencing Commission) urged adoption of sentencing guidelines that would permit only narrow judicial discretion; the letter is reprinted at 8 Federal Sentencing Reporter 196 (1995). See also 8 Federal Sentencing Reporter 197 (1995) (reprinting letter dated November 9, 1994 from Judge Stephen S. Trott to the new Chairman of the Sentencing Commission, explaining that experience under the Guidelines had caused him to conclude that "the cure is worse than the disease.")

2. 28 U.S.C. § 991(b)(1988).

3. See also 28 U.S.C. § 994 (f) (1988) ("The Commission, in promulgating guidelines pursuant

to subsection (a) (1), shall promote the purposes set forth in section 991 (b) (1), with particular attention to the requirements of subsection 991 (b) (1) (B) for providing certainty and fairness in sentencing and reducing unwarranted disparities.")

4. See generally Kate Stith & José A. Cabranes, Fear of Judging: Sentencing Guidelines in the Federal Courts, 38-48 (1998).

5. Before the passage of the Guidelines, these ranges were the only statutory upward bound on sentencing. Under the Guidelines, they continue to serve as a upward bound on the possible length of a sentence -- even if the Guidelines would otherwise mandate a longer sentence.

6. The post-Guidelines era differs from the pre-Guidelines era in two significant respects: the Guidelines themselves (which apply to all crimes committed after November 1, 1987) and statutory minimum sentences (which Congress has enacted with regularity since the mid-1980s and which apply to a significant portion of federal prosecutions).

7. S. Rep. No. 225, 98th Cong. 1st Sess. 52 (1984), reprinted in 1984 U.S. Code Cong. & Admin. News 3182 at 3235.

8. 28 U.S.C. § 991(b)(a)(B).

9. S. Rep. No. 225, *supra* note 7, at 161, 1984 U.S. Code Cong. & Admin. News at 3344.

10. John R. Lott, Jr., Do We Punish High Income Criminals Too Heavily?, 30 Economic Inquiry 583 (1992); Jonathan M. Karpoff & John R. Lott, Jr., Why the Commission's Corporate Guidelines May Create Disparity, 3 Federal Sentencing Reporter 140 (1990). For example, Karpoff & Lott argue that the market imposes significant penalties on corporate and white collar offenders that are not accounted for under the Guidelines. These critics argue that the Guidelines increase "disparity" because judges, hamstrung by the Guidelines, are unable to equalize total punishment (including reputational sanctions) among equally culpable offenders.

11. Under our definition, disparity may exist in law-enforcement, prosecution, probation, and parole stages of criminal justice system. For example, one prosecutor may charge a defendant in such a way that he would face a mandatory minimum of ten years, while another prosecutor would prosecute that same defendant in a way that would result in a sentence of five years. Law enforcement disparity may involve the manner of investigation, such as the amount of drugs offered in a reverse sting. Probation officers may take different approaches to pre-sentence reports, and may implement parole policies differently.

12. U.S. Attorney General Annual Report 5-6 (1940).

13. See Andrew von Hirsch, Doing Justice (1976); Norval Morris, The Future of Imprisonment (1974); Kenneth Culp Davis, Discretionary Justice (1969).

14. Marvin Frankel, Lawlessness in Sentencing, 41 U. Cin. L. Rev. 1 (1972); Marvin Frankel, Criminal Sentences: Law Without Order (1973).

15. Frankel, Criminal Sentences, *supra* note 14, at 5.

16. Frankel, Criminal Sentences, *supra* note 14, at 10.

17. Anthony Partridge & William B. Eldridge, The Second Circuit Sentencing Study, A Report to Judges 1-3, 9 (1974). The study also noted that neither experience with the Eastern District of New York's practice of sentencing councils–whereby a judge confers with two other judges before sentencing–nor time on the bench seemed to increase the likelihood that a particular judge's sentence would be consistent with that of her colleagues.

18. See William Austin & Thomas A. Williams III, A Survey of Judges' Responses to Simulated Legal Cases: Research Note on Sentencing Disparity, 68 J. Crim. L. & Criminology 306 (1977) (forty-seven district court judges reviewed and sentenced five hypothetical cases; wide disparity in sentence lengths noted); Whitney North Seymour, 1972 Sentencing Study for the Southern

District of New York, 45 N.Y. St. B.J. 163 (noting gross sentencing variations in actual cases, not controlling for particular case attributes); Beverly Blair Cook, Sentencing Behavior of Federal Judges: Draft Cases, 42 U. Cin. L. Rev. 597 (1973)

19. Sentence disparity was thought to increase prison unrest. James V. Bennett, a former director of the Federal Bureau of Prisons, explained:

> The prisoner who must serve his excessively long sentence with other prisoners who receive relatively mild sentences under the same circumstances cannot be expected to accept his situation with equanimity. The more fortunate prisoners do not attribute their luck to a sense of fairness on the part of the law but to its whimsies. The existence of such disparities is among the major causes of prison riots and it is one of the reasons why prisons so often fail to bring about an improvement in the social attitudes of their charges.

