British Journal of Criminology Advance Access published July 2, 2013

doi:10.1093/bjc/azt040 BRIT. J. CRIMINOL

SENTENCE CONSISTENCY IN ENGLAND AND WALES

Evidence from the Crown Court Sentencing Survey

Jose Pina-Sánchez and Robin Linacre*

We assess the use of sentencing guidelines for assault issued in England and Wales, and the consistency with which they are applied by judges in the Crown Court. We use data from the Crown Court Sentencing Survey (CCSS), which records data on legal factors considered in the sentencing guidelines. This gives us access to a wide range of explanatory variables, allowing us to produce more robust findings about consistency in sentencing. We first employ a standard regression model to determine how guideline factors affect sentence outcomes empirically. Second, a random slopes multilevel model is used to analyse whether these factors have been consistently applied across different Crown Court centres. Our results point to a substantial degree of consistency in sentencing.

Keywords: sentencing, guidelines, consistency, multilevel modelling

Introduction

Consistency in sentencing, or whether like cases are treated alike, is one of the fundamental principles underpinning a just legal system. As well as being a goal in its own right, consistency in sentencing is also associated with other highly desirable effects. For example, consistency promotes the legitimacy of the criminal justice system, fosters public confidence in sentencing (Council of Europe 1993) and helps to establish a common understanding of the consequences of criminal activity, which may be desirable from the point of view of law and order, and to victims of crime.

The importance of consistency in sentencing has been recognized for many years in England and Wales, and there is a long history of institutions that have helped support this goal. In 1907, the Court of Criminal Appeal was created, partly in response to public and political concern about disparity in sentences (Ashworth 2009). Initially, this court dealt with appeals against undue severity of sentence but not undue leniency.[1] During the twentieth century, the Court's influence grew and, with increasing frequency, its judgments were used to give authority to general sentencing principles (Ashworth 1998).

In recent years, sentencing guidelines have increasingly been used to promote consistency of sentencing. They came initially in the form of guideline judgments from the Court of Appeal, the first of which were issued in the 1970s.[2] However, guideline judgments were relatively rare and only covered a limited range of offences (Ashworth 1998). The Crime and Disorder Act 1998 introduced the Sentencing Advisory Panel to

* Jose Pina-Sánchez, PhD candidate in Social Statistics at the University of Manchester, CCSR, School of Social Sciences, Humanities Bridgeford St Building, University of Manchester, Manchester, M13 9PL, UK, jose.pinasanchez@postgrad.manchester.ac.uk; Robin Linacre is Economic Advisor to the Sentencing Council for England and Wales, Room EB16, East Block, Royal Courts of Justice, The Strand, London, WC2A 2LL, UK.

[1] It was not until 1988 that the remit of the Court was expanded to hear appeals on the basis of undue leniency.

[2] e.g. *Willis* (1974) 60 Cr App R 146.

draft and consult on guidelines, and to advise the Court of Appeal on the form they should take. The Criminal Justice Act 2003 changed this structure so that, rather than advising the Court of Appeal on guidelines, the Sentencing Advisory Panel advised a new body, the Sentencing Guidelines Council, which had sole authority to issue definitive guidelines.

The most recent institutional change was in 2009, when the Coroners and Justice Act established the current Sentencing Council for England and Wales, which replaced its two predecessor bodies. Arguable, the Act its guidelines became more binding on courts. The Act provides that 'Every court must, in sentencing an offender, follow any sentencing guidelines which are relevant to the offender's case [...] unless the court is satisfied that it would be contrary to the interests of justice to do so'[3] (s. 128(1)(a), Coroners and Justice Act, 2009) whilst under previous legislation courts were only required to 'have regard to' sentencing guidelines.[4]

In addition to this change, the Sentencing Council was granted greater analytical duties than its predecessor bodies. Specifically, under s. 128 of the Act, the Council has a duty to 'monitor the operation and effect of its sentencing guidelines'.

In order to fulfil its statutory analytical functions, the Council operates the Crown Court Sentencing Survey (CCSS). The CCSS aims to be a census survey of all cases sentenced in the Crown Court in England and Wales. The survey records in unprecedented detail the factual elements of each case. For example, information is recorded on the seriousness of the offence (as defined in the relevant sentencing guideline), the aggravating and mitigating factors present in the offence, the stage at which a guilty plea was entered and its impact on the sentence, and the number of recent, relevant previous convictions which were taken into account by the sentencer.

These data therefore allow for a more accurate and reliable quantitative investigation into the degree of consistency of sentencing in England and Wales than has previously been possible. Using statistical techniques which are well established in other fields,[5] and which were first applied to the subject of sentencing decisions by Anderson and Spohn (2011), we present results on the use of sentencing guidelines by Crown Court judges, and the consistency with which guideline factors are applied across Crown Court centres in England and Wales.[6] For brevity and clarity, we concentrate our analysis on assault offences. However, we include summary modelling results for burglary and robbery offences in Appendix IV to demonstrate that our main findings on consistency generalize to other offence types.

The structure of this paper is as follows. The first section is a literature review where we consider previous studies that have looked at consistency in sentencing in the judicial system of England and Wales, and assess the adequacy of the research designs that have been used so far. Next, we present our analysis which starts by estimating whether guideline factors have the expected effect on sentences; this will help us understand better how guidelines are being used. The second part of the analysis studies consistency across Crown Court centres,[7] which is the focus of the paper. Specifically, we use a

[3] Coroners and Justice Act, s. 128(1)(a).

[4] Criminal Justice Act 2003, s. 172.

[5] See, e.g. Plewis (1997).

[6] Guideline factors are legal factors which sentencing guidelines state should be taken into consideration when formulating a sentence. These factors are a mixture of statutory legal factors and other factors.

[7] Hereafter, Crown Court centres will be referred to simply as 'courts'.

multilevel model to analyse differences in the effect of guideline factors across courts. In the last part of the paper, we summarize our results, discuss some potential methodological issues and indicate directions which might be taken for further research on the topic.

Literature Review

Consistency has been a relatively difficult concept to operationalize for research, despite the apparent simplicity of the term. How exactly do we investigate whether like cases are sentenced alike? This difficulty can be illustrated by the multitude of different research designs that have been used in the extensive literature of the topic. Figure 1 shows these different research designs, which can be organized in a hierarchical structure with a first partition separating qualitative and quantitative studies.

