Social Science Research xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

Exploring the origin of sentencing disparities in the Crown Court: Using text mining techniques to differentiate between court and judge disparities

Jose Pina-Sánchez[a,*], Diana Grech[a], Ian Brunton-Smith[b], Dimitrios Sferopoulos[c]

[a] School of Law, University of Leeds, Leeds, UK
[b] Department of Sociology, University of Surrey, Guildford, UK
[c] University of Edinburgh, Edinburgh, UK

ARTICLE INFO

Keywords:
Sentencing disparities
Crown Court
Court cultures
Cross-classified multilevel model
Data scraping
Text mining

ABSTRACT

Research on sentence consistency in England and Wales has focused on disparities between courts, with differences between judges generally ignored. This is largely due to the limitations in official data. Using text mining techniques from Crown Court sentence records available online we generate a sample of 7,212 violent and sexual offences where both court and judge are captured. Multilevel time-to-event analyses of sentence length demonstrate that most disparities originate at the judge, not the court-level. Two important implications follow: i) the extent of sentencing consistency in England and Wales has been underestimated; and ii) the importance attributed to the location in which sentences are passed – in England and Wales and elsewhere - needs to be revisited. Further analysis of the judge level disparities identifies judicial rotation across courts as a practice conducive of sentence consistency, which suggests that sentencing guidelines could be complemented with other, less intrusive, changes in judicial practice to promote consistency.

1. Introduction

Consistency in sentencing is a central principle to criminal justice in common law systems. Loosely understood, this principle invokes that only relevant factors should influence the sentence outcome, and that the approach taken in considering those factors ought to show a degree of stability within a given jurisdiction. A system unable to uphold this principle would struggle to ensure equality before the law or to maintain that the punishment fits the crime, two additional key principles lying at the core of most criminal justice systems. From a more instrumental point of view, promoting consistency in sentencing tackles the generation of 'unwarranted disparities'. Unlike legitimate variations in sentencing, which reflect meaningful case or offender differences, unwarranted disparities are rooted in unjustified discrepant approaches to sentencing (Stolzenberg and D'Alessio, 1994; Pina-Sánchez and Linacre, 2016). They can take the form of anything from discriminatory practices against ethnic minorities (see for example Albonetti, 2002; Bushway and Piehl, 2001; Hood, 1992; King and Johnson, 2016) to more random but still illegitimate disparities (see Chen et al., 2016; Danziger et al., 2011; Eren and Mocan, 2018; Heyes and Saberian, 2018; where factors such as the weather, lunch breaks, or local sport results are shown to influence sentencing). Thus, the promotion of consistency helps to foster transparency and predictability in sentencing, which in turn increases public confidence in the criminal justice system (Roberts and

* Corresponding author. School of Law, The Liberty Building, University of Leeds, Leeds, LS2 9JT, UK.
E-mail address: j.pinasanchez@leeds.ac.uk (J. Pina-Sánchez).

https://doi.org/10.1016/j.ssresearch.2019.102343
Received 23 October 2018; Received in revised form 10 July 2019; Accepted 28 August 2019

Plesničar, 2015). It comes as no surprise that many jurisdictions have taken great strides to promote consistency in sentencing. The most relevant experiences are those that have taken place in the United States (both at the Federal and State level) and the United Kingdom (separately for the jurisdictions of England and Wales, and Scotland), which have sought to limit unwarranted disparities by restraining judicial discretion through systems of sentencing guidelines.

The relevance of the concept of sentencing disparities has been matched by the research efforts dedicated to its empirical exploration. However, the vast majority of studies on this topic have focused on the detection of specific illegitimate factors showing an undue influence in sentencing. Studies aiming to measure the overall level of consistency in sentencing - or equivalently, the extent of unwarranted disparities - are scarcer, and have not been able to offer entirely satisfactory answers. Largely this is due to the difficulties at the operationalisation stage (see Cole, 1997; Hofer et al., 1999; Pina-Sánchez and Linacre, 2016). One of the most common approaches pursued in the literature to measure the extent of unwarranted disparities involves the use of multilevel models. This require access to individual records capturing not only details of the case but also identifying the hierarchical nature of sentencing (i.e. sentences clustered within judges, clustered within courts). When sentence data available to researchers includes the locations where each sentence was imposed, multilevel models have commonly been used to measure average disparities between areas while controlling for the different mix of cases processed in each of those locations. These studies have explored differences between districts (Feldmeyer and Ulmer, 2011; Hamilton, 2017; Johnson et al., 2008; Kautt, 2002; Schanzenbach and Tiller, 2006; Ulmer et al., 2011; Ulmer and Johnson, 2017), counties (Britt, 2000; Fearn, 2005; Haynes, 2011; Hester and Sevigny, 2016; Johnson, 2005; Ulmer and Johnson, 2004), or specific courts (Drápal, 2018; Pina-Sánchez et al., 2017, 2018a). Typically, variations in sentence length or type, stemming from any of the levels listed above, which cannot be explained by legitimate case characteristics (e.g., offence severity), are used as measures of inconsistency in sentencing.

These studies are premised (implicitly or explicitly) on the 'court communities' perspective (Hester, 2017). This theory views courts as communities made up of actors such as prosecutors, judges, defence lawyers, and court staff (Eisenstein et al., 1988; Flemming et al., 1992; Ulmer, 1997). Interdependent working relationships emerge between these actors - the 'courtroom workgroup' (Eisenstein and Jacob, 1977) – which produce a distinct social order or 'local legal culture' (Church, 1985) that ultimately shapes case processing and decision-making. The way that the courtroom workgroup functions - and thus, the way a specific culture is formed - is influenced by workgroup interactions (e.g., attitudes, preferences, familiarity of members) as well as environmental factors such as each member's sponsoring organization and the social, economic, or political features of the local court community (Eisenstein et al., 1988; Ulmer and Johnson, 2004). According to this perspective, sentencing disparities are rooted in the unique attributes of each location's respective culture.

The importance placed on location in examining and explaining sentencing consistency has been particularly strong in England and Wales. In this jurisdiction, the principle of consistency is largely interpreted as the absence of between court disparities. This was nowhere better reflected than by Lord Levenson, chairman of the Sentencing Council for England and Wales at the time its first guideline was introduced. In an interview to the BBC Lord Levenson stated, "*[…] the aim [of the new assault guideline] is to increase the consistency of approach to sentencing so that offenders receive the same approach whether they're being sentenced in Bristol, Birmingham, Bolton or Basildon*".[1] Similar interpretations have been made by other government bodies, think tanks,[2] and recent academic research undertaken in this jurisdiction. Using multilevel modelling and Sentencing Council data capturing court locations Pina-Sánchez and Linacre (2013) sought to estimate the level of consistency in sentencing in the Crown Court. Pina-Sánchez and Grech (2017) tested whether court and area characteristics could be behind the observed disparities between courts. Pina-Sánchez et al. (2018a) expanded this work by looking at disparities arising throughout the sequence of steps listed in the sentencing guidelines in England and Wales. Pina-Sánchez (2015) estimated changes in between court disparities following the introduction of the new assault guideline, while Pina-Sánchez et al. (2017) looked at similar changes following the 2011 summer riots. With the exception of Pina-Sánchez et al. (2017), this body of work points at between court disparities in the Crown Court being relatively negligible, which has been interpreted as evidence of high levels of consistency in sentencing.

Such a conclusion is problematic since, even in the presence of complex 'court cultures', the final sentence is up to the discretion of individual judges. Nevertheless, this crucial level of analysis has been ignored in previous multi-level analyses of sentence data in England and Wales. The importance of having missed judge-level disparities cannot be overstated. Nearly a half-century ago Hogarth (1971) famously concluded that, "*one can explain more about sentencing by knowing a few things about a judge than by knowing a great deal about the facts of the case*" (1971: 350), with much subsequent research highlighting the importance of examining judge characteristics more closely when trying to explain unwarranted variations in sentencing (see for example, Croyle, 1983; Divine, 2018; Long and Burnham, 2012; Ulmer, 2012; Yang, 2014, 2015; Zatz, 2000).

The importance of examining the judge's role is highlighted by the 'focal concerns' perspective, which views sentencing as a process based on three primary concerns: blameworthiness, protection of the public, and practical constraints (Steffensmeier and Demuth, 2000; Steffensmeier and Britt, 2001; Steffensmeier et al., 1993; Steffensmeier et al., 1998; Ulmer and Johnson, 2004). Judges are said to use 'perceptual shorthand' (Hawkins, 1981) when assessing these concerns, relying on legitimate offender/offence characteristics as well as non-legal contextual factors, due to a lack of time, resources, and complete background information. Other important judge characteristics that have been shown to lead to betweenjudge disparities in sentencing are gender (Steffensmeier and Hebert, 1999), or time on the bench (Spohn, 1990b), with female and more experienced judges shown to sentence more harshly.

[1] The newsletter can be downloaded from this link: http://www.bbc.co.uk/news/uk-12681250.

