Article:
Pina-Sanchez, J and Linacre, R (2014) Enhancing Consistency in Sentencing: Exploring the Effects of Guidelines in England and Wales. Journal of Quantitative Criminology, 30 (4). pp. 731-748. ISSN 0748-4518

https://doi.org/10.1007/s10940-014-9221-x

ABSTRACT

Objectives: The development and application of methods to assess consistency in sentencing before and after the 2011 England and Wales assault guideline came into force.

Methods: We use the Crown Court Sentencing Survey to compare the goodness of fit of two regression analyses of sentence length on a set of legal factors before and after the assault guideline came into force. We then monitor the dispersion of residuals from these regressions models across time. Finally, we compare the variance in sentence length of equivalent types of offences using exact matching.

Results: We find that legal factors can explain a greater portion of variability in sentencing after the guideline was implemented. Furthermore, we detect that the unexplained variability in sentencing decreases steadily during 2011, while results from exact matching point to a statistically significant average reduction in the variance of sentence length amongst same types of offences.

Conclusions: We demonstrate the relevance of two new methods that can be used to produce more robust assessments regarding the evolution of consistency in sentencing, even in situations when only observational non-hierarchical data is available. The application of these methods showed an improvement in consistency during 2011 in England and Wales, although this positive effect cannot be conclusively ascribed to the implementation of the new assault guideline.

KEYWORDS

Sentencing guidelines, assault, consistency, exact matching.

1. INTRODUCTION

Sentencing consistency is not only a desirable trait of all legal institutions, but a fundamental principle of justice. It generates transparency and predictability in sentencing practice, and as a result promotes the legitimacy of the criminal justice system and fosters public confidence in sentencing.

In recognition of the importance of improving consistency, in the 1970s some states of the US started implementing systems of sentencing guidelines. In recent years similar systems have also been considered by other common law jurisdictions such as Canada, Scotland, England and Wales, New Zealand, Western Australia, and Northern Ireland. Among these countries, England and Wales has made the greatest progress (Roberts, 2013b); currently it is the only jurisdiction outside the US to have developed formal guidelines that are presumptively binding on courts (Roberts, 2012).

The English and Welsh experience differs substantially, though, from the path taken by some jurisdictions of the US such as the State of Minnesota, or Oregon. These states have opted to achieve uniformity through the implementation of a system of grids that places offences into a limited number of categories and links them to narrow ranges of sentence outcomes. Such a formula was rejected by the Sentencing Commission Working Group (2008) as being inappropriately restrictive and contrary to the traditions of English sentencing. In contrast, the system of guidelines developed in England and Wales aims to assist sentencers in determining the most appropriate sentence outcome by reference to a structured decision-making process that incorporates all of the legal factors that need to be considered in each case.

In this research we study the changes in consistency that might have occurred as a result of the implementation of a new sentencing guideline for offences of assault designed by the Sentencing Council for England and Wales.[1] Given the rejection of a guidelines system based on grids, to assess the level of consistency we will not be able to rely on the standard measures of compliance – i.e. statistics denoting the percentages of sentences following the prescribed outcome (Frase, 2005; Kramer and Ulmer, 2002; Minnesota Sentencing Guidelines Commission, 2010; Oregon Criminal Sentencing Commission, 2003; Scott, 2010; Tonry, 1987; and Ulmer et al., 2011). Instead, we implement two original methods which are able to detect and measure

[1] The new assault guidelines can be downloaded from:
http://sentencingcouncil.judiciary.gov.uk/docs/Assault_definitive_guideline_-_Crown_Court.pdf

changes of consistency in sentencing when more complex sentencing guidelines are in use. In doing so we contribute to both the limited literature analysing the English guidelines experience (Roberts, 2012, and Ashworth and Roberts, 2013), and to the underdeveloped debate on the measurement of consistency in sentencing (Tonry, 1996, Hofer et al., 1999, and Pina-Sánchez and Linacre, 2013).

The paper is structured as follows. The next section describes briefly why the old assault guideline was considered inadequate and the changes which were made in the new guideline. This background will help us understand and explain some of our empirical findings. Section 3 describes the methodological challenges of measuring consistency in sentencing and presents two solutions that we will consider in our analysis. Section 4 introduces the Crown Court Sentencing Survey (CCSS) a new dataset capturing most relevant characteristics of the offence and the offender sentenced in the Crown Courts in 2011. In Section 5 we describe the implementation of the two methods that we use to assess changes in consistency across time, the study of the dispersion of residuals and exact matching. Section 6 concludes with a summary of our results, a discussion of the caveats, and future research paths.

2. CHANGES IN THE NEW ASSAULT GUIDELINE

In October 2010 the Sentencing Council published a consultation for the proposed new guideline on assaults with the aim of increasing consistency in sentencing.[2] This consultation document outlined several major rationales for updating the existing Sentencing Guidelines Council guideline.[3]

First, the sentencing scenarios introduced in the guideline were considered to inadequately represent the factual scenarios being dealt with in court. For example, starting point sentences in the guideline were applicable to first-time offenders, however most assault offenders seen in the Crown Courts have relevant previous convictions. Sentencers also felt that the guidelines overemphasized premeditation, which made them difficult to apply in cases where the attack was spontaneous –as it is often the case for assault offences.

[2] The professional consultation can be downloaded from: http://sentencingcouncil.judiciary.gov.uk/docs/ASSAULT_Professional_web.pdf

[3] The old assault guideline can be downloaded from: http://webarchive.nationalarchives.gov.uk/20100305172947/http://www.sentencing-guidelines.gov.uk/guidelines/council/final.html

Second, a related problem of applicability was identified concerning how the guideline combined the multitude of factors present in each case. A generic list of aggravating and mitigating factors was provided, with the caveat that "not all will be relevant to assault and other offences against the person" (Definitive Assault Guideline, 2008, pp. 28). Dhami (2013) also noted that the effect of the legal factors covered by the guidelines was confusing, since neither the stages when the legal factors need to be considered nor the weights to be applied were clearly defined: *"[…] there was a lack of guidance on* what basis adjustments from the starting point *should be made. […] these omissions left potential for double-* or even triple-counting *of aggravating and mitigating factors"* (Dhami, 2013, pp. 172).

Third, the guideline suffered from organizational and formatting imprecision. As noted by Dhami (2013), the old guideline contained too much text, including information that was not relevant and at times reiterative, while the presentation of topics was disorganized and many terms were undefined or open to subjective interpretation.

