Threshold point utilisation in juror decision-making

Lee J. Curley^a, Rory MacLean^a, Jennifer Murray^a, Andrew C. Pollock^b
and Phyllis Laybourn^a

^a*School of Applied Sciences, Edinburgh Napier University, Edinburgh, Scotland;* ^b*1 Pladda Avenue, Irvine, KA11 1DR, Scotland, UK*

This study aims to identify whether a model of juror decision-making (i.e. the threshold point model) that encompasses both rational and intuitive decision-making exists. A total of 60 participants were selected who are eligible for jury duty in Scotland. These individuals read nine vignettes and rated the evidence of each vignette separately by placing the evidence in either a guilty, a not guilty or a not proven (a verdict type specific to Scotland) counter. Participants were asked after being presented with each piece of information to state how likely they thought the suspect was of being guilty, on a scale from 1 to 100. The data are best described using a flexible model (i.e. a diffusion model) that allows for information integration. Future research should examine whether or not the diffusion model can explain cognitive fallacies, such as confirmation bias, that are commonly studied in decision science.

KEY WORDS: jurors; decision-making; stopping rule; information integration; heuristics; cue utilisation; law; psychology; courtroom; not proven verdict.

Introduction

Jurors make potentially life-changing decisions, with relatively little training and potentially no previous experience. They are given the responsibility of delivering justice to both the state and the victim (Roberts & Murray, 2013; Wenzel, Okimoto, Feather, & Platow, 2008). This raises the following question: how do jurors make decisions despite their lack of legal experience? The current study aims to identify the process through which legal laypersons reach decisions when they are presented with legal information, with a specific focus on the Scottish legal system.

Much of the legal and decision-making research focuses on the United States (US) legal system and jurors (Dhami & Ayton,

2001). These investigations cannot easily be generalised to the Scottish legal system and jurors, because the Scottish legal system has unique aspects such as the not proven verdict, an extra acquittal verdict that is uniquely utilised in Scotland and has no legal definition (Duff, 1999).

From the limited research on juror decision-making emerging from Scotland, there are some notable examples. For instance, Smithson, Deady, and Gracik (2007) found task difficulty to be the only variable that consistently forecasts not proven verdicts. In addition, Hope, Greene, Memon, Gavisk, and Houston (2008) focus on how the third verdict of not proven affects conviction rates (i.e. outcome analysis) when the strength of the

evidence varies. It was found that in cases considered to have evidence of 'moderate' strength, the availability of the not proven verdict reduces conviction rates, and that if a not proven verdict is available, the number of not guilty verdicts is also reduced. The research by Smithson et al. (2007) and Hope et al. (2008) focuses on decision outcomes, however, and does not explore the cognitive processes that resulted in certain decisions being reached. To understand these outcomes better, more research is needed that investigates the processes behind them.

One contemporary area of decision-making research which puts the focus on processes rather than outcomes and can be applied to the courtroom is fast and frugal heuristics (or 'cognitive rules of thumb'; Murray & Thomson, 2010). Gigerenzer and Goldstein (1996) found that one of their heuristics – known as 'take the best' (TTB), wherein an option is chosen based on the first cue that allows a binary set of responses to be discriminated – is as accurate as multiple regression analysis when making decisions in regard to city sizes.

Lee and Cummins (2004) used the fast and frugal heuristic metaphor when investigating how individuals decide if a gas is poisonous or not, and evolved it into a more global theory that can explain decision-making data more adequately. Their experiment compares a rational decision-making strategy (called RAT) with the TTB model (where information is searched in the memory of the decision-maker from the most valid to the least valid cue; Gigerenzer & Goldstein, 1996). Both strategies were found to have been used, but not exclusively, across 52.5% of the sample, meaning that some participants used strategies that resembled both RAT and the TTB model (i.e. intra-strategy differences) over the course of the five comparisons (Bergert & Nosofsky, 2007; Lee & Cummins, 2004). Furthermore, a unified threshold point model (which combines RAT and TTB) was found to account for more of the decision-making behaviour than

any single model could alone; it accounts for 84.5% (compared to 64.0% with RAT and 36.0% with TTB) of the decision-making strategies utilised by the participants (Lee & Cummins, 2004). Therefore, a more suitable metaphor for decision-making may be that the participants adapted their stopping rule (which can be referred to as a threshold point) to the task, as this allowed both strategies (i.e. RAT and TTB) to be incorporated into a single model. For instance, a small information search would mirror heuristic decision-making, and a larger information search would mirror rational decision-making (Newell & Lee, 2009).