[2] See for example the following press release from the Howard League for Penal Reform (2013) published by the BBC here, https://www.bbc.co.uk/news/uk-22243680, or Mason et al. (2007) from the Ministry of Justice, where the focus lies on Police Force Areas.

From a methodological standpoint, if disparities originating at the judge level are statistically significant over and above the noted between court disparities, then measures of consistency in sentencing limited to the latter will be biased upwards. That is, it is likely that previous studies – such as the recent work stemming from England and Wales - have overestimated the true extent of sentence consistency or underestimated the true extent of sentence disparities. Evidence supporting this hypothesis can be found in Johnson (2006, 2014). Using data from Pennsylvania the author specified various multilevel models accounting for court and judge disparities simultaneously. In all the models presented, between judge disparities were found significant. For example, looking at sentence length for cases that went to trial Johnson (2014) showed disparities at the judge level to be roughly four times larger than the disparities observed at the court level.

As far as we are aware, Johnson (2006, 2014) and Yang (2014) represent the only studies where between court and judge disparities have been modelled simultaneously.[3] We argue that this is due to the well-documented reticence of the judiciary towards independent research (Ashworth, 2003; Baldwin, 2008; Darbyshire, 2011; Dhami and Souza, 2010; King, 2017; Tata, 2013; Ulmer, 2012), which is manifested in limitations to the recording and publication of official sentence data. One key example is the case of the Transactional Records Access Clearinghouse (TRAC), where researchers based at Syracuse University spent decades in legal disputes before they were able to obtain data identifying the judge who passed each sentence.[4]

Judges have argued that the release of sentencing data could simplify a complex decision-making process that cannot be fully captured quantitatively (Gertner, 2012), as well as compromise judicial independence by drawing judges into policy debates that should be left to executive and legislative branches.[5] Former Chief Justice Rehnquist, who served on the Supreme Court of the United States for 33 years, went as far as to say that collecting data on judges' sentencing practices "*could amount to an unwarranted and ill-considered effort to intimidate individual judges*" and would "*seriously impair the ability of courts to impose just and reasonable sentences*".[6] Whilst these concerns are undoubtedly important, they must be considered alongside the significant impact that refusal to provide this data has had on our understanding of sentencing disparities. "*Often the only viable alternative to the use of official sentencing data is to undertake primary data collection, which requires extraordinary efforts and resources to achieve large enough samples.*"

Here we suggest an alternative approach to access the necessary data based on text mining sentence records available online. We download and code a sample of 7,221 violent and sexual interpersonal offence records from 'The Law Pages', an online case law database from the England and Wales Crown Court. This is a commercial website used by legal practitioners to follow the development of relevant cases, and by members of the general public interested in hiring legal services. The website's archive is updated daily including excerpts from sentence transcripts, certain case characteristics, and crucially, the name of the judge who passed the sentence and the court where the trial took place. We use this new data to estimate and compare the extent of sentencing disparities stemming from the court and judge level in the Crown Court. This allows us to obtain new measures of sentence consistency, test the validity of the conclusions from previous research, and assess the presence of the 'court culture' phenomenon in the Crown Court of England and Wales.

Beyond obtaining specific measurements of between judge disparities, this new dataset offers a unique opportunity to delve into the origin of such disparities. We do so by exploring two unique features of the Crown Court: the variety of judges operating within it, and the practice of judicial rotation. Judicial rotation is a common practice amongst recorders and circuit judges, who represent broadly half of the judges in England and Wales (Ministry of Justice, 2019). Importantly, recent qualitative evidence suggests that the practice of rotation might contribute to foster consistency in sentencing. Specifically, Hester (2017) suggests that the limited sentencing disparities detected in South Carolina (Hester, 2012; Hester and Sevigny, 2016) may be rooted in their practice of rotating judges between courts, which facilitates the spread of norms and encourages judges to harmonize their sentencing practices. Hence, accounting for the identity of the judge responsible for each sentenced case becomes essential to ensure accurate estimations of consistency, while allowing us to explore the role that different types of judges, and the practice of judicial rotation more generally, might have as a buffer or catalyst of unwarranted disparities.

In the next section we outline the nature of the data source and the process undertaken to code it into a series of relevant variables. This is followed by the analytical strategy, where the need for an innovative cross-classified multilevel model differentiating between court and judge disparities that can also account for judicial rotation is explained. The results from our analyses of disparities in sentence length are presented in Section 4. Their implications are discussed in Section 5, followed by a rebuttal of the arguments used by the judiciary to censor their sentence records, and an outline of the new avenues of research that text mining techniques have opened in the field of sentencing.

2. Data

A number of online databases are available to legal practitioners to research existing case law, including the British and Irish Legal Information Institute (BAILII) and Westlaw. However, these databases tend to focus on highly significant cases, sentenced in special

[3] There are other articles where judge and court identifiers were available but they were modelled separately, or simply controlled for using fixed effects, see for example Chatsverykova (2017), Myers and Talarico (1987), Lim et al. (2016), Volkov (2016) or any of the reports from the US Sentencing Commission, such as Saris et al. (2012).

[4] These legal disputes are described in detail here: http://trac.syr.edu/judges/aboutData.html.

[5] See https://www.nytimes.com/2012/03/06/nyregion/wide-sentencing-disparity-found-among-us-judges.html.

[6] See https://www.washingtonpost.com/archive/politics/2003/08/07/ashcroft-orders-tally-of-lighter-sentences/06d0c1f5-0995-441b-9bfe-cd3b0c9639fb/?noredirect=on&utm_term=.bf667353be9f.

courts like the Court of Appeal or the Supreme Court, and as such they are not representative of the average sentencing practice in England and Wales. Recently, however, a commercial website under the domain www.thelawpages.com has been making available information from more common offences sentenced in the Crown Court on a daily basis. Records from this archive, however, can only be accessed individually and much of the relevant information is not available in a way that can be easily transformed into variables amenable to statistical analysis.

We used data scraping techniques to generate an initial database of 18,220 cases that were sentenced in the England and Wales Crown Court between 2007 and 2017. Each sentence record was loaded in a browser using Perl and Selenium with relevant case characteristics automatically extracted and compiled. Special care was taken not to overload 'The Law Pages' server in the case that it could become slow or unresponsive. For this reason HTTP requests were programmed to occur every minute. Once loaded, the HTML source code was saved locally. A parser, written in Perl then looked for specific keywords within the HTML source code, from which relevant variables could be systematically recorded. A number of relevant case characteristics were extracted directly from the source code following this approach (e.g. sentence outcome, type of offence, and age of the defendant).

Variables relating to judge characteristics were generated under a second stage text mining process that focused specifically on the judge's title and name. Clauses including 'Miss', 'Ms', and 'Her' were extracted from the judges title in order to approximate the judges' gender. The position of the judge was also approximated by extracting and analysing specific clauses in their title.[7] Here we distinguish Recorders, Queen's Counsel judges (identified with the prefix 'QC') Circuit judges, (identified using 'HHJ', 'Her Honour' and 'His Honour') and High Court judges (using the clauses 'Justice' and 'Honourable'). The categories generated can be taken as proxies for seniority, with recorders working part-time and often representing the first step on the judicial ladder to appointment to the circuit bench (except when the title is used as "The recorder of [Crown Court location]", where it is indicative of the most senior judge in that court). Circuit judges comprise the biggest bulk of the judiciary in the Crown Court and generally handle more serious or complex cases relative to recorders. Finally, High Court judges represent the most senior group of judges, dealing with more complex and difficult cases. This classification is, however, far from perfect. There are other characteristics besides seniority captured by those judicial titles (for example, QC refer to special appointments that can overlap with the three titles listed above). More importantly, the full title of the judge was only reported in 21.7% of cases. Those for which only the name of the judge was available are used as the reference category in our analysis, under the assumption that they are Circuit judges, the most common category in the Crown Court.

2.1. Current study

This text mining process resulted in the extraction of a total of 52 variables relating to the case, defendant and judge. Two of these refer to the sentence outcome – the length of the prison sentence in months, and whether sentenced to life imprisonment. Two are basic defendant characteristics - age and gender. Seven relate to characteristics of the sentencing judge, including their position, gender, and whether they were found to rotate across different courts. Nine cover general case characteristics, capturing the year when the sentence was passed, the presence of co-defendants, whether multiple offences were considered, or a guilty plea entered before trial, the presence of mitigating factors, whether there was a victim impact statement, injuries were caused, and whether the offender was sentenced to remand or a public protection sentence. Lastly, we were able to differentiate between 32 specific offence types.

For the analyses presented here we make use of a subsample of 7,221 violent and sexual interpersonal offences sentenced to prison in the Crown Court from 2007 to 2017. The focus on a broadly similar group of offences limits the possibility of omitted relevant variable bias, which is a useful strategy given the difficulty of capturing the full list of relevant case characteristics in analyses of sentence data (Anderson et al., 1999; Brantingham, 1985; Hofer et al., 1999; Waldfogel, 1998). In addition, to avoid over-fitting the model and incurring problems of multicollinearity, some of the offence types are grouped together. This process was based on similarities between offence types in their legal definition but also on preliminary analyses of the effect they have on the final sentence. For example, offences coded as 'unlawful wounding', 'grievous bodily harm' and 'conspiracy to commit grievous bodily harm' were grouped into the same category. Table 1 shows descriptive statistics for all the variables that are used in the analyses presented in the next section.