The new assault guideline solved these three problems through the introduction of a nine-step process based on the harm and culpability factors present in a case, rather than specific offending scenarios. For example, in the first step of the process a list of guideline factors relating to the principal elements of the offence is provided to the sentencer in order to determine the category of seriousness of the case, and each of these categories is associated with a starting point and a sentencing range. Premeditation remains an important factor but is no longer so dominant, as it now features as a step one factor, making it one feature amongst many which are used to determine the starting point of the sentence outcome. Previous convictions are now included as a step two factor, which can be used to make adjustments to the starting point decided in the previous step, and the starting point now applies to all offenders rather than to a first time offender. This means that the absence of a relevant criminal record suggests the selection of a sentence below the starting point (Hutton, 2013).

Following the public consultation the Sentencing Council published the Assaults Definitive Guideline in March 2011, after which sentencers were given three months to prepare for the official introduction of the guideline, which finally came into force on 13th June 2011. From this date on the new guideline applied to all offences of assault, irrespective of when the offence was committed.

The resolution of the applicability problems identified in the consultation, and the clearer and more structured stepwise process might be expected to lead to more consistent sentencing practice. This was certainly the objective of the Council, as it was expressed by Lord Justice Leveson, chairman of the Council, in an interview with the BBC, *"[…] 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 Basildo*n".*[4] However, until now no other studies have attempted to assess whether the Council was successful in achieving this objective.

3. THE MEASUREMENT OF CONSISTENCY

In spite of the importance of the principle of consistency, and the great political attention that the sentencing guidelines reformation process has channelled, little is known about their actual effectiveness. We argue that to a great extent this is due to the intrinsic difficulty of measuring consistency in sentencing (Casey and Wilson, 1998; and Hofer et al., 1999).[5],[6] Here we review the major methodological issues with measuring consistency, and in so doing we highlight the relevance of the methods that we implement in our analysis.

Consistency in sentencing is commonly understood as the extent to which like cases are treated alike. This can be formalised mathematically by considering the sources of overall variability in sentencing, $var(S)$, to be composed of legitimate variability, $var(L)$, and variability due to inconsistency, $var(I)$:

$$var(S) = var(L) + var(I) \quad (1)$$

In particular, $var(L)$ reflects the extent to which sentences can -and should- vary to reflect the various legal factors defining each case, such as the type of offence, or the presence of aggravating or mitigating factors, whereas $var(I)$ reflects any other factors that have an undue effect on sentencing, such as differences in the mood of the sentencer that could cause harsher or more lenient sentences.

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

5 *"Students of sentencing reform have recognised the need for more and better research to evaluate* how well these reforms have reduced unwarranted disparity*"* (Hofer et al., 1999, pp. 262).

6 *"[…]* such research has been rife with methodological limitations not least of which is the failure to quantify or appropriately define disparity. This calls into question the true level of disparity within the system*"* (Casey and Wilson, 1998, pp. 237).

A simple way of operationalizing the concept of consistency as defined in Eq. (1) is by calculating the variability of custodial sentence lengths amongst offenders convicted of the same offence. Such design was used by Lovegrove (1984), and Walker and Sager (1991) for the study of the High Courts of Australia and the Federal Courts of the US, respectively.

The validity of this approach is, however, questionable. There will be a substantial degree of variation in the circumstances and severity of offences, even within the same offence type. This variation will lead to legitimate differences in sentences between offenders sentenced for the same offence type. Without additional controls for the type of offending behaviour and the characteristics of the offence, it will not be possible to distinguish between legitimate variability in sentences due to relevant legal factors, and variation due to inconsistency between judges. As a result, findings from this methodology will seriously overstate the level of inconsistency.

When hierarchical data is available an alternative measure of consistency can be obtained by taking the mean sentence outcome for different judges or courts and then assessing their dispersion. Tarling (2006) used both this and the previous design to compare dispersion in disposal types amongst offences of burglary in 1974 and 2000, and differences in disposal types between 30 magistrates' courts of England and Wales between the same years. His findings pointed at substantial disparities in the use of disposal type, although those disparities remained at similar levels between 1974 and 2000. Mason et al., (2007) also combined these two approaches to analyse sentence length variability between magistrates' and Crown Courts of England and Wales controlling for different variables such as type of offence or local crime rates, and found significant disparities in sentence length across the 42 Criminal Justice Areas in England and Wales.

The main problem with comparing sentencing between judges stems from the possibility that the offences sentenced by different judges or courts are systematically different. It certainly helps to condition on local crime or unemployment rate as Mason et al. (2007) did, but it is unlikely that these court or local level factors can adequately control for differences in the caseload experienced by different judges.

To circumvent this problem many studies have relied on a natural experiment taking place in many federal courts in the US, where judges working in the same court

are assigned cases at random (Anderson et al., 1998; Hofer et al., 1999; Scott, 2010; Orchard et al., 1997; Scott, 2010; and Waldfogel, 1991). Measures of consistency can be obtained from simple comparisons of average sentence outcomes between judges. The randomisation process ensures that these comparisons are not biased by confounding effects such as differences in the types of offences sentenced by different judges. However, the statistical power of these techniques is limited since the randomisation process only guarantees caseloads to be similar amongst judges across large samples of cases. As such, the method may not be capable of detecting even moderately large differences between judges in sentencing patterns unless the judges are observed and compared over very long time horizons. In addition, this methodology suffers from a problem of external validity since only a few of the US districts following this practice have made their sentencing data available for research, which limits the comparability of results across jurisdictions.

One last methodology that has been recently used in the literature relies on the application of a type of multilevel models known as random slopes models (Anderson and Spohn, 2011, and Pina-Sanchez and Linacre, 2013). This approach can be used to measure the extent to which legal factors have the same effect on sentence outcomes across courts, and has the advantage of being more robust to the problem of insufficient controls.[7] In consequence such models are particularly useful to infer the level of consistency in specific parts of the sentencing process at one particular point in time. However, as a result of their focus on a number of different legal factors, estimates from such models do not provide a single measure of consistency that is adequately generalizable to the whole of the sentencing process.