Several different threshold point models can potentially explain juror decision-making, each of which fits into two broad categories involving absolute stopping rules (the counter model) and relative stopping rules (the diffusion model). Although there are a number of individual threshold point decision-making models, the current study just concentrates on these two broad categories.

The models mentioned above are mathematical models, and Ratcliff and Smith (2004) should be consulted for a more mathematical description of them. However, the current study is one of the first that tries to map the models onto juror decision-making data and then differentiate between them using inferential statistics and dependent measures, such as likelihood-of-guilt data to assess whether or not threshold points are reached through information integration, and evidence-rating data to investigate whether or not threshold points are reached through the counting of different evidence ratings into specific counters.

Absolute-stopping-rule models propose that when individuals such as jurors are making decisions, they collect evidence on separate counters (alternatives) with separate threshold points (Gold & Shadlen, 2007; Potter, 2011; Smith & Ratcliff, 2004). For instance, in the case of jurors, the two counters could relate to verdict decisions (Walters, 2007), with one counter for a guilty verdict

and the other counter for a not guilty verdict (Smith & Ratcliff, 2004). In these counters, the same mechanism as described by Lee and Cummins (2004) occurs. The threshold point which is reached first leads to an absolute stopping rule, and causes a decision to be made (Ratcliff & Smith, 2004; Rouder, 2001). Consequently, if the threshold point in the guilty counter is reached first then the juror gives a guilty verdict, whereas if the counter in the not guilty verdict is reached first then a not guilty verdict is returned. Further, in race models, where evidence races towards a threshold point, there may be more than two counts, which suggests that Scottish jurors could count not proven evidence until a threshold point is reached (Ratcliff & Smith, 2004). Similarly, some counter models, like the linear ballistic model, have been shown to be an appropriate metaphor for choices that have more than two outcomes (Brown & Heathcote, 2008).

In the current study, count data were measured by asking participants to count information and evidence into one of three counters. Further, the participants were able to place information into either the guilty counter, the not guilty counter or the not proven counter. Previous research provides support for the counter model in the courtroom (Thomas & Hogue, 1976). However, the current study is the first to establish whether or not count information allows threshold points to be reached – thus allowing verdicts to be given – using inferential statistics.

The second theoretical approach relates to relative stopping rules. Ratcliff and Smith (2004) outline that with relative stopping rules, information integration occurs until one alternative is favoured over another, symbolised by a threshold point, at which point the appropriate decision is made. In other words, if the integration of the evidence causes a certain amount of the information to favour a guilty verdict, upon reaching the guilty threshold point, a guilty verdict is chosen. Therefore, the threshold points in the proposed research should be distinct from one another. The

diffusion model is a more dynamic approach than the absolute-stopping-rule approach (Ratcliff & Smith, 2004). This is because in the relative-stopping-rule approach, the two threshold points are on the same continuum and evidence is integrated together, which means that information that supports one outcome has a detrimental impact on the other outcome (Gold & Shadlen, 2007).

The diffusion model fits previous decision-making data better than counter models (Ratcliff & Smith, 2004). Furthermore, the prior point in the diffusion model can vary depending on how early the decision-making process starts, which may fit in well with biased jurors, and jurors who might have been affected by pre-trial information (Laming, 1968; Ratcliff & Smith, 2004; Smith & Ratcliff, 2004). A skewed prior point (i.e., the beginning of the decision making process) may cause the decision making to favour one threshold point over the other, which may explain commonly observed biases. Furthermore, prior information may skew the prior point to be closer to the threshold point that best matches or represents that prior information.

Research on this approach suggests that the further away the threshold point is from the prior point, the slower and more accurate the response will be (Ratcliff & Smith, 2004; Smith & Ratcliff, 2004). This suggests that jurors with higher threshold points (who use more information) are more accurate than jurors with lower threshold points (frugal legal laypersons; Sangero & Halpert, 2007). Diffusion models of decision-making have been supported by neuroscience (Smith & Ratcliff, 2004), electroencephalograms (EEG) and psychophysics (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Ratcliff, Philastides, & Sajda, 2009; Ratcliff & Rouder, 1998; Yeung & Summerfield, 2012).

Within a three verdict-system, the researchers propose that the diffusion threshold model will have two threshold points; one representing a guilty verdict and one representing a not guilty verdict. In addition, there will

be three separate 'zones', each symbolizing one of the three verdict types (guilty, not guilty, and not proven). For example, if a juror has an innocent until proven guilty mindset and does not cross the not guilty threshold point into the not proven zone, they will give a not guilty verdict. If a juror crosses the not guilty threshold point into the not proven zone, a not proven zone will be given. If a juror crosses the guilty threshold point into the guilty zone from the not proven zone, a guilty verdict will be given. The threshold points will be measured using likelihood of guilt ratings.