A majority of studies in the literature have had to rely on using a qualitative approach, perhaps due to the lack of data on the legal factors which were present in cases, or to the difficulty in identifying suitable statistical models to measure consistency. For instance, Davies *et al.* (2002) and Hough *et al.* (2003) ran focus groups with Crown Court judges and magistrates in England and Wales to study their views on how different offences should be sentenced. Both studies found significant divergences in the factors taken into account when imposing custodial sentences, and hence in the potential sentence outcomes for similar offences. For example, Hough *et al.* (2003) identified that, when imposing custodial sentences, most of the sentencers made their decision based on the type of offence but accepted that mitigating factors could change it to a community sentence. However, others relied only on the type of offence to determine whether the custody threshold had been passed, to the exclusion of other factors.

Although qualitative studies such as focus groups or interviews can be extremely valuable to ascertain the underlying processes in judges' deliberations, they suffer from serious design flaws in the form of 'generalizability' and 'replicability' when assessing consistency in sentencing. The number of judges that can be contacted and the number of cases that can be reviewed with their collaboration is very limited due to time constraints. In addition, given the qualitative nature of the results, it is very difficult to make reliable comparisons with other countries or periods of time. Because of these problems affecting qualitative studies, we focus the remainder of the literature review on quantitative designs.

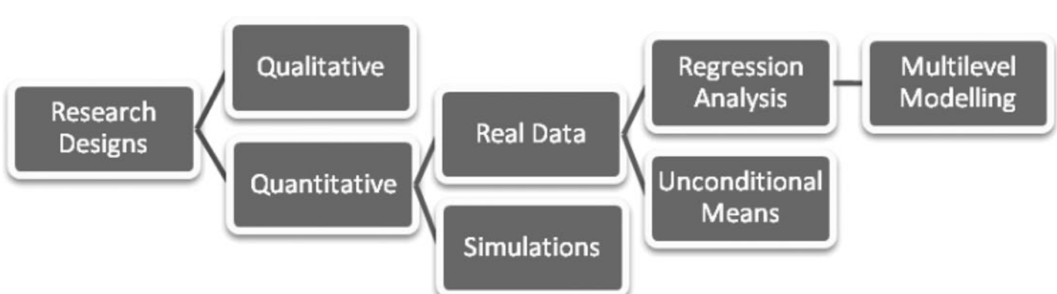

Fig. 1. Research designs used in the literature of consistency in sentencing

Within the family of quantitative studies, we can distinguish between studies relying on simulations from those using real data. In simulation studies, case files describing hypothetical scenarios are distributed to different judges. The judges are asked to consider the scenarios and indicate the sentence that they would impose in each case; consistency is then identified as a measure of dispersion of sentencing outcomes between judges. One study using this design in England and Wales is Davies and Tyrer (2003), where 51 judges from 12 courts were asked to give their views on appropriate sentences for five domestic burglary scenarios. Here again, the findings demonstrated substantial differences in sentence type and length. However, much like qualitative designs, simulation studies are prone to serious limitations. In simulation studies, judges sentence in an artificial environment which may differ from a real case—for instance, they will not have a real defendant in the dock before them. More importantly, it is questionable whether the same amount of attention is dedicated to simulations as it is to real cases (Anderson *et al.* 1999). On the one hand, judges may spend less time on the exercise than on a real case because they know there are no real-life consequences of the sentence. On the other hand, if the simulation is perceived as an assessment of the judge's performance, the opposite could be true.

Studies that use real data can be classified according to the sophistication of their statistical techniques. This would distinguish between those that employ regression analyses or other type of advanced statistical models and those that are limited to analysis of descriptive statistics. Tarling (2006) and Mason *et al.* (2007) represent the only empirical analyses of consistency in sentencing in England and Wales that we are aware of, and they are both based on analyses of descriptive statistics. The former study compares sentence outcomes from 30 magistrates' courts, across all offences in the year 2000. Disparities were found between courts in the use of disposal types; furthermore, the author showed that those disparities had not changed substantially since he first carried out this analysis using data from 1974. The paper also considered sentence practice within offence categories, and found that substantial variations remained in cases of burglary. The latter study analysed custodial sentence lengths imposed in magistrates' courts and the Crown Court for the period 2003–06. Here, the authors compared means and standard deviations controlling for variables such as type of offence or local crime rates and showed significant disparities in sentence length across the 42 Criminal Justice Areas in England and Wales.

A shortcoming of the findings from these two studies stems from their inability to control simultaneously for multiple confounding effects such as type and severity of offences and the demographic characteristics of offenders in different courts. As a result, observed disparities in sentence outcomes do not necessarily imply inconsistent sentencing practice. Differences in sentence outcomes or custodial sentence length could be due to differences in the set of confounding variables between courts. For instance, there may be systematic differences between courts in the mix of aggravating or mitigating factors present in cases, which in turn could be a consequence of factors such as different age structures or levels of economic inequality.

In order to control for these potential confounding effects simultaneously, it is common for studies to use regression modelling. In their methodology section, Ostrom *et al.* (2008) recognize the potential of regression analysis, and draw a distinction between the concepts of macro- and micro-level consistency. Macro-level inconsistency can be defined as any remaining variability in sentence outcomes[8] after controlling for all relevant legal factors considered to have an effect in that outcome (e.g. aggravation/mitigation factors, previous convictions, guilty plea, etc.). The logic behind this definition stems from the understanding that any variability that is not explained by relevant legal factors is due to judges' subjective views. This design is therefore seriously flawed if modelling work omits relevant legal factors. In that case, variability in sentencing that was due to the missing legal factors would be wrongly attributed to inconsistency in sentencing. In practice, it is highly unlikely that researchers will have accurate measures of all relevant legal factors and therefore, using standard regression models, it will not be possible to distinguish between unexplained variability in sentencing arising from imperfect data and variability in sentencing arising from inconsistency.

Micro-level consistency arises when all legal factors which are expected to be relevant to sentence outcomes can be shown to influence sentences empirically. This is a less stringent test than macro-level consistency because it only requires that the legal factors can be shown to have an effect, and does not impose the additional constraint that the size of the effect must be equal amongst courts. In statistical terms, micro-level consistency arises when legal factors that are included as explanatory variables in regression models are shown to be statistically significant and affect sentences in the expected direction.[9] However, findings from these models are also subject to problems derived from omitted relevant variables, which might bias regression coefficients if researchers do not have access to the full set of legal factors that influence sentencing decisions.