In this paper (Section 5), we present two original methods that do not need experimental or hierarchical data for their implementation. Instead they rely on access to a dataset that captures information on the type of offence, sentence outcome, and some of the most relevant legal factors present in the case. In our first method, the study of the "dispersion of residuals", we suggest using a linear regression model of custodial sentence length on a set of guideline factors, and analysing the dispersion of residuals from such model across time. Our second method, "exact matching", does not rely in any model specification; instead it involves matching offences within

[7] See Pina-Sánchez and Linacre (2013) for a description of how measures of inconsistency from random slopes are less prone to problems of omitted relevant variables.

similar groups defined by a set of guideline factors, calculating the variability of sentence outcomes within matched groups, and comparing this variability across time.

These two methods are based on the rationale laid out in Equation 1, which conveyed the idea that to measure inconsistency we must be able to control for legitimate variability in sentencing. The absence of comprehensive statistical controls makes this challenging in practice, and may lead to biased measures of consistency (Brantingham et al., 1984, Anderson et al., 1998, Hofer et al., 1999, Pina Sánchez and Linacre, 2013). To circumvent this problem we appeal to a simple observation. We assume that any bias present in the measures of consistency that we obtain is constant through time; thus, although measures of consistency for one particular time might be biased, changes in those measures could be ascribed to actual changes in consistency. But before we present these two new methods we proceed to introduce the new CCSS dataset.

4. DATA

The CCSS is a dataset created by the Sentencing Council of England and Wales covering all offences sentenced at the Crown Courts of England and Wales in 2011. However, more important than its coverage is its unprecedented level of detail. The CCSS covers most factual elements of each case, including characteristics of the offender (such as the number of relevant previous convictions, or the nature of the plea), the offence (e.g. the seriousness level plus all relevant aggravating and mitigating factors) and the sentence imposed (the disposal type or custodial sentence length).[8] Its uniqueness in terms of coverage and detail makes the CCSS probably the best dataset currently available in any jurisdiction to study sentencing matters, and highly suitable for the two methods that we have devised for the study of sentencing consistency. However, the CCSS also has some weaknesses that need to be noted.

Despite the CCSS's aspiration to be a census, in practice, it suffers from a problem of non-response. The response rate across 2011 was 61%, higher than rates

[8] See Roberts (2013a), and the Guide to Crown Court Sentencing Survey Statistics for more information on the CCSS. The latter can be found here, http://sentencingcouncil.judiciary.gov.uk/docs/Guide_to_CCSS_Statistics.pdf

obtained by most national surveys.[9] However, this percentage varied substantially across courts, with the lowest response rate being 20% and the highest 95%. This should make us aware of the possibility of the missing data being non-ignorable, which in addition to reducing precision of our estimates might bias them.[10]

In addition, the implementation of the new assault guideline brought about a change of format in the CCSS questionnaire for offences of assault, which generated some irregularities. First, some judges were still using the old assault form shortly after the new guideline had come into force.[11] So, in order to make the before and after scenarios fully comparable we decided to drop these cases, 596 in total. Second, some of the questions used a different wording in the new format. To avoid inconsistencies we decided to consider only guideline factors 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. All of these 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.[12] The means and standard deviations for all of the variables used in the analysis are included in Appendix 1.

Finally, we limit our analysis to three of the assault offences covered by sentencing guidelines. These are: assaults occasioning actual bodily harm (ABH), grievous bodily harm (GBH), and grievous bodily harm with intent (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 registered; while offences of common assault were discarded, despite being the third most frequent category, because they are a summary only offence, which means

[9] E.g. In 2012 the British Attitudes Survey and Labour Force Survey achieved a response rate of 54% and 48%, respectively.
[10] See Rubin (1987) for a classification of the implications and possible adjustments for the different missing data mechanisms.
[11] This was due to administrative difficulties in ensuring that introduction of new forms into courts coincided with the date the new guideline came into effect.
[12] The CCSS questionnaire also considers a category for ten or more previous convictions, but the sample used here does not capture subjects with that value. These are more common in more recidivistic offences such as theft.

that cases of common assault are usually limited to being sentenced in the magistrates' court.

5. CHANGES IN CONSISTENCY

In this section we present the application of our two methods and discuss the extent to which the results that we obtain can be used as evidence pointing to an increase of consistency in sentencing after the new assault guideline was implemented.

5.1. Dispersion of residuals

We start our analysis with a comparison of two linear models, before and after the guideline came into force, in which the sentence length is specified on the set of guideline factors listed in Section 4. Sentences shorter than a month were discarded because they seemed to correspond to a different sentencing mechanism, giving the distribution of sentences length a bimodal shape, and the natural logarithm was taken of the sentence duration in days in order to adjust for the right skewness of its distribution (Anderson and Spohn, 2011). In addition, we use robust standard errors to take into account the intra-cluster correlation derived from the hierarchical nature of our dataset (sentences grouped within courts).[13,14] Results from the two models are shown in Table 1 below.

Table 1. Results for the Before and After Models*,**

	Before	After
Constant	**5.78**	**5.52**
	(.03)	(.05)
GBH	**.39**	**.55**
	(.02)	(.03)
Intent	**1.51**	**1.74**
	(.03)	(.03)
Prev. convictions	-.02	**.11**
	(.01)	(.02)
First opportunity	**-.09**	**-.08**
	(.02**)**	(.03)

[13] The statistical modelling presented in this paper was carried out in R. For robust SEs we used the sandwich estimator, from the sandwich package.
[14] This correlation could also be exploited to obtain some first insights into disparities between courts in the sentence length imposed. This method is further explored in Anderson and Spohn (2011) and Pina-Sanchez and Linacre (2013).

Remorse	**-.14**	**-.13**
	(.02)	(.03)
Carer	**-.12**	-.16
	(.04)	(.11)
Gang	.03	.02
	(.02)	(.04)
Vulnerable	**.12**	**.18**
	(.03)	(.04)
Public worker	-.03	-.10
	(.05)	(.06)
Sustained	**.21**	**.20**
	(.02)	(.03)
Drugs	**.06**	.01
	(.02)	(.03)
R2	.55	.62
Sample size	2982	1949

*In bold results which are statistically significant for a 95% confidence level.
**Standard errors are included between brackets.