J. Bennett, Of Prisons and Justice, S.Doc. No. 70, 88th Cong., 2d Sess. 319 (1964).

20. Joseph C. Howard, Racial Discrimination in Sentencing, 59 Judicature 121 (1975-76); Tom Wicker, Judging the Judges, N. Y. Times, Feb. 6, 1976, at A29; Alec Hopkins, Is There a Class Bias in Criminal Sentencing?, 42 Am. Soc. Rev. 176, 176-77 (1977); Frankel, Criminal Sentences, *supra* note 14, at 23-24.

21. See Jonathan D. Casper, Determinant Sentencing and Prison Crowding in Illinois, 1984 U. Ill. L. Rev. 231, 236-37 (explaining that "[c]onservatives and law enforcement interests" desired determinate sentencing because "parole boards seemed often to release prisoners who continued to pose a danger to society" and judges seemed "reluctant to send 'marginal defendants' to prison."); J. Edgar Hoover, The Dire Consequences of the Premature Release of Dangerous Criminals Through Probation and Parole, 27 F.B.I. L. Enforcement Bull 1 (1958); see also Reform of the Federal Criminal Laws: Hearings on S. 1437 Before the Subcomm. on Criminal Laws and Procedures of the Senate Comm. on the Judiciary 95th Cong., 1st Sess.. 8580, 8995 (1977) (statements of Sen. Lloyd Bentsen and Ronald L. Gainer).

22. 28 U.S.C. § 994(b)(2).

23. U.S.S.G. § 5K1.1; see also 28 U.S.C. § 994(m).

24. Melendez v. United States, 116 S.Ct. 2057, 2061 (1996).

25. See 18 U.S.C. 3553(e).

26. U.S.S.G. § 5K2.0. The Sentencing Guidelines themselves also identify a few, preferred bases for departure; for example, the instructions on calculation of the defendant's criminal history "score" advise that the judge should depart up or down depending upon whether the defendant's record of convictions overestimates or underestimates his criminal history. In Koon v. United States, 518 U.S. 81, 116 S.Ct. 2035 (1996), the United States Supreme Court held that federal courts of appeals should generally review the decision of the district judge to depart from the Guidelines under an abuse of discretion standard, rather than review the district court's application of the sentencing guidelines de novo. The Court also noted, "We do not understand it to have been the congressional purpose [for the Guidelines] to withdraw all sentencing discretion from the United States District Judge. Discretion is reserved within the Sentencing Guidelines, and reflected by the standard of appellate review that we adopt." Id. at 2053. It will be interesting to see whether this decision leads to an increase in departures and/or inter-judge sentencing disparity in the future.

27. Molly Treadway Johnson & Scott Gilbert, The U.S. Sentencing Guidelines: Results of the Federal Judicial Center's 1996 Survey (1997). Earlier surveys of federal judges had indicated even less satisfaction with the Guidelines. See Don J. DeBenedictis, The Verdict is In, 79 American Bar Association Journal 78 (1993) (reporting that in poll conducted by ABA, nearly half of federal judges wanted to completely abolish the Guidelines.) As one of the present authors has noted, over half of all active federal judges were appointed since the Guidelines went

into effect, and these judges may be more satisfied with the Guidelines than judges who were appointed in the era of discretionary sentencing. See Stith & Cabranes, *supra* note 4, at 5-6, 143-44.

28. See U.S.S.C., Sourcebook of Federal Sentencing Statistics 39 (1997) (showing also that use of substantial assistance departures has increased markedly over time, from less than five percent in 1989 to nearly 20 percent in 1994, 1995, and 1996).

29. See Francesca D. Bowman, Probation Officers Advisory Group Survey, 8 Federal Sentencing Reporter 303 (1996) (chief probation officers in two-thirds of federal districts responding to survey report that pleas of guilty are often accompanied by an agreement that includes Guideline stipulations or calculations).

30. John Bartolomeo *et al.*, Sentence Decision-making: The Logic of Sentence Decisions and the Extent and Sources of Sentence Disparity (1981).

31. In the Second Circuit survey, the judges were mailed the case packets over the course of six weeks, responding with proposed "sentences" by return mail. In addition to lack of face-to-face contact with the defendant and lack of advocacy by the parties, there was no simulation of the pre-sentencing procedural history of the cases, especially that relating to plea-bargaining. Nor was there any discussion of parole in the study or whether judges should sentence in "real time" or compensate for the effect of parole.

One scholar who led a re-analysis of the data of the Second Circuit study has reported that "much of the seeming difference in sentences . . . was simply a function of different understandings as to . . . how long the judge thought the defendant would actually serve." Milton Heumann, Empirical Questions and Data Sources: Guidelines and Sentencing Research in the Federal System, 6 Federal Sentencing Reporter 15 (1993) (summarizing research by Stanton

Wheeler *et al.*, Sitting in Judgment: The Sentencing of White-Collar Criminals (1988)). See also Jon O. Newman, A Better Way to Sentence Criminals, 63 A.B.A. J. 1562 (1977) (suggesting that disparity in sentencing in Second Circuit Study was increased because of different assumptions about actions of parole board); Jon O. Newman, Forward, Parole Decision-making and the Sentencing Process, 84 Yale L. J. 810, 812 (1975); Shari S. Diamond & Hans Zeisel, Sentencing Councils: A Study of Sentencing Disparity and its Reduction, 43 U. Chic. L. Rev. 109 (1975).

