



# Developing a Complete Sentence Severity Scale using Extended Goodman RC models

Stephanie Wallace<sup>1</sup> · Brian Francis<sup>2</sup>

Accepted: 3 August 2024  
© The Author(s) 2024

## Abstract

**Purpose** The aim of this paper is to construct a single sentence severity scale incorporating the full range of custodial and non-custodial sentences meted out by the courts. Such a scale would allow us to measure and rank the severity of sentences, relative to other sentences.

**Methods** We use disaggregated individual level sentencing data to model the association between offenses and their associated sentences using the Goodman Row Column (RC) Association Model. We then extend this model to control for three legal factors; conviction history, offense plea, and number of offenses, to produce a series of standardised scores. We use linear interpolation and extrapolation to convert the scores to equivalent days in custody.

**Results** The scores from the model enable the sentences to be ranked in order of severity; longer custodial sentences dominate at the severe end whilst non-custodial sentences congregate towards the lower end. In the middle of the scale, non-custodial and shorter custodial sentences interweave. We then demonstrate one use of the scale by applying it to Crown Court data, illustrating change in sentencing severity over time.

**Conclusions** The Goodman RC Association Model provides a suitable methodology for scoring sentence severity on a single scale. The study found that by extending the model, we were also able to control for three legal factors. The sentence severity scale, as a research tool is specific to England and Wales but the method is universal and can be applied in any jurisdiction where the relevant data is available.

**Keywords** Sentencing · Sentence severity scale · Extended Goodman RC association model · Log multiplicative model

---

✉ Stephanie Wallace  
s.j.wallace1@lancaster.ac.uk

<sup>1</sup> Law School, Lancaster University, Lancaster LA1 4YN, UK

<sup>2</sup> School of Mathematical Sciences, Lancaster University, Lancaster LA1 4YF, UK

## Introduction

Criminology has had a long interest in measuring crime seriousness and ranking **crimes** in order (Ramchand et al. 2009). Sellin & Wolfgang's (1964) classic study led the way for measuring and scaling delinquency, and used this scale to weight criminal offenses. Since then a plethora of studies have adopted varying methodologies (for example Nye and Short (1957); Wolfgang et al. (1985); Blumstein (1974); Cohen (1988); Francis et al. (2001); Osgood et al. (2002)) to create similar measures. Traditionally there has been less emphasis on the measurement of sentence severity (Leclerc and Tremblay 2016) or penal metrics (Tremblay 1988). However, in recent years, there has been a renewed level of interest in both scaling crimes (such as the Cambridge Harm Index (Sherman et al. 2016), the UK Crime Severity Index (Office for National Statistics 2016), and the Swedish Crime Harm Index (Kärrholm et al. 2020)), and scaling and measuring sentence severity (Pina-Sánchez et al. (2019); Pina-Sánchez and Gosling (2022); Hartel et al. (2023)). More recently, the potential of AI systems for determining sentences has also generated interest in sentencing severity (Douglas (2024); Ryberg (2024)).

The concept of 'sentence severity' is relatively straightforward to understand when we consider custodial sentences. A custodial sentence of 1 year would always be considered as more severe than a custodial sentence of 6 months, by the nature of the time incarcerated. The concept of sentence severity can be extended to encompass all disposal types and sentences but it becomes more problematic to measure, partly due to their intangible nature (Francis et al. 2001). The main difficulty in measuring sentence severity lies in the different types of disposals and nature of the sentences available to the courts (von Hirsch 1993; Gormley et al. 2022; Pina-Sánchez and Gosling 2022), and the different punitive elements make it difficult to compare. Certain disposals, such as custodial sentences, can be measured in duration by the length of sentence. Fines, measured in currency, can also be ranked (albeit sometimes means tested), but it becomes more difficult to measure sentence severity when we try to compare sentences with different punitive elements. This is the case in England and Wales, where community orders,<sup>1</sup> suspended sentences,<sup>2</sup> and discharges<sup>3</sup> are also thrown into the mix. This also touches on an associated problem of alignment, for example, how do we compare the severity of non-custodial sentences with the severity of custodial sentences? This is the challenge faced by 'penal metric theor[ists]' (Tremblay 1988).

The aim of this paper is therefore to develop a method that can build a single scale of sentence severity encompassing the full range of sentences apportioned by the courts in England and Wales, and which could be applied to any jurisdiction with access to the relevant data. We believe the creation of this type of scale is important for three reasons.

<sup>1</sup> This sentence combines punishment with activities carried out in the community. The offender will receive one or more community order requirements, such as an unpaid work requirement, a curfew requirement, and/or a drug rehabilitation requirement.

<sup>2</sup> When a court imposes a short custodial sentence, the court can opt to suspend that sentence, which means the offender will spend the entirety of their sentence in the community, but are subject to one or more requirements, similar to those on a community order. If the offender does not comply with the requirements, or commits another offense during this time, the offender will likely serve the original custodial sentence in addition to the new sentence for the subsequent offense.

<sup>3</sup> Absolute discharges are for the least serious sentences in which the offender will receive no punishment other than a criminal record. A conditional discharge, will result in a criminal record and no further punishment but if the offender commits another offense within the specified time, they will be sentenced for the first and new offense.

Firstly, it will allow us to measure and rank the severity of sentences relative to other sentences. Secondly, and very much related to the first point, it will allow us to explore the notion of sentence *interchangeability* (Tonry 1994). As the prison population continues to rise (Ministry of Justice 2023), alternative sentences which demonstrate equivalent severity could be considered a viable option in place of a custodial sentence. Thirdly, the existence of such a scale will allow more nuanced research to be carried out on disparity and discrimination in sentencing outcomes, particularly for minority and other groups (see Wallace (2015)).

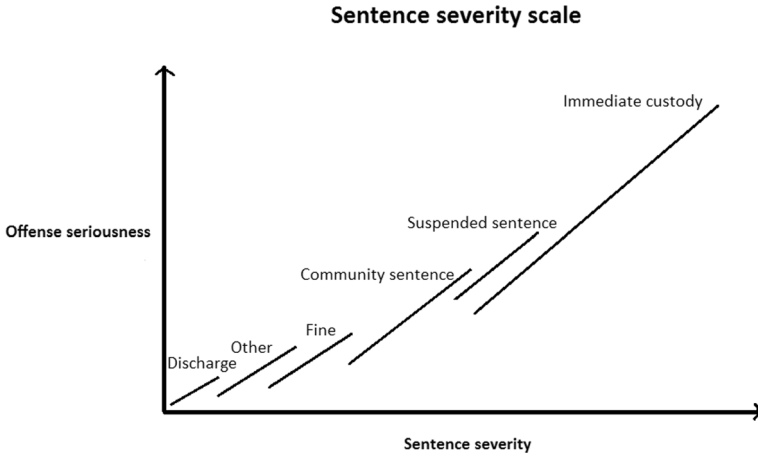
## Sentence Severity and Proportionality

The notion of such a scale derives from the retributivist school of thought, and central to this is the principle of proportionality. Proportionality permits that sentence severity should be proportionate to the seriousness of the offense (von Hirsch 1993). Indeed in practice, many jurisdictions incorporate proportionality into their sentencing framework. A number of US states have employed sentencing grids, which use a complex system to determine the sentencing outcome. For example, the State of Florida scores according to type of offense, any additional offenses brought to the court, victim injury, prior criminal record and a number of other aggravating factors such as motor vehicle theft, gang offense and the use of a firearm. The scale appears to have been determined by expert judgement, and provides sentencing guidelines for length of custodial sentence for high scores, and a choice of sentencing options for low scores (such as county jail or probation).

In England and Wales, in contrast, no such scoring exists. Instead, s.57 of the Sentencing Code states the court must have regard for the following purposes of sentencing:

- the punishment of offenders
- the reduction of crime (including its reduction by deterrence)
- the reform and rehabilitation of offenders
- the protection of the public
- the making of reparation by offenders to persons affected by their offense

While the Code provides no hierarchy or priority in relation to the purposes of sentencing, it does state the court must also follow the relevant sentencing guidelines to ensure a consistent approach to sentencing and to determine an appropriate sentence. The Sentencing Council, an independent non-departmental public body of the Ministry of Justice, responsible for the production of the sentencing guidelines, provide a step-by-step process for sentencers to follow. The initial step is to determine the offense category whereby the court must first assess the seriousness of the offense (s.63) based on the culpability of the offender and harm caused, intended to cause or might foreseeably have caused, by the offense (Sentencing Council 2019). This provides a starting-point sentence based on the categorisation of offense seriousness, as well as a range of possible sentences to take into account the varying circumstances of the case. This starting point can be considered the fulcrum, which is equidistant from the offense and sentencing outcome. Judges must then work through the additional steps outlined by the guidelines, and includes the consideration of possible mitigating and aggravating factors that will decrease or increase the seriousness of the offense. We can consider this as a weight (or effort), and the sentencing outcome (the load) acting as a counterbalance.



**Fig. 1** Sentence severity scale diagram illustrating proportionality

The Sentencing Code (and guidelines) make further reference to the principle of proportionality in relation to the threshold for custodial sentencing (s.230(2)), whereby it should only be used when no other sentence can be justified. If a custodial sentence is the only feasible option, then the length of that sentence, should be for the shortest term possible, to commensurate the seriousness of the offense (s.231(2)). Similarly, this applies to the exercise of powers in relation to imposing a community order (s.204(2)) and fine (s.125(2)). Thereby judges in England and Wales must appropriately balance these elements alongside the other purposes of sentencing to produce a fair and proportionate sentence.

The basic precept of justice in many jurisdictions is to ‘treat like cases alike and different cases differently’ (Walker (2010)<sup>4</sup>). The term ‘alike’ is often taken to involve parity of sentencing in whatever form and one way to achieve this may be to adopt the principle of proportionality. The principle of proportionality requires that sentences are determined by the seriousness of the offense, and therefore in cases where the offender has committed the same gravity of offense should be treated with like severity. Although, as von Hirsch (1993) notes, this does not necessarily mean the same sentence. The diagram in Fig. 1 illustrates this notion of proportionality, whereby sentences are ranked in order of severity, based on the corresponding [ranked] offense seriousness. As the seriousness of an offense increases (y-axis), sentence severity passes through the various thresholds, consistent with the thresholds stated in the Sentencing Code until a sentence that is proportionate to the offense is reached (x-axis). A linear relationship between offense seriousness and sentence severity is assumed, in line with von Hirsch’s (1993) notion of ordinal proportionality. The spacing on the scale then represents the relative severity of each sentence in conjunction with the other sentences in the scale (ibid).

