Journal of Legal Research Methodology

Enhancing the Measurement of Sentence Severity through Expert Knowledge Elicitation

Jose Pina-Sánchez · John Paul Gosling

Abstract Quantitative research on judicial decision-making faces the methodological challenge of analysing disposal types that are measured in different units (e.g. money for fines, days for custodial sentences). To overcome this problem a wide range of scales of sentence severity have been suggested in the literature. One particular group of severity scales that has achieved high validity and reliability are those based on Thurstone's pairwise comparisons. However, this method invokes a series of simplifying assumptions, one of them being that the range of severity covered by different disposal types is constant. We undertook an expert elicitation workshop to assess the validity of that assumption. Responses from the six criminal law practitioners and researchers that participated in our workshop unanimously pointed at severity ranges being highly variable across disposal types (e.g. much wider severity ranges were identified for suspended custodial sentences than for fines). We used this information to re-specify Thurstone's model allowing for unequal variances. As a result, we obtained a new, more robust, scale of sentence severity.

Keywords Sentencing · severity · Thurstone scale · expert elicitation

This work was supported by the National Centre for Research Methods [RIS14241/06]. We are also immensely grateful to the six sentencing experts that participated in our workshop: Julian Berg, Elizabeth Bourgeois, David Hayes, Eleanor Nicholls, Ruth Pope, and Sebastian Walker. This study would not have been possible without the time and unique subject knowledge that they so generously offered.

J. Pina-S´anchez
School of Law, University of Leeds Liberty Building, University Western Campus, Moorland Rd, Leeds LS3 1DB, UK
ORCID: 0000-0002-9416-6022
E-mail: j.pinasanchez@leeds.ac.uk

J.P. Gosling
School of Mathematics, Durham University, UK
ORCID: 0000-0002-4072-3022

1 Introduction

Most quantitative sentencing research seeks to explore the causal mechanisms involved in judicial decision-making, a task that is normally undertaken by studying the variability in the relative severity of large samples of sentences. Compared to other legal or criminal justice research areas, the formal setting in which the sentencing process takes place simplifies its analytical complexity. For example, the outcome variable (i.e. the sentence imposed) is often unequivocally recorded (i.e. measured without error), reverse causality is non-existent since the path from case processing to sentence is unidirectional, and although problems of unobserved confounders are as common as in any other area relying on observational data, sentencing researchers rarely fail to acknowledge this problem and often anticipate its biasing effect. There is, however, one particular methodological challenge affecting sentencing research specifically. This problem stems from the fact that judges can rely on a wide range of disposal types (i.e. sentence types, or sentence outcomes), which are not readily comparable, since they are measured in different units (Freiberg and Fox, 1986). For example, fines can be expressed in pounds, while custodial sentences are measured in days, while community orders are normally based on the completion of certain conditions.

Researchers have often addressed this problem by either restricting their analysis to one disposal type - commonly custodial sentence length - or by specifying the probability of custody compared to other possible outcomes. The former approach induces selection bias (Bushway et al., 2007) since custodial sentences are the most severe of all possible disposal types, representing a small part of all sentences imposed (roughly 7% of the total in England and Wales). The latter involves a substantial loss of information, since it takes all non-custodial outcomes as a homogeneous group, rendering sentencing analyses unduly blunt and leading to forms of measurement error (Berkson, 1950). More methodologically advanced researchers have sought to adjust fore selection bias using Tobit (Albonetti, 1998; King et al., 2010; Kurlychek and Johnson, 2010), hurdle (Hester and Hartman, 2017), or Heckman's two-stage (Feldmeyer and Ulmer, 2011; Steffensmeier and DeMuth, 2001; Ulmer et al., 2010) models. However, these models are also based on different assumptions, e.g. that the type and quantity of the sentence outcome are decided in different steps of the sentencing process, or that the unobserved severity of non-custodial outcomes stems from a hypothetical distribution, which observed, right-hand side is represented by the length of custodial sentences. Pina-S´anchez and Gosling (2020) demonstrated how, at least for the case of England and Wales, those two assumptions are violated. Perhaps more importantly, all of the statistical adjustments that have been suggested in the literature, involve discarding any variability recorded across non-custodial outcomes.

An alternative strategy relies on adopting a scale of sentence severity (Leclerc and Tremblay, 2016; Yan and Lao, 2021). This involves assuming an underlying continuum along which sentence outcomes can be located; so they can all be expressed under the same measurement unit, preventing any loss of information. Unlike the more formally defined statistical adjustments used in the literature, the estimation of scales of severity represents a widely heterogeneous approach, manifested in multiple types of solutions, none of them without limitations. The most common approach is magnitude escalation, which involves establishing a reference sentence, and deriving the relative severity of other sentence outcomes from the subjective comparisons made by a sample of informed participants (Harlow et al., 1995; Spelman, 1995; Tremblay, 1988). In our view, the main problem affecting these types of studies stems from the wide variability in participants' responses, making scales highly sensitive to the composition of the sample studied, which compromises their reliability. Other studies have relied on data-driven methods such as correspondence analysis (Francis et al., 2005; McDavid and Stipack, 1981). Although these methods remove the inherent unreliability of subjective perceptions, they derive the relative severity of different sentence outcomes from the frequency with which they are used as punishments for different crime types. This in turn invokes further assumptions, such as perfect proportionality between crime seriousness and sentence severity, which, if violated can lead to nonsen-

sical severity scores, such as longer suspended sentences deemed more severe than shorter ones; hence, questioning the validity of such data-driven scales.

Our preferred approach for the estimation of sentence severity is Thurstone (1927) pairwise comparisons. This method is based on subjective perceptions (Buchner, 1979; Pina-S´anchez et al., 2019a), but rather than requesting research subjects to report the magnitude by which a sentence outcome is more severe than another, it relies on either ordinal comparisons, or alternatively, on relative comparisons requesting how often - rather than how much - one sentence outcome is more severe than another. That is, the cognitive burden placed on research participants is eased by asking them to identify which of the two outcomes is more severe, or by framing the comparison in terms of frequencies.[1] However, the method is also reliant on a series of parametric assumptions. The potential severity associated with each sentence outcome is assumed to be normally distributed. The means of these distributions vary, reflecting the severity scores attributed to each sentence outcome, but their variances are assumed to be equal. This is a convenient assumption that makes the estimation process more parsimonious. We should nonetheless question its validity, since it involves assuming that the range of severity scores covered by each sentence outcome is equivalent across all of them. For example, using the standard Thurstone model, we might find that community orders are in general more severe than fines, but we will have to assume that the difference between the minimum and maximum severity attributed to each of these outcomes is the same. Importantly, if this assumption is violated, the severity scores estimated for each sentence outcome will be biased, as they won't be accurately reflecting their relative distance in the underlying scale of severity.

In this article we demonstrate how the assumption of equal variances can be relaxed using expert knowledge elicitation techniques. We do so by revisiting the Thurstone scales of severity estimated in Pina-Sanchez et al. (2019a) and Pina-S´anchez and Gosling (2020), using qualitative insights elicited from six criminal law experts and an extended version of the standard Thurstone model. As such, our article contributes to both the measurement and sentencing literature in three important ways: i) we show the importance of testing the underlying assumptions of Thurstone scales; ii) demonstrate the multiple advantages of relying on expert elicitation methods for the estimation of sentence severity; and iii) estimate a more robust scale of severity with which to undertake sentencing research in the jurisdiction of England and Wales. The creation of this new scale of severity represents a timely contribution as studies employing severity scales are becoming more widespread (Roberts and Bild, 2021; Isaac, 2021), but also as new and more sophisticated sentencing datasets that capture differences across non-custodial sentences in unprecedented detail, are becoming increasingly available (Sentencing Council, 2021; Ministry of Justice, 2021).