A comparison of the regression coefficients between the two models show effects that can be explained by some of the adjustments made in the new assaults guideline. Higher coefficients for intent and GBH, compared to the least serious assaults of ABH, could be attributed to the new guideline, which according to the Professional Consultation "maintains the availability of the existing sentences for the most serious offenders while ensuring that sentencing for less serious offences is proportionate" (Professional Consultation, 2011, pp. 5). Similarly, the change of the coefficient for previous convictions, from being non-significant in the before model to the positive and statistically significant effect, could be expected from the substitution of the unrealistic scenarios in the new guideline.

Regarding consistency in sentencing, we can see that the R^2 has increased from 55% to 62%.[15] That is, the percentage of variance in sentence lengths that can be explained by the guideline factors included in the model has increased. However, as we mentioned in Section 3, it is practically impossible to control for all relevant legal factors that explain differences between cases, which might bias measures of consistency such as the ones we obtained from these R^2s. In addition, the R2 statistic, does not adequately allow us to distinguish changes in legitimate variability, var(L), from changes in inconsistency, var(I) because it is not possible to tell whether the change originated from the numerator or denominator of the statistic.

[15] This change of R^2 in addition to the previous changes for the coefficients of GBH, intent, and previous convictions, were found to be statistically significant using a Chow test, with a p-value > .001.

The CCSS helps to minimise bias that may result from inadequate controls. However, even if it had been possible to control for all of them, we might still find other sources of unexplained variability stemming from issues of data quality (missing data or measurement error), or model misspecifications (e.g. omission of quadratic or interaction terms). Hence, it seems almost impossible to quantify accurately the absolute level of inconsistency in sentencing at any point in time when relying on observational data.

This is not necessarily a problem since our interest does not lie in the estimation of a specific coefficient of consistency but in the detection of a relative change between periods. If we can assume that the effect of the confounding elements of unexplained variance (i.e., those stemming from issues of data quality and model misspecifications) remain constant across time we could ascribe changes in the unexplained variance across time to actual changes in consistency, regardless of its absolute level at any particular moment.

We argue that these assumptions are not unrealistic. Both the models presented in Table 1 use the same set of guideline factors, hence the unexplained variability due to the effect of omitted relevant variables should be fairly similar across time. Moreover, since the specification of the models remain constant, we can also discard the hypothesis that changes in unexplained variability are due to problems of misspecification. Issues of data quality can be more problematic as the implementation of the new guideline was accompanied with a change in the CCSS forms. This impact was limited, though, by choosing to use only factors where the wording did not change. In addition, response rates between May and July (the months before and after the guideline came into force) remained practically unchanged at 58.3% and 59.5% respectively.

We can take the statistically significant increase of R^2 from 55% to 62% as an a priori indicator that consistency may have been higher in the period after the new guideline came into force. However, this result only offers us a discrete measure of change, and could be biased by any changes in legitimate variability in sentencing that arose from the new guideline. To see how the change took place across time we proceed now to study the dispersion of the residuals during 2011. Specifically, we use the variance of residuals from offences sentenced in the same week, which can be expressed as,

$$\frac{\sum_{i=1}^{N_j}\left(\mu_{ijl}-\bar{\mu}_l\right)^2}{N_j-1} \qquad (2)$$

where μ represents the residual term of the regression models, i is a subscript identifying the offence so $i = 1,2,...,N$; the formula is applied for all j, which captures the weeks of the year so $j = 1,2,...,52$, l identifies the two regression models, so $l = 1,2$, and $\bar{\mu}$ represents the mean of residuals for each of those models, which by assumption in linear regression models is equal -or very close- to zero.

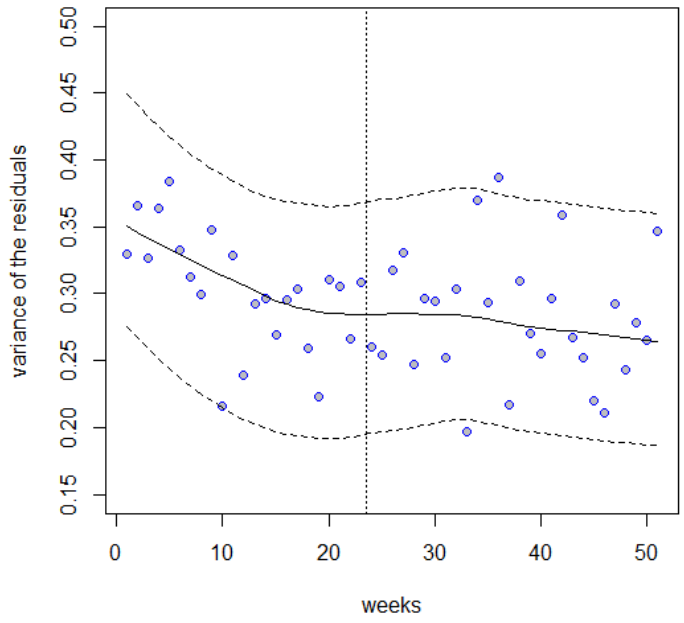
We decided to group residuals by weeks because, given our sample size, it is the time unit that offers the best compromise in terms of making the scale as continuous as possible without incurring problems of volatility derived from small sample sizes and the resultant high sampling errors. In our sample, the lowest number of observations per week is 59 (week 17), which makes estimates of the variance robust enough.

To depict graphically the pattern of these weekly residual variances across 2011, we fit them using a Lowess curve. Lowess is a nonparametric regression method proposed by Cleveland (1979) that stands for local weighted regression.[16] This method allows us to observe changes in the weekly variance patterns without having to rely on a restrictive functional form (e.g. linear, or quadratic).

Results are presented in Figure 1 below, where the weekly variances of residuals across 2011 are illustrated in a scatterplot together with the Lowess curve, its 95% confidence interval, and a vertical dashed line representing the time when the new guideline came into force.

Figure 1. Dispersion of Residuals*

[16] See also Keele (2008) for an excellent review of semi and non-parametric regression methods.



*The Lowess curve is represented by a continuous curve and the confidence interval by the dashed lines around it.