Morris and Tonry (1990) refer to this type of scale as a “continuum of sanctions”, and by arranging sentences in this way, it is possible to determine how sentences relate to each other and in relation to imprisonment. When sentences are thus ordered by their severity, sentences can overlap in terms of their comparative severity. When von Hirsch (1993)

<sup>4</sup> Referring to a Privy Council decision per Lord Hoffmann, (Matadeen v Pointu [1999] 1 AC 98, 109).

refers to parity, he refers to sentences that have the same perceived level of onerousness and not necessarily the same sentence. When two sentences are viewed by those imposing the sentence as comparatively severe, von Hirsch (1993) refers to this as these sentences having an equivalent ‘penal bite’, meaning they have the comparable punity.

In von Hirsch et al.’s (1989) earlier work, they address the ways in which non-custodial sentences can be arranged on a penalty scale and how much substitution is acceptable. The first model they discuss is the ‘No Substitution model’, whereby sentences are arranged on a continuum but the non-custodial and custodial sentences do not overlap. The second is the ‘Full Substitution model’ which operates on a penalty unit basis. Therefore all sentences are given a score, which can then be substituted with a number of non-custodial sentences which equal the same penalty unit amount. However, von Hirsch et al. (1989) suggest this type of model is complex and overly complicated, and therefore may not be useful. The third model is the ‘Partial Substitution model’. Here there are a number of bands and within some of these bands it would be possible to substitute penalties of equivalent severity. This is the most useful of the three. Therefore, we argue that different types of sentences can be used interchangeably as long as they carry the equivalent ‘penal bite’. The idea of interchangeability, meaning that sentences can overlap in terms of their severity can also be seen in the diagram. For example, an onerous community order, can have the same/similar severity as a short-term custodial sentence and can be considered as an equivalent punishment.

Over the years there have been many attempts to create such scales (for example; Sellin and Wolfgang 1964; Tiffany et al. 1975; Erickson and Gibbs 1979; Sebba and Nathan 1984; Harlow et al. 1995; Schiff 1997; Francis et al. 2001; Pina-Sánchez et al. 2019), but there is currently no recognised sentence severity scale in existence. Part of this stems from the conflicting views and purposes of sentencing from a theoretical perspective, but from a methodological perspective the issue lies predominantly with the difficulty of measurement. However, as we will go on to argue, such a scale is important and necessary for the discipline of criminology and criminal justice.

## The Measurement of Sentence Severity

Previous attempts to measure sentence severity have generally used one of two basic approaches, which have remained distinctly separate in the methods used. The first is the data driven approach which tends to use real sentencing data directly in the form of administrative data records and generally adopts one of two methodologies; the “in-out” method or custodial sentence length. The second approach, in contrast, applies methodology to develop a score; this can use either official data, or more commonly, people’s perceptions of sentence severity to create a scale. If opinions on the perceived severity of sentences are gathered, these may be from criminal justice professionals, or the general public. These methods will now be discussed.

### In-Out Method

The in-out method - based on the decision on whether to incarcerate or not - is a simple but crude binary measure of sentence severity used to model the probability that an offender will receive a custodial sentence rather than any other disposal (Ulmer and Johnson 2004). This methodology implicitly assumes that a custodial sentence of any length is more severe than

any non-custodial sentence. Studies which adopt this method, such as those conducted by; Wheeler et al. (1982), Ulmer and Johnson (2004), Merrall et al. (2010), Holleran and Spohn (2004), are primarily interested in the factors which impact on the decision to incarcerate offenders, and also which factors contribute towards sentencing disparity.

### **Custodial Sentence Length**

The second and most popular data driven approach uses custodial sentence length to determine sentence severity; longer sentences are seen as more severe than shorter custodial sentences. This method is used in a number of studies, such as those conducted by: Anderson and Spohn (2010), Britt (2009), Albonetti (1997), Bushway and Piehl (2001), Helms and Jacobs (2002), Ulmer and Johnson (2004), Mueller-Johnson and Dhimi (2009), Wheeler et al. (1982), Miethe and Moore (1986), Mustard (2001), and Pina-Sánchez and Linacre (2013). This method excludes all non-custodial sentences from analysis or the non-custodial sentences were set to zero; and thus is problematic where there are low levels of incarceration.

Studies vary in how sentence severity is determined from these custodial sentences. Studies conducted in England and Wales, such as the study by Pina-Sánchez and Linacre (2013) used the length of custodial sentences as recorded in the Crown Court Sentencing Survey. However, this differs from those studies conducted in the U.S., especially where the state in which the research was conducted use a sentencing grid, such as in Minnesota. Sentencing grids provide a minimum and maximum presumptive sentence range for cases that share the typical criminal history and offense severity characteristics (Minnesota Sentencing Guidelines Commission 2013). For example, Helms and Jacobs (2002) used the awarded minimum and maximum sentence lengths transformed into months, and then averaged them. They then added 1 (to avoid taking the log of 0 in cases where the offender received a non-custodial sentence) and took the natural logarithm. In contrast, Britt (2009) used the minimum number of months the offender was sentenced to custody, and again took the logarithm of that sentence length. Taking the natural logarithm of sentence length is common practice as it reduces the positive skew, due to the fact that the majority of offenders receive relatively short custodial sentences (Britt 2009; Albonetti 1997). Others have used the midpoints of sentence length (Miethe and Moore 1986), and average sentence length (Mustard 2001).

### **Tobit Regression**

A number of studies have opted to combine both the in-out method with the length of custodial sentence to measure severity. Studies that have opted to jointly model the decision to incarcerate in combination with the length of the custodial sentence, such as Bushway and Piehl (2001) and Albonetti (1997) have combined these two measures using tobit analysis. In these cases, left-censoring is generally used to take into consideration the non-custodial sentences which have a sentence length of zero. Tobit analysis therefore allows for left-censoring to generate unbiased results.

### **Limitations of These Methods**

Although the data driven approach uses real sentencing data derived from the courts, the methods employed to measure sentence severity are problematic. The in-out decision to incarcerate creates a very simplified measure of sentence severity and implies that all

custodial sentences are more severe than any non-custodial sentences. It thus dismisses the possibility that two different sentences, for example a short custodial sentence and a community order, can have the same level of severity. This method also omits the magnitude of sentences. It is therefore not possible to determine whether a sentence is, for example, twice as severe as another. Consequently, this method of measuring sentence severity is rather crude.

The length of custodial sentence does give some indication of the magnitude of a custodial sentence relative to another custodial sentence but again excludes all non-custodial sentences. In doing so, Leslie Sebba comments that it “ignores the disparities in terms of severity found among non-custodial penalties” (Sebba 1978, p. 250). This method inevitably leads to bias in the severity measure, as it excludes important sentencing information by omitting the multitude of sentences which do not result in a custodial sentence (Merrall et al. 2010). By omitting all non-custodial sentences, this method, in common with the in-out method, does not allow that certain sentences of different types can be equivalent in terms of their severity, whereby a non-custodial sentence i.e. a community order, could be as severe as certain custodial sentences (von Hirsch 1993). Therefore it rejects the idea that two different disposals can have the same level of severity.

## Severity Scales

An alternative methodology is to construct a scale. The 1960s, 1970s, and 1980s witnessed a number of studies which focused on scaling offense seriousness and sentence severity but by the mid-1980s, there were fewer studies, and the methods and techniques used did not change much over this time, although recently there has been a renewed level of interest in developing these scales, alongside some alternative methods. We will now summarize these.

One method that has been adopted to create severity scales has been to assign arbitrary scores to a range of different sentence categories. For example, Tiffany et al. (1975) assigned a range of scores to 16 sentences, ranging from 0 for a suspended sentence to 50 for custodial sentences over 10 years. Similarly, Malila (2012) assigned arbitrary scores to different types of punishment relative to other punishments using a unit scoring approach. Based on these scores, sentences are then ranked to form a severity scale.

Other studies, such as those conducted by Erickson and Gibbs (1979), Sebba and Nathan (1984), Allen and Anson (1985), and Harlow et al. (1995) used psychometric scaling techniques (sometimes referred to as magnitude estimation) to create sentence severity scales. In this approach members of the public, prisoners, or experts (e.g. judges, probation officers, police officers) were asked to rank the severity of different sentences. Respondents are presented with a list of randomly ordered sentences and one sentence - for example a 1-year custodial sentence - is anchored and given a value of 100, or some such value. Respondents are asked to assign additional values to the remainder of sentence categories relative to this score. The individual scores for each sentence are then usually averaged to produce a final scale with various scores for each sentence.

Harlow et al. (1995) used both line production (LP) and number estimation (NE) which are also types of psychometric scaling. LP involves respondents drawing lines to replicate the severity of a series of sentences relative to one another; and NE is similar to magnitude estimation in which respondents are asked to score sentences relative to the previous sentence through assigning a series of numbers. They acknowledge the scale values are not interpretable on their own because the values only have meaning in comparison to each



other but the ratio of numbers corresponding to different sentences are meaningful. They give the following example; a 1-year probation with a scale value of 54.29 can be considered about half as severe as a 1-year intensive supervision programme, with a scale of 111.79. This then gives the scores far more context than previous studies.

However, Harlow et al. (1995) admit magnitude scaling is time-consuming for both the researcher and respondents; the actual preparation of the surveys and scale validation once the data has been collected is also time-consuming. However, they reason that it allows freedom in determining the adequate range of responses. In other words, respondents are not limited by having to choose from pre-chosen categories, which is likely to restrict their perceptions of sentence severity.

Still in the field of psychometric scaling techniques, Buchner (1979) employed a paired comparisons study based on Thurstone's (1994) Law of Comparative Judgements, in which two stimuli (sentences) are compared. Respondents are presented with two sentences (sentencing outcomes) and asked to select the outcome they perceive to be the most severe. Pina-Sánchez et al. (2019) adopted a similar approach using pairwise comparison data using Thurstone Case V scaling but instead used the number of times a sentence is judged to be more severe as the impetus for the ordering of the severity scale. More recently Pina-Sánchez and Gosling (2022) used a modified version on the Thurstone method to account for the unequal variances across the range of disposals.