2.2. External validity

The generalisability of the sample generated is unknown. 'The Law Pages' does not provide information on their website about the sampling methods used, and they have not replied to our queries on this matter. Nevertheless, examination of the locations of the courts where sentences were passed reveals a good geographical spread across the England and Wales Crown Court locations, with the exception of the Central Criminal Court, which is overrepresented (comprising 9% of the extracted cases, compared to an average of 1.2% for the remaining 86 court locations in the sample). The high concentration of cases from the Central Criminal Court is also reflected in an overrepresentation of serious offences in the sample, with the four most common offences captured being 'murder', 'grievous bodily harm with intent', 'assault occasioning actual bodily harm', and 'rape' (comprising 25.6%, 10.7%, 8.6%, and 8.4% of

[7] The different titles used in England and Wales to distinguish members of the judiciary are listed in the judiciary website (https://www.judiciary.uk/you-and-the-judiciary/what-do-i-call-judge/) where their seniority is also described (https://www.judiciary.uk/about-the-judiciary/who-are-the-judiciary/judicial-roles/judges/).

Table 1
Descriptive Statistics of the sample used in the analysis.

	Mean/Proportion	Range/N
Sample Size		7221
Judge id		1140
Court id		86
Dependent Variables		
Months in custody	126.4	(1, 540)
Indeterminate sentence	0.286	(0, 1)
Judge Characteristics		
Rotating judge	0.631	(0, 1)
Staying judge	0.369	(0, 1)
Female judge	0.087	(0, 1)
Circuit judge	0.006	(0, 1)
High Court judge	0.120	(0, 1)
Queen's Counsel judge	0.296	(0, 1)
Recorder	0.081	(0, 1)
Defendant Characteristics		
Male defendant	0.940	(0, 1)
Defendant's age	33.9	(18, 92)
Case Characteristics		
Year (centered)	1.58	(-4, 6)
Co-defendants	0.360	(0, 1)
Public protection sentence	0.294	(0, 1)
Plead guilty before trial	0.499	(0, 1)
Mitigating factors	0.090	(0, 1)
Sentenced to remand	0.364	(0, 1)
Victim impact statement	0.247	(0, 1)
Injuries caused	0.177	(0, 1)
Multiple offences	0.384	(0, 1)
Specific Type of Offence		
Murder	0.256	(0, 1)
Attempted murder	0.038	(0, 1)
Grievous bodily harm with intent	0.107	(0, 1)
Grievous bodily harm	0.050	(0, 1)
Assault occasioning bodily harm	0.087	(0, 1)
Common assault	0.029	(0, 1)
Affray	0.019	(0, 1)
Arson	0.017	(0, 1)
Robbery	0.091	(0, 1)
Kidnap	0.013	(0, 1)
Rape	0.084	(0, 1)
Rape of a child	0.017	(0, 1)
Sex activity with a child	0.041	(0, 1)
Sexual assault	0.060	(0, 1)
Dangerous driving	0.035	(0, 1)
Causing death by dangerous driving	0.047	(0, 1)

the sample).

'The Law Pages' suggest on their website[8] that their records are acquired from Her Majesty Courts and Tribunals Service (HMCTS). This is confirmed by the inclusion of 'Case Numbers' - a unique identifier used by HMCTS for their internal management of trials[9] - in 'The Law Pages' records. Further reassurance regarding the validity of the data can be obtained from the academic research where this data has been used (Jacobson et al., 2016; Lavorgna, 2015; Pina-Sánchez et al., 2018b).

3. Analytical strategy

Two methodological challenges inform our analytic strategy: i) the correct specification of durations for indeterminate sentences (e.g. those offenders sentenced to life imprisonment), and ii) accounting for the non-hierarchical data structure with some judges operating across multiple courts.

To include cases sentenced to indeterminate custody we use an accelerated-life Weibull model. This treats indeterminate sentences as right censored, with the 'minimum term' duration being the last observed point of a potentially ongoing custodial spell. This specification was chosen in favour of an accelerated-life exponential model, the simplest type of parametric survival models, as it offered a better fit to the monotonically decreasing baseline hazard function observed in our sample of durations. The level-1 residual

[8] See http://www.thelawpages.com/aboutus.php.
[9] See http://xhibit.justice.gov.uk.

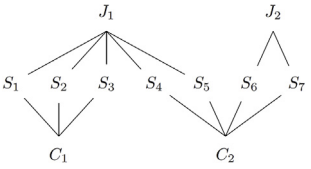

Fig. 1. Representation of the hierarchical structure of sentencing data in the Crown Court.

term, ε, is therefore assumed to follow an extreme value distribution and scaled by the term $\frac{1}{p}$. The 'shape' parameter, p, is used to account for monotonically increasing ($p > 1$) or decreasing ($p < 1$) hazard rates, which the exponential model assumes to be constant.

Existing attempts to account for the clustered structure of sentencing data have typically used hierarchical multilevel models, with the higher level used to account for either the clustering of sentences within specific judges, courts, or geographic areas, depending on the data available. Johnson (2006) demonstrated how this approach may be flawed because it fails to correctly account for the nesting of sentences within specific judges and courts simultaneously, instead advocating the use of three-level models with sentences nested within specific judges, and judges in turn grouped within courts. Here we further extend this model by also allowing for judges who work across multiple courts. This type of cross-classification is represented in Fig. 1, which shows an excerpt of our data composed of seven sentences (S), two judges (J) and two courts (C). Notice how J_2 is perfectly nested within C_2; contrast this with J_1, who has passed sentences in both C_1 and C_2.

To account correctly for judges who work across multiple courts, we adopt a cross-classified multilevel specification (Fielding and Goldstein, 2006; Goldstein, 1994; Johnson, 2012; and Leckie, 2013). Specifically, we differentiate between clustering within courts and judges, while incorporating a complex error structure at the judge level that distinguishes between rotating judges (those that move between courts), and judges that only operate within a single court. Following the specification outlined in Brunton-Smith et al. (2012), this is achieved by specifying a random effect on the binary judge-level variable Z_j that identifies whether a judge rotates across courts, $Z_j = 1$, or operates within a single court, $Z_j = 2$.

The model takes the following form,

$$\log(T_i) = \beta_{0i} + \mathbf{x}_i'\boldsymbol{\beta} + \frac{1}{p}\varepsilon_i \tag{1}$$

$$\beta_{0i} = \beta_0 + \upsilon_{1j} + \upsilon_{0c} \quad \text{if} \quad Z_j = 1$$

$$\beta_{0i} = \beta_0 + \upsilon_{2j} + \upsilon_{0c} \quad \text{if} \quad Z_j = 2$$

$$\begin{bmatrix} \upsilon_{1j} \\ \upsilon_{2j} \end{bmatrix} \sim N(0,\Omega): \ \Omega = \begin{bmatrix} \sigma^2_{\upsilon 1j} & \\ 0 & \sigma^2_{\upsilon 2j} \end{bmatrix}$$

where T_i indicates the duration of the sentence, β_{0i} is the intercept, $\mathbf{x}_i'\boldsymbol{\beta}$ is a vector including the full list of explanatory variables outlined in Table 1, and $\frac{1}{p}\varepsilon_i$ is the scaled level-1 residual. In contrast with a standard cross-classified model, variability in the intercept, β_{0i}, is modelled separately for our two types of judge, with residual errors υ_{1j} and υ_{2j}. These errors are assumed independent and identically distributed with mean zero, variances $\sigma^2_{\upsilon 1j}$ and $\sigma^2_{\upsilon 2j}$, and covariance constrained to zero to reflect the mutually exclusive judge categories. The random part of the model also includes a court-level random effect, υ_{0c}, with variance $\sigma^2_{\upsilon 0c}$.

We specify two nested models and compare the size of their random effects. Model 1 is an empty model, with the random effects capturing 'unconditioned' sentencing disparities, i.e. disparities that could be due to either legal differences between cases or to unwarranted disparities. Model 2 includes the full list of predictors, $\mathbf{x}_i'\boldsymbol{\beta}$, with the random part of the model now capturing the 'net' disparities obtained after controlling for relevant case characteristics. One final model, Model 3, is estimated to explore the extent to which the observed 'net' disparities at the judge level can be explained by the gender, or the title of the judge.