The current study investigates whether or not information integration (using likelihood-of-guilt ratings) allows threshold points to be reached. Previous research has utilised similar methods when studying information integration (see Estrada-Reynolds, Gray, & Nuñez, 2015). The current study investigates whether or not different likelihoods of guilt are associated with different threshold points – that is, once a decision-maker's drift reaches a particular threshold point, does a specific verdict choice become more appropriate? In other words, if the guilty threshold point is reached through information integration, a guilty verdict is given unless further evidence is presented that causes the threshold point to be crossed back over, leading to a change in the verdict (Potter, 2011). This has been confirmed in previous research through mathematical modelling (Ratcliff & Smith, 2004). The current study is the first to establish whether or not the diffusion model fits psychological data from mock jurors using inferential statistics. Traditionally in diffusion models there are two points: the prior point and the threshold point. However, an additional point called the 'last point' is added and used in the current study, in order to investigate whether or not drift in jurors continues or changes once they have reached a threshold point, as – unlike in investigations in other fields – jurors have no control over the information search.

The literature highlights three main issues. First, the threshold point model may be more

suitable for use with legal decisions in comparison to heuristics and more rational models. Second, the Scottish legal system, with its three verdict options, is unique and needs proper scientific enquiry. Third, threshold point models of decision-making may help to explain juror decision-making in the unique three-verdict system. Therefore, the current study assesses the use of likelihood data (to symbolise information integration) and evidence ratings that highlight which verdict (guilty, not guilty or not proven) each particular piece of evidence favours (i.e. to represent count data) in an attempt to find out which model – either the counter model or the diffusion model – best describes the decision-making processes of jurors.

It was expected that the likelihood-of-guilt ratings would differ across verdict types. The reasoning behind this is because the likelihood of guilt symbolises information integration (or drift) in the current study, and previous research shows that drift differs depending on the outcome that is reached (Ratcliff & Smith, 2004). Furthermore, it was believed that the three decision points (prior, threshold and last) would differ in the current study. This is because previous research shows that drift allows decision-makers to move from their prior point to the point at which a decision can be made (i.e. the threshold point). Therefore, it was expected that with the diffusion model explaining the decision-making data, the threshold point would differ from the prior point, thus allowing a decision to be made. Similarly, it was expected that the last point would also differ from the prior point, otherwise a verdict associated with the prior point would be given – that is, the not guilty verdict. As such, a verdict may be initially favoured at the prior point because of the 'innocent until proven guilty' belief.

It was also believed that a significant interaction would be found between the verdict given and the decision points. This is because it was expected that different drifts and threshold points (i.e. guilty and not guilty threshold points) would lead to different legal decisions,

which is confirmed by previous research (Ratcliff & Smith, 2004). Contrarily, it was expected that if the count model was found to best explain the decision-making data then the last piece of count data (evidence ratings) needed would have a strong and significant relationship with the verdict given. This is because the last piece of count information needed should push decision-makers beyond the threshold point, thus allowing a verdict to be given. Therefore, it was expected that the verdict given would correspond to the counter that the piece of evidence was placed in – for example, when the last piece of evidence needed is rated as guilty, a guilty verdict is then given. This expectation can be justified by the findings of Thomas and Hogue (1976), who propose that jurors may make decisions using independent counters and threshold points.

In summary, in the current study, drift – or information integration – was measured by asking the participants to state the likelihood of the suspect's guilt after each piece of evidence was presented (in a similar manner to the method used in Estrada-Reynolds et al., 2015). In addition, count evidence was measured by asking the participants to supply individual ratings (guilty, not guilty and not proven) for each piece of evidence, thus providing the counter into which each piece of evidence was placed. The last piece of evidence needed to make a decision symbolises the threshold point being reached, and with this method it has both a corresponding evidence rating (count data) and a likelihood (drift), which makes it possible to assess whether the threshold point was reached through information integration or through the use of a collection of individual pieces of evidence placed in specific counters. The current study consequently aims to investigate which threshold point model of decision-making best applies to the Scottish juror system, where three verdicts are available. It was hypothesised that:

H1. Significant differences would be found between each of the three points

studied (prior, threshold and last) in relation to likelihood of guilt ratings.

H2. Significant differences would be found between each of the three verdict types (guilty, not guilty and not proven) in terms of likelihood of guilt ratings.