32. See text accompanying notes 9-11 *supra*.

33. U.S. Sentencing Commission, 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 288, 292, 296, 299 (1991).

34. A methodology similar to ours was first used by Frederick Gaudet, George S. Harris, & Charles W. St. John in Individual Differences in the Sentencing Tendencies of Judges, 23 Journal of Criminal Law and Criminology 811 (1933) and Frederick Gaudet, The Differences Between Judges in the Granting of Sentences of Probation, 19 Temp. L. Q. 4 (1946). More recently, it has been used by Joel Waldfogel in Aggregate Inter-Judge Disparity in Sentencing: Evidence from Three Districts, 4 Fed. Sent. R. 151 (1991); Joel Waldfogel, Inter-judge Disparity in Federal Sentencing: Evidence from Three Federal Districts 1984-90, (1992) (unpublished manuscript on file with authors); Abigail Payne, Does Inter-judge Disparity Really Matter? An Analysis of the Effects of Sentencing Reforms in Three Federal District Courts, 17 Int'l J. Law & Econ. 337 (1997); Paul Hofer, Kevin Blackwell, and Barry Ruback, The Effect of the Federal Sentencing Guidelines on Inter-judge Sentencing Disparity (1998) (Unpublished manuscript, U.S. Sentencing Commission).

35. Note that the hypothesis is about the change in overall disparity and not about changes in

behavior by particular judges over time. For example, if the two judges simply reversed roles between the two periods then there would be a change in θ_h and θ_l but no change in ΔD, the expected disparity from the point of view of the defendant. Waldfogel, Inter-judge Disparity in Federal Sentencing, *supra* note 34, tests for a change in disparity by comparing a model in which the judge means are restricted to be the same over time and one in which they are allowed to vary. This is implicitly a test of changes in the behavior of individual judges and not of overall disparity.

36. Analyses based on percentage of variation explained or on F-tests for equality of judge means to evaluate the impact of the Guidelines—as used by Payne and Hofer *et al.*, *supra* note 34—are vulnerable to these potentially confounding factors.

37. By examining the behavior of the same judges over time, the judges are necessarily older in the later periods. Experience may affect behavior, perhaps bringing judges closer together over time as they observe each other's work. We do not expect these changes to be substantial over short periods of time, however, such as the consecutive two year periods that are used in the empirical analysis.

38. The intuition behind this bias is that we don't actually know which judge is truly more harsh and which truly more lenient. By using the absolute difference, we always infer that the one with the higher average sentence length is harsher. In repeated sampling, the truly harsher judge may have a lower mean sentence length in some samples even though the mean sentence length is higher on average. These misclassified instances, where the true difference between the harsh and more lenient judge is negative but measured as positive, are the source of the bias. When the true means are close together and the variance of the underlying distribution is large so that the means are imprecisely estimated, this type of misclassification may occur frequently even though

the estimates of the means themselves are unbiased. Similar intuition applies to other dispersion measures like the variance.

39. A related alternative to the Gini coefficient would be estimation of the variance of a random judge effect if the judge effects are cast in a variance components model, as discussed by Shayle Searle *et al., Variance Components* (1992), at 168-226. When the data are not normal, the asymptotic distribution of the variance of the random effect depends on fourth moments. Our Monte Carlo experiments performed using data distributed like actual sentencing data found that the asymptotic approximation for the standard error on changes between periods in the between-judge component of the variance was far too dispersed to be useful, with two standard deviations covering the truth nearly 100 percent of the time, as opposed to the predicted 95 percent coverage.

40. The negative binomial model in particular is well-suited for the "overdispersed" nature of the sentence length data relative to the more traditional Poisson model. See Jerry Hausman, Bronwyn Hall, & Zvi Griliches, Econometric Models for Count Data with an Application to the Patents-R&D Relationship, 52 Econometrica 909 (1984); A. Colin Cameron and Pravin Trivedi, *Regression Analysis of Count Data* (1998) at 27-37.

41. This augmentation has been previously developed for the Poisson model. See John Mullahy, Specification and Testing of Some Modified Count Data Models, 33 Journal of Econometrics 341 (1986); Diane Lambert, Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing, 34 Technometrics 1 (1992); Shiferaw Gurmu, Paul Rilstone, and Steven Stern, Semiparametric Estimation of Count Regression Models, 88 Journal of Econometrics 123 (1999).

42. Hausman, Hall, & Griliches, *supra* note 40, parameterize the random effect with a beta

distribution, which results in a likelihood function that has a particularly simple form for a single period but is not easily extended to allow correlation in the random effects across periods. Multivariate normal random effects for the Poisson model have been explored by Siddhartha Chib, Edward Greenberg, and Rainer Winkelman, Posterior Simulation and Bayes Factors in Panel Count Data Models, 86 Journal of Econometrics 33 (1995).

43. To see the implication of this, consider a simple example. Say that an office has a superpopulation of potential judges with $f_l(3.5, 0.1)$, and receives two judges h and l, and $\theta_h = 36$ and $\theta_l = 30$. The standard deviation of $\log(\theta)$ in this sample of actual judges in the district is 0.091. This estimation strategy will identify the dispersion in the actual small sample, 0.091, and not the dispersion in the superpopulation of potential judges, 0.1. In general, the dispersion in a small sample will be less than the superpopulation from which it is drawn. In a simulation of 100,000 samples of judge effects from a lognormal distribution $f_l(3.5, 0.1)$, the average standard deviation of the log was 0.056 for a sample of two judges, 0.080 for four judges, and 0.096 for 22 judges. This implies it is important that the number of judges within a district office are the same when making comparisons between time periods using this methodology.