As far as we are aware there are no mainstream statistical packages (commercial or open-source) that can be used to estimate the model suggested here. Hence, we resort to the flexibility afforded by Bayesian statistics and the software WinBUGS (Lunn et al., 2000). We use diffuse prior distributions for all parameter estimates (see Appendix). Two separate MCMC chains are estimated with dispersed starting values, each with a burn-in of 15,000 and a monitoring period of 100,000 iterations. In the following section we report the posterior mean and 95% credible interval for each parameter using the 170,000 monitoring iterations pooled across the two chains. The posterior mean and 95% credible interval are analogous to the point estimate and confidence interval produced by standard statistics packages in a non-Bayesian setting. Visual inspection of the history plots and the Brooks-Gelman-Rubin convergence diagnostic (Brooks and Gelman, 1997) suggest the two chains have reached convergence and achieved good mixing. The posterior distributions are well approximated by the 170,000 MCMC iterations as the Monte Carlo errors are below 5% of the corresponding posterior standard deviations for all parameters (Lunn et al., 2012).

4. Results

Looking first at the empty model (Model 1, Table 2) we can observe the presence of disparities in sentencing operating at both the

Table 2
Results from Model 1, 2 and 3, where sentence length is specified using Weibull models and different set of explanatory variables; 95% credible intervals within brackets.

	Model 1	Model 2	Model 3
Random effects			
Std. dev. intercepts: court	0.43 (0.35, 0.53)	0.21 (0.16, 0.28)	0.20 (0.15, 0.27)
Std. dev. intercepts: rotating judge	1.10 (1.00, 1.21)	0.40 (0.34, 0.47)	0.37 (0.32, 0.45)
Std. dev. intercepts: staying judge	0.69 (0.61, 0.78)	0.45 (0.37, 0.53)	0.44 (0.36, 0.53)
Fixed Effects			
Weibull slope	1.19 (1.16, 1.21)	2.03 (1.98, 2.08)	2.04 (1.99, 2.09)
Intercept: rotating judge	5.94 (5.73, 6.15)	11.85 (11.51, 12.21)	11.70 (11.34, 12.09)
Intercept: staying judge	5.35 (5.17, 5.53)	11.77 (11.44, 12.13)	11.70 (11.34, 12.08)
Judge Characteristics (ref: male, unidentified title)			
Female judge			0.06 (−0.09, 0.22)
Circuit judge			−0.40 (−0.79, 0.02)
High Court judge			0.66 (0.45, 0.86)
Queen's Counsel judge			−0.01 (−0.10, 0.08)
Recorder			−0.33 (−0.44, −0.21)
Defendant Characteristics (ref: female)			
Defendant male		0.39 (0.26, 0.52)	0.40 (0.27, 0.53)
Defendant age		0.01 (0.01, 0.02)	0.01 (0.01, 0.02)
Case Characteristics			
Year		0.04 (0.02, 0.05)	0.04 (0.02, 0.05)
Co-defendants		0.26 (0.19, 0.34)	0.23 (0.16, 0.31)
Public protection sentence		0.93 (0.82, 1.05)	0.89 (0.78, 1.06)
Plead guilty before trial		−0.65 (−0.72, −0.59)	−0.65 (−0.72, −0.58)
Mitigating factors		−0.30 (−0.40, −0.20)	−0.32 (−0.42, −0.22)
Sentenced to remand		0.53 (0.45, 0.61)	0.52 (0.44, 0.60)
Victim impact statement		0.30 (0.23, 0.38)	0.30 (0.22, 0.38)
Injuries caused		0.11 (0.01, 0.22)	0.11 (0.01, 0.21)
Second offence		0.66 (0.58, 0.74)	0.66 (0.58, 0.74)
Third offence		0.43 (0.31, 0.54)	0.41 (0.30, 0.53)
Fourth offence		0.54 (0.39, 0.69)	0.54 (0.39, 0.69)
Offence type (ref: murder)			
Attempted murder		−2.30 (−2.56, −2.05)	−2.26 (−2.52, −2.00)
GBH with intent		−3.69 (−3.91, −3.48)	−3.57 (−3.79, −3.36)
GBH		−5.46 (−5.70, −5.22)	−5.34 (−5.59, −5.11)
ABH		−5.31 (−5.54, −5.09)	−5.21 (−5.44, −4.98)
Common assault		−6.03 (−6.31, −5.76)	−5.93 (−6.20, −5.66)
Affray		−6.37 (−6.66, −6.08)	−6.22 (−6.51, −5.92)
Arson		−4.42 (−4.70, −4.13)	−4.31 (−4.60, −4.03)
Robbery		−3.97 (−4.20, −3.75)	−3.83 (−4.06, −3.60)
Kidnap		−4.09 (−4.40, −3.77)	−3.99 (−4.31, −3.67)
Rape		−3.47 (−3.69, −3.25)	−3.36 (−3.59, −3.14)
Rape of a child under 13		−3.26 (−3.55, −2.98)	−3.14 (−3.43, −2.85)
Sex activity with a child		−4.27 (−4.52, −4.03)	−4.15 (−4.40, −3.91)
Sexual assault		−4.52 (−4.76, −4.28)	−4.41 (−4.65, −4.17)
Dangerous driving		−5.81 (−6.08, −5.55)	−5.70 (−5.96, −5.44)
Causing death by dangerous driving		−4.07 (−4.31, −3.84)	−3.95 (−4.20, −3.72)
Sample Size			
N sentences: 7221			
N rotating judges: 719			
N non-rotating judges: 421			
N courts: 86			

court and judge level. Differences between judges are generally larger than differences between courts, highlighting the importance of correctly accounting for between judge disparities in sentencing data analyses. When considering the two types of judge, greater disparities are evident between judges who rotate across courts (1.10) when compared against those judges who stayed within a single court (0.69). These 'rotating' judges also tend to use harsher sentences, with a higher fixed effect estimate (5.94 compared to 5.35 for judges that work in a single court). These results illustrate the need to go beyond the simpler multilevel models used previously, and justify the cross-classified model specified here. However, no substantive interpretation should be derived from these results since the model does not control for any relevant case characteristics. Consequently, results from Model 1 conflate legitimate with unwarranted disparities in sentencing.

Model 2 (Table 2) replicates Model 1 accounting for the full set of offence and offender characteristics recorded in our sample. All of the explanatory variables included are statistically significant. Male defendants tend to receive longer sentences than female defendants, so do older offenders, although the effect size is almost negligible. At the case-level our findings are consistent with the expected sentencing practice, with longer sentences awarded in cases when the defendant was sentenced for public protection or on remand, as well as cases including a victim statement, where injury was caused, or where the offence was part of a series. In contrast,

shorter sentences were awarded when a guilty plea was entered, or where mitigating factors were noted. We also find evidence of a moderate increase in sentence length over time, which corroborates the increase in sentence severity documented in England and Wales over the last decade (Allen, 2016; Pina-Sánchez et al., 2019; Roberts and Ashworth, 2016; Roberts and Irwin-Rogers, 2015). The bottom part of the table captures the effect of different types of offences compared to cases of murder (the reference category), with all offence categories showing the expected negative effect.

Turning to the random part of the model, and now having accounted for the full range of case characteristics, we observe disparities that are twice as large at the judge level than at the court level. This result illustrates how studies seeking to measure the inconsistency in sentencing using court disparities as a proxy (Drápal, 2018; Pina-Sánchez, 2015; Pina-Sánchez, 2015; Pina-Sánchez and Linacre, 2013; Pina-Sánchez et al., 2017, 2018a,b; Tarling, 2006) are underestimating much of the problem by only being able to detect roughly one third of the total disparities. This result also suggests that the importance of court-level disparities may have been overplayed in existing literature (Ulmer, 2012), and questions the relevance of theories that have leaned on this concept in explaining the origin of disparities in sentencing, such as those focused on court culture.

Redirecting our focus to the judge-level estimates we observe that in Model 2 the typical sentence awarded by judges that rotate is almost identical to the sentence awarded by those that operate in a single court (11.85 and 11.77 respectively). Comparisons of these intercepts to those obtained in Model 1 suggests that 'rotating' judges tend to sentence more serious cases, but once the features of the cases are considered, they do not seem to sentence more harshly than judges who stay in a single court. A more complex picture stems from the random effects observed for the different types of judges, with between judge disparities now lower amongst 'rotating' judges (0.40 compared to 0.45 for judges that work in a single court). Therefore, judges who rotate across courts tend to deal with a more heterogeneous workload (as evidenced by the higher variance estimate in Model 1) but even in spite of that, tend to sentence more consistently than judges who stay in the same court when relevant features of the case have been taken into consideration. This corroborates Hester (2017) hypotheses on the benefits of allowing judges to rotate across different courts.

One last model – Model 3 – is explored to assess whether the observed differences for judges who rotate across courts are not confounded by other judge characteristics. We do not find evidence of significant differences in sentencing based on the judges' gender. By contrast, judicial career stage appears to have an important effect on the relative severity of the punishment, with High Court judges (the most experienced) sentencing significantly more severely than any other type of judges. While this may be a result of the 'hardening' described by Spohn (1990b), whereby judges with more experience become more punitive over time, it might also be a product of the more serious cases that High Court judges are required to handle. In spite of controlling for the specific offence and other important case features, we cannot control for all the relevant case characteristics expected to increase the seriousness of a case. Importantly, we also find that differences in the random effects between 'rotating' and 'staying' judges remain, with the former showing lower disparities in sentencing.