2 Challenging the Assumption of Equal Variances

The Thurstone model works by linking each sentence outcome to a corresponding set of latent normal distributions that have the following property: the probability of a random draw from distribution A exceeding a random draw from distribution B is equal to the proportion of times sentence outcome A will be preferable to sentence outcome B (Thurstone, 1927; Pina-S´anchez and Gosling, 2020). The simplest - and most commonly employed - adaptation of the Thurstone model is in its 'Case V' form (Mosteller, 1951). Under this setting, the latent normal distributions for each sentence outcome are assumed to have equal variance of a half, so that the differences between sentence outcomes have a variance of one. Given a set of judged preference proportions, the Thurstone model is fitted and the means of the resulting latent normal distributions can be used as a ranking or scoring metric. Lastly,

[1] By requesting frequency rather than magnitude comparisons, responses are framed within a 0 to 1 range - as opposed to the 0 to ∞ range involved in magnitude comparisons - which provides intuitive points of reference at 0 (never more severe), 0.5 (as severe) and 1 (always more severe).

because the choice of variance of the latent variables is arbitrary, it is valid - and common - for the metrics to be rescaled to fit a more convenient range (0-100, for instance).

This assumption of equal variances is possibly robust enough in many contexts where Thurstone scaling is employed, but should be questioned when we consider the relative severity of different sentence outcomes. We argue that the range of severity that could be covered by different disposal types is likely proportional to their heterogeneity. For example, fines can only be expressed in pounds, and have specific bands associated to them, whereas community orders or suspended sentences could be composed of a wide range of conditions, such as curfews, completion of rehabilitative programs, unpaid work, and they could also include fines too. Therefore, given that fines are just a subset of the possible sentence outcomes that could make part of a community order sentence, it is not tenable to see the two disposal types as covering equal ranges of potential severity, i.e. the former will always be narrower than the latter. Once established the lack of theoretical soundness of this assumption, it is worth considering how exactly could results from the Thurstone scale (severity scores for different sentence outcomes) be affected.

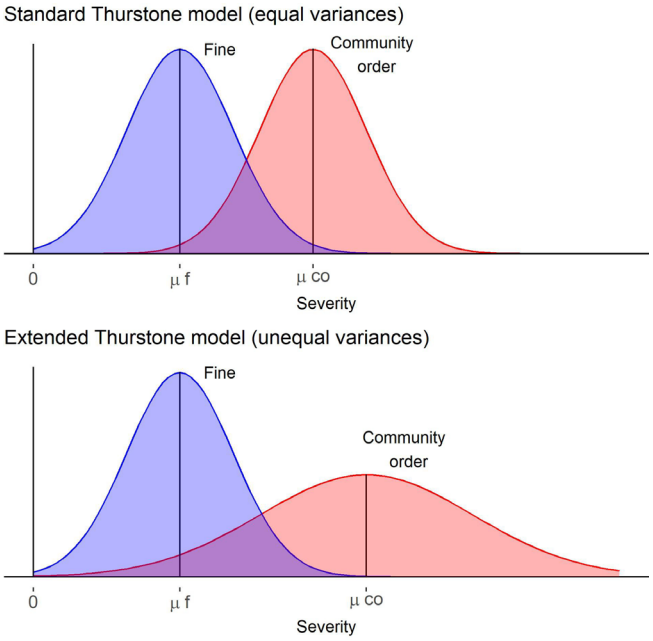

Fig. 1 Simplified representation of the Thurstone scaling method under the assumption of equal variances (top) and allowing for different variances (bottom).

The implications of violating the hypothesis of equal variances is shown visually in Figure 1, based on a simple - made up - example showing the severity scores for two disposal types, fines and community orders, where pairwise comparisons found the latter leading to more severe outcomes than the former in 80% of instances where any of these two disposal are imposed. As previously noted, the severity score for each disposal type (μ_F for fines and μ_{CO} for community orders) is determined by the mean of their respective latent distributions, with those distributions situated as far apart from each other as indicated by the estimates of relative severity derived from their pairwise comparisons. For example, if community orders are considered more severe than fines 50% of times, the two distributions, and their respective means, would be placed at the exact same location (i.e. complete overlap), while a result of

100% would place the minimum severity in the distribution of community orders to follow exactly the maximum severity in the distribution of severity of fines (i.e. no overlap).

Notice as well how it is not just the extent of the overlap between outcomes that determines their mean score, but also the shape of the distributions that they are assumed to follow. The top part of Figure 1 shows the result for the standard Thurstone model, where all outcomes are assumed to be normally distributed with equal variances. The bottom part shows how, for the same level of overlap as on the top graph (20%), the severity score for community order is now higher than before as a result of having considered a latent distribution with a standard deviation twice larger than that used to describe fines. In sum, the severity scores derived from Thurstone scales should be seen - at least in principle - as highly sensitive to the underlying ranges of severity assumed for each of the sentence outcomes considered. We now proceed to investigate such proposal empirically.

3 The Expert Elicitation Workshop

To explore the extent to which ranges of severity vary across different sentence outcomes, we organised an expert elicitation workshop. Simply put, expert elicitation techniques are similar to focus groups with two main distinctions: i) the participants are experts on a given field; and ii) the goal is to retrieve estimates for one or a series of well-defined numerical parameters, normally taking the form of the probability of a given event, or a probability distribution across multiple outcomes (O'Hagan et al., 2006). In addition, to aid repeatability and to help avoid cognitive and social biases, structured elicitation sessions are devised around well-design protocols.

In our case, the Sheffield elicitation protocol was employed, as it is specifically designed to facilitate group judgements about complex quantities whilst recording key reasoning and evidence used by the experts in forming their judgements (Gosling, 2018). The Sheffield elicitation protocol has been used extensively in many areas of science (Dessai et al., 2018; Jansen et al., 2020; Booth and Thomas, 2021) and policy making (Gosling et al., 2012; Usher and Strachan, 2013; Brennan et al., 2017). The outcome of our structured elicitation exercise was a set of judgements on sentencing severity along with documented reasoning, giving ownership to the entire group, i.e. the group discussions and subsequent recording were conducted under the Chatham House rule.

The workshop took place on the 4th of December 2018, at the London Mathematical Society (Morgan House, Russell Sq). All six experts initially approached accepted our invitation. The group of experts was designed to meet two criteria: i) each individual participant should hold expert knowledge about the sentencing process in England and Wales; and ii) the group should reflect different forms of applied and theoretical expertise, including policy-makers, academics, and practitioners. In alphabetical order, the participating experts in our workshop were: Julian Berg (Criminal Law Solicitor's Association), Elizabeth Bourgeois (Bradford and Keighley Magistrates Court), David Hayes (University of Sheffield), Eleanor Nicholls and Ruth Pope (both from the Sentencing Council for England and Wales), and Sebastian Walker (Law Commission). The authors of this article conducted the workshop, which was divided in two parts, for a total of four hours.

The first and longest part of the workshop involved eliciting estimates of the severity overlap across different pairs of sentences. To frame the group discussions we posed specific questions. However, given the complexity of the topic, we formulated the main question in two different ways. We asked: i) *How often can sentence-X be more punitive than sentence-Y?* ; and ii) *What proportion of offenders would prefer sentence-X than sentence-Y?* To provide further context to the discussion, we also asked our participants to consider the following: i) the heterogeneity of sentences possible within a given disposal type (e.g. the different conditions that could be attached to a long community order); ii) that we are not asking whether the different sentences can be used interchangeably, but rather whether there are circumstances when one can have a more punitive effect than the other; and iii) to consider not just the average offender, but the mix of different offenders seen through courts. Discussions for each of the

pairs of sentence outcomes compared were allowed to take as long as necessary, until a consensus was reached and a specific estimate could be derived.

