Sed quis custodet? Sentencing intentions of judges are more variable than those of lay people.

Jonathan Williams

Nuffield Department of Women's & Reproductive Health
University of Oxford
Level 3
Women's Centre
John Radcliffe Hospital
Oxford
OX3 9DU

e-mail: jonathan.williams@wrh.ox.ac.uk

Tel: +1865 221004

Abstract
Background: There is a general assumption that judges assess crimes more reliably than jurors or the general public. Reliability is fundamental for trust and validity. Here, I compare the variability of sentencing intentions of judges and lay people.

Quantifying the variability of sentencing is difficult, because sentencing guidelines impose boundaries on sentence lengths, that cause problems for estimation. I compared two ways to circumvent these problems: (1) by analysing actual sentences as proportions of the guideline range, or (2) by assuming that observed sentences reflect a latent distribution of "true" intentions. In both cases, the primary outcome was the variability of judges' sentences compared with lay people's.

Methods: I analysed published sentencing intentions of professional judges and lay people for the same fictional crime. The source study's sentencing guidelines allowed terms in the range 3-20 years, and over half the participants' sentencing intentions were at these boundaries. I accounted for these boundary effects by transforming the raw sentence intentions to either (1) to proportions of the allowed range, or (2) to an ordinal variable, with 4 levels (0-9, 10-14, 15-19 or 20 years), representing a latent normal distribution. I then analysed the transformed data using distributional regressions, to assess 'dispersion' (reliability) in relation to rater status (judge, lay person)

Results: Accounting for floor and ceiling effects importantly improved estimation of reliability. Judges' sentencing intentions were more variable than those of lay people in both analyses. The proportional (relative) and ordinal (absolute) methods of analysis estimated lower or higher mean sentences for judges.

Conclusion: Distributional regression can assess differential variability, as an indicator of reliability. The present results challenge the assumption that judges assess crimes reliably. This may help to explain the public's lack of trust in the judicial system. It also throws into question the decisions of single expert raters in other high-stakes contexts.
[290 words]

Introduction

"Legal uncertainty" means that the outcomes of court cases are difficult to predict,[1] and is "routinely blamed for undermining the rule of law".[2,3] Judges show "non-trivial" variability in the duration of sentencing, that is attributable partly to their individual characteristics and partly to the context of their work.[4–7] Previous studies have compared the *lengths* of judges' jail sentences with those of jurors,[8,9] but rarely their relative *reliability*,[10] which appears poor.[3,10,11] Reliability is fundamental for trust[3] and, mathematically, constrains validity.[12–14] Here I propose methods to improve the estimation of reliability of sentencing and use them to compare judges' reliability with that of the general public.

The long-term outcomes of jail sentences reflect the combined effects of incapacitation, deterrence, rehabilitation and retribution.[8] Longer sentences prevent further offending, but at the costs of the offender's freedom and the public purse. Shorter sentences reduce those costs, but at the risk of further offending. Hence, it may be possible to define a sentence length, for any given offence, that optimally balances these competing requirements (c.f. [15]). *A priori* we would expect that judges' experience enables them to identify such an optimum more reliably than lay raters. Here I test this hypothesis.

In the judicial setting, raters' sentencing intentions integrate their assessments of (a) the gravity of the offence with (b) their understanding of the effects of retribution, incapacitation, deterrence and rehabilitation, (c) their personal biases and (d) legislative boundaries (see below). Measuring complex assessments of this kind is the realm of Psychometry, and all the principles of psychometry[16] – reliability, validity, standardisation and bias reduction – apply to sentencing.

I re-analysed data from a source study[8] that recorded judges' and laypeople's sentencing intentions for fictional crime, presented as a written vignette to all raters. This design ensures that all candidates have the same information, so that any variation in their responses reflects their personal qualities. The source study estimated the *mean* sentence lengths of judges and lay people (c.f. [9]). But, the paradigm also allows comparison of the *variability* (reliability) of judges with that of lay people. Since judges and lay raters receive identical information, we would expect judges' training and experience to reduce the variability of their sentences.

In line with many jurisdictions, the source study[8] imposed sentencing guidelines that limit the range of possible sentences. Such limitations can introduce floor and ceiling effects that (a) distort the underlying distribution of sentencing intentions,[17] and so (b) make it difficult to estimate the distribution's parameters (including both mean and *variability*) accurately, from a mathematical perspective.[18] I propose two methods to overcome the distorting effects of sentencing guidelines on the variability of prison sentences.

Ideally, we want to model the psychological mechanisms that generate sentencing intentions. To this end, the first method assumes that each rater's sentencing intention is *proportionate* to his/her assessment of the gravity of the crime, *relative* to the sentencing guidelines – and so transforms sentence lengths to proportions of the guideline range. This allows the application of statistical models for proportions that can include floor and ceiling effects.[18–20] The second method assumes that observed sentences reflect a *latent* distribution of *absolute* "true" intentions. It then estimates the parameters of the latent distribution after rank-transforming the observed sentences, which absorbs and removes floor and ceiling effects.[21–24] Both transformations can allow the assessment of sentencing variability, uncontaminated by distorting effects of sentencing guidelines.

Here, I show that the sentencing intentions of professional judges are *less* reliable than those of lay raters, for the same offence, but the latent mean of the judges' absolute "true" sentences is *larger* than that of lay people.

Methods

I analysed publicly-available data from an online study of sentencing intentions for a single fictional crime that was presented to every participant as a written vignette.[8] The study recruited 50 professional judges and 200 lay people from a panel of 2.2 million individuals who may be representative of the general population in Japan. Full details of the recruitment process and other methods are available in the original report.[8]

All data transformations and analyses used the open-source statistical programming language R.[25] First, I transformed actual sentences to proportions of the range allowed by the study guidelines. Specifically, since the guideline imposes a minimum sentence of 3 years and maximum of 20, I transformed the sentences to proportions of the allowed limits, as (raw sentence - 3)/17. Second, I rank-transformed actual sentences by cutting the total distribution into 4 segments: 0-9, 10-14, 15-19 or 20 years. I also standardised participants' ages to decades, centred on 40 years.

I estimated the mean and variability of the proportion- and rank-transformed data using the R package 'brms'. Analysis of the proportion-transformed data used the zero-one-inflated beta family – where zero represents the minimum sentence allowed (3 years) and one represents the maximum (20 years) of the study's sentencing guidelines. Analyses of the rank-transformed data used the continuation ratio family, which assumes that each category of the ordinal response includes the lower categories (see below). Both models fitted main effects for both the means and the variabilities (dispersions) of the study covariates – age, sex and status (judge or layperson).

The goal of psychometric analysis is to represent the form of the underlying psychological mechanisms mathematically, in order to facilitate accurate prediction. I used the continuation ratio family to represent the ordinal categories of sentencing. My reasoning is that raters – both judges and lay people – may increase their sentencing intentions first if they feel that the offence meets (unspecified) fundamental criteria, and increase it further if they feel that the offence additionally meets further criteria. In practice, the continuation ratio family fits the data slightly better than the cumulative family (Bayes factor ~2).

brms allows chaining of equations that can represent path models. I used this approach to pre-adjust rater status (lay/judge) for age and sex. In effect, this approach assumes that age and sex cause the status of the raters in the study and assesses how far rater status mediates and moderates their effects of sentencing.