Although only tangentially related to the construction of sentence severity scales, there has also been a renewed interest in crime severity scales with the development of two methods; the Cambridge Crime Harm Index (CHI) proposed by Sherman et al. (2016), and the Crime Severity Scale proposed by the Office for National Statistics (2016). The CHI uses the sentencing guidelines produced by the Sentencing Council of England and Wales. It uses the starting point sentence for each offense, based an adult offender with no previous convictions and no aggravating or mitigating factors, and the score is expressed as the days in custody (Sherman et al. 2016). Where non-custodial sentences are given, these are converted into days in custody (see Sherman et al. (2016) for the methods used for conversion). The CSS uses a different method in which they calculated the mean number of days in custody for each offense type, based on 5 years worth of conviction data. Similar to the CHI, they also convert non-custodial sentences into nominal days in custody (see Office for National Statistics (2016) for the methods used for conversion). However, Ashby (2018) notes that while both methods have the same objective there are substantial differences in the estimates produced, and therefore analysts should be cautious on the index they use and the conclusions drawn from such analyses.

## Limitations of Severity Scales

Although these severity scales make a step forward in incorporating the non-custodial and custodial sentences, the existing methods used to create these scales also suffer from some limitations. Firstly, the scales are based on what (Tremblay 1988, p. 284) refers to as "subjective perspectives" - whether that of expert or lay persons, people will generally vary in terms of how they view and measure different sentencing outcomes. Furthermore, judgement-based methods are therefore difficult to generalise outside the context of their own studies. McDavid and Stipak (1981) would argue, scales created in this way are highly subjective, and are therefore not useful in terms of measuring sentence severity in the "real world". While others make the case that although these scales are generally reasonable, questions arise as to whether findings are simply an "artefact of an artificial scale"



(Klepper et al. 1983, p.58). Finally, and related to these earlier points, these methods also tend to work with only a limited number respondents; for example, Harlow et al.'s (1995) scale was based on 44 members of the public, and Buchner's (1979) study was based on 51 judges. These samples are therefore not generalisable to any one population and the results are likely to vary by the sample used.

## Creating a New Sentence Severity Scale

As discussed, the methods that have been used up to now remain fairly distinct in their approach. This paper attempts to bridge this gap by firstly, using real sentencing data, and secondly, using statistical modeling to create a single sentence severity scale which includes the full spectrum of sentences used by the courts. This work therefore differs from existing work in the field, as this study is data driven and uses advanced statistical modeling techniques, which have not hitherto been used in sentencing research. While Pina-Sánchez et al. (2019) rightly acknowledge that the use of a severity scale as a research tool is likely to be place (jurisdiction) and time specific, the method employed to create such a scale is more universal.

### Data

The UK Police National Computer (PNC) system (Home Office 2014) is an electronic operational database providing information on arrests, court summons and court disposals. The data is initially deposited by the local police authority and updated with information from the courts following sentencing via the Crest (Crown Court) and Libra (magistrates' court) systems. The data used in this study is a sub-sample of the PNC data and contains information relating to adult offenders aged 18 or over ( $n=29,326$ ) who were sentenced in a court in England and Wales between March 2008 and April 2010, thus excluding non-court based disposals such as police cautions. This sample represents around a quarter of all cases sentenced during this time. Sentencing information relates to 60,662 offenses and their subsequent sentences, based on the fact that offenders can be sentenced for multiple offenses at one given court appearance.

The dataset contains a detailed Home Office offense code for each offense, as well as up to four possible disposal components for that offense, including; the broad disposal type, a detailed Home Office disposal code, disposal duration, and amount of disposal. Disposals such as custodial sentences and fines are relatively straightforward to record, however the complexity of relatively new types of disposals, such as community orders and suspended sentence orders, make the recording of these disposals more difficult. The existing pre-defined fields used in the PNC limit the information that can be recorded, and therefore it is not currently possible to record the range of requirements attached to these types of orders. This is problematic in two respects. Firstly, the range of possible requirements on certain sentence types (for example electronic monitoring, curfew, location restrictions) is not available. Secondly, for suspended sentences, while the length of the suspended sentence is stored, the length of time the offender is subject to the suspended sentence (the suspension period) is not recorded. Both will affect the severity of these sentences. Similarly, the PNC does not record the conditions placed on conditional discharges.

In cases where offenders are sentenced for multiple offenses, official statistical publications in England and Wales will only report on the primary offense. Primary offenses refer

to the most serious offense, and any additional offenses and sentences, will be excluded from any analysis. In contrast, in our sample, 45% of offenders were sentenced for multiple offenses, and therefore we have included all of these offenses and the subsequent sentences in the analysis. However, where multiple disposals were recorded for a single offense, we have only been able to include the primary disposal component for each offense.

## Methods

The aim of this empirical work is to create a single scale measuring sentence severity which includes the full spectrum of sentences meted out by the courts in England and Wales. As already discussed, there are a variety of methods that have been utilised in the past to create a measure of severity, each with their own merits and shortcomings. Fundamentally, these past measures, by omitting information, introduce bias into the resulting scale, and therefore a new, more robust and accurate measure is required.

This work bridges the two approaches of measuring severity: it uses real sentencing data, as opposed to using hypothetical scenarios, and constructs a scale using statistical modelling rather than psychometric scaling. By producing a scale which incorporates both custodial and non-custodial sentences, we are able to create a single measure of sentence severity. This will enable us to identify sentences which are of equivalent severity - the concept of interchangeability. An additional use is the ability to use the scale as the dependent variable within a statistical analysis which can investigate, for example change in sentence severity over time (demonstrated later in the paper) or issues of sentencing disparity (see for example Wallace (2015)).

In this paper, the aim is to use a statistical modelling approach to build a score for sentences using a contingency table approach. We propose the use of an extension to the Goodman RC association model (also known as the log-multiplicative model). First introduced by Goodman (1979) as the association model II, and renamed to the RC association model in 1981 (Goodman 1981), this model has been used extensively for ordering and scaling categorical variables. Thus Clogg (1982) used the model for scaling and ordering attitudes to abortion; and Bergman et al. (2002) have used the model to construct the CAMSIS social stratification scale. Uses of this model in criminology have been limited - a notable exception is by Messner et al. (2002), who used the model to impute unknown victim offender relationships in tabular homicide data.

A crucial component of using the Goodman RC association model for scaling or scoring a categorical variable is choosing a "criterion" variable (Clogg 1982, p.119), which needs to be highly associated with the target variable, in our case, the sentence variable, and can be cross-classified with the target variable to make a two-way contingency table. Clogg (1982) identifies four criteria to assess the suitability of the criterion variable. There must be association between the target and criterion variables; this association must have a good theoretical and substantive rationale; the model using that criterion variable must produce estimated scores that do not violate known ordinality requirements, and the statistical model must fit. We choose offense as our criterion variable as there is clearly a strong substantive association between offense and sentence.

As the model needs a two-way cross-classification of two categorical variables - sentence by offense, it is necessary to group sentences into disjoint categories. This was carried out as follows. Sentences were organised by disposal type and arranged into 29 distinct sentence categories determined by sensible cut-points derived from the frequencies of offenders receiving each sentence, shown in Table 1. Disposals such as

**Table 1** Frequency of sentence categories in the sample

Sentence	Frequencies
Custody 1 Month	2169
Custody 2 Months	1651
Custody 3 Months	1817
Custody 4 Months	1579
Custody 6 Months	1882
Custody 9 Months	1623
Custody 1 Year	2149
Custody 1.5 Years	2169
Custody 2 Years	1948
Custody 3 Years	1915
Custody 4 Years	1128
Custody 5 Years	598
Custody 6 Years	372
Custody 7 Years	174
Custody 8 Years	162
Custody 10 Years	100
Custody 10 Years +	184
Suspended 1 Month	905
Suspended 2 Months	1331
Suspended 3 Months	1979
Suspended 4 Months	1739
Suspended 6 Months	2250
Suspended 1 Years	3738
Community 1 Year	12448
Community 2 Years	3734
Community 3 Years	1121
Other	7362
Fine	1349
Conditional Discharge	1086

confiscation and compensation orders were treated as a single disposal category labelled “other”. Similarly fines were kept as a single category, as the magnitude of the fine does not necessarily indicate severity, due to being means-tested. Conditional discharges were also kept as a single category. Community orders were split into three categories based on the length of the order, suspended sentence orders were split into six categories indicating the length of custodial sentence that was being suspended, and immediate custody grouped by sentence length, forming 17 categories. Thus, these 29 categories cover all possible disposals available to courts in the England and Wales system.

With a large number of distinct offense codes, it was necessary to also group offenses to reduce the size of the contingency table. The offense codes were grouped into major offense code categories, and some offense categories were omitted from the sample due to a very small number of offenders ( $n \leq 5$ ) having committed those types of offenses. This reduced the number of offense categories from 477 to 111. The list of offense codes used in this study can be found in Supplementary Material 3.

**Table 2** Part of the offense by sentence contingency table

	Offenses (and Home Office code)				
	Murder (1)	Man- slaughter (4)	Serious Wounding (5)	Wounding (8)	etc ...
Immediate custody					
10 Years +	3	4	6	0	...
10 Years	0	0	6	1	
7 Years	0	9	6	8	
6 Years	0	2	7	13	
5 Years	0	9	23	14	
4 Years	0	5	38	20	
3 Years	0	19	65	26	
2 Years	0	10	64	99	
1.5 Years	0	5	37	185	
etc					

The resulting contingency table consists of a two-way table of cell counts, cross classified by offense  $O$  and sentence  $S$ , resulting in a  $111 \times 29$  table with 3219 cells.

Table 2 shows a small part of this contingency table, for four offenses and twelve sentence categories. We can observe that there is a strong diagonal in this table indicating a positive association between the four offenses and sentences. We can also observe that some sentences are less common than others - for example, community sentences of more than 2 years are not common.

### Log-Linear Models

Before we describe the Goodman RC association model, we can first demonstrate the association between our target and criterion variables by fitting a simple log-linear model to test for independence. Let  $Y_{os}$  be the observed count for offense  $o$  and sentence category  $s$ . We assume that the cell counts are Poisson distributed with means  $\mu_{os}$

$$Y_{os} \sim \text{Poisson}(\mu_{os}) \tag{1}$$

The log-linear model of independence can be written as

$$\log(\mu_{os}) = \theta + \alpha_o + \beta_s \tag{2}$$

An equivalent way of writing the model is to re-parameterize by removing the  $\theta$  parameter,

$$\log(\mu_{os}) = \alpha_o + \beta_s \tag{3}$$

where the  $\alpha$  parameters have absorbed the  $\theta$  parameter. For identifiability, it is usual to define  $\beta_1 = 0$ , setting the first level to be the reference category.