5. Discussion

In this study, we have employed innovative data scraping and text mining to extract relevant case-level, defendant, judge and court data from an online archive of sentence records from the Crown Court of England and Wales. The dataset generated from this process was subsequently used to model prison sentence durations, investigate the presence of sentencing disparities, and identify whether those disparities are generated at the court or the judge level.

The findings revealed that disparities resulting from Crown Court locations are much less substantial than those stemming from the judge level. That is, it is not so much the location where the case is processed, but the judge presiding over the case within that location that will have the potential for providing an unduly lenient or harsh sentence. This central finding emphasizes the critical role that individual judges have in the sentencing process and lend support to previous calls for a better understanding of the impact these actors have on sentencing (Croyle, 1983; Divine, 2018; Long and Burnham, 2012; Ulmer, 2012; Yang, 2014, 2015; Zatz, 2000). Important methodological and theoretical implications follow.

From a methodological perspective, our findings suggest that studies limited to the analysis of between court disparities are missing much of the picture. We estimate that roughly two-thirds of the sentencing disparities pass unaccounted when the judge level is ignored. This calls into question the validity of most of the recent efforts aimed at estimating the level of consistency in the Crown Court (Pina-Sánchez, 2015; Pina-Sánchez and Grech, 2017; Pina-Sánchez and Linacre, 2013; Pina-Sánchez et al., 2017, 2018a), which have inevitably provided an incomplete view of the problem. This problem is expected to be even more pronounced in studies where measures of consistency are derived from disparities at a higher level than courts, such as counties (Britt, 2000; Fearn, 2005; Haynes, 2011; Hester and Sevigny, 2016; Johnson, 2005; Ulmer and Johnson, 2004), districts (Feldmeyer and Ulmer, 2011; Hamilton, 2017; Johnson et al., 2008; Kautt, 2002; Schanzenbach and Tiller, 2006; Ulmer et al., 2011; Ulmer and Johnson, 2017), or police forces (Mason et al., 2007). Hence, affecting much of the empirical literature on the topic of consistency in sentencing. The more aggregate the units used to assess disparities the larger the extent of disparities taking place within them that will be missed. To borrow a metaphor, much of the research on sentencing disparities has only been able to observe the 'tip of the iceberg'.

Theoretically, these findings suggest a relatively lesser role for court culture in explaining sentencing disparities than previously thought. This is in line with Church (1985), who has argued that court culture has a greater impact on court processes, such as court efficiency, than court outcomes, such as sentencing. Also in line with Ulmer (2012) interpretation of the focal concerns perspective, the social context underlying a court's culture seems to be only one of many non-legal considerations that judge's may take into account when making a sentencing decision.

It is also possible that the relevance of court cultures cannot be detected through the exploration of systematic disparities between courts. Perhaps the concept could be better operationalised by exploring disparities within courts. To do so it would be important to

expand the approach undertaken here and seek to identify not only the judge but as many individual relevant actors operating within the same court (e.g. prosecutors, defence representatives). Such individual identifiers would allow to provide a measure of the influence of the concept of court culture as a whole, and to explore how the specific interactions between those actors influence the final sentence outcome. In so doing, such research could further illuminate the work of Eisenstein et al. (1988), Flemming et al. (1992) and Eisenstein and Jacob (1977), and their understanding of court cultures as complex phenomenon that arises from individual interactions within a workgroup.

Further investigations of between judge disparities in our dataset have provided some interesting insights into distinct sentencing practices within the Crown Court. Comparing the disparities resulting from judges who were observed to rotate across courts against those who stayed in the same court, we found that judges working in multiple courts sentence a more heterogeneous and complex caseload, but do so more consistently than judges who stay in the same court. These results corroborate Hester (2017), who, based on qualitative interviews with thirteen judges from South Carolina, indicates that judicial rotation seems to facilitate the spread of norms and homogenize judicial sentencing practices. Indeed, exposure to other judges and court actors may support consistent sentencing practices.

This finding suggests that consistency in sentencing can be promoted using alternative – or complimentary – strategies to the design of sentencing guidelines. The introduction of guidelines have been historically met with considerable resistance from sentencers, as they can be perceived to undermine judicial autonomy (Sentencing Commission Working Group, 2008). Fostering judicial rotation, on the other hand, represent a less intrusive and prescriptive approach. As such, this approach to the promotion of consistency could simultaneously contribute to rebuild morale across the judiciary, which in England and Wales has been noted to be markedly low (Thomas, 2017), and eliminate some of the negative side-effects that have been attributed to sentencing guidelines, such as the undermining of the principles of individualisation (Cooper, 2013) and proportionality (Alschuler, 1991; Lowenthal, 1993; Schulhofer, 1991).

5.1. Beyond England and Wales

In interpreting our findings, it is important to consider the peculiarities of the Crown Court, one of the Senior Courts in England and Wales, composed of highly experienced judges, specifically dealing with serious offences. Hence, it is worth considering whether the same results pointing at the importance of the judge over the location where the sentence was imposed can be generalised to other jurisdiction; or even to the magistrate's court, a less professional and more geographically dispersed group of courts processing the majority of offences in England and Wales.

One major difference to be noted when considering the generisability of our findings to US jurisdictions, where most of the sentencing literature originates (Johnson et al., 2010), resides on the practice of electing judges. It could be expected that democratically elected judges, as opposed to those who are appointed, are held to account by the electorate directly and are perhaps more influenced by the public. If so, we should expect that between court disparities in the US play a bigger role than what we have noted here for the case of the England and Wales Crown Court. Another important difference stems from the practice of negotiated plea, commonly resorted to in the US but non-existent in England and Wales.

Johnson (2006, 2014) studies of the State of Pennsylvania offer a great opportunity to carry out a comparative assessment since they are based on a relatively similar analytical strategy where judge and court (county in his case) levels where simultaneously modelled. The author found that for cases dealt as negotiated pleas most disparities stem from the area level, however, for those cases ending up in a trial most disparities originate at the judge level. In fact the ratio of judge/court disparities in the latter is more pronounced in Pennsylvania than what we have observed for the Crown Court. This comparison underlines the relevance of the warning we have raised in this study regarding the limitation of much of the literature studying sentencing disparities. Still, Pennsylvania is only one state within the US. The generisability of our findings could be further assessed using datasets from different jurisdictions, where judge and court (or area) identifiers have been obtained (Chatsverykova, 2017; Lim et al., 2016; Volkov, 2016; Yang, 2014).

5.2. A final note on the need for better sentencing data

Progress in sentencing research has been stalled by data limitations. Collecting primary data[10] is both expensive and time-consuming, with studies typically based on relatively small samples[11] that are restricted to one or a few courts (see Brown and Hullin, 1992; Davies and Tyrer, 2003; Dhami, 2005; Frazier and Bock, 1982; Hester, 2017; Jacobson and Hough, 2007). In most jurisdictions secondary data is published only at an aggregate level (e.g. number of prison sentences handed out per year). Certain jurisdictions have provided access to datasets composed of individual cases, but with the exception of Johnson (2006, 2014) such datasets are limited to a small number of courts (Anderson and Spohn, 2010; Fearn, 2005; Gruhl et al., 1981; Johnson, 2005, 2012; Kautt, 2002; Kritzer and Uhlman, 1977; Spohn, 1990a, 1990b; Steffensmeier and Britt, 2001; Steffensmeier and Hebert, 1999; Uhlman, 1978; Welch et al., 1988), while those offering wider coverage provide no judge identifiers (Drápal, 2018; Hamilton, 2017; Pina-Sánchez and Linacre, 2013; Ulmer et al., 2011).

[10] Normally in the form of observations, file reviews, interviews, and questionnaires (including simulations).

[11] There are some important exceptions to this claim. See for example Myers and Talarico (1987) where the authors gathered over 27,720 offenders, or Hood (1992), who compiled a sample of 3,317 defendants.

Restrictions to the publication of sentencing records are often justified as a requirement to safeguard offenders' anonymity and comply with data protection acts.[12] Yet a review of some of the most commonly used sentencing datasets in the US and the UK (e.g. the Crown Court Sentencing Survey, and the Court Proceedings Database from England and Wales,[13] or the United States Sentencing Commission, the Pennsylvania Commission on Sentencing, and the Minnesota Sentencing Guidelines Commission in the US[14]) shows that it is details of the judge that are most often missing, with offenders' gender, age, and ethnic background generally provided. As noted in the introduction, we believe that the reasons are more likely to lie on the reluctance of the judiciary to release data that allows researchers to examine their practice empirically.