In conclusion, standard regression analyses can unveil interesting dynamics in testing whether legal factors are functioning as expected, they can detect whether there are systematic differences due to non-legal features and ultimately they can be used to assess consistency in sentencing if they are carefully designed. However, they are prone to a serious methodological flaw derived from the omission of relevant variables, which can mean that findings from these types of analyses may be unreliable.

An interesting extension to standard regression analysis is multilevel modelling. These models are well established in other fields[10] and have been progressively introduced in the literature of criminology during the last decade (e.g. Anderson and Spohn 2011; Fearn 2005; Johnson 2006; Kautt 2002; Ulmer *et al.* 2011, etc.). They could now be considered a mainstream tool in the field.

These models provide a statistically sound method of accounting for the effects of hierarchical groupings within data sets. In our case, sentences can be grouped by court. These models are also attractive because, as we will see, they offer a method of assessing consistency which is less sensitive to the problem of omitted relevant variables.

The simplest form of multilevel model is a random intercept model, where the model's constant (intercept) is allowed to vary across judges and/or courts.[11] This type of model allows the average sentence length to vary across courts and therefore can

[8] Sentence outcome could be either the disposal type used, the custodial sentence length or a combination of the two. The remaining variability could be measured by the standard deviation of residuals.

[9] Two examples where this type of design has been used are Hola *et al.* (2009) and Kempf-Leonard and Sample (2001). The former analyses how predictable sentences from the International Criminal Tribunal for the Former Yugoslavia were; the latter included a set of non-legal factors to identify whether disparities in sentencing were based on demographic factors and personal circumstances, which were understood as measures of discrimination.

[10] Such as the study of education attainment.

[11] See Snijders and Bosker (1999) for a good introduction to multilevel modelling.

be used to obtain additional insights regarding consistency in sentencing. Under a random intercept model, unexplained variability in sentencing can be partitioned for each of the levels considered in the model. It is therefore possible to separate variability of sentences that arises due to differences between courts and variability in sentencing resulting from other confounding factors such as differences in the mix of offences amongst courts. Anderson *et al.* (1999) analysed 'inter-judge' disparity in the average length of prison sentences in the United States before and after the Federal Sentencing Guidelines were implemented. The authors used a random intercept model and found that the unexplained variability at the judge level fell from 17 to 11 per cent and, as a result, they concluded that the implementation of the guidelines improved consistency. Fearn (2005) also used a random intercept model to study whether there are community effects (such as unemployment rate or economic inequality) that determine incarceration, which are independent of characteristics of either the defendant or the case in the United States. The author found that about 25 per cent of the unexplained variation is located at the county level, even after controlling for legal and extra-legal factors, and concludes that this variability is evidence of disparities in sentencing.

In our view, although random intercept models can be used to obtain new insights in the study of consistency in sentencing, they are equally susceptible to the problem of omitted relevant variables as more traditional regression models. In particular, variability in the residuals at the judge or court level might be due to unexplained differences between the cases being seen by judges or courts. The researcher therefore cannot fully separate genuine inconsistency in sentencing from variation in sentencing arising from differences in the caseload of the court or the judge. For example, it might be the case that those located in urban areas have to deal with cases with more aggravating factors than those in rural areas. If data on these aggravating factors are unavailable, a random intercepts model may attribute the resultant differences in sentencing between urban and rural courts to the court level of the model, giving a misleading picture of the sources of variability in sentencing.

The random intercept model can be extended to include random slopes for the legal factors, which proves to be remarkably useful in the assessment of consistency. This extension removes the simplifying assumption that legal factors must have a constant effect on sentences across courts. Legal factors are thus permitted to have varying effects on sentences in different courts and the researcher can test whether this is the case empirically. If this test points to statistically significant variability in the random slope, this can be interpreted as evidence of inconsistency in sentencing practice. Importantly, the results of this test are more resilient to the problem of omitted relevant variables than random intercept or standard regression models. When using the random slopes model to assess consistency, the interest does not lie in the value of the regression coefficients per se, but on whether the effect of each specific legal factor on sentence length is consistent across courts. This means that the possibility that the coefficients of a legal factor may be biased due to omission of relevant variables does not directly affect the findings of consistency from the random slopes model.[12] These

[12] Omitted variable bias may still be a problem but the situations in which it affects the findings on consistency from the random slopes model are more unusual. An example could be a model in which we omit a relevant legal factor, where there exists multicolinearity amongst legal factors and where the nature of multicolinarity varies between courts.

models allow us to assess whether legal factors are treated consistently between courts, but it should be noted that this does not offer a complete picture of consistency in sentencing. Even if all legal factors are treated the same between courts, it remains possible that there could be systematic differences is sentence levels if different courts start from a different 'baseline' sentence length.

We are only aware of one study using random slopes to analyse consistency in sentencing: Anderson and Spohn (2011). Here, the authors assessed whether the Federal Sentencing Guidelines have led to uniformity in sentence lengths but, instead of using a before-and-after comparison of residuals (as in Anderson *et al.* 1999), random slopes were included for the legal factors considered in the guidelines. In particular, three out of the five legal factors included in the model were found to have a significant random slope term, meaning that these legal factors were treated inconsistently.

To sum up, different research designs have been used in the study of consistency in sentencing. However, most of these designs suffer from serious methodological limitations. Here, we have introduced them in order of methodological integrity. We conclude that a quantitative approach using real data is to be preferred due to reasons of robustness, generalizability and replicability; unfortunately, many of the quantitative studies on the topic have relied on inadequate statistical techniques. In this regard, we have identified multilevel modelling, and the random slopes model in particular, as the best method to study consistency in sentencing, since it can be used to test whether relevant legal factors are being applied consistently between judges or courts and is less sensitive to problems of omitted relevant variables than direct assessments of residuals.

A second conclusion from this literature review relates to the lack of empirical analyses in England and Wales. Findings collected so far for the level of consistency in courts of England and Wales has relied on either qualitative or simulation studies. The only exceptions are Tarling (2006) and Mason *et al.* (2007), although the methods used in these papers (descriptive statistics by groups) were only exploratory. Thanks to the recently available data from the CCSS, we are able to run a first methodologically robust empirical analysis of consistency in sentencing in England and Wales.