The deviance, (or minus twice the log-likelihood ratio of the model compared to the saturated model) of this model has a chi-squared distribution with  $(O - 1) \times (S - 1)$  degrees of freedom if there no evidence of association in the table. This is known as  $G^2$  in some texts. When applying this model to the data, we obtain a deviance of 61837 on 3080 degrees of freedom, which is highly significant ( $p < 0.0001$ ).

A second model is the saturated model. This models the association between the row and column variables by adding in a large number of interaction parameters, one for each cell. This model totally reproduces the observed cell counts and has a deviance of 0 on 0 degrees of freedom.

$$\log(\mu_{os}) = \alpha_o + \beta_s + \gamma_{os} \quad (4)$$

The Goodman RC association model and other association models aim to fit a model which explains the interaction structure between offense and sentence with a relatively small number of parameters - a model between the independence model and the saturated model.

### Score Association and Linear by Linear Models

The simplest way of modelling an association in a two way table is to give each row and column a fixed score. For example, if we knew a seriousness score for offenses, and a severity score for sentences, then these scores could be used directly in the model. If these scores are called  $score_o$  and  $score_s$  then the product of these scores can be formed, and make an additional explanatory variable in the model. The score association model is then:

$$\log(\mu_{os}) = \alpha_o + \beta_s + \lambda \times score_s \times score_o \quad (5)$$

A special form of this model is the linear by linear score association model, where the scores for the offenses are simply the number  $1, 2, \dots, O$  for each row, and scores for the sentences are similarly  $1, 2, \dots, S$  for each column. This model assumes that the categories are in some natural order, and that therefore it makes sense to assign a rank order. For example, this model might be considered if the columns were categories of a Likert response and the rows of the table were social class categories.

However, we do not know either a seriousness score for offenses or a severity score for sentences, and therefore these need to be estimated from the model. We replace  $score_o$  and  $score_s$  by  $\gamma_o$  and  $\delta_s$  with the Greek letter notation indicating that the scores now need to be estimated from the model. This motivates the Goodman RC association model.

### The Goodman RC Association Model

The model is written as

$$\log(\mu_{os}) = \alpha_o + \beta_s + \lambda \gamma_o \delta_s \quad (6)$$

Without the last term, the model is a main effects Poisson model representing independence between offenses and sentences; with  $\alpha_o$  and  $\beta_s$  modelling the row and column marginal totals; in other words how “popular” a particular offense or sentence category is. The Goodman RC association model contains an additional multiplicative association term between the set of scaling parameters for  $o$  and the set for  $s$ . The  $\lambda$  parameter can be absorbed into either the  $\gamma_o$  or the  $\delta_s$  parameters. Without loss of generality, this gives the model

$$\log(\mu_{os}) = \alpha_o + \beta_s + \gamma_o \delta_s \quad (7)$$

The Goodman RC association model does not assume that categories are correctly ordered: instead the model reveals the ordering of the categories through estimating the two sets of unknown parameters  $\gamma_o$  and  $\delta_s$  (Powers and Xie 2000). This requires an iterative procedure

at the model building stage where, beginning with a random set of starting parameter values, one set of parameters are treated as known and the other set of parameters are estimated; these are then fixed and the other parameters are updated until they converge. The final estimates for  $\delta_s$  provide the latent sentence severity scale.

To summarize, the  $\beta_s$  are modeling the marginal sentence totals and represent the popularity of each sentence - and are not of primary interest in this paper.  $\delta_s$ , in contrast, provides a linear score for sentencing severity which, when combined with the linear score for offense, explains as much of the association as possible between offense and sentence.

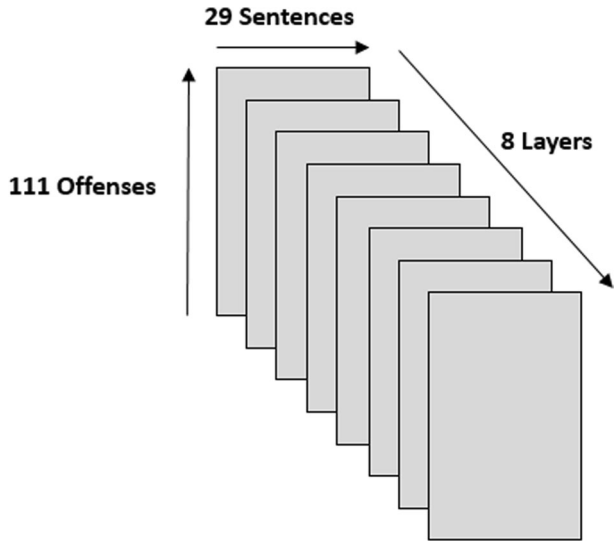
The model can be intuitively explained in the following way. The independence model will not fit. This is because serious offenses have high severity sentences and will be very unlikely to have low severity sentences. In modelling terms, a serious offense and a severe sentence will happen relatively often, and the count for these cells will be larger than expected under the model of independence, giving a large positive deviance residual. This can be reproduced in the Goodman RC model when  $\gamma_o$  (for offenses) and  $\delta_s$  (for sentences) are both large and positive. If, on the other hand we have a serious offense and a low severity sentence, then this would be relatively unusual, and the observed count will be smaller than expected under the model of independence and the deviance residual will be large and negative. This can be fitted in the Goodman RC model with a large positive value for  $\gamma_o$  and a large negative value for  $\delta_s$ . Note that the signs of the estimates of  $\gamma_o$  and  $\delta_s$  are not guaranteed to be correct, as the large positive residual can also be fitted with large negative values for both  $\gamma_o$  and  $\delta_s$ . The practitioner can easily reverse signs if the estimates for both the offense score and the sentence severity score both have the wrong sign.

## Extending the Goodman RC Association Model

While the simple Goodman RC association model can be used to estimate a basic sentence severity score, the estimates will not be optimal. This is because judges in England and Wales are required to consider both legal factors, and aggravating and mitigating factors before reaching a final decision. The number of factors are large, and includes whether the offense was committed when the offender was on bail, whether the offense was aggravated in relation to race, religion, disability or sexual orientation of the victim, and whether the offense involved a terrorist connection. While information on most aggravating and mitigating factors is not available in our dataset, we can include some legal factors which will affect sentencing.

We limit consideration of factors to three variables which are consistently coded in the data and contain small amounts of missing information: whether the offender has a history of offending (acting as a proxy for previous convictions and discretized as prior convictions or no prior convictions), the offender's plea (guilty or not guilty), and number of offenses brought before the court (discretized as a single offense or multiple offenses). The first variable reflects criminal history; an offense which is a debut offense for offenders will attract a lower sentence compared to the same offense of equal seriousness that is committed later in a criminal career. The second variable is plea; a guilty plea is a mitigating factor for sentencing, offender's that plead guilty are entitled to a reduction in the severity of their sentence depending on at what stage they admit their guilt. Finally, offenses brought before the court indicates whether the offender was sentenced for a single offense or as part of a number of offenses to reflect the totality

**Fig. 2** Diagram representing the Extended Goodman RC association model



principle. The aim is therefore to extend the Goodman RC association model to control for such factors which will have an impact on sentencing decisions.

Layers are conceptualized in the following way. There are eight possible combinations of these three variables (for example, single offense, guilty plea, no prior is one combination; multiple offenses, not guilty plea, prior is another combination) and each combination now has its own offense by sentence table. These eight combinations are *layers* in this new terminology. Thus, conceptually, the whole data now consists of a set of eight separate layers of 29 by 111 tables, as illustrated in Fig. 2 - or a three-way table of size  $8 \times 29 \times 111$  with 25752 cells in total.

### Existing Extensions of Two-way Contingency Table Models with Layers

Other authors have extended the Goodman RC association model into three dimensions to model data of the form row \* column \* layers. Probably the best used extension is by Xie (1992) which is known as the log-multiplicative layer effect model. This fits a full interaction between layers and rows, and between layers and columns, and assumes a common association between the rows and the columns of the table for each layer, but multiplied by a different constant for each layer. Such models have been used by Erikson and Goldthorpe (1992) to model social mobility tables and how they change over time.

A different approach was suggested by Goodman and Hout (1982), who proposed a 'modified regression-type approach' for modelling the association between rows and columns in the presence of layers. Essentially, this approach decomposes the interaction into a common interaction term present for all layers, and a second common term which is multiplied by a layer-specific constant. This method has been used to examine how the association between job satisfaction and self-employment varies across occupational groups.

Neither of these approaches allow for covariates to be placed on the estimated association parameters. We therefore proceed to develop a new model.



## A New Model for Incorporating Covariates into the Goodman RC Association Model

In this section, we develop a model which allows covariates to be placed on the sentence severity scale, building on the concept of layering introduced by Xie (1992).

Let  $Y_{os\ell}$  be the three way table of counts with

$$Y_{os\ell} \sim \text{Poisson}(\mu_{os\ell}) \quad (8)$$

Then we propose

$$\log(\mu_{os\ell}) = \alpha_o + \beta_s + \tau_\ell + \gamma_o \delta_{s\ell} \quad (9)$$

The  $\delta_s$  in the standard Goodman model is replaced by  $\delta_{s\ell}$  with the sentence scores now additionally depending on the layer. Thus  $\ell$  is an index for the layers. We also include the main effect of  $\tau_\ell$  in the model, which represents the marginal effect of each of the layers. This model however has a very large number of parameters, with  $\ell$  sets of sentence severity scores, one for each layer. We seek to simplify this model, and instead specify a simple main effects model for the  $\delta_{s\ell}$ :

$$\delta_{s\ell} = \delta_s + \sum_{p=1}^P \psi_p X_{ps} \quad (10)$$

The  $\delta_{s\ell}$  is decomposed into a baseline sentencing severity score  $\delta_s$  and the additive effects of the  $P$  covariates. The  $X_{ps}$  represent dummy variables for plea, prior court appearance, and number of offenses and  $\psi_p$  are a set of  $P = 3$  unknown regression coefficients. As the  $\psi_p$  parameters do not depend on  $s$ , they simply apply an additive offset to the baseline sentencing scale, either positive or negative. The adjusted sentence severity scores are given by the  $\delta_s$  parameters, which will now control for the legal covariates. We can determine which of the legal variables are important by the use of information criteria, such as AIC or BIC.