44. To enhance the comparability between results for the overall mean and those for accounting for offense types, we use direct weighting in equation (9) so that the estimates are sensitive to the sentence length in the same way as the estimates for the unweighted data. Another way to account for offense types would be to include covariates directly in the model, such as including indicators for offense types in μ. If there were differences in disparity by offense type, this strategy would instead essentially weight by the number of cases for each offense type without regard to the average sentence length for that offense type -- complicating the comparison between estimates adjusted to account for offense types and unadjusted estimates.

45. The data used in this paper are the same as the Cases Terminated files of the Administrative Office of the U.S. Courts (AOUSC), with documentation available from the Federal Justice Statistics Program data in the National Archive of Criminal Justice Data, except that the judge code is not suppressed as it is in the public-use file. By agreement with the AOUSC, we cannot provide any information that would result in the disclosure of data about specific judges. Therefore, we do not report the identity of specific district offices used in this analysis.

46. In order to be included in a sample for analysis, a judge had to have been assigned a proportion of cases within a certain range of the maximum assigned to any judge in that office in any one year period. A chi-square test statistic was then computed for the null hypothesis that the cases were distributed between with equal probability in the final sample of cases used for analysis. The p-values from this test should be uniformly distributed between zero and one. The size of the range used was 3.9 standard errors of the maximum proportion, which was calibrated so that the median was 0.50 for the 304 test statistics for offices and periods used in the final sample. It appears the most judges excluded by this rule either were not in the district for the entire period or were likely on senior status and consistently receiving a reduced caseload. The identity of the judges is masked in our data, however, so senior status cannot be directly verified.

47. The exact time periods cover one calendar year, with the following exceptions. To increase the sample size for the two earliest periods, the 1982 period is actually from July 1981 to September 1982, and the 1983 period is from October 1982 to December 1983. The 1988 period includes November 1987 through January 1989, including cases from the initial promulgation of the Guidelines through the ruling on their constitutionality. (In Mistretta v. United States, 488 U.S. 361 (1989), the United States Supreme Court upheld the constitutionality of both the Sentencing Guidelines and the Commission in the face of challenges under the nondelegation

doctrine and the separation of powers.)

48. The particular methods used to assign cases to judges vary between districts. In some districts, judge's names are drawn from a deck of cards—sometimes shuffled in the dark and/or sealed with wax to prevent tampering. In others, envelopes with the judge's names are placed in a box or circular hopper that is spun before the clerk pulls an envelope out of the box. In the 1990s, many districts have adopted computer software to implement the random assignment. Also, in order to more evenly distribute potentially lengthy cases, random assignment is sometimes stratified by projected length of the case.

Paul Hofer at the U.S. Sentencing Commission also provided information about case assignment practices based on interviews conducted during his own research, and from interviews previously conducted by the Federal Judicial Center as part of the FJC Time Study (on file at the Federal Judicial Center). Personal communication with Hofer, December 4, 1997.

49. One model is that the true judge means are identical within each office. To simulate the bias when the true g is zero, we created data where each judge was randomly matched with the sentencing outcomes of N_{jt} cases from the district office. An alternative model is that the estimated judge mean is an unbiased estimate of the true mean. Treating the observed distribution as the true distribution, an estimate of the bias can then be computed from the difference between the mean of the bootstrap estimates and the original estimate. We computed bootstrap estimates of the bias for this model by drawing bootstrap samples for each judge from the observed distribution of sentencing outcomes for that judge. In 1986-87, for example, the estimated g was .154. In 500 replications, the bias was estimated to be .122 from the true mean zero simulation and .038 from the bootstrap procedure. The middle ground between these two models would allow differences in the judge means while incorporating the intuition that an

estimated judge mean particularly different from the district office mean is more likely to have come from an unusual sample of cases (rather than treating these extremes as unbiased estimates of the true mean for that judge).

50. Note that the bias in absolute measures of disparity, such as the variance or the mean absolute deviation, is increasing over time as the underlying variance of the data is increasing. The mean of the data is also increasing over time, however, so that ratio of the standard deviation to the mean in Table I has actually decreased slightly over time, implying slightly less bias for relative disparity measures in later periods. A counteracting factor, however, is that the true means appear to be closer together in later periods, which increases the bias as discussed in note 40.

51. The weights and abscissae used were from the 16 point Gauss-Hermite rule in Gwynne Evans, *Practical Numerical Integration* (1993) at 308. 25 point and 40 point rules were also tried, and resulted in almost identical point estimates. (Estimation programs in MATLAB are on file with the authors).

52. In order to assess the usefulness of these asymptotic approximations for inference in finite samples, we conducted Monte Carlo simulations where the true parameters of the model were known. The asymptotic standard errors reported in the paper appear to be slightly too small, with the confidence interval of 1.96 times the standard error including the truth about 90 percent of the time as opposed to the asymptotic prediction of 95 percent coverage.