Exploratory Analysis of Crown Court Sentencing Survey Data Set

In this first part of the analysis, we use a standard regression model of custodial sentence lengths, including guideline factors derived from sentencing guidelines as explanatory variables. This is an exploratory analysis designed to test whether the empirical relationships we would expect to find in the data are present and to detect any possible problems with the data set. As such, we test whether the guideline factors can be shown to influence sentence lengths, which helps us understand the empirical functioning of guidelines factors. Due to the problems of omitted variables discussed in the first section, we do not use this regression analysis part to draw conclusions about sentence consistency. Instead, the regression analysis sets the foundations for moving on to random slopes model, which we use to derive our substantive findings on sentence consistency.

The response variable in our regression analysis captures custodial length for assault offences imposed in 2011. In particular, we consider offences of assault covered by sentencing guidelines which resulted in periods of custody longer than a month. These offences are:[13] actual bodily harm[14] (ABH), grievous bodily harm[15] (GBH) and grievous bodily harm with intent[16] (GBH with intent). Other offences covered by the guidelines such as assault with intent to resist arrest and assault on a police constable were not considered because of the small number of cases that fit the criteria of our sample. Offences of common assault were discarded, despite being the third most frequent category, because they are a summary-only offence, which means that cases of common assault are usually limited to being sentenced in the magistrates' court. Sentences of lower than a month were discarded because they seemed to correspond to a different sentencing mechanism, giving the distribution of sentences length a bimodal shape. Finally, the natural logarithm was taken of the sentence duration in days in order to adjust for the right skewness of its distribution.[17]

The explanatory variables of primary interest were the guideline factors relating to the offence and the offender, as recorded in the CCSS. Due to the implementation of a new sentencing guideline for assault offences in June 2011, new items were included in the CCSS form[18] and some of the old ones were reworded. In order to avoid inconsistencies, we decided to consider only items that could be found in both guidelines and which were left unchanged on the CCSS forms. These are: whether the offender pled guilty at the *first reasonable opportunity*, whether he or she showed *remorse*, was the main *carer* of a dependent person, was a member of a *gang* and whether the assault was perpetrated on a *vulnerable* person, on a *public worker*, under the effect of *drugs* or *sustained* in time. In addition to these guideline factors, we added another set of variables with the intention of controlling for confounding effects. These are: 75 dummy variables capturing court location at which the sentence was passed and two dummy variables capturing the different types of assault.

All the explanatory variables are binary with the exception of *previous convictions*, which is a three-level ordinal variable indicating: none, one to three, and four to nine convictions.[19] The final sample size is 5,527 sentences of assault imposed in 2011 at the Crown Court in England and Wales. The means and standard deviations for all of the variables used in the analysis are included in Appendix I.

In our first model, ordinary least squares (OLS hereafter) is used to regress the log of sentence length on all the explanatory variables described above. Full specifications of the statistical modelling presented in this paper can be found in Appendix II. Results from the OLS model are shown in Table 1, with the exception of the coefficients for the different courts, which have been included in Appendix III for reasons of space.

The strongest predictors of long sentence length are the binary variables relating to the type of assault, especially *GBH with intent*, which increases average sentence length by 894 days (2.5 years) relative to the reference case.[20] On the other hand, of the 75

[13] Terms in italics will be used throughout the remainder of the paper to refer to the variables used in our models.

[14] Offences Against the Person Act 1861, s. 47.

[15] Offences Against the Person Act 1861, s. 20.

[16] Offences Against the Person Act 1861, s. 18.

[17] This process of using the natural logs of sentences longer than a month in order to normalize the response variable was first suggested by Anderson and Spohn (2011).

[18] The CCSS form is the Crown Court Sentencing Survey form that judges fill out which records details of the case.

[19] The CCSS questionnaire also considers a category for ten or more previous convictions, but we did not capture subjects with that value in our sample.

[20] The reference case is an assault occasioning body harm committed by a man with no previous convictions, no aggravation or mitigating factors and sentenced in Aylesbury.

Table 1 *Results from the OLS Model**

Variable	Regression coefficient	Standard error
Intercept	**5.407**	0.010
Previous conviction	**0.428**	0.061
(Previous convictions)^2	**–0.114**	0.015
First opportunity	**–0.097**	0.017
Remorse	**–0.116**	0.017
Carer	**–0.124**	0.038
Gang	0.015	0.021
Vulnerable	**0.140**	0.022
Public worker	**–0.076**	0.037
Sustained	**0.205**	0.017
Drugs	**0.056**	0.016
GBH	**0.465**	0.017
GBH with intent	**1.612**	0.021
Adjusted R^2	0.564	
Sample size	5,527	

* Estimates that are statistically significant at the 5 per cent level appear in bold.

courts, only Bournemouth, Wood Green, Merthyr Tydfil, Norwich and Woolwich were significant at the 5 per cent confidence level.[21] Specifically, those courts differ from the sentence length of the reference case (averaging 223 days) by –43, –57, –59, –69 and –52 days, respectively.

Regarding the explanatory variables of primary interest, the guideline factors, we find that most of them are having the expected effect, with *first opportunity*, *remorse* and *carer* reducing the length of the sentence, and *previous convictions*,[22] *vulnerable*, *sustained* and *drugs* increasing it. The only exceptions are *gang* and *public worker*. The coefficient for gang is positive, as would be expected from an aggravating factor; however, we find that it is not statistically significant, which raises questions on the extent to which this factor influence custodial sentence length, after other guideline factors have been taken into account. The case of *public worker* is more surprising, since it is statistically significant but has a downward influence on sentence length. However, we suspect that this is a manifestation of the problem of omitted relevant variables that we mentioned in the previous section, but further research is required.

Despite the observed problems with *gang* and *public worker*, overall, this analysis provides statistical evidence that sentencing guidelines for assault are used by judges in the ways anticipated by the Sentencing Council. The ratio of legal factors which have the expected effect[23] on sentence lengths (seven of nine) is greater than in the other two studies that look at this specific question using a similar research design and a similar response variable. In particular, Hola *et al.* (2009), in a study on the predictability of the International Criminal Tribunal for the Former Yugoslavia, found that three out of eight of the legal factors included in their model on sentences worked as expected, while Anderson and Spohn (2011), using federal data from the United States, found that the same ratio was two out of five.