These concerns reflect an important degree of misunderstanding on how sentencing data is used in social research, and seem unaware of recent developments in secure data sharing procedures. In the UK there are multiple examples of secure portals like the Ministry of Justice Data Lab, as well as repositories managing access of sensitive data like the UK Data Service. Other options that could be explored are robust data anonymisation techniques, such as those based on the incorporation of simulated errors (Biewen et al., 2008). All of these approaches have been successfully applied to various other fields of research, and could be equally well applied to sentencing research. Furthermore, it is important to highlight how, for the most part, fears expressed by the judiciary are unfounded, and the above mentioned secure approaches unnecessary. Quantitative sentencing research seeks to detect and explain underlying broad patterns, its goal is not the study of specific cases or individuals. As a matter of fact, the identity of the judge or the location of the court are completely unnecessary, all that is needed is a way to tell them apart and in so doing be able to reproduce the hierarchical dimension observed in sentencing data. To do so numerical - or otherwise coded – 'ids', keeping no relation to the judge name or the court where they operate,[15] suffice.

We have shown how such a lack of judge id has limited the study of sentencing disparities in England and Wales, a central question in assessing the impact of their new sentencing guidelines. One would expect the judiciary to be the first interested party in understanding the impact that such process of reform has had on sentencing practice, which makes it difficult to comprehend their reticence towards the disclosing of sentencing records. Yet, the implications of this problem extends beyond the study of sentencing disparities, or the jurisdiction of England and Wales. The incorrect assumption of sentencing data being independently distributed leads to biased estimations of standard errors and other measures of uncertainty (Dhami and Belton, 2016; Johnson, 2006). With almost all quantitative research in the field failing to account for clustering at the court and judge level simultaneously, we should expect this problem to be ubiquitous.

The growing availability of administrative data (Connelly et al., 2016) has made it possible for us to circumvent the limitations of official data, and to generate a new dataset reconstructing the main features of a sample of cases sentenced in the Crown Court, including the identity of the judge and the location of the court. We believe the application of a similar research design in some of the US jurisdictions offers a great opportunity to overcome some of the key issues that have affected sentencing research in that country too. Official State and Federal sentencing records do not offer adequately fine grained distinctions between cases, leading to problems of omitted relevant variable bias (Baumer, 2013; Ulmer, 2012). Using sentence transcripts and text mining techniques new sentencing datasets could be created aiming at capturing the presence of aggravating and mitigating factors, key case characteristics that have been commonly missed in most sentencing research in the US.

Recent developments undertaken at the Finish Ministry of Justice (Frosterus et al., 2014; Oksanen et al., 2017) demonstrate how advanced natural language processing algorithms can be used to generate more detailed sentencing datasets than the one we have generated here. The same process could be replicated in England and Wales - and elsewhere – if the judiciary decides to provide access to their court records. This is clearly in their own interest since not only would it generate new datasets with which to carry out more robust sentencing research, but it would also eliminate the need for expensive court surveys, such as those currently commissioned by the Sentencing Council for England and Wales to assess the impact of their guidelines.

Acknowledgement

We would like to thank John Gannon, Ian Marder, Javier Velásquez, Lyndon Harris and our anonymous reviewers.

This work was supported by National Centre for Research Methods (RIS14241/06).

[12] Such as the new General Data Protection Regulation in operation in the whole of the European Union since the 25th-May-2018.

[13] Available online here https://www.sentencingcouncil.org.uk/analysis-and-research/crown-court-sentencing-survey/record-level-data/ and here https://www.gov.uk/government/collections/criminal-justice-statistics.

[14] Available online here https://www.ussc.gov/, here http://pcs.la.psu.edu/, and here https://mn.gov/sentencing-guidelines/.

[15] This process should also be applied to record other actors that have been shown to influence sentencing decisions such as prosecutors and probation officers (Goulette et al., 2015; Leifker and Sample, 2011; Spohn et al., 2017; Sutton, 2013; Wooldredge, 2012).

Appendix. WinBUGS Code

```
model{
#Model (accelerated-life Weibull models with a complex error structure to account for court-judge cross-classification)
for(i in 1:7221) {
months[i] ~ dweib(shape,mu[i])I(cens[i],)
log(mu[i]) <- -(b01[judge[i]]*rotate[i] + b02[judge[i]]*stay[i] + b2[court[i]] + A[i] + B[i])
A[i] <- b3*d.male[i] + b4*d.age[i] + b5*codef[i] + b6*protection[i] + b7*onplea[i] + b8*miti[i] +
        b9*remand[i] + b10*off2[i] + b11*off3[i] + b12*off4[i] + b13*victimpct[i] + b14*injuries[i] +
        b15*circuit[i] + b16*HighCourt[i] + b17*QC[i] + b18*recorder[i] + b19*j.male[i] +
B[i] <- b20*ABH[i] + b21*attempted[i] + b22*GBH[i] + b23*affray[i] + b24*common[i] +
        b25*intent[i] + b26*arson[i] + b27*robbery[i] + b28*kidnap[i] + b29*rape[i] + b30*sexchild[i] +
        b31*sexaslt[i] + b32*rapechild[i] + b33*dangedrive[i] + b34*dthdngdrv[i] + b35*yearcntrd[i]
}

#Random effects
#Court
for(j in 1:86){
b2[j] ~ dnorm(0,tau.2)
}
#Rotating judges
for(l in 1:1140){
beta01[l] ~ dnorm(b01,tau.01)
}
#Staying judges
for(k in 1:1140){
beta02[k] ~ dnorm(b02,tau.02)
}

#Priors
shape ~ dgamma(1,.001)
b01 ~ dnorm(0,0.001)
b02 ~ dnorm(0,0.001)
b3 ~ dnorm(0,0.001)
…
B35 ~ dnorm(0,0.001)
#Hyperpriors
tau.2 ~ dgamma(0.001,0.001)
tau.01 ~ dgamma(0.001,0.001)
tau.02 ~ dgamma(0.001,0.001)
}
```

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ssresearch.2019.102343.

References

Albonetti, C.A., 2002. The joint conditioning effects of defendant's gender and ethnicity on length of imprisonment under the federal sentencing guidelines for drug trafficking/manufacturing offenders. J. Gend. Race Justice 6, 39–60.

Allen, R., 2016. The Sentencing Council for England and Wales: Brake or Accelerator on the Use of Prison? Transform Justice. London. http://www.transformjustice.org.uk/wp-content/uploads/2016/12/TJ-DEC-8.12.16-1.pdf.

Alschuler, A., 1991. The failure of sentencing guidelines: a plea for less aggregation. Univ. Chic. Law Rev. 58 (3), 901–951.

Anderson, J., Kling, J., Stith, K., 1999. Measuring inter-judge sentencing disparity: before and after the Federal sentencing guidelines. J. Law Econ. 42, 271–307.

Anderson, A.L., Spohn, C., 2010. Lawlessness in the federal sentencing process: a test for uniformity and consistency in sentence outcomes. Justice Q. JQ 27 (3), 362–393.

Ashworth, A., 2003. Sentencing and sensitivity: a challenge for research. In: Zedner, L., Ashworth, A. (Eds.), The Criminological Foundations of Penal Policy. Oxford University Press, New York, pp. 295–332.