The format used represents a substantial departure from the typical approach followed in applications of Thurstone's scaling. In its original form, overlaps between categories are estimated by repeating the same pairwise comparison over a large sample of participants using questionnaires. Here, we are asking our participants to identify not only the most severe of two outcomes, but we also request them to estimate how often they think that would be the case. As a result, the questions asked are more cognitive demanding, which is why they are ideally explored under a format that allows for in depth discussion. There is therefore a trade-off between external and internal validity. Such format cannot be easily scaled up to obtain a large sample of expert views. However, given the complexity of the questions asked, we follow Bolger (2018) and O'Hagan (2019) in setting a low number of experts, since the possibly lower reliability associated with a small sample is more than offset from the gains in focus and accuracy made possible through an expert elicitation workshop.

The number of pairs of sentence outcomes compared expanded those considered in Pina-S´anchez et al. (2019a). This was to capture the higher granularity with which non-custodial sentences are recorded by the latest data releases from the Sentencing Council for England and Wales, which disaggregates fines within six bands (A, B, C, D, E, and F), reflecting different quantities, and community orders in three categories (low, medium and high), based on the conditions that could be imposed.[2] This gave us eighteen outcomes to be compared, meaning 153 potential pairwise combinations. To limit participants' fatigue, we restricted our discussion to comparisons of 30 pairs of sentences where it could not be ruled out that one of the sentences would always be deemed more severe than the other, i.e. a severity overlap could be theoretically possible.

In the second part of the workshop we explored the extent to which severity ranges covered by each of the eighteen sentence outcomes explored could be considered equivalent across all of them, and if not, how different could they be. This is an even more complex question, hence, we approached it in two stages. First, we asked introductory questions requesting to identify the sentence outcome with a wider severity range out of a series of pairwise comparisons. After corroborating our working hypothesis (i.e. severity ranges vary widely across different sentences), we proceeded to estimate the relative range for each of the sentences considered. To do so we asked two questions: i) *Do you think [introduce specific disposal type] has a different 'severity spread' than the average disposal type in this list (e.g. than a medium community order)?* ; and ii) *Roughly, how much do you think the spread of severity differs from the average?*; to which the following list of answers were presented [3]: i) *A quarter*

[2] The quantity associated for each of the six fine bands can be found here: https://www.sentencingcouncil.org.uk/explanatory-material/magistrates-court/item/fines-and-financial-orders/approach-to-the-assessment-of-fines-2/2-fine-bands/; the types of conditions that could be attached to different forms of community orders are explained here: https://www.sentencingcouncil.org.uk/droppable/item/community-orders-table/. For a more detailed description of the different disposal types available to sentencers in England and Wales see Harris and Walker (2021).

[3] By comparing relative spread across sentence outcomes we are implicitly assuming that the underlying severity scores for each outcome are part of a ratio-level - rather than interval-level - variable. This implies considering our severity scale as bounded on the left by zero (representing 'no punishment'), but also that all levels of severity along the scale can be objectively defined. The latter is clearly not a realistic assumption since levels of severity are subjectively defined, and as such, direct comparisons of their intensity (e.g. outcome A is twice as severe as outcome B) are bound to be unstable across subjects. See for example Thomas et al. (2018), who demonstrated how, in similar workshops seeking to elicit arrest risks across crime scenarios, participants provided stable probabilistic perceptions of risk that were rank-stable within participants, but were also simultaneously arbitrary in the sense that the specific risk of arrest for each scenario was neither stable between individuals nor meaningful, a set of measurement properties that they deemed to reflect *coherent arbitrariness*. We argue, however, that to assume our severity scale possesses the characteristics of a ratio-level variable is well justified, as the reason we do so is to be able to relax an even less tenable assumption, namely that of equal variances. Put differently, while the assumption of seeing a severity index as a ratio-level variable is theoretically questionable, and the elicitation of comparisons of severity spread methodologically challenging, the assumption of equal variances is not just wrong but directly leads to biased estimates of severity derived from the Thurstone method.

of the average spread (25%); ii) *Half of the average spread (50%)*; iii) *About equal spread to the average (100%)*; iv) *Fifty percent wider than the average spread (150%)*; v) *Twice the average spread (200%)*; vi) *Three times the average spread (300%)*; vii) *Other*.

4 Results

Results from the pairwise severity comparisons are summarised in Table 1. Each cell reports the proportion of instances in which the severity of the sentence outcome on the top of the column could be deemed more severe than the corresponding sentence outcome on the left margin. As could be expected, most sentence outcomes at the two extremes of the distribution (such as discharges, or immediate custodial sentences) show little to no overlap with others, whereas the picture is muddier when we focus on sentences lying on the middle of the distribution, such as fines, community orders and suspended sentences. Notice for example how a band-F fine was identified to be, in some instances, more punitive than a suspended sentence, or how high community orders could similarly be judged, sometimes, more punitive than an immediate custodial sentence. These overlaps resonate well with the concept of 'penal exchangeability' (Freiberg and Fox, 1986; Lovegrove, 2001; Sebba and Nathan, 1984), which sees different disposal types not as discrete steps but as a range of choices that are not always differentiated by their punitive effect, but potentially by other sentencing goals such as restitution, rehabilitation, or incapacitation. Similarly, such overlaps in severity allow acknowledging the concept of 'penal subjectivity', that is, the variability in between-subject experiences of punitiveness resulting from identical sentences (Ginneken and Hayes, 2017; Hayes, 2016, 2018; Padfield, 2011).

Estimates of the variability of severity ranges across sentence outcomes are shown in Table 2. Compared to the reference case (medium community orders), our participants manifested that fines and immediate custodial sentences comprise a much narrower severity range, from around a third to a fourth of the range encompassed by medium community orders. Contrary to that, high community orders and suspended sentences where deemed to cover wider severity ranges, reflecting the ample discretion that judges could apply to configure such disposal types, and the much wider range of punitive outcomes that could arise as a result, e.g. from relatively lenient suspended sentences that will never be activated, to those composed of multiple punitive conditions that will potentially become activated, involving prison time. These results corroborate our suspicion regarding the implausibility of the equal variances assumption invoked in the standard Thurstone model when applied to the estimation of sentence severity.

To assess the implications of violating the equal variances assumption we compare two scales of severity; both of them derived from the pairwise severity comparisons shown in Table 1. However, one scale relies on the standard Thurstone model, assuming equal variances in the severity distributions of each sentence outcome considered, while the other incorporates the different expert elicited variances reported in Table 2.[4] The two scales are reported in Table 3. In both of them, severity scores for immediate custody sentences longer than three months were estimated as a second step by extrapolating linearly from the severity scores obtained from the Thurstone model for one, two, and three months custodial sentences.[5]

The most noticeable effect appears to be the wider range of severity now covered by community orders and suspended sentences, reflecting the substantial heterogeneity that characterises many of

[4] To allow for unequal variances the Thurstone model was estimated from scratch using R, as opposed to relying on built-in functions (e.g. *thurstone*, available in the package *psych* (Revelle, 2018)). The code used to account for unequal variances has been included in the Technical Appendix: R Code at the end of this article.