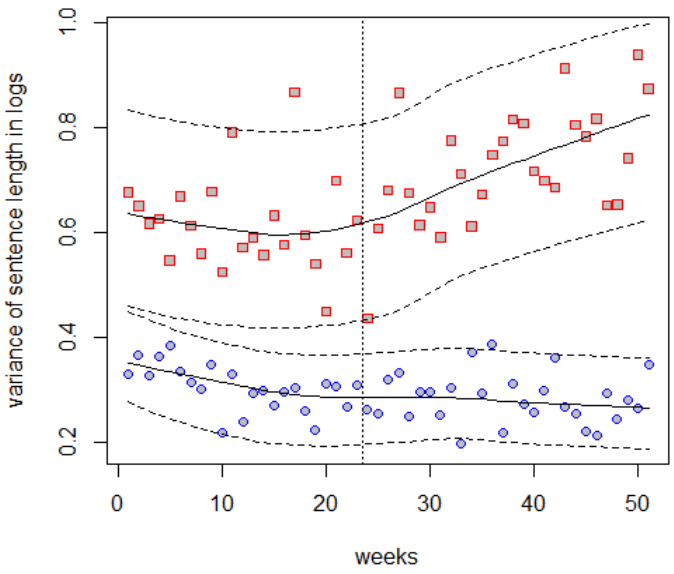
We can observe a monotonic decrease in dispersion of residual variability by week throughout the entire year, which can be understood as a decrease in the sentencing variability observed after controlling for legitimate legal factors, i.e. a decrease in var(I). However, the reduction does not correspond closely to week 23, the week in which the new assault guideline came into effect. In fact, the fastest reduction occurred before that, casting doubt over whether the increase in consistency happening across 2011 was caused by the new assaults guideline.

One possibility is that we are observing an anticipatory effect consequence of the familiarisation and training process that judges went through during consultation and after the guideline was published. After all, the publication of the professional consultation (October 2010) included a draft guideline, which was similar to the definitive guideline that was eventually published the 16th March 2011. In addition, it is interesting to note that the smooth transition between the weeks immediately before and after the guideline came into force refutes the hypothesis that the new form brought a substantial change in terms of data quality, in the form of either measurement error or non-ignorable missing data. If that would have the case we would have expected to see an abrupt change exactly at that time point.

It might also be argued that the apparent increase in consistency during 2011 could be the result of sentences for assaults simply becoming more uniform. In

Equation 1, this would correspond to a reduction in legitimate variability, rather than any change in inconsistency. Uniformity implies a lack of nuance in sentencing decisions, which results in dissimilar offences being treated alike, hence causing a deterioration in both proportionality and consistency. This is exactly why the Sentencing Commission Working Group rejected the application of a US-style grid system in England and Wales. In order to improve the robustness of our analysis we proceed to explore this possibility by plotting the weekly variability in overall sentence lengths for cases of assault across 2011 in Figure 2, and to facilitate comparison we have also superimposed results from Figure 1. In terms of Equation 1, the overall variability var(S) corresponds to the upper line, the lower line approximates inconsistency, var(I), and $var(L)$ can be thought of as the difference between the two.

Figure 2. Dispersion of Sentence Length vs Guidelines Residuals*



*The upper continuous line represents dispersion in sentence length, and the lower continuous line dispersion in the residuals. The dashed lines represent the 95% confidence bands around the Lowess curves for both the residuals and sentence length.

We see that the variability of sentence length increases after the guideline came into force, which refutes the hypothesis that the observed decrease in variance of the residuals is simply a result of increased uniformity. In fact, at a time where overall variation in sentencing is on the increase, variation in sentencing amongst cases with similar legal factors is decreasing. Hence, we can deduce that the variability due to legitimate reasons has increased, which could also be interpreted as an improvement in proportionality during 2011.

5.2. Exact matching

Our second approach to investigate changes in consistency is both conceptually and methodologically simpler. It can be implemented in three steps using descriptive statistics: 1) offences are matched into groups with the same mix of legal factors, before and after the new guideline came into force; 2) the variance in sentence length by group is calculated; and 3) the difference of variances between same groups before and after is interpreted as a measure of the change in consistency.

This approach could be considered an extension of the simple study of the variability within same types of offences mentioned in Section 3, with the difference that now we control simultaneously for most of the relevant guideline factors, including type of offence. In addition, unlike the dispersion of residuals, this methodology imposes no restriction on the functional form of the relationship between legal factors and sentence length, which delivers a more straightforward operationalisation of consistency.[17]

Despite the level of precision that the CCSS affords us in generating homogeneous groups of offences, exact matching suffers from the same main limitation as the dispersion of residuals. We can only match cases based on the guideline factors to which we have access, and so matched cases may differ in ways which are unobserved. Hence, some variability would be expected in the sentence outcomes amongst matched cases, even if sentencing was perfectly consistent. As before, we circumvent this problem by invoking assumptions that allow us to make temporal comparisons. Specifically, we assume that sentence variability due to the unobserved guideline factors remained constant in the before and after periods.

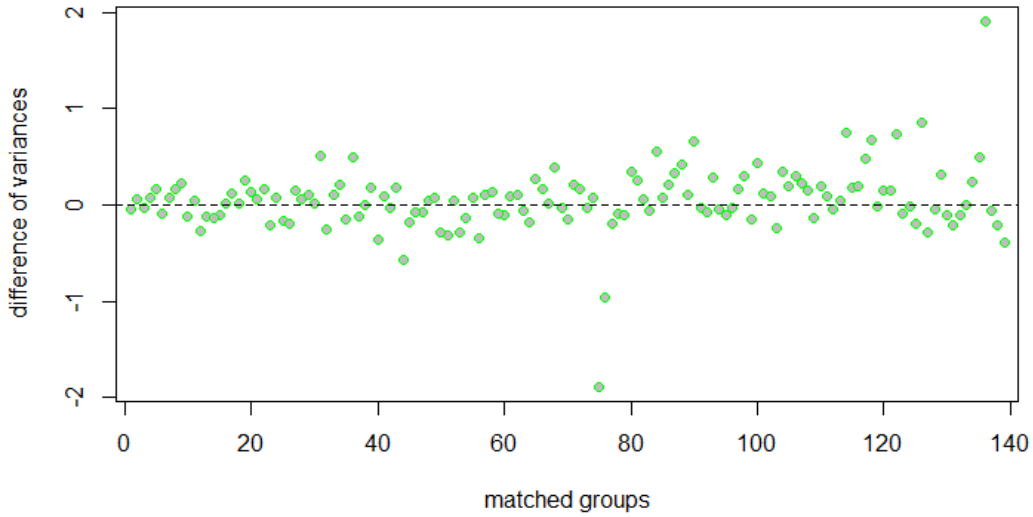
After matching offences into groups we excluded those that have fewer than two observations either before or after because of the impossibility of calculating their variance. Under this constraint we achieved a match of 70% cases from our sample (3467 cases in total), classified in 139 different groups.