Baldwin, J., 2008. Research on criminal courts. In: King, R., Wincup, E. (Eds.), Doing Research on Crime and Justice. Oxford University Press, Chippenham, pp. 375–398.
Baumer, E.P., 2013. Reassessing and redirecting research on race and sentencing. Justice Q. JQ 30 (2), 231–261.
Biewen, E., Nolte, S., Rosemann, M., 2008. Perturbation by multiplicative noise and the Simulation Extrapolation method. AStA Adv. Stat. Anal. 92 (4), 375–389.
Brantingham, P., 1985. Sentencing disparity: an analysis of judicial consistency. J. Quant. Criminol. 1, 281–305.
Britt, C.L., 2000. Social context and racial disparities in punishment decisions. Justice Q. JQ 17, 707–732.
Brooks, S.P., Gelman, A., 1997. General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. 7, 434–455.
Brown, I., Hullin, R., 1992. A study of sentencing in the Leeds Magistrates' Courts: the treatment of ethnic minority and white offenders. Br. J. Criminol. 32 (1), 41–53.
Brunton-Smith, I., Sturgis, P., Williams, J., 2012. Is success in obtaining contact and cooperation correlated with the magnitude of interviewer variance? Public Opin. Q. 76 (2), 265–286.
Bushway, S.D., Piehl, A.M., 2001. Judging Judicial Discretion: Legal Factors and Racial Discrimination in Sentencing. Law Soc. Rev. 733–764.
Chatsverykova, I., 2017. Severity and leniency in criminal sentencing in Russia: the effects of gender and family ties. Int. J. Comp. Appl. Crim. Justice 41 (3), 185–209.
Chen, D.L., Loecher, M., Barry, N., Buchanan, L., Bakhturina, E., 2016. Events unrelated to crime predict criminal sentence length. (Available at:). https://users.nber.org/~dlchen/papers/Events_Unrelated_to_Crime_Predict_Criminal_Sentence_Length.pdf.
Church, T., 1985. Examining local legal culture. Am. Bar Found Res. J. 10 (3), 449–518.
Cole, K., 1997. The Empty Idea of Sentencing Disparity. Northwest University Law Review, pp. 91.
Connelly, R., Playford, C.J., Gayle, V., Dibben, C., 2016. The role of administrative data in the big data revolution in social science research. Soc. Sci. Res. 59, 1–12.
Cooper, J., 2013. Nothing personal: the impact of personal mitigation at sentencing since the creation of the Council. In: Ashworth, A., Roberts, J.V. (Eds.), Sentencing Guidelines: Exploring the English Model. Oxford University Press, Oxford, pp. 157–164.
Croyle, J.L., 1983. Measuring and explaining disparities in felony sentences: courtroom work group factors and race, sex, and socioeconomic influences on sentence severity. Political Behav. 5 (1), 135–153.
Danziger, S., Levav, J., Avnaim-Pesso, L., 2011. Extraneous factors in judicial decisions. Proc. Natl. Acad. Sci. 108, 6889–6892.
Darbyshire, P., 2011. Sitting in Judgement. Hart Publishing, Oxford.
Davies, M., Tyrer, J., 2003. 'Filling in the gaps' a study of judicial culture: views of judges in England and Wales on sentencing domestic burglars contrasted with the recommendations of the Sentencing Advisory Panel and the Court of Appeal guidelines. Crim. Law Rev. 243–265.
Dhami, M.K., 2005. From discretion to disagreement: explaining disparities in judges' pretrial decisions. Behav. Sci. Law 23 (3), 367–386.
Dhami, M.K., Belton, I., 2016. Statistical analyses of court decisions: the example of multilevel models of sentencing. Law Method 10, 247–266.
Dhami, M.K., Souza, K., 2010. Breaking into court. In: Streiner, D., Souraya, S. (Eds.), When Research Goes off the Rails. Guildford Press, New York, pp. 81–87.
Divine, J.M., 2018. Booker disparity and data-driven sentencing. Hastings Law J. 69 (3), 771–834.
Drápal, J., 2018. Sentencing disparities in the Czech Republic: empirical evidence from post-communist Europe. Eur. J. Criminol.
Eisenstein, J., Flemming, R., Nardulli, P., 1988. The Contours of Justice: Communities and Their Courts. Little, Brown and Co., Boston.
Eisenstein, J., Jacob, H., 1977. Felony Justice: an Organizational Analysis of Criminal Courts. Little, Brown and Co., Boston.
Eren, O., Mocan, N., 2018. Emotional judges and unlucky juveniles. Am. Econ. J. Appl. Econ. 10 (3), 171–205.
Fearn, N.E., 2005. A multilevel analysis of community effects on criminal sentencing. Justice Q. JQ 22 (4), 452–487.
Feldmeyer, B., Ulmer, J.T., 2011. Racial/ethnic threat and Federal sentencing. J. Res. Crime Delinq. 48, 238–270.
Fielding, A., Goldstein, H., 2006. Cross-classified and Multiple Membership Structures in Multilevel Models: an Introduction and Review. Department of Education and Skills, Birmingham(Available at:)http://webarchive.nationalarchives.gov.uk/20130321021729/https://www.education.gov.uk/publications/eOrderingDownload/RR791.pdf.
Flemming, R.B., Nardulli, P.F., Eisenstein, J., 1992. The Craft of Justice: Politics and Work in Criminal Court Communities. University of Pennsylvania Press, Philadelphia.
Frazier, C.E., Bock, E.W., 1982. Effects of court officials on sentence severity: do judges make a difference? Criminology 20 (2), 257–272.
Frosterus, M., Tuominen, J., Hyvönen, E., 2014. Facilitating re-use of legal data in applications: Finnish law as a linked open data service. JURIX 115–124.
Gertner, N., 2012. TRAC data on federal sentencing: judge identifiers, TRAC, and a perfect world. Fed. Sentencing Report. 25 (1), 46–49.
Goldstein, H., 1994. Multilevel cross-classified models. Sociol. Methods Res. 22 (3), 364–375.
Goulette, N., Wooldredge, J., Frank, J., Travis III, L., 2015. From initial appearance to sentencing: do female defendants experience disparate treatment? J. Crim. Justice 43 (5), 406–417.
Gruhl, J., Spohn, C., Welch, S., 1981. Women as policymakers: the case of trial judges. Am. J. Pol. Sci. 25, 308–322.
Hamilton, M., 2017. Sentencing disparities. Br. J. Am. Legal Stud. 6 (2), 177–224.
Hawkins, D., 1981. Causal attribution and punishment for crime. Deviant Behavior 2, 207–230.
Haynes, S.H., 2011. The effects of victim-related contextual factors on the criminal justice system. Crime Delinq. 57 (2), 298–328.
Hester, R., 2012. Criminal Sentencing in the Court Communities of South Carolina: An Examination of Offender, Judge, and County Characteristics. University of South Carolina (PhD dissertation).
Hester, R., 2017. Judicial rotation as centripetal force: sentencing in the court communities of South Carolina. Criminology 55 (1), 205–235.
Hester, R., Sevigny, E.L., 2016. Court communities in local context: a multi-level analysis of felony sentencing in South Carolina. J. Crime Justice 39 (1), 55–74.
Heyes, A., Saberian, S., 2018. Temperature and decisions: evidence from 207,000 court cases. Am. Econ. J. Appl. Econ. 11 (2), 238–265.
Hogarth, J., 1971. Sentencing as a Human Process. University of Toronto Press.
Hofer, P.J., Blackwell, K.R., Ruback, R.B., 1999. The effect of the federal sentencing guidelines on inter-judge sentencing disparity. J. Crim. Law Criminol. 90 (1), 239–322.
Hood, R.G., 1992. Race and Sentencing: A Study in the Crown Court - A Report for the Commission for Racial Equality. Clarendon Press.
Jacobson, J., Hough, M., 2007. Mitigation: the Role of Personal Factors in Sentencing. Prison Reform Trust, London (Available at:). http://www.prisonreformtrust.org.uk/uploads/documents/FINALFINALmitigation%20-%20small.pdf.
Jacobson, J., Kirby, A., Hunter, G., 2016. Joint Enterprise: Righting a Wrong Turn? Report of an Exploratory Study. Prison Reform Trust, London (Available at:). http://www.prisonreformtrust.org.uk/Portals/0/Documents/Joint%20Enterprise%20Writing%20a%20Wrong%20Turn.pdf.
Johnson, B.D., 2005. Contextual disparities in guidelines departures: courtroom social contexts, guidelines compliance, and extralegal disparities in criminal sentencing. Criminology 43 (3), 761–796.
Johnson, B.D., 2006. The multilevel context of criminal sentencing: integrating judge- and County-level influences. Criminology 44 (2), 259–298.
Johnson, B.D., 2012. Cross-classified multilevel models: an application to the criminal case processing of indirect terrorists. J. Quant. Criminol. 28, 163–189.
Johnson, B.D., 2014. Judges on trial: a reexamination of judicial race and gender effects across modes of conviction. Crim. Justice Policy Rev. 25 (2), 159–184.
Johnson, B.D., Ulmer, J.T., Kramer, J.H., 2008. The social context of guidelines circumvention: the case of Federal District Courts. Criminology 46 (3), 737–783.
Johnson, B.D., Van Wingerden, S., Nieuwbeerta, P., 2010. Sentencing homicide offenders in The Netherlands: offender, victim, and situational influences in criminal punishment. Criminology 48 (4), 981–1018.
Kautt, P.M., 2002. Location, location, location: interdistrict and intercircuit variation in sentencing outcomes for Federal drug-trafficking offenses. Justice Q. JQ 19 (4), 633–671.
King, K., 2017. Lawyers' Anger as Students Told They Can't Make Notes during Court Visits. Legal Cheek Available at: https://www.legalcheek.com/2017/12/lawyers-anger-as-students-told-they-cant-make-notes-during-court-visits/.
King, D.K., Johnson, B.D., 2016. A punishing look: skin tone and Afrocentric features in the halls of justice. Am. J. Sociol. 122, 90–124.
Kritzer, H.M., Uhlman, T., 1977. Sisterhood in the courtroom: sex of judge and defendant in criminal case disposition. Soc. Sci. J. 14, 77–88.
Lavorgna, A., 2015. The online trade in counterfeit pharmaceuticals: new criminal opportunities, trends and challenges. Eur. J. Criminol. 12 (2), 226–241.
Leckie, G., 2013. Module 12: cross-classified multilevel models. LEMMA VLE Module 12, 1–60. Available at: http://www.bristol.ac.uk/cmm/learning/course.html.