53. Waldfogel, Inter-judge Disparity in Federal Sentencing, *supra* note 34, analyses three districts (CT, SDNY, NDCA) from 1984-90 and finds an increase in inter-judge disparity in two of the three districts during 1988-90, although no standard errors for these estimates are reported.

Payne, *supra* note 34, analyses three districts (EDNY, SDNY, EDPA) from 1980-91. For

property crimes, she finds that inter-judge disparity declines for one of three districts. For drug crimes, she finds declines for all three districts. Payne emphasizes that the fraction of variation explained by mean judge effects is small relative to the total variation. This is true and tells us that there are many additional factors that drive differences in sentences, but it does not lead us to conclude that inter-judge disparity itself is small or unimportant. Just as with other empirical relationships, such as the amount of variation in wages that is explained by differences in education levels, the fact that the percentage of explained variation is small is often less important than the magnitude of the coefficients associated with the variable of interest, such as differences in average wages between education levels or (as we believe in this case) with differences in average sentences between judges.

Hofer *et al.*, *supra* note 34, compare 1984-85 to 1994-95 using the same 42 judges in nine district offices in the part of their analysis most comparable to ours. They also focus on percentage of variation explained, and report that the partial R-squared drops from 2.32 in 1984-85 to 1.08 in 1994-95.

As pointed out in Section IV, there are several methodological differences between our analysis and those by Waldfogel, Payne, and Hofer *et al.* Most importantly, we measure the magnitude of inter-judge disparity directly, and provide a confidence interval to assess the statistical precision of the estimate of the change before and after the Guidelines, so the null hypothesis of no change can be tested. Our data is also more complete, covering 26 district offices over up to twelve and a half years. In addition to the advantages of sample size and representativeness, our data have been constructed with criteria selecting districts using random assignment of cases to judges that appears to be more stringent. Of the 13 districts in these three

studies, only three had procedures and caseloads that appeared to us to consistently use random assignment in the time periods under study here and was included in our analyses.

54. The conversion of percentage differences into months depends in part upon the treatment of acquittals and dismissals. Cases (and not convictions) are randomly assigned to judges, and we use all cases assigned in computing our estimates. An argument can be made that acquittals and dismissals are independent of the judge assigned to the case, based on the logic that judges have much more discretion over sentencing for convicts than for other dispositions. Statistical tests for independence analogous to those for offense types described in Section V indicate that the combined fraction of acquittals and dismissals is roughly balanced across judges. Under true independence, with a mean p-value of the 606 chi-square statistics for independence of judge and acquittal/dismissal should be .5, and the actual value is .46.

Inclusion of acquittals and dismissals appear to have a negligible effect on the estimates of the Gini coefficient, since this is roughly equivalent to re-scaling the judge means by a constant factor. If we were to use the mean sentence for convictions (33 in 1986-87 and 39 in 1988-93) instead of the mean for all cases, the results would be an expected difference of 5.7 months in 1986-87 and 4.4 in 1988-93.

55. Since the data are weighted to reflect the 1986-87 case mix within each district office, the roughly 10 percent reduction in the magnitude of point estimates for the 1986-87 period is not due to changes in the offense mix over time. Instead, it reflects the fact that different offense types are weighted exactly equally in modeling the overall distribution for each judge, instead of the approximate equality of random assignment.

56. For context, note that the consistency in public behavior over time is even greater among representatives in the U.S. Congress, where the correlation in indices of voting patterns is 0.78 to

0.95 (depending on the index) when examining the voting by the same representative separated in time by two terms. See John Lott & Stephen Bronars, Time Series Evidence on Shirking in the U.S. House of Representatives, 76 Public Choice 125 (1993).

57. Our data only record date of case filing and termination, and not date of offense or use of the Guidelines in sentencing. As shares of cases filed in 1988-89, .3 were terminated in 1988, .43 in 1989, and .27 in 1990 or later. The United States Sentencing Commission, 1990 Annual Report at 39 tabulates the fractions of terminated cases sentenced under the Guidelines to be .18 in 1988, .55 in 1989, and .70 in 1990. As a rough estimate of Guideline application, we use the shares of cases filed as weights to infer that the average fraction of Guidelines application was about 48 percent for cases filed in 1988-89.

58. Before the Sentencing Guidelines were even promulgated, Congress enacted the Anti-Drug Abuse Act 1986, Pub. L. No. 99-570, 100 Stat. 3207, which provided an array of mandatory minimum penalties for narcotics offenses and violent crimes. Most significantly, the Act "set up a new regime of non-parolable, mandatory minimum sentences for drug trafficking offenses that tied the minimum penalty to the amount of drugs involved in the offense." U.S. Sentencing Commission, Special Report to Congress: Mandatory Minimum Penalties in the Federal Criminal Justice System 8 (1991). These mandatory minimum sentences were severe for all drugs, but especially for offenses involving crack cocaine, and they applied to nearly all drug offenses prosecuted in the federal courts. Because statutory changes in sentencing law are not applied retroactively to crimes before the enactment of the statutes, and because there is a substantial delay between the commission of a crime and sentencing, the effect of mandatory minimums (like the Sentencing Guidelines) phases in gradually over several years.