[5] Non-linear functions were also considered to reflect the marginally diminishing returns of severity that could be expected for every additional month in prison (Leclerc and Tremblay, 2016; Spelman, 1995). To do so, different rates of decay were considered based on insights elicited from our experts. We found that severity scores for sentences longer than five years varied widely depending on the rate of decay considered. Hence, to avoid introducing such a potential source of unreliability, we decided to employ somehow less realistic, but possibly more robust, linear extrapolations.

Table 1 Pairwise severity comparisons (each cell reports how often the sentence at the top of the column could be deemed more severe than the corresponding sentence on the left margin).

	absolute discharge	cond. discharge	fine A	fine B	fine C	fine D	fine E	fine F	low community order	medium community order	high community order	1m custody 6m suspended	1m custody 12m suspended	6m custody 6m suspended	12m custody 24m suspended	1m immediate custody	2m immediate custody	3m immediate custody
absolute discharge	0.50	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
conditional discharge	0	0.50	0.65	0.80	1	1	1	1	1	1	1	1	1	1	1	1	1	1
fine A	0	0.35	0.50	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
fine B	0	0.20	0	0.50	1	1	1	1	1	1	1	1	1	1	1	1	1	1
fine C	0	0	0	0	0.50	1	1	1	0.65	1	1	1	1	1	1	1	1	1
fine D	0	0	0	0	0	0.50	1	1	0.55	0.75	1	0.75	1	1	1	1	1	1
fine E	0	0	0	0	0	0	0.50	1	0.40	0.60	0.95	0.70	0.75	1	1	1	1	1
fine F	0	0	0	0	0	0	0	0.50	0.20	0.50	0.90	0.60	0.65	0.80	1	1	1	1
low community order	0	0	0	0	0	0.45	0.60	0.80	0.50	1	1	1	1	1	1	1	1	1
medium community order	0	0	0	0	0	0.25	0.40	0.50	0	0.50	1	0.55	1	1	1	1	1	1
high community order	0	0	0	0	0	0	0.05	0.10	0	0	0.50	0.40	0.45	0.60	1	0.60	0.70	0.75
1m custody 6m suspended	0	0	0	0	0	0.25	0.30	0.40	0	0.45	0.60	0.50	1	1	1	1	1	1
1m custody 12m suspended	0	0	0	0	0	0	0.25	0.35	0	0	0.55	0	0.50	0.95	1	1	1	1
6m custody 6m suspended	0	0	0	0	0	0	0	0.20	0	0	0.40	0	0.05	0.50	1	0.85	1	1
12m custody 24m suspended	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.50	0.30	0.35	0.40
1m immediate custody	0	0	0	0	0	0	0	0	0	0	0.40	0	0	0.15	0.70	0.50	1	1
2m immediate custody	0	0	0	0	0	0	0	0	0	0	0.30	0	0	0	0.65	0	0.50	1
3m immediate custody	0	0	0	0	0	0	0	0	0	0	0.25	0	0	0	0.60	0	0	0.50

Table 2 Severity ranges covered by each sentence outcome relative to a medium community order.

sentence outcome	relative range
conditional discharge	62.5%
fine A	37.5%
fine B	37.5%
fine C	37.5%
fine D	37.5%
fine E	37.5%
fine F	37.5%
low community order	100%
medium community order	100%
high community order	125%
1 month custody 6 months suspended	100%
1 month custody 12 months suspended	125%
6 months custody 6 months suspended	125%
12 months custody 24 months suspended	175%
1 month custody	25%
2 months custody	25%
3 months custody	25%

Table 3 Severity scales considering equal and unequal variances.

sentence outcome	equal variances	unequal variances	unequal variances rescaled
absolute discharge	0	0	0
conditional discharge	2.75	1.67	1.47
fine A	3.02	1.83	1.59
fine B	4.32	2.35	2.03
fine C	7.08	4.10	3.55
fine D	8.44	4.79	4.15
fine E	9.26	10.42	9.01
fine F	9.98	10.88	9.42
low community order	8.59	4.86	4.21
medium community order	9.66	10.65	9.22
high community order	11.06	12.26	10.61
1 month custody 6 months suspended	9.90	10.87	9.40
1 month custody 12 months suspended	10.69	11.65	10.08
6 months custody 6 months suspended	11.40	12.37	10.70
12 months custody 24 months suspended	16.56	14.09	12.2
1 month custody	12.01	12.97	11.22
2 months custody	15.65	13.41	11.60
3 months custody	16.59	13.82	11.96
12 months custody	37.65	17.69	15.31
5 years custody	147.58	38.30	33.14
10 years custody	284.98	64.05	55.43
20 years custody	559.80	115.55	100

them, which under the standard form of the Thurstone model appears to be oversimplified. Similarly, we can also observe how the scale accounting for unequal variances show a much narrower range of severity scores across the less punitive fines (fines A, B, and C, those which were never deemed to overlap with more punitive disposal types like community orders), and also across custodial sentences. Specifically, severity scores for fines A (1.83) to C (4.10) are limited to a range of 2.27 when considering unequal variances, but that range expands to 4.06 if equal variances are assumed. Comparisons for the range of severity scores for the first three months of custodial sentences are even starker, 0.85 when allowing for unequal variances, compared to 4.58 when equal variances are assumed. These

relative changes in severity scores across different sentence outcomes reflect well the insights about the variability in severity ranges elicited from our sample of experts.

The much narrower ranges of severity estimated for the three immediate custodial sentences (one to three months) is particularly relevant, since all other custodial sentences are extrapolated from the trend seen in the first three months. As such, being able to estimate the marginal increase in severity associated with an additional month in custody is key in order to obtain valid estimates of severity for all other custodial sentences This effect can be noticed by contrasting the divergence in severity scores between the scales based on equal and unequal variances as custodial sentences grow longer. To facilitate interpretations of the relative severity for different sentence outcomes, Table 3 also includes a rescaled version (ranging from 0 to 100) of the unequal variances scale of severity. For example, using that rescaled index, we can see that the highest possible fines (fine-F), medium community orders, and the shortest possible suspended sentences (1 month custody, 6 months suspended), are all three considered of similar severity, which is also roughly equivalent to twice the severity attributed to low community orders.

5 Discussion

The new scale of severity presented here expands those recently developed by Pina-S´anchez et al. (2019a) and Pina-Sanchez and Gosling (2020) in three important ways.[6] First, we have contemplated a wider range of non-custodial sentences, including different types of fines and community orders. This new capacity to discriminate between non-custodial sentences in higher detail matters, since they represent roughly 93% of sentences imposed. Yet, they rarely feature in quantitative sentencing research, leading to a deeply partial understanding in relation to many of the key questions explored in the discipline, such as: i) what was the impact of key sentencing reforms (Fleetwood et al., 2015; Pina-S´anchez and Linacre, 2014; Roberts and Pina-S´anchez, 2021; Ulmer et al., 2011); ii) the effect of specific aggravating or mitigating factors (Irwin-Rogers and Perry, 2015; Lightowlers and Pina-S´anchez, 2017; Lightowlers et al., 2020; Kane and Minson, 2022); iii) changes in trends of sentence severity (Allen, 2016; Doob and Webster, 2003; Hindelang et al., 1975; Pina-S´anchez et al., 2016); iv) consistency in sentencing (Barbora et al., 2012; Isaac, 2020; Brunton-Smith et al., 2020; Dr´apal, 2020); v) or disparities associated with different offender's demographic characteristics (Baumer, 2013; Yan and Lao, 2021; Isaac, 2020; Pina-Sanchez et al., 2019b); to name a few.