The final model is

$$\log(\mu_{os\ell}) = \alpha_o + \beta_s + \tau_\ell + \gamma_o \left( \delta_s + \sum_{p=1}^P \psi_p X_{ps} \right) \quad (11)$$

The final sentence severity score estimates are given by  $\delta_s$ , now controlling for the effects of the legal factors. Aggravating factors should increase the sentence severity, whereas the mitigating factors should decrease the sentence severity. Therefore by controlling for the legal covariates (aggravating and mitigating factors) we are able to control for the imbalance these factors have on various offenses. For example, 81% of those charged for burglary at the Crown Court pleaded guilty in 2009 whereas only 39% of those charged for a sexual offense similarly pleaded guilty (Ministry of Justice 2010).

## Model Fitting and Selection

All models were fitted using the Generalized Nonlinear Models (gnm) (Turner and Firth 2023) package in R (R Core Team 2023). As well as the standard Goodman RC model, we fitted a sequence of extended Goodman RC models allowing us to investigate using AIC which legal covariates were necessary to include in the model. Table 3 shows the AIC

**Table 3** AIC values for the main effects severity score models

Severity model	AIC
Plea + Previous convictions	71196.75
Plea + Number of offenses	72780.58
Previous convictions + Number of offenses	71032.88
<b>Plea + Previous convictions + Number of offenses</b>	<b>70302.99</b>

values for a variety of severity score models. The full main effects model has the lowest AIC and we report on this model.

### Linear Interpolation and Extrapolation

The final extended Goodman RC model produced by the software provided estimated sentence severity scores. Although the scores produced by this method provide a measure of sentence severity in their own right, they are not anchored or standardised in any way. We therefore standardized the scores using Agresti's technique (Agresti 2013; Kateri 2014) to have a weighted mean of zero and weighted variance of 1 (with the weights being the column proportions in the two-way table).

Once the standardized scores have been produced, the final task is to convert them to an equivalent number of days of custodial sentence so that we can interpret them from a 'real-world' perspective. We have adopted the following procedure. For custodial sentences, we use the observed mean number of days in each sentencing category.<sup>5</sup> For all other sentence types, we used the standardized score estimate and linear interpolation to place the category on the custodial sentence scale, using the two custodial standardised scores which were closest and their mean number of days in custody. This gave an equivalent number of days in custody for each of the non-custodial sentence types. For some sentencing categories, we used linear *extrapolation* rather than interpolation as the standardized estimates were less than any of the standardized scores for custodial sentences. Extrapolation will produce negative estimates for the equivalent number of days in custody, and are therefore less severe than any custodial sentence. This is a natural byproduct of placing all sentencing types on the "days of custodial sentence" scale.

It is important to emphasize that linear interpolation and extrapolation only yields an approximate measure. However, it provides a sentence severity score which can be interpreted in a meaningful way. These sentence severity scores - the equivalent number of days in custody - then provide us with a new and improved measure of sentence severity, which can then be used to investigate sentencing disparity. It also provides us with a scale which allows us to assess the interchangeability of sentences, which Sebba (1978), von Hirsch (1993), and Morris and Tonry (1990) view as of primary importance.

<sup>5</sup> Sentences over 10 years include a small number of life and indeterminate sentences, and for this category we used the model sentence length.

**Table 4** Sentence severity scores

Sentence	Score	Lower CI	Upper CI	Approx Days <sup>a</sup>
Other	-0.654	-0.674	-0.634	<b>-11</b>
Fine	-0.625	-0.645	-0.605	<b>-1</b>
Custody 1 Month	-0.551	-0.569	-0.533	23
Suspended 1 Month	-0.526	-0.546	-0.506	<b>31</b>
Conditional Discharge	-0.476	-0.498	-0.454	<b>47</b>
Custody 2 Months	-0.456	-0.474	-0.438	54
Suspended 2 Months	-0.421	-0.443	-0.399	<b>70</b>
Community Order 1 Year	-0.417	-0.427	-0.407	<b>72</b>
Community Order 2 Years	-0.403	-0.417	-0.389	<b>79</b>
Custody 3 Months	-0.386	-0.406	-0.366	86
Suspended 3 Months	-0.375	-0.395	-0.355	<b>111</b>
Custody 4 Months	-0.351	-0.373	-0.329	117
Suspended 4 Months	-0.345	-0.367	-0.323	<b>118</b>
Suspended 6 Months	0.002	-0.027	0.031	<b>167</b>
Custody 6 Months	0.014	-0.019	0.047	169
Suspended 1 Year	0.440	0.401	0.479	<b>239</b>
Custody 9 Months	0.486	0.431	0.541	247
Custody 1 Year	0.719	0.666	0.772	351
Custody 1.5 Years	0.905	0.85	0.96	493
Community Order 3 Years	1.181	1.108	1.254	<b>686</b>
Custody 2 years	1.209	1.15	1.268	706
Custody 3 Years	1.440	1.381	1.499	996
Custody 4 Years	1.783	1.714	1.852	1367
Custody 5 Years	2.184	2.102	2.266	1773
Custody 6 Years	2.354	2.256	2.452	2154
Custody 7 Years	2.453	2.316	2.59	2497
Custody 8 Years	2.489	2.348	2.63	2888
Custody 10 Years	3.143	2.945	3.341	3522
Custody 10 Years +	3.792	3.535	4.049	4380

<sup>a</sup> Days highlighted in bold in the table indicate the estimated equivalent number of days in custody calculated from linear interpolation, or extrapolation

## Results

### Sentence Severity Scores

The standardised raw scores, 95% confidence levels of the scores, and equivalent number of days in custody for each of the sentence categories are displayed in Table 4, and displayed in graphical form in Fig. 3. The days highlighted in bold in the table indicate the estimated equivalent number of days in custody calculated from linear interpolation, or extrapolation.

The scores provide an assessment of severity, which enables the categories to be ranked in order. Given the different punitive elements of sentences meted out by the courts, the

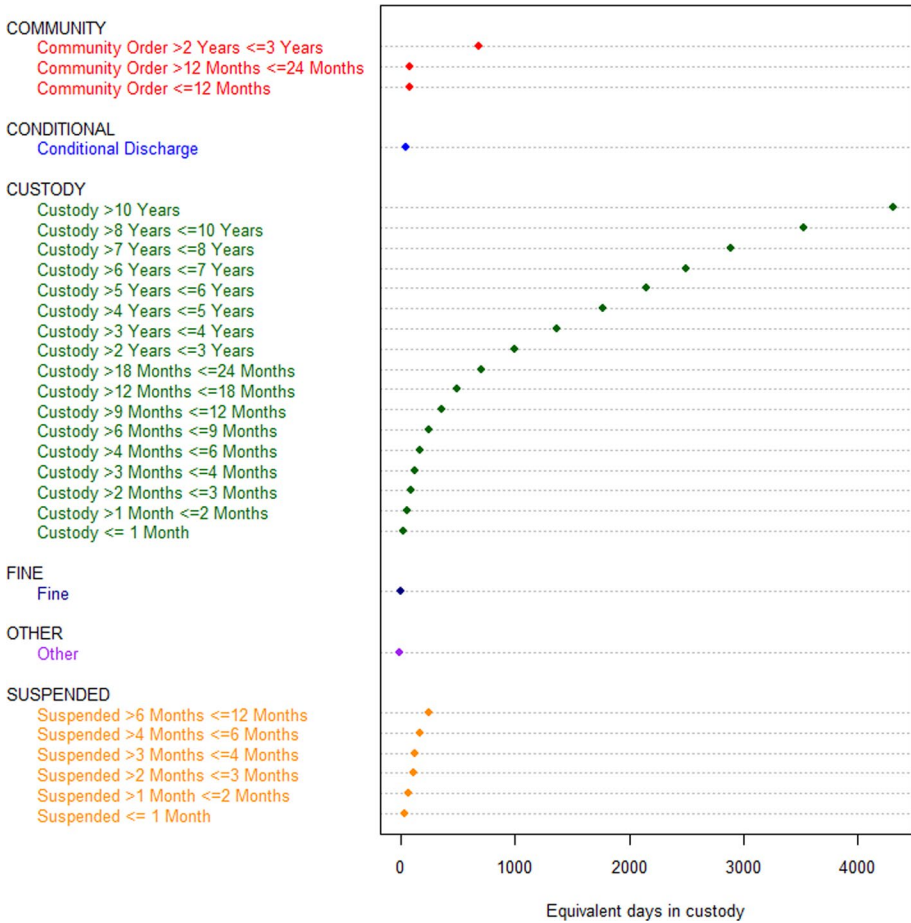


Fig. 3 Dotplot of sentence severity scores from the Extended Goodman RC model

results of the Extended Goodman RC association analysis allow us to align these sentences on a single scale. The severity scores range from least severe  $-0.654$  (0.010) for the group of sentences we have referred to as ‘other’, and range up to  $3.792$  (0.131) - the most severe score for custodial sentences over 10 years. The remaining sentence categories are calibrated between these two scores at either end of the sentencing severity spectrum. As expected, custodial sentences dominate at the severe end of the scale, whilst non-custodial sentences congregate at the lower end of the severity scale.

Whilst the standardised scores provide the ordering of the scale, we will generally refer to the equivalent days in custody, although these should be treated with some caution when interpreting the results. Fines and the ‘other’ sentence categories are estimated through linear extrapolation resulting in minus days ( $\approx -1$  and  $\approx -11$  respectively), as these are estimated to be less severe than any custodial sentence. The remaining non-custodial sentences can also be estimated in equivalent days following linear interpolation. We can see that, for example, a conditional discharge ( $\approx 47$  days) falls somewhere between a custodial sentence of less than 1 month ( $\approx 23$  days) and a custodial sentence of between one and 2 months

( $\approx 54$  days). Community orders of less than 1 year ( $\approx 72$  days) and between 1 and 2 years ( $\approx 79$  days) fall in line with relatively short custodial sentences, whereas community orders of between 2 and 3 years ( $\approx 686$  days), which we know are used much less frequently, are almost equivalent to a custodial sentence of between 18 and 24 months ( $\approx 706$  days). Suspended sentences closely follow the order and magnitude of their comparative custodial sentence. In England and Wales, the Sentencing Council (2017) state that suspended sentences are custodial sentences, and can be used as an alternative where the power to suspend is available. Thus slight differences in estimates between custodial and suspended sentences of the same length are simply a reflection of random variation. If desired, it is possible to equate estimates using the methods described in Goodman (1986) (discussed later).