Leifker, D., Sample, L.L., 2011. Probation recommendations and sentences received: the association between the two and the factors that affect recommendations. Crim. Justice Policy Rev. 22 (4), 494–517.
Lim, C.S.H., Silveira, B.S., Snyder, J.M., 2016. Do judges' characteristics matter? Ethnicity, gender, and partisanship in Texas state trial courts. Am. Law Econ. Rev. 18 (2), 302–357.
Long, S.B., Burnham, R., 2012. TRAC report: examining current federal sentencing practices: a national study of differences among judges. Fed. Sentencing Report. 25 (1), 6–17.
Lowenthal, G., 1993. Mandatory sentencing laws: undermining the effectiveness of determinate sentencing reform. Calif. Law Rev. 81 (1), 61–123.
Lunn, D.J., Jackson, C., Best, N., Thomas, A., Spiegelhalter, D., 2012. The BUGS Book – A Practical Introduction to Bayesian Analysis. CRC Press, London.
Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. Stat. Comput. 10 (4), 325–337.
Mason, T., de Silva, N., Sharma, N., Brown, D., Harper, G., 2007. Local Variation in Sentencing in England and Wales. Ministry of Justice (Available at:). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/217971/local-variation-sentencing-1207.pdf.
Ministry of Justice, 2019. Judicial Diversity Statistics 2019. Available at: https://www.judiciary.uk/wp-content/uploads/2019/07/Judicial-Diversity-Statistics-2019.pdf.
Myers, M.A., Talarico, S.M., 1987. The Social Contexts of Criminal Sentencing. Springer, New York.
Oksanen, A., Tuominen, J., Mäkelä, E., Tamper, M., Hietanen, A., Hyvönen, E., 2017. Law and justice as a linked open data service (Available at:) https://seco.cs.aalto.fi/publications/submitted/law-justice-linked.pdf .
Pina-Sánchez, J., 2015. Defining and measuring consistency in sentencing. In: Roberts, J.V. (Ed.), Sentencing Guidelines: Exploring Sentencing Practice in England and Wales. Palgrave Macmillan, Basingstoke, pp. 76–92.
Pina-Sánchez, J., Brunton-Smith, I., Guangquan, L., 2018a. Mind the Step: A More Comprehensive Empirical Study of the Functioning of the England and Wales Sentencing Guidelines. Criminology and Criminal Justice.
Pina-Sánchez, J., Gosling, J.P., Chung, H., Bourgeois, E., Geneletti, S., Marder, I.D., 2019. Have the England and Wales guidelines affected sentencing severity? An empirical analysis using a scale of severity and time-series analyses. Br. J. Criminol. 59 (4), 979–1001.
Pina-Sánchez, J., Grech, D., 2017. Location and sentencing: to what extent do contextual factors explain between court disparities? Br. J. Criminol. 58 (3), 529–549.
Pina-Sánchez, J., Linacre, R., 2013. Sentence consistency in England and Wales: evidence from the Crown court sentencing survey. Br. J. Criminol. 53 (6), 1118–1138.
Pina-Sánchez, J., Linacre, R., 2016. Refining the measurement of consistency in sentencing: a methodological review. Int. J. Law Crime Justice 44, 68–87.
Pina-Sánchez, J., Lightowlers, C., Roberts, J., 2017. Exploring the punitive surge: Crown Court sentencing practices before and after the 2011 English riots. Criminol. Crim. Justice 17 (3), 319–339.
Pina-Sánchez, J., Roberts, J., Sferopoulos, D., 2018. Does the Crown court discriminate against Muslim-Named offenders? A novel investigation based on text mining techniques. Br. J. Criminol. 59 (3), 718–736.
Roberts, J.V., Ashworth, A., 2016. The evolution of sentencing policy and practice in England and Wales, 2003–2015. Crime Justice 45 (1), 307–358.
Roberts, J.V., Irwin-Rogers, K., 2015. Sentencing practices and trends: 1999-2013. In: Roberts, J.V. (Ed.), Sentencing Guidelines: Exploring Sentencing Practice in England and Wales. Palgrave Macmillan, Basingstoke, pp. 76–92.
Roberts, J.V., Plesničar, M.M., 2015. Sentencing, legitimacy, and public opinion. In: Mesko, J., Tankebe, J. (Eds.), Trust and Legitimacy in Criminal Justice. Springer, pp. 33–51.
Saris, P.B., Carr, W.B., Jackson, K.B., Hinojosa, R.H., Howell, B.A., Friedrich, D.L., Fulwood, I., Wroblewski, J.J., 2012. Report on the Continuing Impact of the United Stats V. Booker on Federal Sentencing. United States Sentencing Commission Available at: https://www.ussc.gov/research/congressional-reports/2012-report-congress-continuing-impact-united-states-v-booker-federal-sentencing.
Schanzenbach, M.M., Tiller, E.H., 2006. Strategic judging under the US sentencing guidelines: positive political theory and evidence. J. Law Econ. Organ. 23 (1), 24–56.
Schulhofer, S., 1991. Assessing the Federal sentencing process: the problem is uniformity, not disparity. Am. Crim. Law Rev. 29 (3), 833–873.
Sentencing Commission Working Group, 2008. Sentencing Guidelines in England and Wales: an Evolutionary Approach. Courts and Tribunals Judiciary, London.
Spohn, C., 1990a. The sentencing decision of black and white judges: some expected and unexpected similarities. Law Soc. Rev. 24, 1197–1216.
Spohn, C., 1990b. Decision making in sexual assault cases: do black and female judges make a difference? Women Crim. Justice 2 (1), 83–105.
Spohn, C., Brennan, P.K., Kim, B., 2017. Explicating the patterns of disparities using a path model. In: Ulmer, J.T., Bradley, M.S. (Eds.), Handbook on Punishment Decisions: Locations of Disparity. Routledge, pp. 211–238.
Steffensmeier, D., Britt, Ch, 2001. Judge's race and judicial decision making: do black judges sentence differently? Soc. Sci. Q. 82 (4), 749–765.
Steffensmeier, D., Demuth, S., 2000. Ethnicity and sentencing outcomes in US Federal Courts: Who is punished more harshly? Am. Sociol. Rev. 65, 705–729.
Steffensmeier, D., Hebert, Ch, 1999. Women and men policymakers: does the judge's gender affect the sentencing of criminal defendants? Soc. Forces 77 (3), 1163–1196.
Steffensmeier, D., Kramer, J.H., Streifel, C., 1993. Gender and imprisonment decisions. Criminology 31, 411–446.
Steffensmeier, D., Ulmer, J., Kramer, J., 1998. The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, black, and male. Criminology 36 (4), 763–798.
Stolzenberg, L., D'Alessio, S.J., 1994. Sentencing and unwarranted disparity: an empirical assessment of the long-term impact of sentencing guidelines in Minnesota. Criminology 32 (2), 301–310.
Sutton, J.R., 2013. Structural bias in the sentencing of felony defendants. Soc. Sci. Res. 42 (5), 1207–1221.
Tarling, R., 2006. Sentencing practice in magistrates' courts revisited. Howard J. 45, 29–41.
Tata, C., 2013. The struggle for sentencing reform. In: Ashworth, A., Roberts, J.V. (Eds.), Sentencing Guidelines. Oxford University Press, Croydon, pp. 236–256.
Thomas, C., 2017. 2016 UK Judicial Attitude Survey: Report of Findings Covering Salaried Judges in England and Wales Courts and UK Tribunals. UCL Judicial Institute(Available at:)https://www.judiciary.uk/wp-content/uploads/2017/02/jas-2016-england-wales-court-uk-tribunals-7-february-2017.pdf.
Uhlman, T.M., 1978. Black elite decision-making: the case of trial judges. Am. J. Pol. Sci. 22 (4), 884–895.
Ulmer, J.T., 1997. Social Worlds of Sentencing: Court Communities under Sentencing Guidelines. State University of New York Press, Albany.
Ulmer, J.T., 2012. Recent developments and new directions in sentencing research. Justice Q. JQ 29, 1–40.
Ulmer, J.T., Johnson, B.D., 2004. Sentencing in context: a multilevel analysis. Criminology 42, 137–178.
Ulmer, J.T., Johnson, B.D., 2017. Organizational conformity and punishment: federal court communities and judge-initiated guidelines departures. J. Crim. Law Criminol. 107, 253–292.
Ulmer, J., Light, M.T., Kramer, J., 2011. The “liberation” of Federal judges' discretion in the wake of the Booker/Fanfan decision: is there increased disparity and divergence between courts? Justice Q. JQ 28, 799–837.
Volkov, V., 2016. Legal and extralegal origins of sentencing disparities: evidence from Russia's criminal courts. J. Empir. Leg. Stud. 13 (4), 637–665.
Waldfogel, J., 1998. Does inter-judge disparity justify empirically based sentencing guidelines? Int. Rev. Law Econ. 18, 293–304.
Welch, S., Combs, M., Gruhl, J., 1988. Do black judges make a difference? Am. J. Pol. Sci. 32, 126–136.
Wooldredge, J., 2012. Distinguishing race effects on pre-trial release and sentencing decisions. Justice Q. JQ 29, 41–75.
Yang, C.S., 2014. Have inter-judge sentencing disparities increased in an advisory guidelines regime? Evidence from Booker. N. Y. Univ. Law Rev. 89, 1268–1342.
Yang, C.S., 2015. Free at last? Judicial discretion and racial disparities in Federal sentencing. J. Leg. Stud. 44 (1), 75–111.
Zatz, M., 2000. The convergence of race, ethnicity, gender, and class on court decision making: looking toward the 21st century. Natl. Inst. Justice J. 3, 503–552.