Second, we have relaxed the assumption of equal variances invoked in the standard Thurstone model. That is, we have acknowledged that the ranges of severity which could be attributed to different sentence types are not uniform. This provides a sounder theoretical foundation to the new index of severity presented here, since we can now reflect not only the highly variable conditions that could be attached to different disposal types (e.g. community orders being much more heterogeneous in their potential composition than fines, which are limited to prescribed financial amounts), but we can also reflect the variability in the punitive effect that a given sentence can have across different subjects. Importantly, by improving the theoretical foundations of the scale of severity in such way, we have further enhanced its capacity to accurately discriminate across non-custodial sentences.

Third, to derive the expected differences in severity ranges across sentence outcomes we have employed expert elicitation techniques, a new and promising research design in the penal metric literature. Specifically, we conducted a four hours workshop with six sentencing experts following the Sheffield elicitation protocol (Gosling, 2018). The small sample size required to conduct an expert elicitation workshop of these characteristics effectively has undoubtedly affected the reliability of our findings, however, we believe this is a price worth paying given the higher validity afforded by this research

[6] The latest of which has also been adopted by the Sentencing Council for England and Wales in impact assessments of their sentencing guidelines (Isaac, 2021).

design. In contrast with pairwise comparisons of sentence severity elicited from questionnaire (Pina-S´anchez et al., 2019a; Buchner, 1979; Spelman, 1995), expert elicitation techniques can provide the necessary context, discussion, focus, and time for reflection. As a result, responses are substantially more thoughtful, which, for a subject as complex and subjective as that of sentence severity, represents a much desirable trait.

We also believe that our choice of sentencing experts, which combined academics, members of the bar, sentencers, and policy-makers, is particularly fitting to the types of research questions contemplated during the elicitation workshop. In particular, the group of experts demonstrated ample knowledge regarding the typical conditions that are commonly used for each of the sentence outcomes contemplated. In the literature on penal metric theory, it is more common to see perceptions of sentence severity derived from samples of offenders (McClelland and Alpert, 1985; Petersilia and Deschesnes, 1994; Spelman, 1995) or members of the general public (Erickson and Gibbs, 1979; Tremblay, 1988). We believe, however, that the general population is not sufficiently informed to provide valid answers when queried about sentencing. For example, 17% of respondents in a recent opinion poll conducted for the Scottish Sentencing Council (Black et al., 2019) claimed that the adequate response for an early guilty plea should be an increase in severity as it represents an admission of guilt, while the latest opinion poll conducted by the Sentencing Academy (Roberts et al., 2022) reported that 56% of participants thought the average custodial sentence length in England and Wales was shorter in 2021 than two decades before, even though the average sentence length had actually grown by roughly 50% during that period. We could also expect offenders' perspectives to be biased. Offenders found guilty will know better than anyone else the true severity of the sentence that was imposed on them, but their understanding will likely be limited to that specific sentence - or range of sentences in the case of recidivist offenders. Hence, we would not expect them to be more knowledgeable than other members of the general public when it comes to assess the severity that could be attributed to other sentence types. Nor will they be better equipped to assess the severity which that same sentence could exert on others.

There is, however, one group of experts that we believe are particularly well equipped to provide highly meaningful views in relation to the specific concept of penal subjectivism (i.e. how personal context can make a given sentence more or less severe). These are criminal solicitors and barristers. Their unique understanding of the question at hand stems from their experience in discussing potential sentence outcomes with their clients, which will often involve an explicit revelation of preferences. For example, during discussions on whether to plead guilty. Furthermore, criminal lawyers, especially those that are more senior, will likely have become aware of a higher number of cases, contexts and offenders, providing them a unique perspective in relation to the varying punitive effect of different disposal types. As such, we believe it would be interesting to assess whether the results obtained here would replicate if the expert elicitation workshop was to be conducted with a sample entirely composed of criminal lawyers. Future replications would also provide useful insights into the generalisability of the responses elicited from our sample, and in so doing assess the reliability of the severity scale derived from those responses. Undertaking such work will be especially meaningful following processes of sentencing reforms, or similar structural changes within the criminal justice system, such as the renationalisation of probation services, or any further deterioration of prison conditions. All of them instances where the relative severity of disposal types available to sentencers should be reassessed.

Considering avenues of research with which to continue enhancing the robustness of future severity scales, we identify two areas that ought to receive further attention. First, because longer custodial sentences will always be more severe than shorter ones, there is a limit on the range of severity scores that can be directly estimated using Thurstone pairwise comparisons. In our view, after the assumption of equal variances, this represents the next major limitation affecting the estimation of severity scores using Thurstone scaling. The solution adopted here, based on the linear extrapolation of severity scores of one, two and three months custodial sentences, makes the estimation of severity scores for custodial sentences of four months or longer, more unreliable than for the rest of outcomes directly

estimated through Thurstone pairwise comparisons. Moreover, this uncertainty is not uniform across custodial sentences, but proportional to their length, making long sentences particularly unreliable. This limitation could be potentially overcome by combining more than one estimation method for different parts of the severity scale, e.g. Thurstone scaling up to three months in custody combined with magnitude escalation beyond that.

Lastly, we also intend to explore the use of interactive visual apps. We believe expert elicitation techniques to be the most appropriate approach for the exploration of sentence severity, given the discursive yet highly focused setting that they can facilitate, both features equally necessary to structure discussions around a subject as complex as this. However, the ultimate purpose of expert elicitation is the translation of subjective perceptions into specific numerical parameters, which is something that legal experts are not adequately trained to do. Visual aids could be employed to facilitate that process. In particular, it would be most useful to design an app that could show visually the magnitude of proportions and standard deviations being considered in the discussions around the relative severity of different sentence outcomes. By projecting such visualisations in real-time, as the discussion takes place, participants could obtain a more intuitive appreciation of their suggested parameters, which should also contribute to frame the discussions and enhance the consistency of the process. Furthermore, this app could not only help visualise the numerical parameters to be elicited - i.e.1 the input information that will be fed into the Thurstone model, i.e.2 the elicited views in relation to how often a given sentence is thought to be more severe, or the relative range of severity that could be attributed to such sentence - but also how those views will be translated into severity scores - i.e. the output of the Thurstone model.

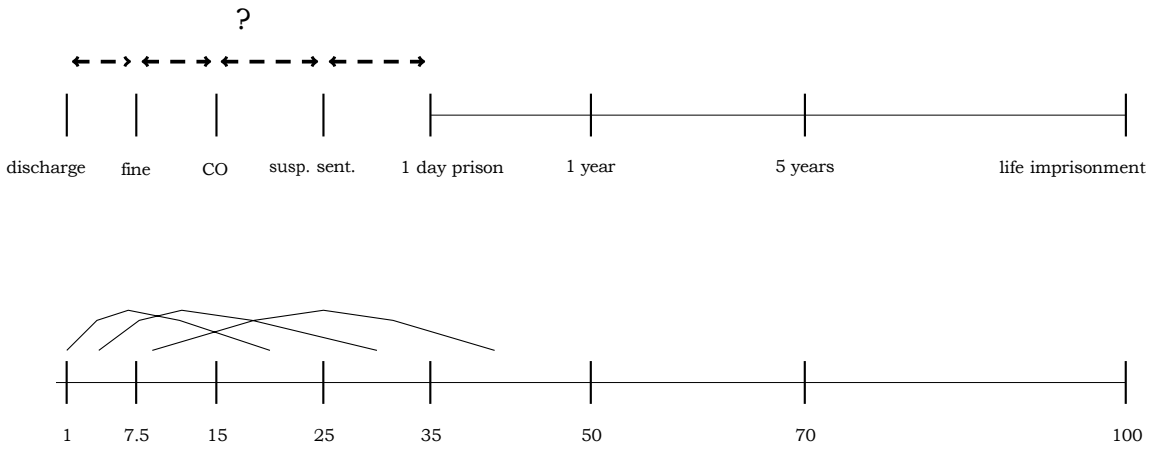

Fig. 2 Conceptual representation of a visual app designed to facilitate the interpretation and discussion of the concepts of penal exchangeability, unequal ranges of severity, and their translation into severity scores, for a selection of sentence outcomes.