Analyses of both forms of transformed data used the same model for the mean sentence length:-

1) ***status ~ age + sex + error***, family = bernoulli
2) ***t(sentence) ~ age + sex + status + error***, family = zero_one_inflated beta, or continuation ratio

- where "t()" means the applied transform – to proportions or ranks. Additionally, (a) the beta regression included estimation of the dispersion coefficient (φ) according to the raters' age, sex and status, and of one-inflation (coi) according to status.; (b) the ordinal regression included estimation of the discrimination parameter (θ) according to raters' age, sex and status.

I compared the distributional regressions of proportions and ordinal categories of sentences using simple linear regression and using distributional linear regression that can analyse differential dispersion in relation to design factors.

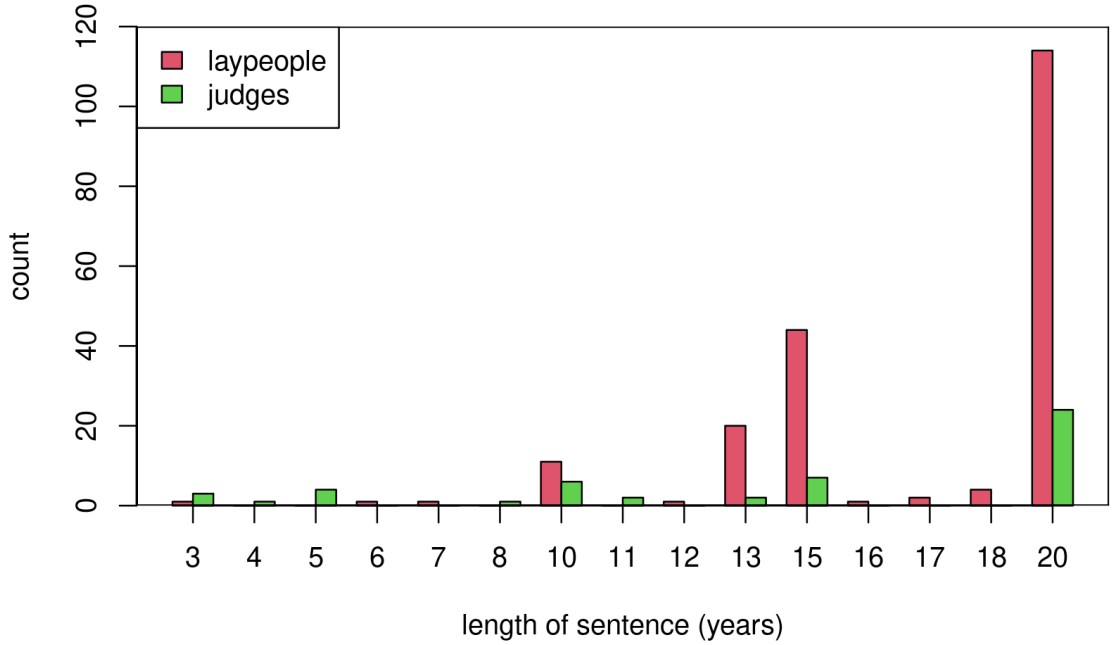
The Supplementary Information includes all the R code to perform the analyses and generate the figures and tables. The data are available from the source study.[8]

Results

Characteristics of the raw data
The source report shows the demographic characteristics of the samples of lay people and judges. Judges were 7 years older, on average (t=4.41, 112df, $p=2e^{-5}$) and 90% were male, compared with 50% of the lay people OR = 8.9, Fisher exact $p=1e^{-7}$). Figure 1 shows the raters' raw sentencing intentions. Half of the raters (55%) chose the maximum sentence that the study's sentencing guidelines allowed (Figure 1). The distributions of the raw sentences from lay people and judges differ importantly (Figure 1; $\chi^2=40.0$, 14df, $p<0.001$).

Simple linear regression
Using ordinary linear regression (that does not account for either the ceiling and floor effects caused by sentencing guidelines, or for differential variability of sentencing between judges and members of the public), rater status accounted for 3.93% of the variance of sentencing intentions (due to the difference of mean sentences). Nevertheless, a simple variance ratio indicated that judges' sentencing intentions are more variable than those of lay raters (F=2.91, 49/249df, $p\approx10^{-8}$).

Figure 1: the counts of the raw sentencing intentions of laypeople and judges



Legend: pairs of bars show the raw counts of each sentence length from laypeople and judges. The study's sentencing guidelines generates marked ceiling effects in the sentencing intentions of both laypeople and judges.

Distributional linear regression
Distributional regression indicated that judges' mean sentences were shorter and their variance greater than those of lay people (elpd difference = 8.2±4.3). This linear regression accounted for 8.42% of the total variance of sentencing (95% credibility interval 2.12 – 16.72%).

Analysis of proportions.
Judges gave shorter sentences than laypeople, expressed as a proportion to the sentencing guidelines (see Figure 2 and Table 1) and adjusting for differential ceiling effects due to rater status.

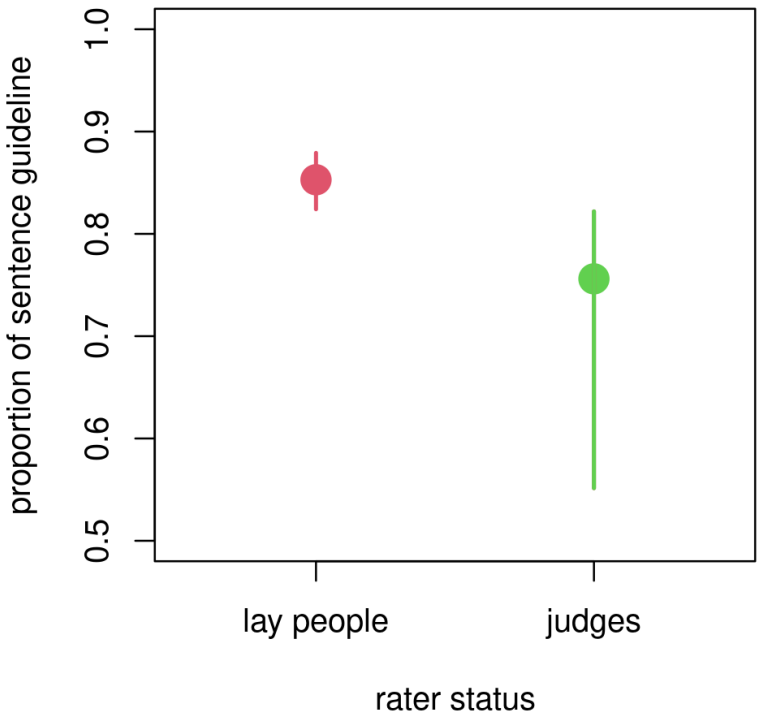
The dispersion of the judges' sentences was *larger* than that of the laypeople's (see Figure 2). The beta regression that accounted for ceiling and floor effects accounted for 11.75% of the variance in sentencing – an improvement of 39% over distributional linear regression.