[21] Five per cent is the significance level to determine statistical significance used across this article.

[22] A quadratic effect was found for this variable, indicating that the increase in sentence length due to higher previous convictions is not linear.

[23] That is, having a significant effect in the expected direction.

Our modelling provides interesting statistical evidence in its own right, but it also suggests there is scope for further study that may yield insights into the consistency with which guideline factors are applied across courts. In what follows, we extend the standard regression model used here to a multilevel model which allows for random effects at the court level.

Assessment of Consistency in Sentencing

We have argued that multilevel modelling provides a more robust way of analysing consistency in sentencing than standard regression analysis. In this section, we estimate two multilevel models, using the same data set as in the previous section. First, we consider the simplest type of multilevel model, an empty random intercept model (also known as variance component model). This allows us to identify whether there exists variability in sentencing between courts. Without such variability, more complex multilevel modelling would be unlikely to yield interesting results. This simple model also gives us some initial findings about how sentencing varies amongst courts and how it compares to other jurisdictions. Having shown that variability exists at the court level, we then develop a more sophisticated multilevel model which allows for random intercepts and slopes, and which uses the set of explanatory variables from the CCSS. The results from this model represent our primary findings on consistency of sentencing across courts in England and Wales.

The CCSS survey can be viewed as a hierarchical data set with two levels. Individual offences (level one of the data set) are nested within courts (level two). There are currently 76 Crown Court centres in England and Wales,[24] each of them having many judges. The CCSS does not collect judge identifiers which could have allowed us to estimate a three-level model. Hence, the analysis presented in this section refers to sentencing consistency between courts, not between judges.

Variance component model

We start by specifying the simplest type of multilevel model, an empty random intercept (also known as variance component) model,[25] to explore the variability of sentence means between courts. The differences with the OLS model (see Appendix II, Equation A1) are twofold: all the explanatory variables are taken out and a new residual term is included to allow different intercepts (random intercepts) for each of the courts. So, now we have two residuals terms, one capturing within-court variability, which we denote by σ^2,and another capturing between-court variability, denoted by τ^2. In particular, the random intercept term could be thought as a vector capturing the difference between the court's mean sentence length and the mean sentence length for the entire population. Using the variances of these two residual terms, we can calculate the intra-cluster correlation (ICC hereafter) as follows:

$$ICC=\tau^2/\tau^2+\sigma^2$$

[24] www.judiciary.gov.uk/you-and-the-judiciary/going-to-court/crown-court Last accessed 20/June/2013.

[25] Estimation of the multilevel models included here was conducted using SAS 9.2. In addition, the analysis was also replicated using MLwiN 2.25 to test the robustness of our results.

The ICC is used to measure the proportion of total variability in outcomes which arises at the higher level (in this case, at the court level). The results from our analysis show that the variance of the random intercept term is statistically significant, but the estimate of the ICC is just 0.018, which means that only 1.8 per cent of the overall variance in sentence length is due to differences between courts.

Variance component models do not include explanatory variables, and it would be expected that the inclusion of guideline factors as explanatory variables would change the ICC. The direction of change would depend on the source of variability in sentencing between courts. If variability is primarily due to differences between courts in the explanatory variables (the offence mix), then the ICC would be expected to fall when these explanatory factors were included. If variability is primarily due to inconsistency, then the ICC would be expected to rise. However, 0.018 is a low figure, which could be understood as preliminary evidence in support of a substantial degree of consistency in sentence lengths across courts.

Random slopes model

We have shown that the degree of variability in sentence lengths between courts is low in England and Wales. However, there are many reasons for variability in sentencing between courts: for instance, it could arise as a result of variability in the mix of offences (or offenders) appearing before different courts, as much as it could be a result of differing sentencing practices. This means that the conclusions we can draw about consistency from the random intercept model are limited. We now extend the model to allow for random slopes. This model allows us to consider differences in how individual guideline factors are treated across courts. It therefore has the potential to identify not only the presence of inconsistency, but also its possible sources.

By using this model, we accept that it is unrealistic to explain all variability in sentencing levels using our data set.[26] Instead, a different research goal is used: we study differences in the effect of guideline factors for which we do have data. We therefore have a realistic methodology that can be applied with our data set, which is more resilient to the inevitable problems of omitted relevant variables than other methodologies like standard regression analysis. This will lead to more robust insights about consistency in sentencing in England and Wales.

We extend the variance component model (see Appendix II, Equation A2) by including the set of the explanatory variables used in the standard regression model (Appendix II, Equation A1), with the exception of the dummy variables for the different courts, which are now accounted for by the random intercepts. In addition, we allow for all of the coefficients for the legal factors to be formed by a fixed and a random part. The latter represents the random slopes terms, which could thought as a matrix where each random slope forms a column capturing the difference between the mean of the slope at the population level and the slope for that particular court.

The random slopes component is thus a measure of whether the explanatory variables have different effects on sentence outcome in different courts. As such, it can be

[26] We do not see this as a shortcoming of our data set; rather, it is difficult to imagine that any data set could be collected in sufficient detail to explain all variability in sentencing levels arising from legal factors.

interpreted as a measure of consistency in sentencing: if the variance of a random slope is statistically significant, this suggests that the explanatory variables (e.g. guideline factors) may be treated inconsistently across courts.

Results from the random slopes model are shown in Table 2. We started by allowing for random slopes for each of the explanatory variables, but deviance tests showed that only the variances of two of those were statistically significant: the ones for *vulnerable* and *sustained*. The rest of the coefficients were fixed in the final model, meaning that the statistical evidence was consistent with the hypothesis that there was no difference in the way these factors contributed to sentence length across courts. The 'fixed effects'—that is, the estimates of the mean of the slopes and intercept—are very similar to the ones observed in the OLS model, with the exception of *public worker*, which is no longer statistically significant. Compared to the variance components model, the variance of the random intercept term is still statistically significant and the ICC is little changed (it is now 0.021).

The statistically significant random slopes terms for *vulnerable* and *sustained* indicate that these factors have different effects on sentence lengths in different courts. The specific magnitude of these differences can be expressed using a 95 per cent confidence interval. Following this approach, we find that the base case of an assault of *ABH* committed against a *vulnerable* victim is 8.6 months, but may vary by ±1.3 months. A similar base case of an offence of *ABH* where the attack has been perpetrated in a *sustained* manner is 8.1 months and it can vary by ±1.7 months.