59. Frankel, Criminal Sentences, *supra* note 14, at 29.

60. *Id.*

61. Officially, offenders are to be charged with the most serious offense which can be proved at trial with certain limited exceptions. In practice, however, charging policies vary widely. See 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; Ahmed Taha, The Effect of the Federal Sentencing Guidelines on the Disposition of Criminal Cases, (1998) (Unpublished manuscript, U.S. Department of Justice) (provides evidence that prosecutors filed less serious charges against the average defendant, defendants pleaded guilty to charges that were closer to the charges prosecutors filed, more defendants initially pleaded guilty); Joe Brown, Quo Vadis? What Congress and the Department of Justice should do in Response to the Justice Department's Analysis of Non-Violent Drug Offenders with Minimal Criminal Histories, 7 Fed. Sent. R. 25, 26-27 (1994) (former U.S. Attorney criticizes disparity among federal prosecutors in use of 5K1.1 motions and in charging to avoid mandatory minimums: "popular way to avoid mandatory sentences entirely [is] by charging a telephone count under 21 U.S.C. § 843(b), which does not involve a mandatory sentence."); Robert H. Edmunds Jr., Guidelines Sentencing and Department of Justice Policies Under the Reagan-Bush Administrations, 6 Fed. Sent. R. 306 (1994) (noting that in the district in which he was U.S. Attorney, the Middle District of North Carolina, an illegal alien who was convicted of reentry after deportation could expect a "term of years" while in a California border district, illegal immigrations were common and the aliens would receive far lighter sentences.)

There is an enormous range in policies among U.S. Attorney's offices for making a motion for downward departure on the basis of substantial assistance (§5K1.1 motion), which allows a judge to sentence below the Guidelines range. In the District of Connecticut, any

§5K1.1 motion must be approved by a committee including the U.S. Attorney for the District. The §5K1.1 departure rate in the district was 8.5 percent in 1994. In contrast, judges in the Eastern District of Pennsylvania departed downward on § 5K1.1 grounds in 49.3 percent of cases in 1994.

62. Gerald H. Heaney, The Reality of Guidelines Sentencing: No End to Disparity, 28 Am. Crim. L. Rev. 161, 200 (1991) (role of probation officers varies widely by district; one federal defender describes probation as more adversarial than prosecution.).

63. See, for instance, United States v. Giles, 768 F. Supp. 101 (S.D.N.Y.), aff'd, 953 F.2d 636 (2d Cir. 1991), cert. denied, 503 U.S. 949 (1992).

64. This result may reflect legislative intent. Sponsors of Sentencing Reform Act included both liberals primarily concerned with disparity in sentencing and conservatives primarily concerned with leniency in sentencing.

65. See Stith & Cabranes, *supra* note 4, at 78-103.

66. See Lott, *supra* note 10; Karpoff & Lott, *supra* note 10. These critics have argued that the Guidelines have wrongly eliminated sentencers' ability to consider the reputational effects of a conviction for white-collar and corporate offenders. They argue that as a result of this inability to consider significant extra-legal sanctions, variation in total penalties (incarceration, fines, reputational penalties, and lost earnings) have increased.

67. Albert Alschuler, The Failure of Sentencing Guidelines: A Plea For Less Aggregation, 58 U. Chicago L. R. 901 (1991); Stephen Schulhofer, Assessing the Federal Sentencing Process: The Problem is Uniformity, Not Disparity, 29 Am. Crim L. R. 833 (1992); Daniel Freed, Federal Sentencing in the Wake of the Guidelines: Unacceptable Limits on the Discretion of Sentencers, 101 Yale L. J. 1682 (1992).

TABLE I

DESCRIPTIVE STATISTICS OF PRISON SENTENCES IN MONTHS

	1982-83	1984-85	1986-87	1988-89	1990-91	1992-93
% 0 Months (Acquittals)	3	2	2	2	2	1
% 0 Months (Dismissals)	13	13	11	10	10	10
% 0 Months (Convictions)	32	35	33	27	28	24
% 1-24 Months	26	25	24	29	27	28
% 25-48 Months	12	11	11	11	9	12
% 49-96 Months	9	8	10	12	12	14
% 97-540 Months	6	6	8	9	11	11
Mean	24	24	29	32	36	38
Standard Deviation	50	53	57	60	66	67
Offices	23	26	26	25	26	26
Judges	139	148	154	157	161	159
Cases	11997	12007	12575	13689	13604	13329

Note: Data are from Cases Terminated files of the Administrative Office of the U.S. Courts. For sample creation details, see Section V.

TABLE II

QUANTILES OF P-VALUES FROM

CHI-SQUARE TESTS OF INDEPENDENCE BETWEEN JUDGES AND OFFENSE TYPES

Quantile	.10	.25	.50	.75	.90
P-value	.09	.23	.47	.76	.92

Note: 606 Chi-square statistics were computed for 26 offices and for up to 25 six month periods. Quantiles of p-values are weighted by number of cases in that office and time period. The offense types were grouped into four categories: violent & weapons, drug, embezzlement & fraud, and other cases.