Table 1: Effects of age sex and type of rater on proportions of sentences

	Mean			Dispersion			One-inflation		
term	Coef.	2.5%CI	97.5%CI	Coef.	2.5%CI	97.5%CI	Coef.	2.5%CI	97.5%CI
Intercept	0.69	0.51	0.88	2.60	2.13	3.03	8.29	4.30	15.27
age	-0.05	-0.13	0.04	-0.06	-0.24	0.13	**2.12**	0.64	4.04
sex	*-0.21*	-0.46	0.04	-0.11	-0.67	0.46	*-2.90*	-9.75	0.96
Rater judge	**-0.72**	-1.15	-0.31	**-1.12**	-1.77	-0.51	**-5.17**	-9.57	-1.64

Legend: values are cofficients from the Bayesian analysis of sentences as a proportion of the range of the sentencing guidelines, that used the zero-one-inflated beta family. The coefficients for the mean (μ) and for One-inflation ('coi') are on the logit scale; those for Dispersion (φ) are on the log scale. Values in bold differ from zero ($p<0.05$); those in italics are almost different ($0.10>p>0.05$). Note that *smaller* values of the dispersion parameter equate to *greater* variability – see text and Figure 2. Older people were more likely to choose the maximum sentence allowed (20 years), but male raters were marginally less likely to choose longer sentences.

Figure 2: conditional effects of rater status on sentences as a proportion of the sentencing guideline



Legend: The figure shows the conditional effects from the Bayesian model of sentencing intentions transformed to a proportion of the range of the sentencing guidelines. Values are the sentencing intentions of average raters, adjusted for age and sex effects on both the probability of being a judge and on sentencing itself, as well as for differential one-inflation (ceiling effects) for all factors (see Table 1).Error bars show the 95% Credibility Intervals (CIs) from the Bayesian analysis. Note that the mean value for the average judge (green) is lower and the range (variability – 95%CI) of sentencing intentions is much larger, compared with the average layperson (red).

Table 2: Effects of age, sex and type of rater on probabilities of categories or sentence length

	Mean			Discrimination		
term	Coef.	2.5%CI	97.5%CI	Coef.	2.5%CI	97.5%CI
age	-0.02	-0.80	1.05	0.10	-0.47	0.54
sex	0.10	-0.91	1.34	-0.56	-1.41	0.29
Rater judge	*7.09*	-0.25	20.81	**-2.16**	-3.63	-0.56

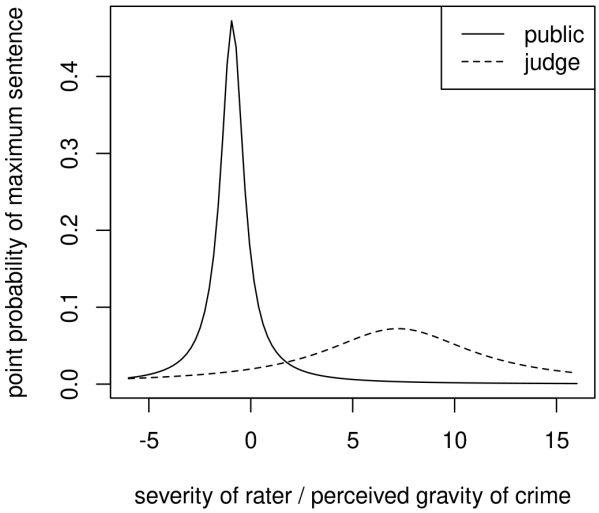
Legend: values are coefficients from the Bayesian analysis of sentences as ordinal categories, that used the continuation ratio family. The coefficients for the mean (μ) are on the logit scale; those for Discrimination (disc) are on the log scale. Values in bold differ from zero ($p<0.05$); those in italics almost differ from zero ($p<0.06$).

Analysis of ranks

The *latent* mean of judges' sentences tended to be *longer* than that of laypeople (Table 2 & Fig. 3).

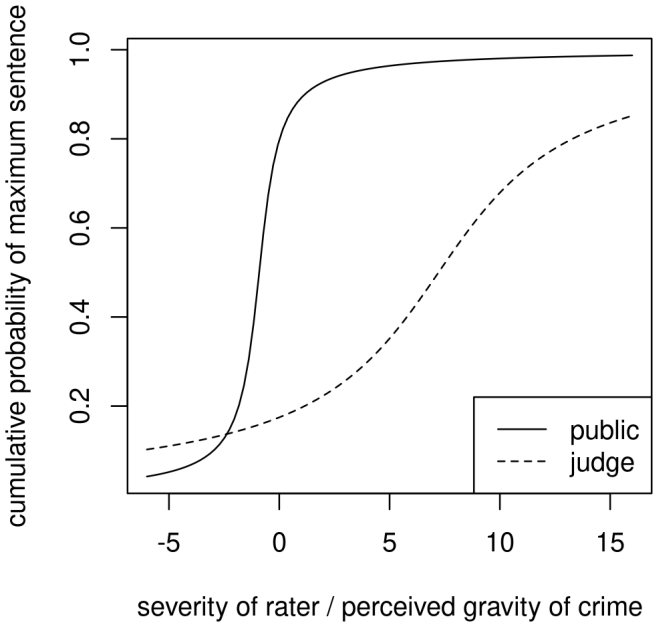
Judges assessed the gravity of the index crime *less* reliably (Figures 3-4). Statistically, the discrimination parameter of judges' sentencing intentions was *less* than that of laypeople, (Figure 4) – i.e.. The distributional ordinal regression accounted for 15.37% of the variance in sentencing – an improvement of 79% over distributional linear regression.

Figure 3: The probability distributions of the estimated latent "true" sentencing intentions



Legend: the figure shows the probability density functions for the latent distribution of "true" intentions to apply a sentence up to the maximum, for lay raters who were members of the public (solid line) and judges (dashed line). Note that (a) the variability of the judges' "true" intentions is much greater than that of lay raters; (b) the mean value of the judges' "true" intentions is higher than that of the lay raters – which contrasts with the results from the simple and zero-one-inflated beta regressions. See the Discussion for an explanation of this contrast.

Figure 4: The estimated cumulative distribution functions of the latent "true" sentencing intentions



Legend: The curves show the latent cumulative functions of the latent distributions of "true" intentions to apply the maximum sentence of 20 years, for members of the public and for judges, derived from the mean and discrimination parameters from the continuation ratio model with cauchit link. The x-axis shows the 'severity' of each rater. The relative steepness of the curve for lay raters indicates low variability in their sentencing intentions as severity increases; in contrast, the shallow slope of the judges' latent sentencing intentions is indicates that their sentences varied little over a wide range of severity (i.e their perception of the gravity of the crime).

Discussion

Judges' assessments of a single fictional crime were less reliable than those of lay people, after accounting for mathematical distortions due to sentencing guidelines. This implies that training and experience do not improve the accuracy of judges' assessments of a crime – but may impair it. Reliability is crucial because it underpins trust and, mathematically, constrains validity.[12,14] The present study provides more reliable ways to measure reliability. In summary, adopting a psychometric approach to the analysis of sentencing may help improve the administration of justice.

There is much interest in developing mathematical methods to analyse sentencing patterns.[26–28] However, few previous studies have compared the dispersion of sentencing intentions of judges with those of lay people. Kääriäinen[29] analysed sentences from 192 judges and 1251 lay people in response to vignettes of 7 crimes. They concluded that judges' sentences generally showed less dispersion than those of lay people, but performed no statistical tests to confirm this. The present study provides methods to disentangle the estimation of the mean and dispersion of sentencing intentions, while accounting for statutory limits on sentences.