Figure 2, shows a conceptual representation of how such visual app could operate, illustrating in different planes the positions of the sentence outcomes being discussed, their potential severity ranges, or their degree of overlap, while simultaneously estimating the position of each outcome in the severity scale according to the information provided. In addition to making some of the numerical elicitations more intuitive, or providing a better framing for the group discussions, being able to see the final product - i.e. the severity scores estimated for each of the sentences considered - would provide an additional layer of robustness to the process. Specifically, participants will be able to assess the face validity of the severity scores to be derived, which would provide them with the opportunity

to reconsider any of the parameters discussed during the workshop. In the event that any issues are raised, the parameter in question could be modified as part of an iterative process, until an agreement is reached and participants are satisfied with the final scale of severity.

6 Conclusion

In this study we have used expert elicitation techniques to explore the validity of a key assumption invoked in studies estimating the relative severity of different sentence types. Namely, we have investigated the assumption of equal variances underpinning the Thurstone scaling method (Thurstone, 1927; Mosteller, 1951; Buchner, 1979; Kwan et al., 2000). An assumption that is rarely stated explicitly, but one we have demonstrated how, when applied to the estimation of sentence severity, is clearly violated. Based on insights elicited from six sentencing experts, we noted wide differences in the range of severity covered by some of the main disposal types used in England and Wales. Our experts agreed that the intervals of severity that could be associated with different types of fines (according to the amount imposed), or custodial sentences (according to their duration), are much narrower than previously considered, while community orders and suspended sentences could encompass much wider severity ranges, depending on the conditions attached to them.

Further, we have demonstrated how accounting for the unequal variances seen across disposal types matters, as the estimated severity scores for specific sentence outcomes using Thurstone scaling vary substantially depending on whether the equal variances assumption is invoked or not. Using the elicited insights into the exchangeability and unequal severity ranges characterising eighteen sentence outcomes, and a modified version of the Thurstone method allowing for unequal variances, we have been able to estimate a new scale of sentence severity.

Ultimately, the goal of the new scale of severity presented here is to be used as an analytical tool to facilitate more robust and insightful quantitative sentencing research. Research that will be better equipped to shed much needed new light on the study of differences across non-custodial outcomes. A highly heterogeneous group of disposal types, representing the vast majority of the sentencing practice, which, for methodological reasons, have remained largely overshadowed by studies focusing on more technically convenient sentence outcomes, such as the probability or duration of custodial sentences.

References

Albonetti CA (1998) The role of gender and departures in the sentencing of defendants convicted of a white-collar offense under the Federal Sentencing Guidelines. Sociology of Crime, Law and Deviance 1:3–48

Allen R (2016) The Sentencing Council for England and Wales: Brake or accelerator on the use of prison? Tech. rep., Transform Justice, URL http://www.transformjustice.org.uk/wp-content/uploads/2016/12/TJ-DEC-8.12.16-1.pdf

Barbora H, Bijleveld C, Smeulers A (2012) Consistency of international sentencing: ICTY and ICTR case study. European Journal of Criminology 9(5):539–552

Baumer EP (2013) Reassessing and redirecting research on race and sentencing. Justice Quarterly 30(2):231–261

Berkson J (1950) Are there two regressions? Journal of the American Statistical Association 45(250):164–180

Black C, Warren R, Ormston R, Cyrus T (2019) Public perceptions of sentencing: National survey report. Tech. rep., Scottish Sentencing Council, URL https://www.scottishsentencingcouncil.org.uk/media/1996/20190902-public-perceptions-of-sentencing-report.pdf

Bolger F (2018) The selection of experts for (probabilistic) expert knowledge elicitation. In: Dias LC, Morton A, Quigley J (eds) Elicitation, pp 393–443

Booth C, Thomas L (2021) An expert elicitation of the effects of low salinity water exposure on bottlenose dolphins. Oceans 2(1):179–192

Brennan A, Pollard D, Coates L, Strong M, Heller S (2017) Expected value of sample information for individual level simulation models to inform stop/go decision making by public research funders: A methodology for the dafneplus diabetes education cluster rct. Value in Health 9(20)

Brunton-Smith I, Pina-Sanchez J, Li G (2020) Re-assessing the consistency of sentencing decisions in cases of assault: Allowing for within court inconsistencies. British Journal of Criminology 60(6):1438–1459

Buchner D (1979) Scale of sentence severity. The Journal of Criminal Law and Criminology 70(2):182–187

Bushway SD, Johnson BD, Slocum LA (2007) Is the magic still there? The use of the Heckman two-step correction for selection bias in criminology. Journal of Quantitative Criminology 23(2):151–178

Dessai S, Bhave A, Birch C, Conway D, Garcia-Carreras L, Gosling J, Mittal N, Stainforth D (2018) Building narratives to characterise uncertainty in regional climate change through expert elicitation. Environmental Research Letters 13(7)

Doob AN, Webster CM (2003) Sentence severity and crime: Accepting the null hypothesis. Crime and Justice 30:143–195

Dr´apal J (2020) Sentencing disparities in the Czech Republic:Empirical evidence from post-communist Europe. European Journal of Criminology 17(2):151–174

Erickson ML, Gibbs JP (1979) On the perceived severity of legal penalties. The Journal of Criminal Law and Criminology 70(1):102–116

Feldmeyer B, Ulmer JT (2011) Racial/ethnic threat and federal sentencing. Journal of Research in Crime and Delinquency 48(2):238–270

Fleetwood J, Radcliffe P, Stevens A (2015) Shorter sentences for drug mules: The early impact of the sentencing guidelines in England and Wales. Drugs: Education, Prevention and Policy 22(5):428–436

Francis B, Soothill K, Humphreys L, Cutajar Bezzina A (2005) Developing measures of severity and frequency of reconviction. Tech. rep., Lancaster University, URL https://eprints.lancs.ac.uk/id/eprint/50137/1/seriousnessreport.pdf

Freiberg A, Fox R (1986) Sentencing structures and sanction hierarchies. Criminal Law Journal 10:216–235