When modelling sentence severity at the individual level, either the original scores or the equivalent number of days in custody could be used. If a log transformation is desired, and the negative values present problems, then adding a suitable constant to the scores before taking logs will provide a solution.

Figure 3 shows that, sentences within a disposal type, have estimates which maintain the ordering of the sentences. Thus the score estimates for custodial sentences increase with custodial sentence length; as do estimates for suspended sentences and community orders. We can also comment briefly on the  $\gamma_o$  estimates, which represent the seriousness of the offenses. It is reassuring to observe that these estimates are in a sensible order. Thus, murder has a standardised estimate of 0.67, robbery 0.48, arson 0.34, shoplifting  $-0.16$  and bail offenses  $-0.76$ . Overall, we can say that ordinality is preserved, and Clogg's third principle for the choice of a good criterion variable is satisfied.

### Applying the Scores to Crown Court Sentencing Survey Data to Assess Changes in Sentencing Severity

To illustrate just one of the practical capabilities of the new scale, we apply the scale to data from the England and Wales Crown Court Sentencing Survey (CCSS). We do this to investigate change in the severity of sentencing for two broad offense types - theft and sexual offending. Using four complete years of sentencing data (2011–2014), we are able to determine if sentencing for these offenses has remained constant over time.

The CCSS was commissioned by the Sentencing Council of England and Wales to monitor the use of guidelines and thereby fulfil its obligations under section 128 of the Coroners and Justice Act 2009. Between 1 October 2010 and 31 March 2015, sentencing information was recorded by the presiding judge during sentencing, including all the factors taken into account when passing sentence. This makes the survey one of the most detailed datasets available in England and Wales (available on the Sentencing Council website), despite it being affected to a degree by non-response (Pina-Sánchez and Grech 2017). While the dataset contains detailed information relating to the characteristics of the case, such as number of aggravating and mitigating factors taken into consideration at sentencing, the sentencing outcomes are less detailed and we lose some of the finer gradation of sentencing outcomes. For example, custodial sentences are disaggregated into only nine distinct categories, while suspended sentence orders and community orders remain single categories.

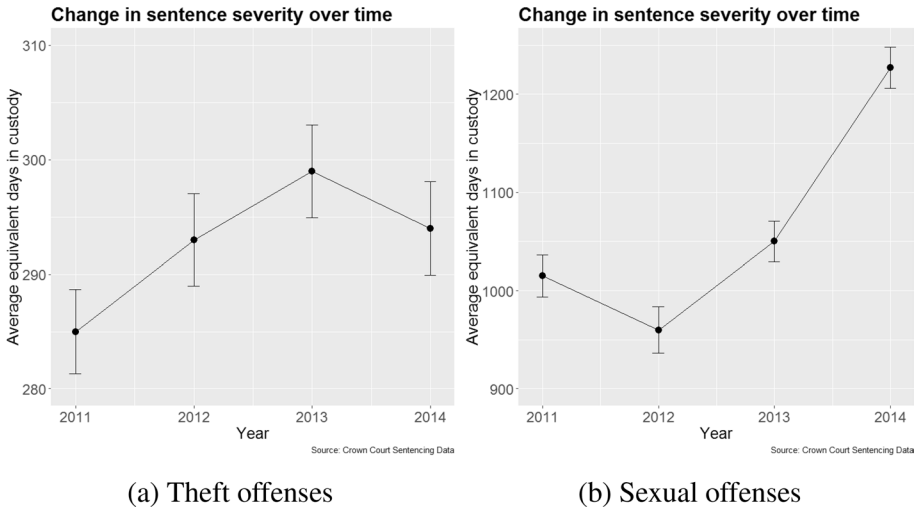
Crown court data will capture the more serious offenses, as relatively minor offences will be dealt with at the magistrates court. Both offences average into the custodial scale,

**Table 5** Reclassification of sentence categories for CCSS analysis

Sentence severity scale	CCSS	Weighted equivalent days in custody
Immediate custody	Immediate custody	
1 month	≤ 3 months	52
2 months		
3 months		
4 months	> 3 months ≤ 6 months	145
6 months		
9 Months	> 6 months ≤ 12 months	396
1 Year		
1.5 Years	> 12 months ≤ 18 months	493
2 Years	> 18 months ≤ 3 years	850
3 years		
4 years	> 3 years ≤ 4 years	1367
5 years	> 4 years ≤ 5 years	1773
6 years	> 5 years ≤ 10 years	2544
7 years		
8 years		
10 years		
10 + years including life	> 10 years including life	4380
Suspended sentence		
1 Month	Suspended sentence	152
2 Months		
3 Months		
4 Months		
6 Months		
1 Year		
Community Order		
1 Year	Community Order	113
2 Years		
3 Years		
Fine	Fine	-1
Conditional Discharge	Conditional Discharge	47
Other	Other	-11

but there is still a substantial number of cases who do not receive an immediate custodial sentence. For sexual offences, only 64% receive a custodial sentence; for theft only 47%.

To apply the severity scale to the CCSS data, we had to reconfigure the severity scale categories to be consistent with the CCSS data, reducing the number of sentence categories from 29 to 14. For each of the CCSS sentence categories, we calculated the average equivalent number of days in custody weighted by the number of cases (individuals) receiving that sentence in the original PNC dataset for which the scale was computed (see Table 5). Then for each of the 4 years, for both theft and sexual offenses, we calculated the average number of days in custody (including the standard errors of the mean). The results are



**Fig. 4** Change in average sentence severity over time (bars represent  $\pm$  standard error of the mean)

displayed graphically in Fig. 4 which allow us to assess any change in sentence severity over the 4-year period.

For theft offenses, the change in average sentence severity can be seen in Fig. 4a. In 2011, the average sentence stood at the equivalent of 285 (SE=3.69) days in custody, rising to a peak of 299 (SE=4.05) days in 2013, and then falling slightly in 2014— 294 (SE=4.09) days. This slight positive trend for theft would suggest that sentence severity is incrementally on the increase.<sup>6</sup> We see a similar trend for sexual offenses, shown in Fig. 4b. Despite a slight fall between 2011 (1015 days, SE=21.34) and 2012 (960 days, SE=23.47), the average sentence for sexual offending has increased over this period, rising to 1227 days (SE=21.08) in 2014.

Of course, changes in sentence severity over time can be attributed to a number of factors, such as change in the crime mix or seriousness of the offenses committed, and/or change in the legal characteristics of the offenders (such as, offenses committed by more repeat than first time offenders, or more offenders pleading not guilty, etc). Alternatively, it may be that judges are sentencing offenders more harshly, and thus sentences are becoming more severe as a result. We briefly examined crime mix over the 4 year period for sexual offences using ONS figures (Office for National Statistics 2023) and found that in 2011/12 adult rapes among all police reported sexual offences were 9773 out of 16,038 (61%). In 2014/15, they were 18,338 out of 29,420 (62%). Thus the proportion of reported rapes among all sexual offences does not change to any great extent. Further enquiry is required to begin to disentangle some of these factors. However, this example has allowed

<sup>6</sup> To note, offenses sentenced in the Crown Court in England and Wales from which the CCSS is drawn, are likely to be more serious in nature than those sentenced in the magistrates' court, and are thus subject to greater sentencing powers of the Crown Court. Alternatively, offenders can opt for a trial by pleading not guilty, in which their case will be referred to the Crown Court for trial, in which case they cannot receive any reductions in their sentence if found guilty. Therefore, the average sentences reported here are likely to be more severe than those obtained for all theft offenses, and sexual offenses respectfully.



us to illustrate just one way in which the sentence severity scale can be used, which up to now could only be carried out with those offenses that received a custodial sentences.

## Discussion and Conclusions

In this section, we first discuss the proposed methodology, and then discuss the potential uses for the scale within criminal justice. Use of the Goodman RC association models for scaling is well established but perhaps unexplored in criminology and criminal justice. Clogg (1982) established four principles for the model to be used for scaling, and we have shown that our analysis satisfies these principles. We have extended the model to allow for covariates on the severity scale; work that we believe is new in the field. This method allows us to control for certain legal factors in the development of the scale and could be expanded to control for further legal factors considered to be relevant (Ashworth and Kelly 2021), and where the data is available. This could be in relation to the offense and the offender (Guilfoyle et al. 2024). A possible disadvantage of the method is that it works on a categorical scale, but this is not a serious problem.

Through extending the RC association model to control for a number of legal factors taken into consideration at sentencing, we were able to create a sentence severity scale that includes the broad range of sentences available to the courts in England and Wales. We were thus able to align and calibrate all sentences onto a single scale measuring sentence severity and converting the scores into their equivalent number of days in custody. The results obtained from the Extended Goodman RC model provide sensible and compelling ordering of the different sentencing categories used within the England and Wales court system. As we would expect, non-custodial sentences, particularly fines, 'other', and conditional discharges dominate at the less severe end of the severity spectrum, while lengthy custodial sentences dominate at the more severe end of the scale. Community orders and suspended sentences interweave with custodial sentences consistent with von Hirsch's (1993) ordinal proportionality and Partial Substitution Model and Tonry's (1994) notion of interchangeability. Of course, more detailed sentencing information relating to the requirements attached to these orders would enable further refinement of the scale, particularly for the community orders but also for the suspended sentences. However, it is reassuring to observe that suspended sentences are being used appropriately and in line with the equivalent custodial sentences.

This particular methodology not only allows the different sentences to be incorporated into a single scale, it also allows for the overlapping of sentences, akin to von Hirsch's (1993) Partial Substitution Model. This enables us to address and explore the interchangeability of sentences, which suggests that different sentence types can be compared and even used interchangeably providing they carry an equivalent 'penal bite' or severity as viewed by the judge. Methods such as Tobit regression are unable to account for this overlapping of sentences, specifically non-custodial and custodial sentences. In this vein, there is the potential for such a scale to be utilised by, and as an aid to judges, to identify alternative non-custodial sentences that are the equivalent (severity) to custodial sentences, and thus may facilitate and justify the use of non-custodial sentences for appropriate defendants in an age of forever rising prison populations.