We found that in 57% of these matched groups, our estimate of variance was lower for the period after the new guideline came into force. Figure 3 shows the

[17] Under the dispersion of residuals methodology a linear regression model has to be formulated to specify the mathematical relationship between legal factors and sentence length. Although this model allows considerable flexibility in specifying functional form, it can only ever provide an approximation to the true empirical relationship.

individual differences between the before and after variances; with each of the 139 groups plotted in the x-axis and ordered in descending frequency from left to right.[18] The fact that more than half of the groups show a positive change (lie above $y = 0$), indicates that for a small majority of groups of offences consistency has improved. This is especially noticeable for the last groups (roughly those beyond group 80), which are the least common and therefore the ones that might be expected to be more complex to sentence.

Figure 3. Variance Reduction in Matched Groups



However, the percentage of cases showing a reduced dispersion using these one-to-one comparisons can be a misleading estimate of the overall change in consistency. Groups' sample size vary substantially (ranging from 4 to 230 observations), hence reduced dispersion in the majority of groups does not necessarily mean reduced dispersion in the majority of the offences sentenced. For an estimate to properly account for the relative frequency of different types of cases, a weighting adjustment needs to be implemented so group differences in sample size are taken into account.

Equation 3 shows the statistic that we have designed to achieve this goal,

$$\frac{\sum_k \frac{n_k}{N} s^2_{A,k}}{\sum_k \frac{n_k}{N} s^2_{B,k}} \qquad (3)$$

[18] The number of matched groups and their sample size are shown in Appendix 2.

Sample variances are represented by the s^2 terms which are calculated for before and after groups, denoted by the subscripts B and A, with the subscript k indexing the different groups, so $k = 1,2,...,K$ (in this case K being 139). Weights are defined as the sampling fraction of each group, that is the ratio of the sample size of one group, n_k, over the total sample size, N. The summation term over k means that the individual weighted differences in sample variance are summed together to produce an overall measure of the change in variance.

For our sample, the statistic yields a value of .923, which could be interpreted as a 7.7% reduction in the weighted variance of the groups. To determine whether this reduction is statistically significant we would normally use an F-test. However, because of the weighting adjustment and the summation of different groups, the test statistic that could be derived from the ratio presented in Eq. (3) does not follow a standard F distribution. To circumvent this problem we decided to produce the sampling distribution of the test statistics using Monte Carlo simulations –described in Appendix 4- which indicated that the reduction of the groups weighted variance is statistically significant at a five percent significance level.

We can also use results from exact matching to explore within group changes in variability.[19] Standard F tests of equality variance could be used for this purpose. Unfortunately, due to the small sample size of groups we only found a statistically significant change in the variance for one matched group. This was the group of cases of GBH with intent with 1-3 previous convictions and no aggravating or mitigating factors, where we observe a 12.5% reduction in their variance.

Therefore, in spite of having demonstrated that overall consistency has increased, we cannot assess whether individual group-by-group comparisons are significantly more or less dispersed in the majority of those groups because of their small sample size. This is a paradoxical feature of exact matching. If we want to obtain a more precise estimate of changes in overall consistency the more variables we have to match on the better, although at the cost of ending up with only a few cases per group. If the interest lies in observing changes in consistency for specific

[19] The characteristics used to define the 10 matches, and their before and after variances are presented in Appendix 3.

groups, we would recommend using a limited number of controls. Here we have opted for the former.

6. CONCLUSION

Sentencing guidelines have been introduced gradually in the jurisdiction of England and Wales since 2003, with the aim of increasing sentencing consistency. However, due to methodological difficulties involving the measurement of consistency (Casey and Wilson, 1998; Hofer et al., 1999; and Pina-Sánchez and Linacre, 2013) and the lack of adequate data, the empirical explorations of the success of sentencing guidelines have been practically non-existent (Roberts, 2012, and Ashworth and Roberts, 2013).

In this paper we have used the recently published CCSS to assess changes in consistency in England and Wales before and after the 2011 assault guideline came into force. This has been achieved through development of two original methods, the dispersion of residuals and exact matching, hence, contributing both to the methodological debate on the measurement of consistency in sentencing, and to the understanding of the effect of the sentencing guidelines in England and Wales.

In the application of exact matching we defined 139 specific groups of assault offences by conditioning on eleven of the most relevant guideline factors, and observed that in 2011, the variability of sentence length within matched groups decreased by a statistically significant 7.7% after the new guideline came into force. In the application of the dispersion of residuals we studied how this increases in consistency developed throughout the year by monitoring the unexplained variability derived from regression models that specify sentence length on the same eleven guideline factors. Here, we found that increases in consistency were monotonic, but did not correspond in any obvious way to the date the new assaults guideline came into effect. We were therefore unable to prove a definitive causal link between the new assaults guideline and the observed improvements in consistency.

Although no causal link could be established, the ability to analyse this link demonstrates the strengths of combing exact matching and the dispersion of residuals. In particular, the former offers a direct –and statistically testable- measure for the change of consistency between two periods; while the latter allows the visualisation of

how those changes took place through time, which helps to investigate whether they respond to known events such as the introduction of a sentencing guideline.

In addition, through the application of these two methods we have also shed light on other aspects of the new sentencing guidelines. Our regression analyses showed that the effects of GBH and GBH with intent on sentence length –compared to the reference case of ABH- were stronger following the implementation of the guideline. These effects align with the Sentencing Council's aim of increasing the proportionality of sentencing for assaults, which was further underlined when we observed that the overall dispersion of sentence lengths for assaults increased after the new guideline was introduced. Similarly, we also found that previous convictions -an area of the old guideline that had been heavily criticised- had a more predictable effect on sentencing after the new guideline came into force.