Ginneken E, Hayes D (2017) "just" punishment? offenders' views on the meaning and severity of punishment. Criminology and Criminal Justice 17(1):62–78
Gosling J (2018) Shelf: the Sheffield elicitation framework. In: Dias L, Morton A, Quigley J (eds) Elicitation, Springer, pp 61–93
Gosling J, Hart A, Mouat D, Sabirovic M, Scanlan S, Simmons A (2012) Quantifying experts' uncertainty about the future cost of exotic diseases. Risk Analysis: An International Journal 5(32):881–893
Harlow RE, Darley JM, Robinson PH (1995) The severity of intermediate penal sanctions: A psychological scaling approach for obtaining community perceptions. Journal of Quantitative Criminology 11(1):71–95
Harris L, Walker S (2021) Sentencing Principles, Procedure and Practice 2022. Sweet and Maxwell
Hayes D (2016) Penal impact: Towards a more intersubjective measurement of penal severity'. Oxford Journal of Legal Studies 36(4):724–750
Hayes D (2018) Proximity, pain, and State punishment. Punishment and Society 20(2):253–254
Hester R, Hartman T (2017) Conditional race disparities in criminal sentencing: A test of the liberation hypothesis from a non-guidelines state. Journal of Quantitative Criminology 33:77–100
Hindelang MJ, Dunn CS, Sutton LP, Aumick AL (1975) Sourcebook of criminal justice statistics, 1974. Tech. rep., US National Criminal Justice Information and Statistics Service, URL https://www.ojp.gov/ncjrs/virtual-library/abstracts/sourcebook-criminal-justice-statistics-1974
Irwin-Rogers K, Perry TH (2015) Exploring the impact of sentencing factors on sentencing domestic burglary. In: Roberts JV (ed) Sentencing Guidelines: Exploring Sentencing Practice in England and Wales, Palgrave, Basingstoke, pp 213–239
Isaac A (2020) Investigating the association between an offender's sex and ethnicity and the sentence imposed at the Crown Court for drug offences. Tech. rep., Sentencing Council for England and Wales, URL https://www.sentencingcouncil.org.uk/wp-content/uploads/Sex-and-ethnicity-analysis-final-1.pdf
Isaac A (2021) Estimating the changes in sentencing severity and requirements for prison places associated with the Sentencing Council's guidelines. Tech. rep., Sentencing Council for England and Wales, URL https://www.sentencingcouncil.org.uk/wp-content/uploads/Changes-in-sentencing-severity-and-prison-places-associated-with-SC-guidelines.pdf
Jansen J, Wang H, Holcomb J, Harvin J, Richman J, Avritscher E, Stephens S, Troung V, Marques M, DeSantis S, Yamal J, Pedroza C (2020) Elicitation of prior probability distributions for a proposed Bayesian randomized clinical trial of whole blood for trauma resuscitation. Transfusion 6(3):498–506
Kane E, Minson S (2022) Analysing the impact of being a sole or primary carer for dependent relatives on the sentencing of women in the Crown Court, England and Wales. Criminology and Criminal Justice
King RD, Johnson KR, McGeever K (2010) Demography of the legal profession and racial disparities in sentencing. Law and Society Review 44(1):1–32
Kurlychek MC, Johnson BD (2010) Juvenility and punishment: Sentencing juveniles in adult criminal court. Criminology 48(3):725–758
Kwan YK, Ip WC, Kwan P (2000) A crime index with Thurstone's scaling of crime severit. Journal of Criminal Justice 28(3):237–244
Leclerc C, Tremblay P (2016) Looking at penalty scales: How judicial actors and the general public judge penal severity. Canadian Journal of Criminology and Criminal Justice 58(3):354–384
Lightowlers C, Pina-Sanchez J (2017) Intoxication and assault: an analysis of Crown Court sentencing practices in England and Wales. The British Journal of Criminology 58(1):132–154
Lightowlers C, Pina-Sanchez J, Watkins E (2020) Contextual culpability: How drinking and social context impact upon sentencing of violence. Criminology and Criminal Justice
Lovegrove A (2001) Sanctions and severity: To the demise of von Hirsch and Wasik's sanction hierarchy. The Howard Journal of Crime and Justice 40(2):126–144

McClelland KM, Alpert GP (1985) Factor analysis applied to magnitude estimates of punishment seriousness: Patterns of individual differences. Journal of Quantitative Criminology 1(3):307—-318
McDavid JC, Stipack B (1981) Simultaneous scaling of offense seriousness and sentence severity through canonical correlation analysis. Law and Society Review 16(1):147–162
Ministry of Justice (2021) Data first: An introductory user guide. Tech. rep., Ministry of Justice, URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/984510/data-first-user-guide-v5.pdf
Mosteller F (1951) Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. Psychometrika 16(1):3–9
O'Hagan A (2019) Expert knowledge elicitation: Subjective but scientific. The American Statistician 73(1):69–81
O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) Uncertain Judgements: Eliciting Expert Probabilities. John Wiley, Chichester
Padfield N (2011) Time to bury the "custody threshold"? Criminal Law Review 8:593–612
Petersilia J, Deschesnes E (1994) Perceptions of punishment: Inmates and staff rank the severity of prison versus intermediate sanctions. Prison Journal 74(3):306–328
Pina-S´anchez J, Gosling JP (2020) Tackling selection bias in sentencing data analysis: a new approach based on a scale of severity. Quality and Quantity
Pina-S´anchez J, Linacre R (2014) Enhancing consistency in sentencing: Exploring the effects of guidelines in England and Wales. Journal of Quantitative Criminology 30(4):731–748
Pina-S´anchez J, Lightowlers C, Roberts JV (2016) Exploring the punitive surge: Crown Court sentencing practices before and after the 2011 English riots. Criminology and Criminal Justice 17(3):319–339
Pina-S´anchez J, Gosling JP, Chung H, Geneletti S, Bourgeois E, Marder I (2019a) Have the England and Wales guidelines influenced sentence severity? An empirical analysis using a scale of sentence severity and time-series analyses. British Journal of Criminology
Pina-S´anchez J, Roberts JV, Sferopoulos D (2019b) Does the Crown Court discriminate against Muslim-named offenders? A novel investigation based on text mining techniques. British Journal of Criminology 59(3):718–736
Revelle W (2018) psych: Procedures for personality and psychological research,. Tech. rep., Northwestern University, URL https://www.scholars.northwestern.edu/en/publications/psych-procedures-for-personality-and-psychological-research
Roberts JV, Bild J (2021) Ethnicity and custodial sentencing in England and Wales: A review of trends, 2009-2019. Tech. rep., Sentencing Academy, URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3881930
Roberts JV, Pina-Sanchez J (2021) Sentence reductions for a guilty plea: The impact of the revised guideline on rates of pleas and 'cracked' trials. Journal of Criminal Law
Roberts JV, Bild J, Pina-Sanchez J, Hough M (2022) Public knowledge of sentencing practice and trends. Tech. rep., Sentencing Academy, URL https://sentencingacademy.org.uk/wp-content/uploads/2022/01/Public-Knowledge-of-Sentencing-Practice-and-Trends.pdf
Sebba L, Nathan G (1984) Further explorations in the scaling of penalties. The British Journal of Criminology 24(3):221–249
Sentencing Council (2021) Background quality report. Tech. rep., Sentencing Council for England and Wales, URL https://www.sentencingcouncil.org.uk/wp-content/uploads/Background-Quality-Report-Theft-from-a-shop-or-stall-data.pdf
Spelman W (1995) The severity of intermediate sanctions. Journal of Research in Crime and Delinquency 32(2):107–135
Steffensmeier D, DeMuth S (2001) Ethnicity and judges' sentencing decisions: Hispanic-Black-White comparisons. Criminology 39:145—-176
Thomas KJ, Hamilton BC, Loughran TA (2018) Testing the transitivity of reported risk perceptions: Evidence of coherent arbitrariness. Criminology 56(1):59–86

Thurstone LL (1927) A law of comparative judgement. Psychological Review 34:273–286
Tremblay P (1988) On penal metrics. Journal of Quantitative Criminology 4(3):225–245
Ulmer J, Light MT, Kramer J (2011) The “liberation” of federal judges’ discretion in the wake of the Booker/Fanfan decision: Is there increased disparity and divergence between courts? Justice Quarterly 28(6):799–837
Ulmer JT, Eisenstein J, Johnson BD (2010) Trial penalties in federal sentencing: extra-guidelines factors and district variation. Justice Quarterly 27(4):560–592
Usher W, Strachan N (2013) An expert elicitation of climate, energy and economic uncertainties. Energy Policy 61:811–821
Yan S, Lao J (2021) Sex disparities in sentencing and judges’ beliefs: A vignette approach. Victims and Offenders