These findings show that, even amongst the two legal factors that were found to have statistically significant random slopes terms, the variability in sentencing amongst

Table 2 *Results from the Random Slopes Model**

Variables	Regression coefficient	Standard error
Fixed effects		
Intercept	**5.356**	0.056
Previous conviction	**0.429**	0.060
(Previous convictions)^2	**−0.114**	0.015
First opportunity	**−0.094**	0.017
Remorse	**−0.119**	0.016
Carer	**−0.125**	0.037
Gang	0.014	0.021
Vulnerable	**0.139**	0.028
Public worker	**−0.078**	0.036
Sustained	**0.201**	0.020
Drugs	**0.053**	0.016
GBH	**0.467**	0.017
GBH with intent	**1.619**	0.021
Random effects		
Intercept	**0.005**	0.002
Vulnerable	**0.015**	0.008
Sustained	**0.007**	0.004
Level 1 residuals	**0.293**	0.006
−2 log likelihood	9,078.3	
Level 1 units	5,527	
Level 2 units	76	

* Estimates that are statistically significant at the 5 per cent level appear in bold.

courts is moderate. It is also important to recognize that the bulk of sentences exhibits lower variability than these confidence intervals may suggest: for instance, if we consider only the most similar 50 per cent of courts, we find that the variability in the effect of a *sustained* attack is ±0.6 months. It should also be noted that, although inconsistency in sentencing is one reason why such variability may exist, there may be other reasons. In particular, the CCSS survey uses a binary measure of *vulnerable* and *sustained*, but no specific definition is given of the extent to which these factors must be present to warrant ticking the box on the form. It is therefore possible that there is variability between courts in the degree of vulnerability amongst cases where *vulnerable* was ticked or the nature and severity of the *sustained* factor.

In summary, the random slopes model for assaults shows that the majority of legal factors seem to be treated consistently across courts and, even for the two variables where there is statistical evidence of differences in treatment, the magnitude of these differences is moderate. Furthermore, the number of random slopes which were found to have significant variances is relatively less substantial when comparing it to results from Anderson and Spohn (2011), which used a similar design but related to Federal Courts in the United States. In their model, three out of the five legal factors were found to have random slopes with statistically significant variances, whereas, in our model, that ratio was only two out of nine. However, the higher variability observed in Anderson and Spohn (2011) might be a result of having used judges instead of courts as the second-level unit.

The results we have presented are limited to offences of assault; in order to test their external validity, we replicated the same analysis for robbery and burglary offences, which are also well suited to modelling due to the high volumes of custodial sentences each year. The results for these two other offences are very similar to what we have seen for offences of assault. For burglary, we find that, of the nine legal factors included in the model, only the aggravating factor of being a member of group or gang had a significant random slope. For robbery, of the nine legal factors included in the model, two had significant random slope terms. These were the aggravating factor of *gang* and perpetrating the offence on a vulnerable victim. More detailed results are included in Appendix IV.

Conclusion

In this paper, multilevel modelling techniques have been applied to the problem of assessing consistency in sentencing in England and Wales for the first time. This has been made possible by new data from the CCSS, which records detailed information on the factors present in each case sentenced at the Crown Court in England and Wales. Specifically, we have analysed a sample of 5,527 real cases of assault sentenced in 2011 and the use of a random slopes model for a set of guidelines factors has allowed us to assess consistency in sentencing using a more robust methodology than traditional regression models.

Our findings support the view that, for cases of assault, guideline factors are being taken into account in the way intended by sentencing guidelines and that there is consistency in the way that guideline factors are applied across courts. The first part of our analysis confirms that a majority of guideline factors are statistically significant in regression models and show effect in the expected direction. The second part starts by illustrating how there are statistically significant differences in mean sentence length by courts. However, we have shown that, in an empty model which does not account for differences in the offence mix and other explanatory variables across courts, the percentage of the variability in sentence lengths attributed to differences between courts is only 1.8 per cent of the total. Finally, we considered whether guideline factors are treated differently across courts. We found that the majority of guideline factors are treated consistently amongst courts. Statistically significant differences in the treatment of two guideline factors were found, but the absolute differences in the treatment of these factors were found to be moderate.

Compared to other studies using a similar research design but based on federal sentences from the United States, our analysis points to greater consistency in sentencing. We have found both a higher ratio of legal factors which are statistically significant in a standard regression model and point in the expected direction and a lower ratio of legal factors with statistically significant random slopes. Furthermore, by replicating the analysis to other types of offences, we have found that our findings can also be generalized to offences of burglary and robbery.

Our findings of substantial consistency contrast with the findings from previous studies on the case of England and Wales. This discrepancy is likely to be due to differences in research designs, with the previous studies relying on qualitative methods, simulations or using comparisons of unconditional means. Another possibility is that there may have been changes in sentencing consistency in England and Wales from the time these studies were published (2000–05) to 2011.

In the literature review, we have argued that our methodology is more robust than any other that has been used in this area; however, a number of caveats must be observed. First, it is important to be precise about how our findings relate to the assessment of consistency. We are able to show that guideline factors have a similar influence on sentence lengths across courts, which is evidence in favour of consistency of sentencing practice. However, this is a necessary but not sufficient condition to prove that sentencing consistency exists. It leaves open the possibility that courts treat guideline factors in a similar way, but start the sentencing exercise from a different baseline sentence length. We were unable to assess whether this is the case because of the lack of comprehensive and detailed data on all legal factors that apply in each case. Second, unlike the studies from the United States that we have used to compare our levels of consistency, we use courts instead of judges as level two units in our hierarchical model. An assessment of consistency across judges would have probably showed more discrepancies, since the aggregation of judges at the court level will tend to attenuate some of the discrepancies between judges. Third, we are aware that some relevant legal factors could not be included in the analysis due to unavailability of data (the adjusted R^2 of our OLS model is 0.564). As a result, there might be a problem of omitted relevant variables. This is potentially serious because our regression coefficients might have been biased, which would affect the assessment of the degree of influence guideline factors have on sentence length. Fourth, our findings are limited to consistency in custodial sentence length. In order to obtain a full picture of consistency in sentences from the Crown Court, we should extend our analysis to consistency in determining the disposal types used.[27]

Finally, the CCSS suffers from a 39 per cent non-response rate and there is the possibility that these missing data are not ignorable.[28] For example, there may exist systematic

[27] That is assessing consistency in location, not only in duration (Ostrom *et al.* 2008).