H3. A significant interaction would be found between verdict type and diffusion points in relation to likelihood of guilt ratings.

H4. A significant association would be found between the ratings of the last cue/count needed and the verdict given (for each of the different verdict types).

Method

Design

A mixed factorial quasi-experimental design is used, in which verdict (guilty, not guilty, not proven) is the between-subjects factor and decision point (prior, threshold, last) is the within-subjects factor. These three likelihoods were applied because it is important to see if they differ across and between the different verdict types that were given. The prior point is the decision point that existed after the context of the case had been provided but before any evidence had been presented; therefore, it shows whether or not participants had pre-trial verdict preferences before seeing any relevant information. The threshold point is the decision point associated with the stopping rule. The last point is the decision point related to giving a rating of guilty to the last piece of evidence shown in the case.

The first dependent variable utilised in the current study is the likelihood-of-guilt rating, which ranges from 1 to 100 where 1 = *definitely not guilty*, 50 = *not proven*, and 100 = *definitely guilty* (Price & Dahl, 2014; Thomas & Hogue, 1976; Windschitl, Scherer, Smith, & Rose, 2013). This dependent variable highlights whether or not information integration was occurring and allows small changes to be tracked, similar to traditional drift models that use logarithmic odds (Bitzer, Park,

Blankenburg, & Kiebel, 2014; Lee & Cummins, 2004; Vandekerckhove & Tuerlinckx, 2008).

The participants were also asked to rate each piece of evidence as either guilty, not guilty, or not proven. This was to mirror the counter model, wherein pieces of evidence are counted separately. The participants were also asked to state the last piece of evidence they needed to make a decision at the end of each vignette. This symbolises when a threshold point was reached. This information was also used for the dependent measure of cue utilisation (how much evidence was used by the participants). The participants also stated, after each vignette, which verdict they thought was the most appropriate.

Participants

A total of 60 participants (31 females; 21 students and 39 non-students with a range of professional and manual employment) were recruited via opportunity sampling. A post hoc power analysis was run using the software G*power (Faul, Erdfelder, Lang, & Buchner, 2007). The alpha level utilised was $p < .05$, and a large effect size of $F = 0.59$ was chosen because of the uniqueness of this study. The analysis of variance (ANOVA) for the diffusion analysis was found to be adequately powered, with an actual power of .98.

Recruitment posters were placed around university campuses and on social media sites (such as Facebook) and a snowballing technique was also employed by asking participants to pass on the study contact information to others. Individuals were only accepted if they were eligible to serve as Scottish jurors (Scottish Court Service, 2015). The mean age of participants was 26.8 years ($SD = 9.6$).

Materials

Standardised information sheets, consent forms and debriefing sheets were employed with all participants. A legal inventory was used to identify individuals who could not take

part in a real-life jury, and a demographics questionnaire was used to collect participant details such as age and gender.

Vignettes

Nine vignettes were shown using the software package Superlab (Cedrus Corporation, 2014). The presentation order of the vignettes was changed after every 20 participants to reduce order effects. The vignettes are drawn from real-life cases collected from media articles. All vignettes are based on murder trials. Three of the real-life trials resulted in guilty verdicts, three in not proven verdicts, and three in not guilty verdicts.

Literature on vignette development was consulted when designing the vignettes (i.e. Ashill & Yavas, 2006; Heverly, Fitt, & Newman, 1984). All nine vignettes were designed to be of similar length in order to reduce attentional biases. The information in each of the cases falls under a similar theme, and the consistency of the information and evidence across the vignettes was evaluated and compared by two of the authors (LC and JM). The evidence from the prosecution was presented prior to the evidence from the defence in every trial. This was done so that consistent evidence was presented in comparable rankings, thus decreasing primacy and recency effects. The vignettes are all of modest length in order to prevent length from acting as a confounding variable. On average the vignettes are 484 words long. The victims and suspects across the vignettes are comparable in terms of gender and age. Piloting was conducted to guarantee that the familiarity, realism and severity of the trials are alike throughout the nine vignettes.

Each stance (prosecution vs defence) presents 5 to 9 pieces of evidence (Appendix 1 gives more details on the counterbalancing). This range was chosen for two reasons: 1) so that memory constraints would not affect which cues were remembered within each of the stances when making a final verdict, which links with the 7 ± 2 rule (Miller 1956), and 2) for generalisability, since some real-life trials

Table 1. Descriptive statistics for the verdict types.