TABLE III

OBSERVED VS. PREDICTED CELL PROBABILITIES OF 3 PARAMETER MODEL

Months of Prison Term	Observed	Predicted
0	.464	.464
1-12	.154	.135
13-24	.088	.084
25-48	.107	.110
49-96	.106	.113
97-192	.059	.072
193-540	.020	.020

Note: Estimates are based on equation (4), assuming that the three parameters δ, γ, and θ are constant across all cases, using pooled data for 1986-87.

TABLE IV

SHARE OF CASES BY OFFENSE TYPE

	1982-83	1984-85	1986-87	1988-89	1990-91	1992-93
Drug	.21	.25	.27	.32	.30	.33
Violent & Weapons	.16	.14	.14	.15	.18	.17
Embezzlement & Fraud	.26	.28	.31	.28	.29	.26
Other	.37	.33	.28	.25	.23	.24

Note: Shares based on data described in Table I.

TABLE V

CHANGES OVER TIME IN GINI COEFFICIENT ESTIMATES
OF INTER-JUDGE SENTENCING DISPARITY

	unweighted			weighted for offense type comparability		Judges
1982-83	1984-85	Change	1982-83	1984-85	Change	
.079	.107	.028	.092	.102	.011	118
(.009)	(.008)	(.011)	(.008)	(.009)	(.012)	
1984-85	1986-87	Change	1984-85	1986-87	Change	
.102	.092	-.010	.097	.080	-.017	126
(.009)	(.009)	(.012)	(.009)	(.008)	(.011)	
1986-87	1988-89	Change	1986-87	1988-89	Change	
.090	.039	-.051	.079	.042	-.038	133
(.008)	(.010)	(.013)	(.008)	(.010)	(.012)	
1988-89	1990-91	Change	1988-89	1990-91	Change	
.055	.057	.002	.057	.053	-.004	128
(.010)	(.008)	(.013)	(.008)	(.009)	(.014)	
1990-91	1992-93	Change	1990-91	1992-93	Change	
.059	.046	-.012	.045	.041	-.004	120
(.008)	(.010)	(.012)	(.008)	(.011)	(.014)	

Note: Estimates are for Gini coefficients γ (with standard errors in parentheses) from equations (7) and (8) using data summarized in Table I. Unweighted estimates are based on w_t in (4), while weighted estimates are based on w_{ijt} from equation (9). As described in the text, the weights statistically adjust each judge's caseload over time to reflect the district office's mix of offense types in 1986-87. The four offense types used are violent/firearms, drug, embezzlement and fraud, and other.

TABLE VI

CORRELATION OF JUDGE EFFECTS BETWEEN TIME PERIODS

1982-83 & 1984-85	.73 (.09)
1983-84 & 1985-86	.70 (.09)
1984-85 & 1986-87	.75 (.07)
1985-86 & 1987-88	.22 (.14)
1986-87 & 1988-89	.68 (.13)
1986-87 & 1988-89	.34 (.17)

Note: Estimates of ρ (with standard errors in parentheses) are from maximum likelihood estimation of equation (7), using data summarized in Table I.