[28] See Rubin (1987) for a classification of the implications and possible adjustments for the different missing data mechanisms.

differences in the use of sentencing guidelines between judges who fill in the forms and judges who choose not to. It could be that judges who fill out CCSS forms may also have a more favourable view of the Sentencing Council, and adhere to sentencing guidelines more closely. If this was the case, our results would be biased towards showing higher levels of consistency than actually exist.

In a future paper, we intend to make use of the 2011 data from the CCSS in order to monitor the evolution of consistency in sentencing across time and assess the impact of new sentencing guidelines. In order to do that, we are going to explore the data set using more advanced statistical techniques such as local regression or matching techniques, which we suspect can also offer interesting insights into the study of consistency in sentencing. This will also allow us to compare results drawn from different methods to the ones obtained here, which would allow us to offer additional insights in the robustness and adequacy of multilevel modelling to assess consistency in sentencing.

Funding

Economic and Social Research Council; the Sentencing Council.

Acknowledgements

The views expressed in this paper are those of the authors and do not necessarily represent the views of the Sentencing Council for England and Wales.

References

Anderson, A. and Spohn, C. (2011), 'Lawlessness in the Federal Sentencing Process: A Test for Uniformity and Consistency in Sentence Outcomes', *Justice Quarterly*, 27: 362–93.

Anderson, J., Kling, J. and Stith, K. (1999), 'Measuring Inter-Judge Sentencing Disparity: Before and After the Federal Sentencing Guidelines', *Journal of Law & Economics*, 42: 271–307.

Ashworth, A. (2009), 'Four Techniques for Reducing Sentence Disparity', in A. Ashworth, A. Von Hirsch and J. Roberts, eds, *Principled Sentencing*, 227–39. Hart Publishing.

Council of Europe (1993), 'Consistency in Sentencing', *Legal Issues*, 92, available online at http://book.coe.int/EN/ficheouvrage.php?PAGEID=36&=EN&produit_aliasid=809 Last accessed 20/June/2013.

Davies, M. and 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', *Criminal Law Review*, April: 243–65.

Davies, M., Takala, J. P. and Tyrer, J. (2002), 'Sentencing Burglars in England and Finland: A Pilot Study', in C. Tata and N. Hutton, eds, *Sentencing and Society: International Perspectives*. Ashgate.

Fearn, N. (2005), 'A Multilevel Analysis of Community Effects on Criminal Sentencing', *Justice Quarterly*, 22: 452–87.

Hola, B., Smeulers, A. and Bijleveld, C. (2009), 'Is ICTY Sentencing Predictable? An Empirical Analysis of ICTY Sentencing Practice', *Leiden Journal of International Law,* 22: 79–97.

Hough, M., Jacobson, J. and Millie, A. (2003), *The Decision to Imprison: Sentencing and the Prison Population* Prison Reform Trust.

Johnson, B. (2006), 'The Multilevel Context of Criminal Sentencing: Integrating Judge and County Level Influences in the Study of Courtroom Decision Making', *Criminology*, 44: 259–98.

Kautt, P. (2002), 'Location, Location, Location: Interdistrict and Intercircuit Variation in Sentencing Outcomes for Federal Drug-Trafficking Offenses', *Justice Quarterly*, 19: 633–71.

Kempf-Leonard, K., and Sample, L. (2001), 'Have Federal Sentencing Guidelines Reduced Severity? An Examination of One Circuit', *Journal of Quantitative Criminology*, 17: 111–144.

Mason, T., de Silva, N., Sharma, N., Brown, D. and Harper, G. (2007), *Local Variation in Sentencing in England and Wales.* Ministry of Justice.

Ostrom, B., Ostrom, C., Hanson, R. and Kleiman, M. (2008), *Assessing Consistency and Fariness in Sentencing: A Comparative Study in Three States.* Washington: National Institute of Justice.

Plewis, I. (1997), *Statistics in Education*. Edward Arnold.

Rubin, D. (1987), *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

Snijders, T. and Bosker, R. (1999), *An Introduction to Basic and Advanced Multilevel Modelling.* Sage.

Tarling, R. (2006), 'Sentencing Practice in Magistrates' Courts Revisited', *Howard Journal*, 45: 29–41.

Ulmer, J., Light, M. and 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 Quarterly*, 28: 799–837.

Appendix I:

Descriptive Statistics of the Variables Used

Variable	Mean	Standard deviation
Log sentence length	6.171	0.818
Previous convictions	1.867	0.683
First opportunity	0.249	0.432
Remorse	0.317	0.465
Carer	0.041	0.199
Gang	0.162	0.368
Vulnerable	0.138	0.345
Public officer	0.043	0.203
Sustained	0.321	0.467
Drugs	0.356	0.467
ABH	0.515	0.500
GBH	0.308	0.462
GBH with intent	0.176	0.381

Appendix II:

Statistical Modelling

OLS model

This model can be expressed as follows:

$$Y_i = \alpha + B_j X_{ij} + \varepsilon_i. \qquad \text{(A1)}$$

The response variable is denoted as a vector Y where i is a subindex differentiating sentences within the range i = 1, 2, …, 5527. The regressors are denoted by the matrix X with the subscript j used to differentiate between each of them, so j = 1, 2, …, 93; α is a scalar denoting the constant term, B is a vector representing the slope coefficients for each of the regressors included and ε is a vector denoting the error term, which is assumed to be normally distributed.

Variance component model

This model is specified as follows:

$$Y_{il} = \alpha + \mu_1 + \varepsilon_{il}. \qquad \text{(A2)}$$

Notice that the response variable has now two subscripts, one for each of the two levels. As before, i differentiates sentences, but l = 1, 2,…,76 is also introduced to index courts. As a result, the residual term has now been split in two: ε captures within-court residuals, which is the distance of individual sentences from the mean duration of their court, and μ is the random intercepts term, a vector capturing the difference between the court's mean sentence length and the mean sentence length for the entire population. In other words, ε captures within-court variability while μ captures between-court variability. Each of these residuals is assumed to be normally distributed with a mean of 0 and variance σ^2 and τ^2 respectively.