The creation of guidelines in England and Wales is an ongoing process and to inform that process more empirical analyses will be necessary. In this respect, the new CCSS represents an extremely valuable resource that has not yet been discovered by most of academics studying the jurisdiction of England and Wales. In addition to its remarkable coverage and depth of detail, the CCSS will release new waves of data for the years following 2011. This will add a greater longitudinal dimension to the dataset, which would markedly increase the research possibilities of the dataset. For example, the bigger sample sizes afforded by new releases of data could be used to replicate the exact matching design presented here. This would allow observing the dispersion within matched groups with greater accuracy, which could be used to identify the specific offences that are most problematic to sentence, and this way obtain a more detail insight into the functioning of sentencing guidelines. The study of the result of guidelines released in 2012 and beyond is also likely to provide a more robust answer to the question of whether new sentencing guidelines cause increases in consistency. A longer time series would allow anticipatory effects to be pinned down with greater accuracy, and comparison between offences would allow the researcher to distinguish more clearly between general patterns in sentencing and the effects of specific sentencing guidelines.

REFERENCES

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.

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--393.

Ashworth, A., and Roberts, J. (2013). The Origins and Nature of the Sentencing Guidelines in England and Wales. In Ashworth, A. and Roberts, J. (eds.), Sentencing Guidelines: Exploring the English Model. Oxford University Press, Oxford, pp. 1--12.

Brantingham, P. (1985). Sentencing Disparity: An Analysis of Judicial Consistency. Journal of Quantitative Criminology. 1: 281--305.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association. 74: 829--836.

Dhami, M. K. (2013). A "Decision Science" Perspective on the Old and New Sentencing Guidelines in England and Wales. In Ashworth, A. and Roberts, J. (eds.), Sentencing Guidelines: Exploring the English Model, Oxford University Press, Oxford, pp. 165--181.

Frase, R. S. (2005). Sentencing Guidelines in Minnesota, 1978-2003. Crime and Justice. 32: 131--219.

Hofer, P., Blackwell, K., and Ruback, R. B. (1999). The Effect of the Federal Sentencing Guidelines on Inter-Judge Sentencing Disparity. The Journal of Criminal Law and Criminology. 90: 239--321.

Hutton, N. (2013). The Definitive Guideline on Assault Offences Issued by the Sentencing Council for England and Wales. In Ashworth, A. and Roberts, J. (eds.), Sentencing Guidelines: Exploring the English Model, Oxford University Press, Oxford, pp.86--103.

Keele, L. (2008). Semiparametric Regression for the Social Sciences, John Wiley and Sons, Chichester.

Kramer, J. and Ulmer, J. (2002). Downward Departures for Serious Violent Offenders: Local Court "Corrections" to Pennsylvania's Sentencing Guidelines. Criminology. 40: 807--932.

Lovegrove, A. (1984). An Empirical Study of Sentencing Disparity Among Judges in an Australian Criminal Court. International Review of Applied Psychology. 33: 161--176.

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

Minnesota Sentencing Guidelines Commission. (2012). Sentencing Practices: Controlled Substance Offenses Sentenced in 2010. (available at: http://www.leg.state.mn.us/docs/2013/other/130860.pdf)

Orchard, N., Howlett, J., Davies, E., Pearson, G., Payne, A. (1997). Does Inter Judge Disparity Really Matter? An Analysis of the Effects of Sentencing Reforms in Three Federal District Courts. International Review of Law and Economics. 17: 337-366.

Oregon Criminal Sentencing Commission. (2003). Sentencing Practices: Summary Statistics for Felony Offenders Sentenced in 2001. (available at: http://www.oregon.gov/CJC/docs/SG01v2.pdf)

Pina-Sanchez, J, and Linacre, R. (2013). Sentence Consistency in England and Wales: Evidence from the Crown court Sentencing Survey, British Journal of Criminology. 53: 1118--1138.

Roberts, J. (2011). Sentencing Guidelines and Judicial Discretion: Evolution of the Duty of Courts to Comply in England and Wales. British Journal of Criminology. 51: 997--1013.

Roberts, J. (2012). Structured Sentencing: Lessons from England and Wales for Common Law Jurisdictions. Punishment & Society. 14: 267--288.

Roberts, J. (2013a). Complying with Sentencing Guidelines: Latest Findings from the Crown Court Sentencing Survey. In Ashworth, A. and Roberts, J. (eds.), Sentencing Guidelines: Exploring the English Model, Oxford University Press, Oxford, pp. 104--121.

Roberts, J. (2013b) Sentencing Guidelines in England and Wales: Recent Developments and Emerging Issues. Law and Contemporary Problems, 76: 1-26.

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

Sentencing Commission Working Group. (2008). Sentencing Guidelines in England and Wales: An Evolutionary Approach, SCWG, London.

Sentencing Council of England and Wales. (2011). Assault: Definitive Guideline. (available at: http://sentencingcouncil.judiciary.gov.uk/docs/Assault_definitive_guideline_-_Crown_Court.pdf)

Sentencing Guidelines Council of England and Wales. (2011). Guide to Crown Court Sentencing Survey Statistics. (available at: http://sentencingcouncil.judiciary.gov.uk/docs/Guide_to_CCSS_Statistics.pdf)

Sentencing Guidelines Council of England and Wales. (2011). Resource Assessment – Assault Guideline. (available at: http://sentencingcouncil.judiciary.gov.uk/docs/Assault_definitive_guideline_-_Resource_assessment.pdf)

Sentencing Guidelines Council of England and Wales. (2008). Assault and other offences against the person. Definitive Guideline. (available at http://webarchive.nationalarchives.gov.uk/20100305172947/http://www.sentencing-guidelines.gov.uk/docs/assault-against-the%20person.pdf)

Sentencing Guidelines Council of England and Wales. (2011). Assault Guideline - Professional Consultation (available at: http://sentencingcouncil.judiciary.gov.uk/docs/ASSAULT_Professional_web.pdf)

Scott, R. (2010). Inter-Judge Sentencing Disparity After Booker: A First Look, Express. (available at http://works.bepress.com/ryan_scott/2/)

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

Tonry, M. H. (1987). Sentencing Reform Impacts, National Institute of Justice, Rockville.

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.