The present study used two methods to eliminate distortion due to sentencing guidelines, when assessing the reliability (variability, or dispersion) of raters' sentencing intentions. The two methods present reciprocal advantages and disadvantages. The advantage of the beta regression is that it uses all the information in the raw data; its disadvantage is that it cannot estimate the parameters of the unobserved distribution of "true" sentencing intentions. In contrast, the ordinal regression coarsens the data and so loses some information, but it can help to quantify the latent distribution of "true" sentencing intentions. The ordinal regression has the further advantage that it can flexibly fit the ordinal categories to the data, whereas the beta regression is more constrained. This may explain why the ordinal regression explained more variance than the beta regression. In the limit, however, the two approaches relate closely.[18]

Every analysis in the present study found that the variability of judges' sentencing intentions was greater than that of lay people. This is surprising, because judges' long experience and training should guide them towards more uniform assessments of any given offence. However, one possible explanation of the judges' variability is that they over-complicated the case (c.f.[30]). That is, each judge engaged in hidden legal reasoning about details that lay raters thought unimportant. If legal reasoning about details is error-prone, then the sum of such errors could reduce judges' sentencing reliability. Further psychometric studies should test this possibility.

Two analyses – the simple linear and beta regressions – indicated that judges give *shorter* sentences than those of lay people. In contrast, the ordinal regression indicated that the *latent* mean of judges' sentences is *longer* (see Fig. 3). The reason for this contrast is that the ordinal regression estimates the unobserved distribution of latent "true" sentencing intentions ([23,24] and see Figure 4; *assuming* that the "true" distribution is Normal), but the linear and beta regressions aim to explain the observed data. Reversal of means can occur when comparing "metric" and ordinal data (see detailed explanation in Figure 4 of [23]). Therefore, the result of the ordinal regression perhaps indicates that a substantial proportion of judges favour "retributive" justice (see [8]). Further psychometric studies should test this possibility.

The present results have implications for the standardisation of sentencing. Standardisation may be either relative or absolute.[16] Ordinal regression can, in principle, estimate the absolute parameters of the "true" distribution of sentencing intentions – and defining these parameters should inform legislation. Once legislators have defined the limits of sentences, beta regression may be optimal for calibrating the 'severity' of individual judges (and other raters).

The main limitations of the present study are those of the source study.[8] Specifically, (1) this is a small online study of a single fictional crime. The ecological validity and generalisability of this design are open to question. Additionally, (2) the source study recruited only 50 judges, many of whom specialised in non-criminal branches of law. (3) The age and sex distributions of the judges and lay people differed. However, the present study adjusted rater status (lay person or judge) for age and sex. (4) The study did not match social class or educational levels between the judges and lay raters (although it is likely that these factors were more homogeneous among the judges and so were unlikely to explain the greater heterogeneity of judges' sentences). (5) The source study only allowed custodial sentences and the present methods focus on these; many further complications may accompany the analysis of sentencing for offences that allow a range of different *types* of penalty.[31] Further studies should aim to test the present methods over a range of crimes and sentence types (e.g. [32]) in real-life settings.

No psychometric enterprise can be 100% reliable.[16] However, it is disquieting that judges are *less* reliable that *lay* people in the source study's sample.[8] Similar inverted variability has occurred in medical examinations.[33] Therefore, the present results may generalise to other complex assessments – such as the UK GMC's strategy of delegating assessment of doctors' "fitness to practice" to individual "Responsible Officers" (ROs).[34,35] The present results also raise the question whether, in principle, it is reasonable for society to delegate full responsibility for complex assessments to individuals – no matter how experienced and well-trained they may be (c.f. [36]).I argue elsewhere that assessing fitness to practice may be best achieved via statistical assessment of routine outcome data that link individual doctors and patients (c.f.[37,38]; Williams, in preparation). By analogy, the present study suggests a means to 'calibrate' judges (and other people who rate complex problems, such as ROs), by asking them to rate a range of scenarios and comparing their ratings with those of lay people (c.f. [37,38]). The present methods would allow both relative and absolute calibration of each rater.

The real-world consequences of poor calibration of raters in high-stakes scenarios – such as judges and ROs – are very important. In the legal setting, judges must balance the needs for incapacitation, retribution and rehabilitation. The present findings suggest that individual judges' balance these needs very differently. Potentially, therefore, some judges may place the public at risk, if they under-weight incapacitation, while other judges may over-weight retribution and/or rehabilitation – which may waste public money and offenders' lives by keeping them incarcerated for too long (c.f. [15]). Relating raters' calibrations with the long-term outcomes of real cases that they rate (e.g. [38]) would then enable regulators and legislators to refine their policies and find the best balance between public expense and public safety.

==========

References

1. Robbennolot, J. K. Evaluating Juries by Comparison to Judges: A Benchmark for Judging? *Fla. State Univ. Law Rev.* **32**, 462–509 (2005).
2. Johnson, B. D. THE MULTILEVEL CONTEXT OF CRIMINAL SENTENCING: INTEGRATING JUDGE- AND COUNTY-LEVEL INFLUENCES. *Criminology* **44**, 259–298 (2006).
3. Diamond, S. & Zeisel, H. Sentencing Councils: A Study of Sentence Disparity and Its Reduction. *Univ. Chic. Law Rev.* **109**, (1975).
4. Silver, E., Ulmer, J. T. & Silver, J. R. Do moral intuitions influence judges' sentencing decisions? A multilevel study of criminal court sentencing in Pennsylvania. *Soc. Sci. Res.* **115**, 102927 (2023).
5. Cook, D. A. & Beckman, T. J. Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *Am. J. Med.* **119**, 166.e7-166.e16 (2006).
6. Sullivan, G. M. A Primer on the Validity of Assessment Instruments. *J. Grad. Med. Educ.* **3**, 119–120 (2011).
7. Reliability vs. Validity: What Matters More in Research? - Socio.Health. https://socio.health/research-methodology-population-family-health/reliability-vs-validity-research-importance/ (2024).
8. Rust, J. Discussion piece: The psychometric principles of assessment. *Res. Matters* **3**, (2007).
9. Watamura, E. & Ioku, T. Comparing sentencing judgments of judges and laypeople: The role of justifications. *PLOS ONE* **17**, e0277939 (2022).
10. Brinkman, N., Looman, R., Jayakumar, P., Ring, D. & Choi, S. Is It Possible to Develop a Patient-reported Experience Measure With Lower Ceiling Effect? *Clin. Orthop.* **483**, 693–703 (2025).
11. Tamhane, A., Ankenman, B. & Yang, Y. The beta distribution as a latent response model for ordinal data (I): Estimation of location and dispersion parameters. *J. Stat. Comput. Simul.* **72**, 473–494 (2002).
12. Bilcke, J., Hens, N. & Beutels, P. Quality-of-life: a many-splendored thing? Belgian population norms and 34 potential determinants explored by beta regression. *Qual. Life Res. Int. J. Qual. Life Asp. Treat. Care Rehabil.* **26**, 2011–2023 (2017).
13. Shanks, M. Winning and Losing at the Federal District Courts. (South Carolina, 2024).
14. Wallace, A. & Goodman-Delahunty, J. Measuring Trust and Confidence in Courts. *Int. J. Court Adm.* **12**, 3 (2021).