The scale can be used for numerous purposes. Firstly, change in sentence severity, as in the above example we gave for theft and sexual offenses over a 4-year period. In addition to this, the scale, can also be used as a tool to measure disparity in sentencing (see Wallace

2015). Building on the work of, for example; Britt (2009); Engen and Gainey (2000); Stefensmeier et al. (1993); Ulmer and Johnson (2004); Johnson (2006); Pina-Sánchez and Linacre (2013); and Weidner et al. (2005), it is possible to extend this research to also include non-custodial sentences by way of the sentence severity scale. In this case, the scale could be used as a dependent variable in a regression or multi-level model, in which control covariates can also be added to account for any sentencing variation.

*Algorithmic and Modelling Issues:* We have presented an example with 29 sentence categories and 111 offense types from just over 60,000 proven offenses. A larger sample of offenses would help to improve the severity scale by increasing the number of sentence and offense categories. For example, information relating to the type and number of requirements attached to community orders would help to increase and refine the number of sentence categories. Similarly, it would also be possible to extend the scale to incorporate multiple sentences and combinations of sentences. This additional information would enable finer gradation of the scale. There are no software restrictions on fitting these models, which are limited only by the internal memory of the computer.

Additionally, the Extended Goodman RC association model could be extended further to control for additional covariates if available, such as information relating to any aggravating or mitigating factors taken into consideration, or more specific information relating to the offense. Adding more covariates to the model would create additional layers, and with that, additional cells within the table, making the model more complex. The upside of this would be to create a model that more closely mimics the judges decision-making process in coming to a final sentencing decision.

There is also scope for algorithmic development. One problem identified is that the parameters are unconstrained, and so the score for (say) a custodial sentence of 2 years could be higher than a score for a custodial sentence of 3 years. This issue is discussed in the context of the Goodman RC model by Bartolucci and Forcina (2002) who discuss modelling with order restrictions. A Bayesian approach with order constraints approach is described by Ben-Shachar (2023) in the context of archaeological data. Using a constrained maximum likelihood fitting procedure would also be a route around this problem, although a large number of constraints would be necessary and convergence properties of the likelihood would need to be investigated. An alternative is to apply equality constraints to the sentence severity scores in the extended Goodman model. For the two-way model, Goodman refers to this as the  $RC^{(0)}$  association model (Goodman 1986). This can be carried out easily by collapsing sentencing categories in the non-linear part of the model.

It needs to be emphasised that the sentencing scale calibrated here is appropriate for England and Wales, and therefore any scale of this nature would need to be recalibrated for any new jurisdiction. This is because other jurisdictions have different ideas about the relative importance of fines, community sentences and custody. Additionally, change in sentencing behaviour by judges over time (perhaps driven by societal changes and changes in government guidelines) may mean that as time passes the method needs to be recalibrated to account for these changes, and we would suggest that a period of 10 to 15 years should be appropriate. To recalibrate the model is straightforward: new data would be collected from sentencing databases, then model would be rerun with available control variables, numeric scores extracted, and the jurisdiction-specific scale to the sentences would be applied (see Supplementary material 2).

One referee raised the possibility of estimating more than one dimension of sentencing. While the major dimension is severity (as it is aligned with offence seriousness), an additional dimension might be “degree of offender control”, where fines have the least control over an offender, and custody provide most control. The Goodman RC model can easily be

extended to multiple dimensions (see Goodman 1986, Section 4) and can be fitted using the same R package. This is additional work beyond the scope of this paper.

In summary, the Extended Goodman RC association model with covariates provides a suitable method to solve the thorny issue of placing court sentences onto a common scale. Different jurisdictions will have different types of sentence (for example the US distinguishes between state jails and county jails) and the method will need to be adapted for the specific issues in these jurisdictions. However, the method has great potential and we encourage its further use and development.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10940-024-09591-6>.

**Acknowledgements** The authors wish to thank the referees, associate editors and editors for numerous insightful and helpful comments which have improved this paper.

**Data Availability** The dataset of conviction records used to calibrate the sentencing scale is not publicly available for data confidentiality reasons. However, a synthetic version of the data constructed using the method of Jackson et al. (2022) has been built and is available at <https://doi.org/10.17635/lancaster/researchdata/670>. The method and R code used to synthesise the data is fully documented in Supplementary material 1 and the R code for constructing the sentencing scores from this data is available in Supplementary Material 2.

## Declarations

**Conflict of interest** Partial financial support was received from the Economic and Social Research Council (ESRC).

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Agresti A (2013) *Categorical data analysis*, 3rd edn. Wiley, Hoboken
- Albonetti CA (1997) Sentencing under the federal sentencing guidelines: effects of defendant characteristics, guilty pleas, and departures on sentence outcomes for drug offenses, 1991–1992. *Law Soc Rev* 31(4):789–822. <https://doi.org/10.2307/3053987>
- Allen RB, Anson RH (1985) Development of a punishment severity scale: the item displacement phenomenon. *Crim Justice Rev* 10(2):39–44. <https://doi.org/10.1177/073401688501000207>
- Anderson AL, Spohn C (2010) Lawlessness in the federal sentencing process: a test for uniformity and consistency in sentence outcomes. *Justice Q* 27(3):362–393. <https://doi.org/10.1080/07418820902936683>
- Ashby MP (2018) Comparing methods for measuring crime harm/severity. *Polic: J Policy Pract* 12(4):439–454. <https://doi.org/10.1093/police/pax049>
- Ashworth A, Kelly R (2021) *Sentencing and criminal justice*. Bloomsbury Publishing, London
- Bartolucci F, Forcina A (2002) Extended RC association models allowing for order restrictions and marginal modeling. *J Am Stat Assoc* 97(460):1192–1199. <https://doi.org/10.1198/016214502388618988>
- Ben-Shachar MS (2023) Order constraints in Bayes models (with brms). <https://blog.msbstats.info/posts/2023-06-26-order-constraints-in-brms/>

- Bergman MM, Lambert P, Prandy K, Joye D (2002) Theorization, construction, and validation of a social stratification scale: Cambridge social interaction and stratification scale (CAMISIS) for Switzerland. *Schweiz Zeitschrift für Soziol* 28(1):7–26
- Blumstein A (1974) Seriousness weights in an index of crime. *Am Sociol Rev* 39(6):854–864. <https://doi.org/10.2307/2094158>
- Britt CL (2009) Modeling the distribution of sentence length decisions under a guidelines system: an application of quantile regression models. *J Quant Criminol* 25(4):341–370. <https://doi.org/10.1007/s10940-009-9066-x>
- Buchner D (1979) Scale of sentence severity. *J Crim Law Criminol* 70(2):182–187
- Bushway SD, Piehl AM (2001) Judging judicial discretion: legal factors and racial discrimination in sentencing. *Law Soc Rev* 35(4):733–764. <https://doi.org/10.2307/3185415>
- Clogg CC (1982) Association models in sociological research: some examples. *Am J Sociol* 88(1):114–134. <https://doi.org/10.1086/227636>
- Cohen MA (1988) Some new evidence on the seriousness of crime. *Criminology* 26(2):343–353. <https://doi.org/10.1111/j.1745-9125.1988.tb00845.x>
- Crown court sentencing survey. <https://www.sentencingcouncil.org.uk/research-and-resources/data-collections/crowncourt-sentencing-survey/>. Online; Accessed 17-July-2024
- Douglas T (2024) Criteria for assessing AI-based sentencing algorithms: a reply to Ryberg. *Philos Technol*. <https://doi.org/10.1007/s13347-024-00722-2>
- Engen RL, Gainey RR (2000) Modeling the effects of legally relevant and extralegal factors under sentencing guidelines: the rules have changed. *Criminology* 38(4):1207–1230. <https://doi.org/10.1111/j.1745-9125.2000.tb01419.x>
- Erickson ML, Gibbs JP (1979) On the perceived severity of legal penalties. *J Crim Law Criminol* 70(1):102–116
- Erikson R, Goldthorpe JH (1992) *The constant flux: a study of class mobility in industrial societies*. Clarendon Press, Oxford
- Francis B, Soothill K, Dittrich R (2001) A new approach for ranking ‘serious’ offences. The use of paired-comparisons methodology. *Br J Criminol* 41(4):726–737
- Goodman LA (1979) Simple models for the analysis of association in crossclassifications having ordered categories. *J Am Stat Assoc* 74(367):537–552. <https://doi.org/10.1080/01621459.1979.10481650>
- Goodman LA (1981) Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *J Am Stat Assoc* 76(374):320–334. <https://doi.org/10.1080/01621459.1981.10477651>
- Goodman LA (1986) Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *Int Stat Rev* 54(3):243–270. <https://doi.org/10.2307/1403053>
- Goodman LA, Hout M (1982) Statistical methods and graphical displays for analysing how the association between two qualitative variables differs among countries, or over time. Part II: some explanatory techniques, simple models and simple examples. *Sociol Methodol* 31:189–221. <https://doi.org/10.1111/0081-1750.00095>
- Gormley J, Roberts J, Pina-Sánchez J, Tata C, Navarro A (2022) The Methodological challenges of comparative sentencing research: literature review. scottish sentencing council. <https://www.scottishsentencingcouncil.org.uk/media/nvniyfn/20220519-research-comparing-sentencing-final-as-published.pdf>
- Guilfoyle E, Pina-Sánchez J (2024) Racially determined case characteristics: exploring disparities in the use of sentencing factors in England and Wales. *Br J Criminol*. <https://doi.org/10.1093/bjc/azae039>
- Harlow RE, Darley JM, Robinson PH (1995) The severity of intermediate penal sanctions: a psychophysical scaling approach for obtaining community perceptions. *J Quant Criminol* 11(1):71–95. <https://doi.org/10.1007/BF02221301>
- Hartel P, Wegberg R, van Staaldouin M (2023) Investigating sentence severity with judicial open data. *Eur J Crim Policy Res* 29:579–599. <https://doi.org/10.1007/s10610-021-09503-5>
- Helms R, Jacobs D (2002) The political context of sentencing: an analysis of community and individual determinants. *Soc Forces* 81(2):577–604. <https://doi.org/10.1353/sof.2003.0012>
- Holleran D, Spohn C (2004) On the use of the total incarceration variable in sentencing research. *Criminology* 42(1):211–240. <https://doi.org/10.1111/j.1745-9125.2004.tb00518.x>
- Home Office (2014) police national computer (PNC). <https://www.gov.uk/government/publications/police-national-computer-pnc>
- Jackson J, Mitra R, Francis B, Dove I (2022) Using saturated count models for user-friendly synthesis of large confidential administrative databases. *J R Stat Soc Ser A Stat Soc* 185(4):1613–1643. <https://doi.org/10.1111/rssa.12876>