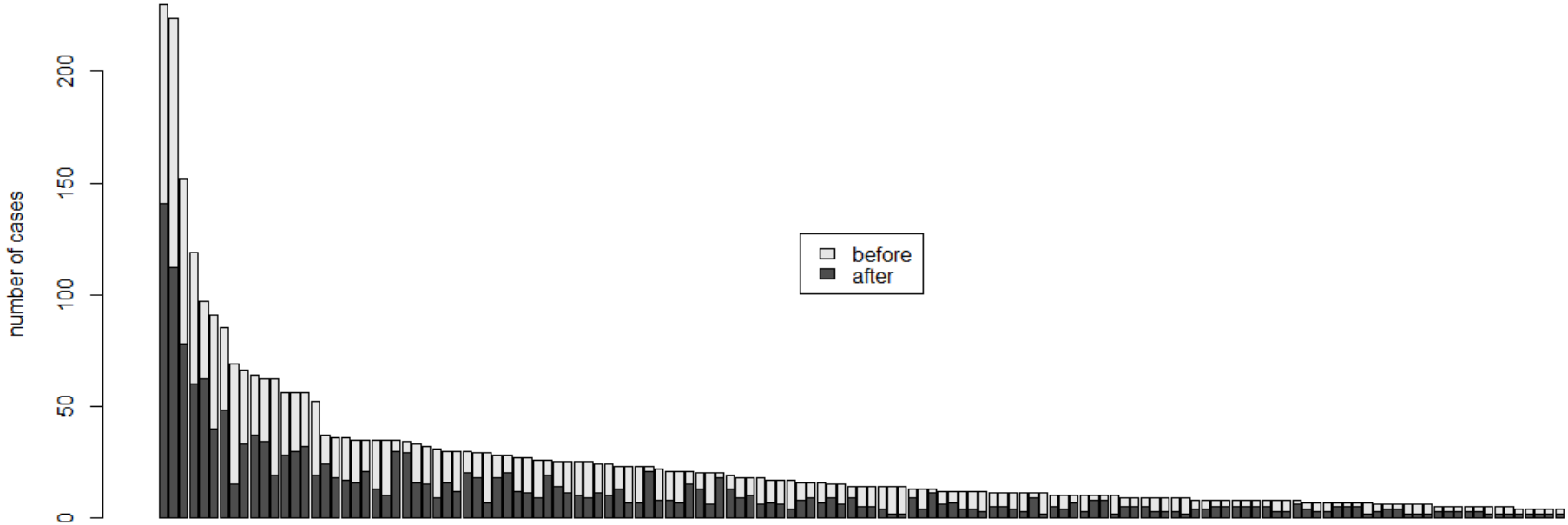
Waldfogel, J. (1991). Aggregate Inter-Judge Disparity in Federal Sentencing: Evidence from Three Districts. Federal Sentencing Reporter. 4: 151--154.

Walker, T., and Sager, T. (1991). Are the Federal Sentencing Guidelines Meeting Congressional Goals?: An Empirical and Case Law Analysis. Emory Law Journal. 40: 393-444.

APPENDIX 1. DESCRIPTIVE STATISTICS OF THE VARIABLES USED

Variable	Mean: Before	Mean: After	Std dev: Before	Std dev: After
Log sentence length	6.11	6.25	.78	.85
Previous convictions	2.04	1.55	.71	.50
First opportunity	.22	.30	.42	.46
Remorse	.34	.29	.47	.45
Carer	.05	.02	.22	.14
Gang	.20	.08	.40	.27
Vulnerable	.15	.10	.36	.30
Public officer	.04	.04	.21	.20
Sustained	.30	.32	.46	.47
Drugs	.35	.35	.48	.48
ABH	.53	.48	.50	.50
GBH	.31	.31	.46	.46
GBH with intent	.15	.21	.36	.41

APPENDIX 2. MATCHED GROUPS ORDERED BY SAMPLE SIZE



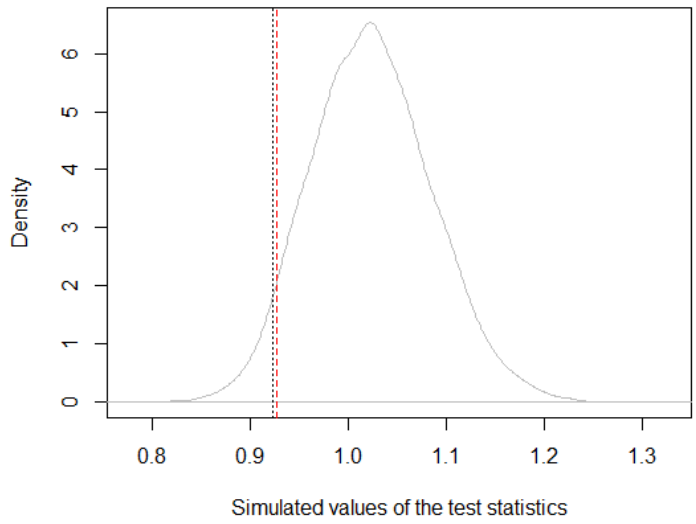
APPENDIX 3. TOP TEN LARGEST MATCHED GROUPS

Type of Offence	Previous Convictions	Aggravating / Mitigating	Group Size: Before	Group Size: After	Variance: Before	Variance: After	Variance Difference
ABH	0	-	112	112	.37	.42	-.05
ABH	1-3	-	141	89	.36	.30	.05
GBH	1-3	-	78	74	.21	.24	-.03
GBH	0	-	60	59	.32	.24	.08
ABH	1-3	sustained	40	51	.50	.34	.16
GBH	1-3	drugs	48	37	.23	.32	-.08
ABH	1-3	drugs	62	35	.28	.20	.08
Intent	1-3	-	33	33	.30	.14	.16
ABH	1-3	first op.	34	28	.55	.32	.23
GBH	1-3	remorse	28	28	.13	.24	-.11

APPENDIX 4. MONTE CARLO SIMULATION

Our approach to determine whether the change in the groups weighted variance is statistically significant involved four steps: 1) the simulation of a new dataset of sentence lengths with the number of groups and cases per group corresponding to our CCSS data. Simulations are drawn from different normal distributions for each group with mean of zero and variance determined by $s^2_{B,k}$; 2) the statistic presented in Eq. (3) is calculated using the simulated dataset; 3) steps 1 and 2 are iterated 10,000 times so a sampling distribution of test statistics can be constructed; 4) the estimate of from Eq. (3) using the real data is compared to the 5th percentile of the simulated sampling distribution, and if the former is smaller we can say that the reduction in the variance after the new guideline came into force is statistically significant.

Figure A1. Simulated sampling distribution



In Figure A1 above we show the simulated sampling distribution together with two vertical lines: a red line signalling the 5th percentile at 0.927 and a black line indicating the value of our test statistics at 0.923. Since the latter is smaller than the former we can claim that the observed reduction of the aggregated group variance after the new guideline came into force is statistically significant.