Technical Appendix: R Code

```
################################
# Thurstone - Unequal Variances #
################################

###### Enter number of outcomes under consideration
num_outcomes = 18

###### Enter customised standard deviations
latent_sds = sqrt(1/2) * c(0.01, 0.625, 0.375, 0.375, 0.375, 0.375, 0.375, 0.375,
                1, 1, 1.25, 1, 1.25, 1.25, 1.75, 0.25, 0.25, 0.25)

###### Enter matrix of preferences
elicited_proportions = matrix(c(
0.5,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,
0.001,0.5,0.65,0.8,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,
0.001,0.35,0.5,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,
0.001,0.2,0.001,0.5,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,
0.001,0.001,0.001,0.001,0.5,0.999,0.999,0.999,0.65,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,
0.001,0.001,0.001,0.001,0.001,0.5,0.999,0.999,0.55,0.75,0.999,0.75,0.999,0.999,0.999,0.999,0.999,0.999,
0.001,0.001,0.001,0.001,0.001,0.001,0.5,0.999,0.4,0.6,0.95,0.7,0.75,0.999,0.999,0.999,0.999,0.999,
0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.5,0.2,0.5,0.9,0.6,0.65,0.8,0.999,0.999,0.999,0.999,
0.001,0.001,0.001,0.001,0.001,0.45,0.6,0.8,0.5,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,0.999,
0.001,0.001,0.001,0.001,0.001,0.25,0.4,0.5,0.001,0.5,0.999,0.55,0.999,0.999,0.999,0.999,0.999,0.999,
0.001,0.001,0.001,0.001,0.001,0.001,0.05,0.1,0.001,0.001,0.5,0.4,0.45,0.6,0.999,0.6,0.7,0.75,
0.001,0.001,0.001,0.001,0.001,0.25,0.3,0.4,0.001,0.45,0.6,0.5,0.999,0.999,0.999,0.999,0.999,0.999,
0.001,0.001,0.001,0.001,0.001,0.001,0.25,0.35,0.001,0.001,0.55,0.001,0.5,0.95,0.999,0.999,0.999,0.999,
0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.2,0.001,0.001,0.4,0.001,0.05,0.5,0.999,0.85,0.999,0.999,
0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.5,0.3,0.35,0.4,
0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.4,0.001,0.001,0.15,0.7,0.5,0.999,0.999,
0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.3,0.001,0.001,0.001,0.65,0.001,0.5,0.999,
0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.25,0.001,0.001,0.001,0.6,0.001,0.001,0.5),
num_outcomes, num_outcomes, byrow=TRUE)

###### Function for creating preference matrix from latent parameters
create_matrix = function(latent_means){
# Empty matrix to store
pref_probs = matrix(0, num_outcomes, num_outcomes)
# Calculate probability of one latent random variable being greater than another
for (i in 1:num_outcomes){
for (j in 1:num_outcomes){
pref_probs[i,j] = pnorm(0,
latent_means[i] - latent_means[j],
sd = sqrt(latent_sds[i]^2 + latent_sds[j]^2))
}
}
# Return matrix
return(pref_probs)
}

###### Function for measuring distance between created matrix and elicited proportions
matrix_distance = function(latent_means){
# set first latent mean to be 0
latent_means = c(0,latent_means)
# return distance based on Frobenius norm
return(Matrix::norm(create_matrix(latent_means)-elicited_proportions, type = 'F'))
}

###### Find the latent means
optim_out <- optim(1:(num_outcomes-1), # We are trying to find num_outcomes - 1 because the first is fixed at 0
```

```
matrix_distance, method = 'L-BFGS-B', lower = 0, control = list())
# Has it converged?
ifelse(optim_out$convergence == 0, TRUE, FALSE)
# Found latent means
c(0,optim_out$par)
# Fitted preference matrix
round(create_matrix(c(0,optim_out$par)),3)
# Original matrix
elicited_proportions


#####################
# Custodial scores  #
#####################

#Immediate custody
month = c(1, 2, 3)
sev = optim_out$par[15:17]
#A linear function to estimate severity of immediate custody > 3 months
linear_imm = lm(sev ~ month)
summary(linear_imm)
#We can now predict severity scores for different immediate custody sentences using the above function, i.e.:
#severity = 12.54 + 0.43*custody_months
#Predicted score for a 12 months immediate custody
linear_imm$coefficients[1] + 12*linear_imm$coefficients[2]
#Predicted score for a 24 months immediate custody
linear_imm$coefficients[1] + 24*linear_imm$coefficients[2]
#Predicted score for a 60 months immediate custody
linear_imm$coefficients[1] + 60*linear_imm$coefficients[2]
#Predicted score for a 120 months immediate custody
linear_imm$coefficients[1] + 120*linear_imm$coefficients[2]
#Predicted score for a 240 months immediate custody
linear_imm$coefficients[1] + 240*linear_imm$coefficients[2]

#Suspended sentences
month_cust = c(1, 1, 6, 12)
month_susp = c(6, 12, 6, 24)
sev = optim_out$par[11:14]
#A linear function to estimate severity of immediate custody > 3 months
linear_susp = lm(sev ~ month_cust + month_susp)
summary(linear_susp)
#We can now predict severity socres for different suspended sentences using the above function, i.e.:
#severity = 10.73 + 0.20*custody_months + 0.04*suspended_months
#3 months custody suspended for 12 months - predicted score
linear_susp$coefficients[1] + 3*linear_susp$coefficients[2] + 12*linear_susp$coefficients[3]
#9 months custody suspended for 12 months - predicted score
linear_susp$coefficients[1] + 9*linear_susp$coefficients[2] + 12*linear_susp$coefficients[3]
#9 months custody suspended for 24 months - predicted score
linear_susp$coefficients[1] + 9*linear_susp$coefficients[2] + 24*linear_susp$coefficients[3]


##########################################
# Constraining the scale to a 0-100 range #
##########################################

#First we need to contemplate the maximum severity score possible
#Here we have assumed that is a 20 years custodial sentence, which was estimated at 115.56 severity
#Then we need to re-scale accordingly by multiplying severity scores by the following: 100/115.56
#So, a low community order with severity equal to 8.066 will be rescaled as:
4.86 * 100/115.56

#A table with the severity scores for the main sentence outcomes considered could be presented as follows:
```

```
outcomes = c("absolute discharge", "conditional discharge", "fine-A", "fine-B", "fine-C", "fine-D",
"fine-E", "fine-F", "community order-low", "community order-medium", "community order-high",
"1-month custody suspended for 6 months", "1-month custody suspended for 12 months",
"6-month custody suspended for 6 months", "12-month custody suspended for 24 months",
"1-month immediate custody", "2-month immediate custody", "3-month immediate custody",
"12-month immediate custody", "60-month immediate custody", "120-month immediate custody",
"240-month immediate custody")
severity = c(0, optim_out$par[1:17],
linear_imm$coefficients[1] + 12*linear_imm$coefficients[2],
linear_imm$coefficients[1] + 60*linear_imm$coefficients[2],
linear_imm$coefficients[1] + 120*linear_imm$coefficients[2],
linear_imm$coefficients[1] + 240*linear_imm$coefficients[2])
severity_rescaled = severity * 100/115.56
cbind(outcomes, severity_rescaled)
```