- Johnson BD (2006) The multilevel context of criminal sentencing: integrating judge and country-level influences. *Criminology* 44(2):259–298. <https://doi.org/10.1111/j.1745-9125.2006.00049.x>
- Kärholm F, Neyroud P, Smaaland J (2020) Designing the Swedish crime harm index: an evidence-based strategy. *Camb J Evid-Based Polic*. <https://doi.org/10.1007/s41887-020-00041-4>
- Kateri M (2014) Contingency table analysis: methods and implementation using R. Birkhauser, Secaucus
- Klepper S, Nagin D, Tierney L (1983) Discrimination in the criminal justice system: a critical appraisal of the literature. In: Blumstein A (ed) *Research on sentencing: the search for reform*. National Academy Press, Washington
- Leclerc C, Tremblay P (2016) Looking at penalty scales: how judicial actors and the general public judge penal severity. *Can J Criminol Crim Justice* 58(3):354–384. <https://doi.org/10.3138/cjccj.2014.E31>
- Malila IS (2012) Severity of multiple punishments deployed by magistrate and customary courts against common offences in Botswana: a comparative analysis. *Int J Crim Justice Sci* 7(2):618–634
- McDavid JC, Stipak B (1981) Simultaneous scaling of offense seriousness and sentence severity through canonical correlation analysis. *Law Soc Rev* 16(1):147–162. <https://doi.org/10.2307/3053555>
- Merrall EL, Dhami MK, Bird SM (2010) Exploring methods to investigate sentencing decisions. *Eval Rev* 34(3):185–219. <https://doi.org/10.1177/0193841X10369624>
- Messner SF, Deane G, Beaulieu M (2002) A log-multiplicative association model for allocating homicides with unknown victim-offender relationships. *Criminology* 40(2):457–480. <https://doi.org/10.1111/j.1745-9125.2002.tb00963.x>
- Miethe TD, Moore CA (1986) Racial differences in criminal processing: the consequences of model selection on conclusions about differential treatment. *Sociol Q* 27(2):217–237. <https://doi.org/10.1111/j.1533-8525.1986.tb00258.x>
- Ministry of Justice (2010) Criminal statistics: England and Wales 2009. <https://www.gov.uk/government/statistics/criminal-statistics-annual-report-ns>
- Ministry of Justice (2023) Prison population projections 2022 to 2027, England and Wales. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/1138135/Prison\\_Population\\_Projections\\_2022\\_to\\_2027.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1138135/Prison_Population_Projections_2022_to_2027.pdf). Online; Accessed 03-July-2023
- Minnesota Sentencing Guidelines Commission (2013) Minnesota sentencing guidelines & commentary. [http://mn.gov/sentencing-guidelines/images/2014\\_2520Guidelines.pdf](http://mn.gov/sentencing-guidelines/images/2014_2520Guidelines.pdf). Minnesota Sentencing Guidelines Commission, Minnesota
- Morris N, Tonry M (1990) *Between prison and probation: intermediate punishments in a rational sentencing system*. Oxford University Press, New York
- Mueller-Johnson KU, Dhami MK (2009) Effects of offenders' age and health on sentencing decisions. *J Soc Psychol* 150(1):77–97. <https://doi.org/10.1080/00224540903365315>
- Mustard DB (2001) Racial, ethnic, and gender disparities in sentencing: evidence from the US federal courts. *J Law Econ* 44(1):285–314. <https://doi.org/10.1086/320276>
- Nye FI, Short JF (1957) Scaling delinquent behavior. *Am Sociol Rev* 22(3):326–331. <https://doi.org/10.2307/2088474>
- Office for National Statistics (2016) Research outputs: developing a crime severity score for England and Wales using data on crimes recorded by the police. ONS, London
- Office for National Statistics (2023) Crime in England and Wales: appendix tables. Year ending March 2023, Table A4a. <https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/datasets/crimeinenglandandwalesappendixtables>. Online; Accessed 30-May-2024
- Osgood DW, McMorris BJ, Potenza MT (2002) Analyzing multiple-item measures of crime and deviance i: item response theory scaling. *J Quant Criminol* 18(3):267–296. <https://doi.org/10.1023/A:1016008004010>
- Pina-Sánchez J, Gosling JP (2022) Enhancing the measurement of sentence severity through expert knowledge elicitation. *J Legal Res Methodol*. <https://doi.org/10.19164/jlrm.v2i1.1241>
- Pina-Sánchez J, Gosling JP, Chung H-I, Bourgeois E, Geneletti S, Marder ID (2019) Have the England and Wales guidelines affected sentencing severity? An empirical analysis using a scale of severity and time-series analyses. *Br J Criminol* 59(4):979–1001. <https://doi.org/10.1093/bjc/azz005>
- Pina-Sánchez J, Grech DC (2017) Location and sentencing: To what extent do contextual factors explain between court disparities? *Br J Criminol* 58(3):529–549. <https://doi.org/10.1093/bjc/azx033>
- Pina-Sánchez J, Linares R (2013) Sentence consistency in England and Wales evidence from the crown court sentencing survey. *Br J Criminol* 53(6):1118–1138. <https://doi.org/10.1093/bjc/azt040>
- Powers DA, Xie Y (2000) *Statistical methods for categorical data analysis*. Academic Press, London
- R Core Team (2023) R: a language and environment for statistical computing [Computer software manual]. Vienna, Austria. Accessed from <https://www.R-project.org/>
- Ramchand R, MacDonald JM, Haviland A, Morral AR (2009) A developmental approach for measuring the severity of crimes. *J Quant Criminol* 25(2):129–153. <https://doi.org/10.1007/s10940-008-9061-7>

- Ryberg J (2024) Criteria for assessing AI-based sentencing algorithms: a reply to Douglas. *Philos Technol*. <https://doi.org/10.1007/s13347-024-00726-y>
- Schiff MF (1997) Gauging the intensity of criminal sanctions: developing the criminal punishment severity scale (CPSS). *Crim Justice Rev* 22(2):175–206. <https://doi.org/10.1177/073401689702200204>
- Sebba L (1978) Some explorations in the scaling of penalties. *J Res Crime Delinq* 15(2):247–265. <https://doi.org/10.1177/002242787801500209>
- Sebba L, Nathan G (1984) Further explorations in the scaling of penalties. *Br J Criminol* 24:221–249. <https://doi.org/10.1093/oxfordjournals.bjc.a047451>
- Sellin T, Wolfgang ME (1964) *The measurement of delinquency*. Sentencing council. Wiley, Oxford, p 2014
- Sentencing Council (2017) *Imposition of community and custodial sentences*. <https://www.sentencing-council.org.uk/overarching-guides/magistrates-court/item/imposition-of-community-and-custodial-sentences/>. Online; Accessed 04-April-2024
- Sentencing Council (2019) *General guideline: overarching principles*. <https://www.sentencingcouncil.org.uk/overarching-guides/crown-court/item/general-guideline-overarching-principles/>. Online; Accessed 04-April-2024
- Sherman L, Neyroud PW, Neyroud E (2016) The Cambridge crime harm index: measuring total harm from crime based on sentencing guidelines. *Polic: J Policy Pract* 10(3):171–183. <https://doi.org/10.1093/police/paw003>
- Steffensmeier D, Kramer J, Streifel C (1993) Gender and imprisonment decisions. *Criminology* 31(3):411–446. <https://doi.org/10.1111/j.1745-9125.1993.tb01136.x>
- Thurstone LL (1994) A law of comparative judgment. *Psychol Rev* 101(2):266–270. <https://doi.org/10.1037/h0070288>
- Tiffany LP, Avichai Y, Peters GW (1975) Statistical analysis of sentencing in federal courts: defendants convicted after trial, 1967–1968. *J Leg Stud* 4:369–390. <https://doi.org/10.1086/467537>
- Tonry M (1994) Proportionality, parsimony, and interchangeability of punishments. In: Duff A, Garland D (eds) *A reader on punishment*. Readings on socio-legal studies. Oxford University Press, Oxford
- Tremblay P (1988) On penal metrics. *J Quant Criminol* 4(3):225–245. <https://doi.org/10.1007/BF01072451>
- Turner H, Firth D (2023) *Generalized nonlinear models in R: An overview of the gnm package [Computer software manual]*. <https://cran.r-project.org/web/packages/gnm/vignettes/gnmOverview.pdf>. (R package version 1.1-5)
- Ulmer JT, Johnson B (2004) Sentencing in context: a multilevel analysis. *Criminology* 42(1):137–178. <https://doi.org/10.1111/j.1745-9125.2004.tb00516.x>
- von Hirsch A (1993) *Censure and sanctions*. Clarendon Press, Oxford
- von Hirsch A, Wasik M, Greene J (1989) Punishments in the community and the principles of desert. *Rutgers Law J* 20(3):595–618
- Walker R (2010) *Treating like cases alike and unlike cases differently: Some problems of antidiscrimination law*. [https://www.supremecourt.uk/docs/speech\\_100809.pdf](https://www.supremecourt.uk/docs/speech_100809.pdf). ([Speech given in Otago, August 2010. Online; Accessed 17-December-2017])
- Wallace SJ (2015) *Developing methods to measure sentence severity and assess disparity in sentencing in the criminal courts of England and Wales (Doctoral dissertation, Lancaster University)*. (Accessed 22 May 2024) <https://doi.org/10.17635/lancaster/thesis/2334>
- Weidner RR, Frase RS, Schultz JS (2005) The impact of contextual factors on the decision to imprison in large urban jurisdictions: a multilevel analysis. *Crime Delinq* 51(3):400–424. <https://doi.org/10.1177/001128704271467>
- Wheeler S, Weisburd D, Bode N (1982) Sentencing the white-collar offender: rhetoric and reality. *Am Sociol Rev* 47:641–659. <https://doi.org/10.2307/2095164>
- Wolfgang ME, Figlio RM, Tracy PE, Singer SI (1985) *The national survey of crime severity*. US Government Printing Office, Washington
- Xie Y (1992) The log-multiplicative layer effect model for comparing mobility tables. *Am Sociol Rev* 57(3):380–395. <https://doi.org/10.2307/2096242>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.