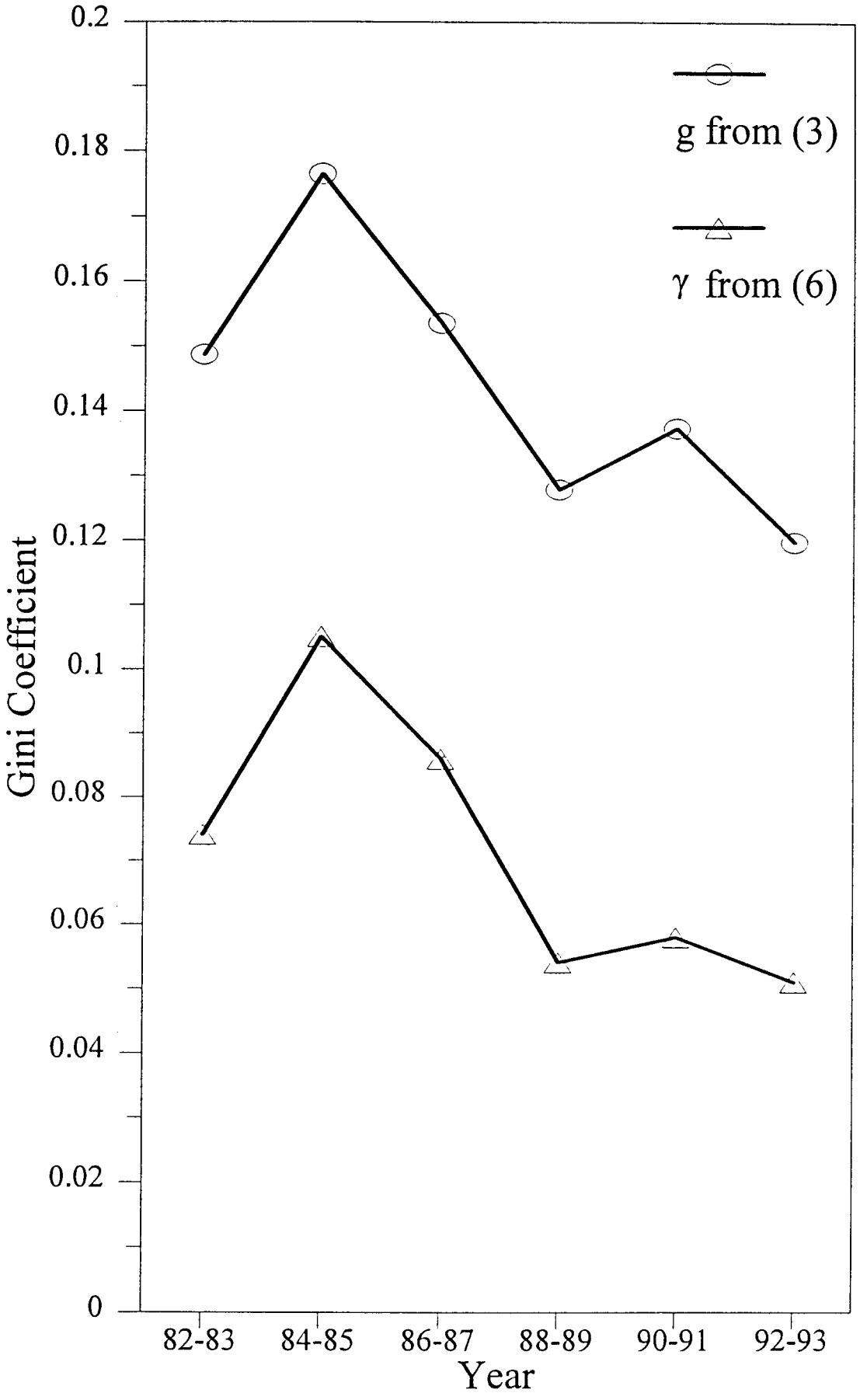

Figure 1: Interjudge Disparity for All Judges

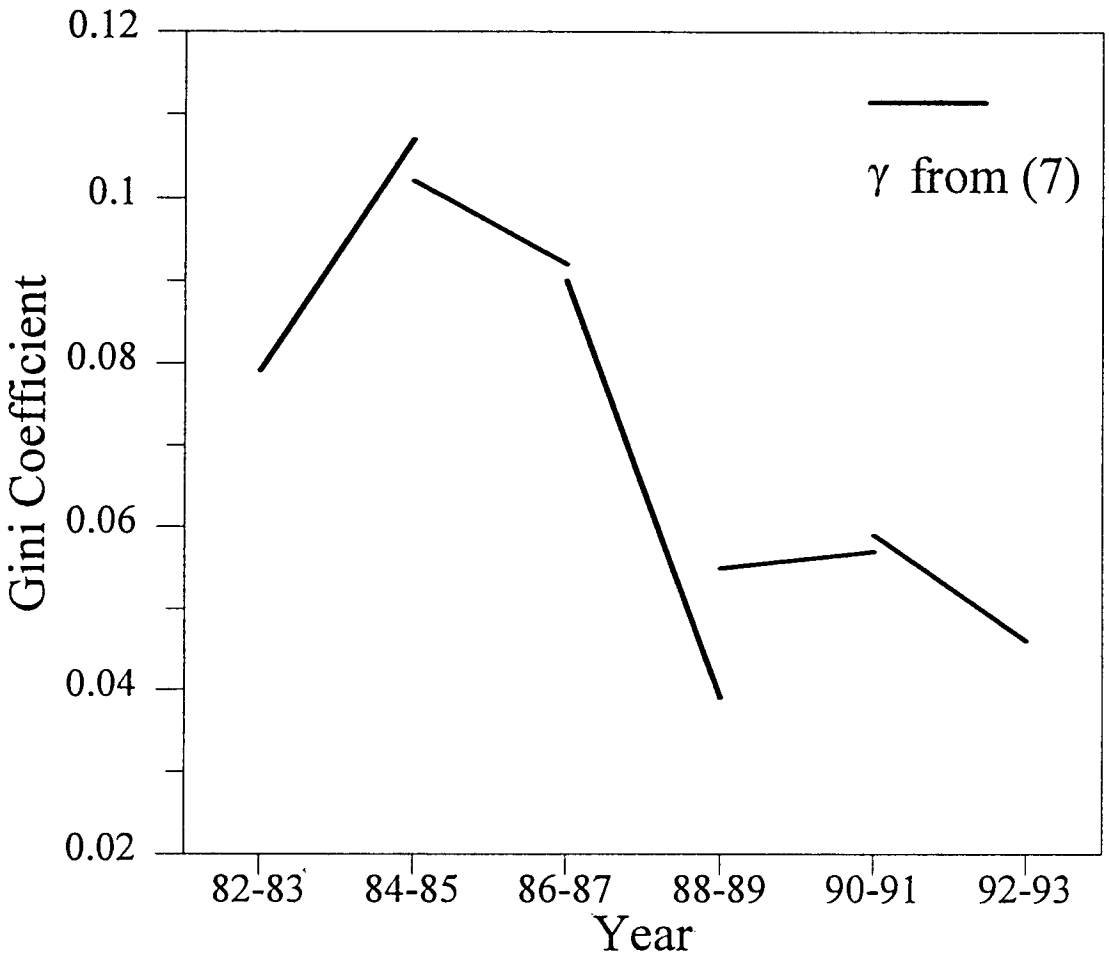

Figure 2: Interjudge Disparity for the Same Judges in Two Periods