Random slopes model

This random slopes model takes the following form:

$$Y_{il}=\alpha+ B^*_j X_{ij}+\mu_1+\varepsilon_{il}. \tag{A3}$$

Notice that this equation expands A2 by including a set of explanatory variables (X) and their respective coefficients (B*). In addition, it is now formed of a fixed and a random part, B_j and v_{j1} respectively. The latter represents the random slopes terms. Here, they are presented by a matrix which, for each random slope, captures the difference between the mean of the slope at the population level and the slope for that particular court.

Appendix III:

*Regression Coefficients Including Courts**

Court	Regression coefficient	Standard error
Bradford	−0.077	0.099
Hull	−0.111	0.110
Birmingham	0.036	0.09
Bournemouth	**−0.214**	0.107
Dorchester	−0.121	0.137
Bristol	−0.008	0.103
Burnley	−0.122	0.112
Cambridge	0.003	0.134
Cardiff	−0.056	0.097
Carlisle	−0.159	0.107
Central Criminal Court	0.172	0.139
Chelmsford	−0.007	0.105
Chester	−0.006	0.105
Chichester	−0.078	0.134
Coventry	−0.009	0.118
Croydon	0.046	0.126
Derby	0.010	0.095
Doncaster	−0.022	0.126
Wolverhampton	0.041	0.096
Durham	−0.043	0.116
Exeter	0.065	0.107
Gloucester	−0.203	0.119
Great Grimsby	−0.056	0.109
Ipswich	0.028	0.113
Kingston upon Thames	−0.055	0.124
Blackfriars	−0.037	0.130
Leeds	0.013	0.096
Leicester	0.070	0.096

Appendix III: Continued

Court	Regression coefficient	Standard error
Lewes	0.030	0.103
Lincoln	−0.028	0.106
Liverpool	−0.118	0.089
Maidstone	0.003	0.104
Manchester Crown Sq.	0.008	0.102
Manchester Minshull St	−0.133	0.094
Merthyr Tydfil	**−0.308**	0.097
Mold	0.162	0.101
Newcastle-upon-Tyne	−0.094	0.091
Inner London Sessions House	−0.115	0.124
Northampton	0.067	0.134
Norwich	**−0.368**	0.114
Nottingham	−0.078	0.096
Oxford	0.038	0.118
Plymouth	−0.114	0.109
Portsmouth	−0.173	0.116
Preston	−0.122	0.098
Reading	−0.088	0.111
St Albans	0.012	0.113
Sheffield	0.025	0.093
Shrewsbury	0.029	0.134
Snaresbrook	−0.032	0.091
Southampton	0.032	0.143
Stafford	−0.107	0.104
Stoke-on-Trent	−0.163	0.105
Swansea	−0.136	0.098
Swindon	0.123	0.136
Taunton	0.030	0.222
Teesside	0.123	0.122
Basildon	−0.001	0.115
Warwick	0.066	0.121
Winchester	−0.145	0.150
Worcester	−0.088	0.104
York	−0.016	0.115
Harrow	0.067	0.112
Wood Green	**−0.295**	0.101
Bolton	0.028	0.101
Southwark	−0.151	0.143
Woolwich	**−0.267**	0.118
Peterborough	−0.018	0.115
Guildford	−0.097	0.125
Isleworth	−0.094	0.095
Luton	0.095	0.104
Truro	0.038	0.115
Newport (I.o.W.)	−0.215	0.139
Canterbury	−0.019	0.117
Salisbury	−0.203	0.552

* Estimates that are statistically significant at the 5 per cent level appear in bold.

Appendix IV: Results for Offences of Robbery and Burglary

Burglary

The ICC went from 2.2 per cent in the variance component model to 2.5 per cent for the random intercept model including legal factors as fixed effects. Table A1 shows the results from the random slopes model, where only the random slope for the aggravating factor of *gang* was found significant. In particular, the mean sentence for an *ABH* with *gang* present was 11.2 months, with a 95 per cent confidence interval of ±4.5 months.

TABLE A1 *Results from the Random Slopes Model on Burglary**

Variables	Regression coefficient	Standard error
Fixed effects		
Intercept	**5.757**	0.034
Previous conviction	**0.259**	0.044
(Previous convictions)^2	**−0.183**	0.024
First opportunity	−0.021	0.021
Age	**−0.210**	0.025
Remorse	0.023	0.026
In dwelling	**0.523**	0.026
Pre-planning	**0.189**	0.024
Gang	0.062	0.034
Under the influence	0.020	0.025
On bail	**0.080**	0.024
High value	**0.185**	0.025
Random effects		
Intercept	**0.009**	0.003
Gang	**0.029**	0.013
Level 1 residuals	**0.442**	0.009
−2 log likelihood	9,861.7	
Level 1 units	4,803	
Level 2 units	76	

* Estimates that are statistically significant at the 5 per cent level appear in bold.

Robbery

The ICC went from 6.3 per cent in the variance component model to 4.8 per cent for the random intercept model including legal factors as fixed effects. Table A2 shows the results from the random slopes model, where only the random slopes for the aggravating factor of *gang* and *vulnerable* were found significant. In particular, the mean effect the mean sentence for an *ABH* with *gang* present was 25.8 months with a 95 per cent confidence interval of ±7.7 months, and the mean the mean sentence for an *ABH* with *vulnerable* present was 26.1 months with a 95 per cent confidence interval of ±6.6 months.

Table A2 *Results from the Random Slopes Model on Robbery**

Variables	Regression coefficient	Standard error
Fixed effects		
Intercept	**6.590**	0.034
Previous conviction	**0.139**	0.048
(Previous convictions)^2	**−0.081**	0.023
First opportunity	**−0.106**	0.025
Age	**−0.229**	0.024
Remorse	**−0.091**	0.026
Group or gang	**0.063**	0.030
Vulnerable victim	**0.073**	0.028
Use of a weapon	**0.494**	0.023
Committed at night	−0.011	0.025
Degree of force	**0.247**	0.026
Random effects		
Intercept	**0.012**	0.005
Gang	**0.018**	0.008
Vulnerable	**0.013**	0.007
Level 1 residuals	**0.342**	0.009
−2 log likelihood	5250.0	
Level 1 units	2879	
Level 2 units	76	

* Estimates that are statistically significant at the 5 per cent level appear in bold.