Supplementary Information

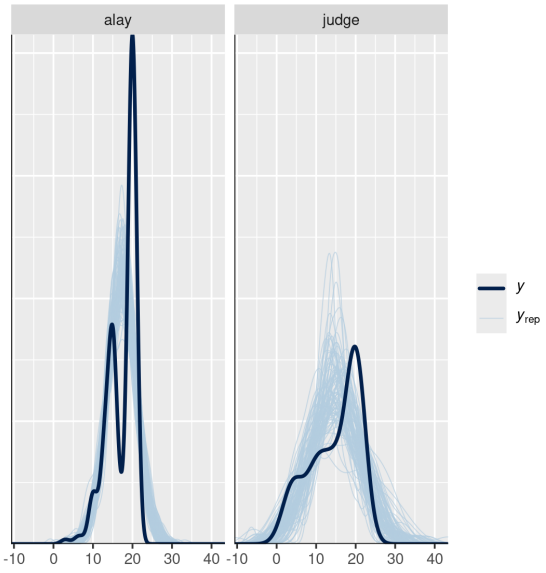
Methods

I used the cauchit link for the ordinal model because it may be optimal for noisy data with outliers and may be able to account for 'guessing' effects and can provide reliable results with much smaller samples[39].

Results

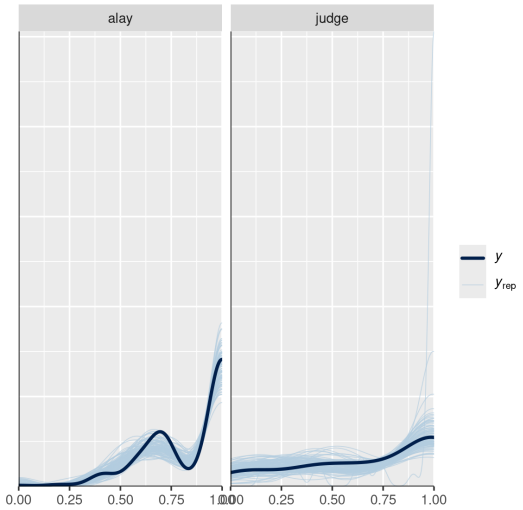
Figures S1-S3 show the posterior-predictive checks for the simple regression (S1), the zero-and-one-inflated beta regression (S2) and the ordinal regression (S3). In each case, the left-hand panel shows the distributions of the original and fitted data for the lay raters ('alay') and the right-hand panel shows the same data for judges ('judge').

Figure S1: the posterior predictive check for the simple linear regression



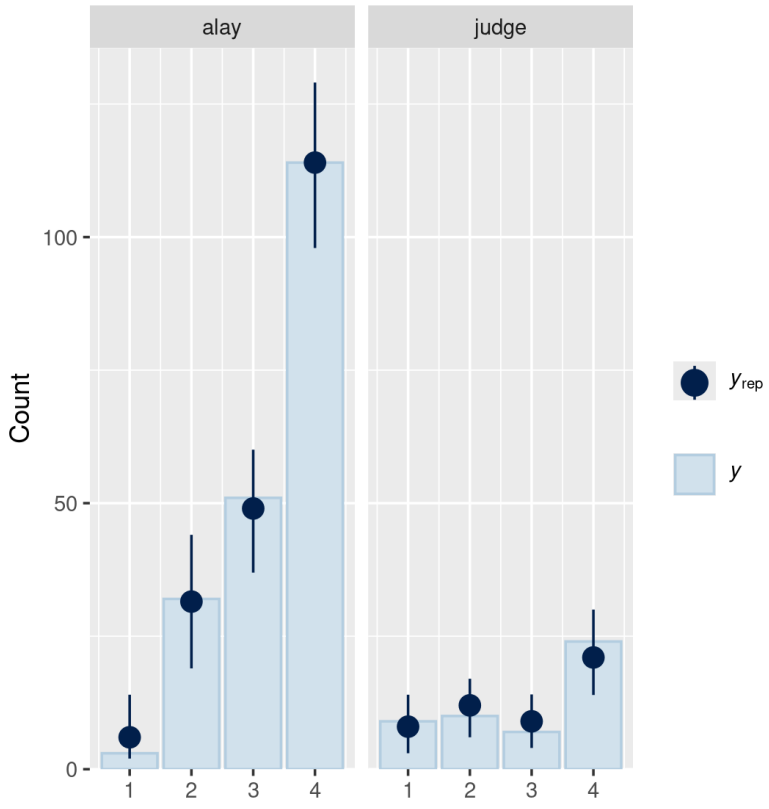
Legend: The figure shows the distribution of the raw sentences for each type of rater (lay, or judge, in black) and the distributions of 100 random draws from the modeled posterior distribution from the Bayesian simple regression (light blue). Note that the model-generated predictions bear little resemblance to the distributions of the raw data for each group

Figure S2: the posterior predictive check for the zero-one-inflated beta regression



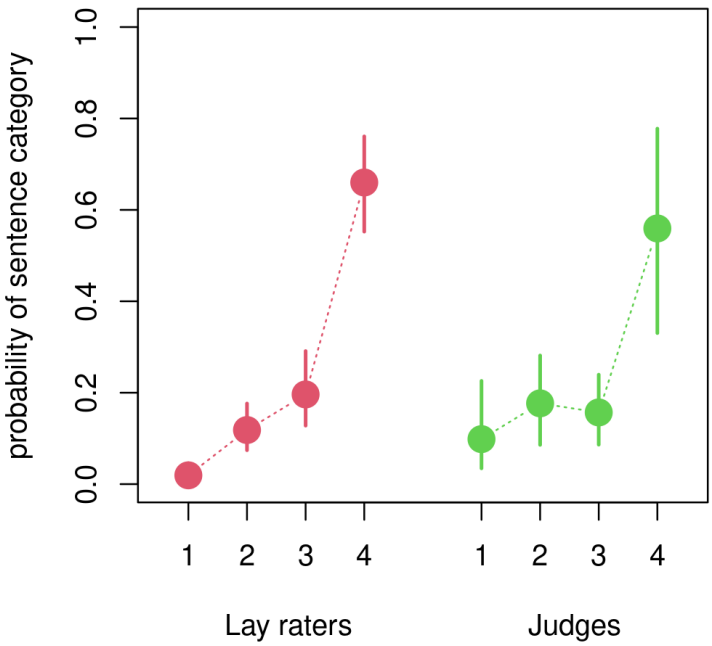
Legend: The figure shows the distribution of the raw sentences, expressed as a proportion of the range allowed by the sentencing guidelines, for each type of rater (lay, or judge, in black) and the distributions of 100 random draws from the modeled posterior distribution from the Bayesian zero-one-inflated beta regression (light blue). Note that the model-generated predictions resemble the distributions of the raw proportions for each group quite closely.

Figure S3: the posterior predictive check for the ordinal regression



Legend: The figure shows the distribution of the rank-transformed sentences for each type of rater (lay, or judge, in black) and the distributions of 100 random draws from the modeled posterior distribution from the Bayesian zero-one-inflated beta regression (light blue). Note that the model-generated predictions resemble the distributions of the raw proportions for each group quite closely.

Figure S4: conditional effects of rater status on sentences as categories of the sentencing guidelines



Legend: The figure shows the conditional effects from the Bayesian model of sentencing intentions cut into four categories (3-9, 10-14, 15-19 or years). Values are the sentencing intentions of average raters, adjusted for age and sex effects on both the probability of being a judge and on sentencing. Error bars show the 95% Credibility Intervals (CIs). Note that the slope across the four categories is less steep for average judges (green) than for average lay people (red), so that judges 'discriminate' less clearly between categories of sentence length (Table 2).

1. Shim, H., Bonifay, W. & Wiedermann, W. Parsimonious item response theory modeling with the cauchit link: Revisiting the rationale of the four-parameter logistic model. *Behav. Res. Methods* **57,** 176 (2025).

Annotated R code to generate all analyses, figures and tables

```
# Watamura E, Ioku T (2022) Comparing sentencing judgments of judges and laypeople: The role of justifications.
# PLoS ONE 17(11): e0277939. https://doi.org/10.1371/journal.pone.0277939
# They present 2 hypotheses:-
# Hypothesis 1: Judges are not biased toward retribution but give importance to other justifications as well when sentencing.
# Hypothesis 2: Judges' sentences are shorter than those of laypeople.
#
# My hypothesis is that judges' sentences are less variable than the general public's
# due to judges' long training and experience. But to show this requires distributional regressions

# dispersion parameter for the beta [distribution] is phi with variance equal to mean*(1-mean)/(1+phi)
# ... "a high dispersion parameter value results in smaller variance".
# https://stats.stackexchange.com/questions/459460/confused-about-over-dispersion-for-my-beta-distribution

# one way to tackle the sentencing is to treat the values as %s of the allowed range (3-20 years)
# -> can estimate dispersion (phi), and zero- and one-inflation, according to group etc.

# another way is to assume that observed sentences represent an underlying latent normal distribution of opinions
# and cut the actual values into ordinal ratings, to allow estimation of the latent normal distribution
# -> can estimation discrimination according to group, etc.

# read in the data
# download the data file 'Comparing Sentencing Judgments of Judges and Laypeople.xlsx' from Watamura (reference above)
library(xlsx) # set 'mydir' to your data directory
dat=read.xlsx('/mydir/Comparing Sentencing Judgments of Judges and Laypeople.xlsx', sheetIndex=1, startRow=2)
# I then saved the relevant data sheet as a csv file and used its data
dat=read.csv('/mydir/Comparing Sentencing Judgments of Judges and Laypeople.csv',skip=1)
names(dat); dim(dat)
# I renamed the variables in the file.
names(dat)=c('id','status','pretrib','pincap','pdeter','prehab','rretrib1','rretrib2','rodeter1','rodeter2',
'rincap1','rincap2','rrehab1','rrehab2','rgdeter1','rgdeter2','rposprev1','rposprev2','sentence','gender','age')
# The only variables relevant here are status (judge/lay), sentence (years), gender and age.

# centre age on 40 years and divide by 10
if (mean(dat$age)>30) dat$age=(dat$age-40)/10; quantile(dat$age)
# create new variable "sent5" of rank-transformed sentences
dat$sent5=ordered(cut(dat$sentence,c(0,9,14,19,30),labels=F)); table(dat$sent5)
# create new variable "psent" of sentences transformed to proportion of sentencing guidelines
dat$psent=((dat$sentence-3)/17) # (guidelines are 3 - 20 years - so the range is 17)

# recode gender and rename as sex
dat$sex=''; dat$sex[dat$gender==2]='F'; dat$sex[dat$gender==1]='M'; dat$sex=factor(dat$sex)
# rename status as "rtype" - i.e. rater type - with levels "judge" or "alay"
# (the preceding 'a' causes brms to treat "alay" as the baseline)
```

```
dat$rtype=''; dat$rtype[dat$status==1]='judge'; dat$rtype[dat$status==3]='alay';
dat$rtype=factor(dat$rtype)
table(dat$sex,dat$rtype)

# quod est explicandum - in fact, this is a coarsened version of Figure 1 in the report
xl='proportion of sentencing guideline'; m=''
plot(density(dat$psent[dat$rtype=='alay'],adjust=.5),xlab=xl,main=m)
lines(density(dat$psent[dat$rtype=='judge'],adjust=.5),lty=2,col=2)
legend('topleft',lty=1:2, col=1:2, legend=c('laypeople','judges'))

t1=table(dat$rtype,dat$sentence); t1; chisq.test(t1)
# X-squared = 48.026, df = 14, p-value = 1.302e-05 - quoted

# Figure 1
barplot(t1, beside=T, col=2:3, xlab="length of sentence (years)", ylab='count', axis.lty=1,
ylim=c(0,120))
legend('topleft', fill=2:3, legend=c('laypeople','judges'))
abline(h=c(0,120),v=c(-.75,46.75), lty=1,lwd=1.5)

jok=dat$rtype=='judge'; t1=table(cok,jok); t1; fisher.test(t1) # NS
var(dat$sentence[jok])
var(dat$sentence[jok])/var(dat$sentence[!jok]) # 2.91 - quoted in report
1-pf(var(dat$sentence[jok])/var(dat$sentence[!jok]),49,299) # 1e-8 - quoted in report

# same pattern of results if I omit the people who give ceiling sentences of 20 years
cok=dat$sentence<20
var(dat$sentence[cok&jok])/var(dat$sentence[cok&!jok]) # 2.93 - not quoted in report
1-pf(var(dat$sentence[cok&jok])/var(dat$sentence[cok&!jok]),25,85) # 0.0001 - not quoted

# now perform simple linear regression
r0=lm(sentence~age+sex, data=dat); summary(r0)
r1=lm(sentence~age+sex+rtype, data=dat); summary(r1)
# compute variance attributable to status
summary(r1)$adj.r.squared-summary(r0)$adj.r.squared # 0.03929997 - quoted
# note that Rsq is 0.066 - which is identical to that from Bayes_R2 for j01a, below

# now use distributional regression
a0=brmsformula(rtype~age+sex, family=bernoulli)
m00=brmsformula(sentence~age+sex+rtype) # ordinary Normal distribution for sentence
# estimate the model
f00=brm(a0+m00, data=dat, backend='cmdstanr', seed=1234, cores=4, chains=4,
save_pars=save_pars(all=T), iter=1e4)
# now with variability of sentencing - sigma - predicted by age and sex
m0=brmsformula(sentence~age+sex+rtype, sigma~age+sex)
f0=brm(a0+m0, data=dat, backend='cmdstanr', seed=1234, cores=4, chains=4,
save_pars=save_pars(all=T), iter=1e4)
# finally, with variability regressed on age, sex and rtype
m1=brmsformula(sentence~age+sex+rtype, sigma~age+sex+rtype)
f1=brm(a0+m1, data=dat, backend='cmdstanr', seed=1234, cores=4, chains=4,
save_pars=save_pars(all=T), iter=1e4)
loo(f00,f0,f1); bayes_R2(f00); bayes_R2(f0); bayes_R2(f1)
# Supplementary Figure S1
pp_check(f1, type='dens_overlay_grouped', resp='sentence',ndraws=100, group='rtype')
# Model comparisons:
#     elpd_diff se_diff
# f1    0.0       0.0
# f0   -8.2       4.3
```

```
# f00 -14.4      6.2
#     Estimate  Est.Error      Q2.5     Q97.5
# R2  0.0856655 0.03049427 0.03215322 0.1506324 # excluding modelling sigma (variance)
#     Estimate  Est.Error      Q2.5     Q97.5
# R2  0.0924529 0.03231211 0.03585434 0.1607784 # excluding sigma~rtype
#     Estimate  Est.Error      Q2.5     Q97.5
# R2  0.0845655 0.03858961 0.02131361 0.1691483 # including sigma~rtype

bayes_factor(f1,f0,log=T)
#Estimated log Bayes factor in favor of f1 over f0: 7.56766 (raw = 1934.608)
# how does it happen that the best-fitting model has a lower R-square?
# I believe the Bayes Factor in preference to the R-square

a0=brmsformula(rtype~age+sex, family=bernoulli)
a1=brmsformula(psent~age+sex+rtype)
j1a=brm(a0+a1, data=dat, backend='cmdstanr', seed=1234, cores=4, chains=4,
save_pars=save_pars(all=T), iter=1e4)
summary(j1a); loo(j1a)
pp_check(j1a, resp='psent', group='rtype', type="dens_overlay_grouped", ndraws=100)
br2a=bayes_R2(j1a); br2a
#         Estimate  Est.Error      Q2.5     Q97.5
# R2rtype 0.15410625 0.03896368 0.08116516 0.2321279
# R2psent 0.08580302 0.03101540 0.03106753 0.1517093

r000=brmsformula(psent~age+sex, phi~age+sex, zoi~1, coi~age+sex,
family=zero_one_inflated_beta())
j000=brm(a0+r000, data=dat, backend='cmdstanr', seed=1234, cores=4, chains=4,
save_pars=save_pars(all=T), iter=1e4)
summary(j000); loo(j1a, j000)
r001=brmsformula(psent~age+sex+rtype, phi~age+sex+rtype, zoi~1,
family=zero_one_inflated_beta())
j001=brm(a0+r001, data=dat, backend='cmdstanr', seed=1234, cores=4, chains=4,
save_pars=save_pars(all=T), iter=1e4)
summary(j001); loo(j000, j001); bayes_R2(j001)
bayes_factor(j001,j000,log=T)
# Estimated **log** Bayes factor in favor of j001 over j000: 11.13485
# raw Bayes factor = 68517.88

# re-ran using cauchit link (because of wide scatter perhaps representing outliers)
# Estimated Bayes factor in favor of j000c over j000: 0.66549 -> cauchit not helpful
# Estimated Bayes factor in favor of j000c over j000: 0.23461 -> probit not helpful
# Estimated Bayes factor in favor of j000c over j000: 0.26566 -> cloglog not helpful
# Estimated Bayes factor in favor of j000c over j000: 0.48435 -> loglog not helpful
# inverse of cloglog, derived from analysing 1-psent, with zoi~age+sex, coi~1

r002=brmsformula(psent~age+sex+rtype, phi~age+sex+rtype, zoi~1, coi~age+sex+rtype,
family=zero_one_inflated_beta())
j002=brm(a0+r002, data=dat, backend='cmdstanr', seed=1234, cores=4, chains=4,
save_pars=save_pars(all=T), iter=1e4)
summary(j002); loo(j000, j001, j002)
# Supplementary Figure S2 - with fitted rtype
pp_check(j002, resp='psent', type="dens_overlay_grouped", group='rtype', ndraws=100)

bayes_factor(j002,j001, log=T) # Estimated **log** Bayes factor in favor of j002 over j001:
10.84141
br22=bayes_R2(j002); br22; br22[1,1]/br2a[1,1]; br22[2,1]/br2a[2,1] # 1.36897
# R2rtype 0.1543306 0.03870164 0.08207899 0.2328526
```

```
# R2psent 0.1174617 0.04800560 0.03638763 0.2166732
# [1] 1.001456
# [1] 1.36897

lay_phi=exp(-fixef(j002)['phi_psent_Intercept',1]); lay_phi
judge_phi=exp(-fixef(j002)['phi_psent_Intercept',1]-fixef(j002)['phi_psent_rtypejudge',1]);
judge_phi
judge_phi/lay_phi # judges 3x more variable than lay people
plogis(fixef(j002)['coi_psent_Intercept',1]) # .999748 - is this p(laypeople who want higher max)?
plogis(fixef(j002)['coi_psent_Intercept',1]+fixef(j002)['coi_psent_rtypejudge',1]) # 0.9574

r0020=brmsformula(psent~age+sex+rtype, phi~age+sex, zoi~1, coi~age+sex+rtype,
family=zero_one_inflated_beta())
j0020=brm(a0+r0020, data=dat, backend='cmdstanr', seed=1234, cores=4, chains=4,
save_pars=save_pars(all=T), iter=1e4)
summary(j0020); loo(j000, j001, j002, j0020)
bayes_factor(j002,j0020) # Estimated Bayes factor in favor of j002 over j0020: 662.74820
bayes_R2(j002); bayes_R2(j0020)
#         Estimate  Est.Error      Q2.5     Q97.5
# R2rtype 0.1543306 0.03870164 0.08207899 0.2328526
# R2psent 0.1174617 0.04800560 0.03638763 0.2166732
#         Estimate  Est.Error      Q2.5     Q97.5
# R2rtype 0.1538590 0.03848880 0.08137537 0.2315116
# R2psent 0.1212331 0.04771908 0.04039607 0.2194634

round(fixef(j002)[,c(1,3,4)],2) # Table 1 in report
#                  Estimate  Q2.5 Q97.5
# rtype_Intercept       -3.44 -4.52 -2.54
# psent_Intercept        0.69  0.51  0.88
# phi_psent_Intercept    2.60  2.13  3.03
# zoi_psent_Intercept    0.27  0.03  0.52
# coi_psent_Intercept    8.29  4.30 15.27
# rtype_age              0.44  0.17  0.74
# rtype_sexM             2.25  1.33  3.36
# psent_age             -0.05 -0.13  0.04
# psent_sexM            -0.21 -0.46  0.04
# psent_rtypejudge      -0.72 -1.15 -0.31
# phi_psent_age         -0.06 -0.24  0.13
# phi_psent_sexM        -0.11 -0.67  0.46
# phi_psent_rtypejudge  -1.12 -1.77 -0.51
# coi_psent_age          2.12  0.64  4.04
# coi_psent_sexM        -2.90 -9.75  0.96
# coi_psent_rtypejudge  -5.17 -9.57 -1.64

set.seed(123); p1=predict(j002)[,,2]; str(p1)
boxplot(p1[,1]~dat$rtype, varwidth=T, notch=T, xlab='rater status', ylab='proportion of sentence
guideline',xaxt='n', col=2:3)
axis(side=1, at=1:2, labels=c('lay person','judge'))

# Figure 2 in report - uses brms estimates of conditional effects
c1=conditional_effects(j002, resp='psent')[[3]]; str(c1)
plot(1:2,c1$estimate,ylim=c(.5,1),pch=19,cex=2,xlim=c(.5,2.5),xaxt='n',xlab='rater
status',ylab='proportion of sentence guideline',col=2:3)
lines(x=c(1,1),y=c(c1$lower__[1],c1$upper__[1]),lwd=2,col=2)
lines(x=c(2,2),y=c(c1$lower__[2],c1$upper__[2]),lwd=2,col=3)
axis(side=1, at=1:2, labels=c('lay people','judges'))
```

```
# now assess latent sentencing intentions
a0=brmsformula(rtype~age+sex, family=bernoulli)
r00=brmsformula(sent5~age+sex, disc~0+age+sex, family=cumulative)
j00=brm(r00+a0, data=dat, backend='cmdstanr', seed=1234, cores=4, chains=4,
save_pars=save_pars(all=T),
control=list(adapt_delta=.9999, max_treedepth=20), iter=1e4)
pp_check(j00, type='bars', categorical=T, resp='sent5')
pp_check(j00, type='bars_grouped', categorical=T, group='rtype', resp='sent5')
# continuation ratio is where lower levels must be completed to achieve current level
# e.g. GCSE -> A-levels -> BA -> MA -> PhD
# not sure how cumulative includes this assumption -> better to use continuation ratio
# continuation ratio may in fact model rationale of increasing severity

r00c=brmsformula(sent5~age+sex, disc~age+sex, family=cratio)
j00c=brm(r00c+a0, data=dat, backend='cmdstanr', seed=1234, cores=4, chains=4,
save_pars=save_pars(all=T),
control=list(adapt_delta=.9999, max_treedepth=20), iter=1e4)
summary(j00c); loo(j00, j00c); bayes_factor(j00c, j00)
# Estimated Bayes factor in favor of j00c over j00: 2.86075

r00c2=brmsformula(sent5~age+sex, disc~age+sex, family=cratio(link='cauchit'))
j00c2=brm(r00c2+a0, data=dat, , backend='cmdstanr', seed=1234, cores=4, chains=4,
save_pars=save_pars(all=T), control=list(adapt_delta=.9999, max_treedepth=20), iter=1e4)
summary(j00c2); loo(j00, j00c, j00c2)
pp_check(j00c, type='bars', resp='sent5')
pp_check(j01, resp='sent5', type="rootogram")
bayes_factor(j00c2, j00c) # Estimated Bayes factor in favor of j00c2 over j00c: 8.89280
bayes_factor(j00c2, j00)  # Estimated Bayes factor in favor of j00c2 over j00: 26.06198
# so cratio with cauchit link is best - and importantly better than cumulative with logit link

# now estimate judge vs lay person
r01=brmsformula(sent5~age+sex+rtype, disc~age+sex+rtype, family=cratio(link='cauchit'))
j01=brm(r01+a0, data=dat, , backend='cmdstanr', seed=1234, cores=4, chains=4,
save_pars=save_pars(all=T), control=list(adapt_delta=.9999, max_treedepth=20), iter=1e4)
summary(j01); loo(j00, j00c, j00c2, j01)
bayes_factor(j01, j00c2) # Estimated Bayes factor in favor of j01 over j00c2: 449.63610
# bayes_R2(j01) - probably not valid for ordinal responses
pp_check(j01, type='bars', resp='sent5', ndraws=100) # looks good
# Supplementary Figure S3
pp_check(j01, type='bars_grouped', resp='sent5', group='rtype', ndraws=100) # still looks good
round(fixef(j01)[,c(1,3,4)],2) # Table 2 in report
#                   Estimate   Q2.5 Q97.5
# sent5_Intercept[1]     -12.69 -34.56 -3.51
# sent5_Intercept[2]      -1.93  -5.12 -0.47
# sent5_Intercept[3]      -0.91  -2.67 -0.17
# disc_sent5_Intercept     0.40  -0.80  1.64
# rtype_Intercept         -3.43  -4.49 -2.54
# sent5_age               -0.03  -0.82  1.01
# sent5_sexM               0.11  -0.88  1.35
# sent5_rtypejudge         7.22  -0.09 21.42
# disc_sent5_age           0.11  -0.47  0.54
# disc_sent5_sexM         -0.57  -1.41  0.26
# disc_sent5_rtypejudge   -2.16  -3.63 -0.60
# rtype_age                0.44   0.17  0.74
# rtype_sexM               2.24   1.32  3.32

# in brms:
```

```
# "disc is not the variance itself, but the inverse of the standard deviation, "s".
# That is, s = 1/disc. Further, because disc must be strictly positive,
# it is by default modeled on the log-scale" (Buerkner & Vuorre, 2019)
# so s=1/exp(disc_estimate)

# use "distributions3" package for estimation of Cauchy distributions
library(distributions3)
# Figure S4 in Supplementary Information
yl='point probability of maximum sentence'
xl='severity of rater / perceived gravity of crime'
delta=location=fixef(j01)['sent5_Intercept[3]',1]
alpha=1/exp(fixef(j01)['disc_sent5_Intercept',1])
lay=Cauchy(delta,alpha); curve(pdf(lay,x),xlim=c(-6,16),xlab=xl,ylab=yl)
alphaj=1/exp(alpha+fixef(j01)['disc_sent5_rtypejudge',1])
deltaj=fixef(j01)['sent5_rtypejudge',1]
judge=Cauchy(deltaj,alphaj); curve(pdf(judge,x),add=T,lty=2)
legend('topright',lty=1:2, legend=c('public','judge'))

# Figure S4 in report
delta=location=fixef(j01)['sent5_Intercept[3]',1]
alpha=1/exp(fixef(j01)['disc_sent5_Intercept',1])
yl='cumulative probability of maximum sentence'
xl='severity of rater / perceived gravity of crime'
lay=Cauchy(delta,alpha); curve(cdf(lay,x),xlim=c(-6,16),xlab=xl,ylab=yl)
alphaj=1/exp(alpha+fixef(j01)['disc_sent5_rtypejudge',1])
deltaj=fixef(j01)['sent5_rtypejudge',1]
judge=Cauchy(deltaj,alphaj); curve(cdf(judge,x),add=T,lty=2)
legend('bottomright',lty=1:2, legend=c('public','judge'))

# Figure 3 in report - conditional effects of ordinal regression
c3=conditional_effects(j01, resp='sent5', categorical=T)[[3]]
xl='Lay raters        category        Judges'
xl='Lay raters                  Judges'
plot(1:4,c3$estimate__[seq(1,7,2)],ylim=c(0,1),xlim=c(.5,9.5),pch=19,col=2,cex=2,
xlab=xl,ylab='probability of sentence category',xaxt='n', type='b', lty=3)
points(6:9,c3$estimate__[seq(2,8,2)],pch=19,col=3,cex=2, type='b', lty=3)
j=0; for (i in seq(1,7,2)) {j=j+1
lines(x=c(j,j),y=c(c3$lower__[i],c3$upper__[i]),lwd=2,col=2)
} # end for i
j=5; for (i in seq(2,8,2)) {j=j+1
lines(x=c(j,j),y=c(c3$lower__[i],c3$upper__[i]),lwd=2,col=3)
} # end for i
axis(side=1, at=c(1:4,6:9), labels=c('1','2','3','4','1','2','3','4'))
```