	Likelihood-of-guilt rating		
	Prior point	Threshold point	Last point
Guilty			
Mean	47.2	83.9	81.8
Median	50.0	85.0	82.5
SD	14.1	9.6	11.5
Not guilty			
Mean	41.5	49.7	39.2
Median	50.0	50.0	40.0
SD	20.2	18.2	15.1
Not proven			
Mean	44.9	62.6	58.0
Median	50.0	63.5	59.0
SD	12.1	11.2	9.9

last longer and involve the presentation of more information than others.

Procedure

All participants read an information sheet and completed a consent form. They were told that their responses would be used for court appeals (a mild deception) so that they would think that their responses had real-world consequences. They then completed the legal inventory and, if found eligible to take part, were given standardised instructions on-screen using Superlab. The procedure for each vignette is identical. The participants were provided with an opening statement to give context to the case. They were then asked to give a prior likelihood of guilt, ranging from 1 to 100. They were also reminded of how to rate the evidence (either guilty, not guilty or not proven) and which keys to press (G, N or P, respectively) prior to the evidence being presented. The participants were then shown the pieces of evidence, with the prosecution’s evidence presented first, followed by the defence’s evidence (Englich, Mussweiler, & Strack, 2005; Justice Education Society of BC, 2016). Cross-examinations are not included in order to avoid problems that could arise from the materials becoming too

complex. The participants rated each piece of evidence as either guilty, not guilty or not proven. They were also asked to give a likelihood-of-guilt rating after each piece of evidence was presented, again ranging from 1 to 100, which symbolises information integration. Once each piece of evidence had been read and evaluated, the participants were asked to give a final verdict: guilty, not guilty or not proven. They were then asked to identify the last piece of evidence that they had needed in order to make their decision. This response symbolises their threshold point being reached, and allows cue utilisation to be quantified. This process was repeated until all nine vignettes had been completed, after which the participants read through the debrief sheet.

Results

Descriptive statistics for the verdict types

Of the 60 participants, 59 gave a guilty verdict at least once, 58 gave a not proven verdict at least once and 39 gave a not guilty verdict at least once. The verdict groupings are therefore, relatively, similar. Data for each of the points (prior, threshold and last) were averaged within the verdict types given by the participants. For example, if an individual gave four

Figure 1. Interaction between likelihoods and verdict types.

not proven verdicts, the data (prior, threshold and last points) that were analysed are an average gathered from these four verdicts. Likewise, if the same participant gave two guilty verdicts and three not guilty verdicts, each of the decision points were averaged within their verdict type, and this average was used for the analysis. Table 1 sets out the descriptive statistics for the data points that were averaged across verdict types.

As shown in Table 1, for guilty verdicts the mean increases from the prior point to the threshold point, and then minimally decreases from the threshold point to the last point. The median shows a similar pattern. The standard deviation seems to decrease from the prior point to the threshold point, which may highlight greater variation in the guilty prior point in comparison to the guilty threshold point. Similarly, for not guilty verdicts there is an increase in the mean between the prior point and the threshold point, and a decrease between the threshold point and the last point. The median remains relatively stable throughout the decision-making process, with a dip between the threshold point and the last point. The standard deviation shows a similar pattern. The standard deviation is higher for not guilty verdicts at all the stages of the decision-making process, thus highlighting a higher

variation in drift within not guilty verdicts. The mean and the median for not proven verdicts change throughout the three points, increasing between the prior point and the threshold point and then decreasing between the threshold point and the last point. The standard deviation decreases through each of the three points. Technically, the not proven verdict does not have a threshold point. However, in order not to confuse the reader, the authors have used the term not proven threshold point to convey when the jurors can make a decision. When reading this, keep in mind that the not proven threshold point is the point where a not proven threshold can be reached, but has been reached by the juror either crossing the guilty or not guilty threshold point.

Investigating verdict choices across decision points

A two-way mixed-groups ANOVA was carried out, with verdict given (guilty, not guilty, not proven) as the between-subjects factor and decision point (prior, threshold, last) as the within-subjects factor. There is a significant main effect of verdict given, $F(2, 153) = 110.5$, $p < .001$, $\eta_p^2 = .59$, large effect size. Tukey's post hoc tests highlight that the guilty

Table 2. Descriptive statistics of cue utilisation between verdict types.

Verdict	Mean	Median	SD
Guilty	6.2	5.0	3.4
Not proven	11.4	11.0	3.3
Not guilty	11.7	11.0	3.0

verdicts have significantly higher likelihood ratings ($M=71.0$) than the not guilty ($M=43.4$, $p=.001$) and not proven ($M=55.2$, $p<.001$) verdicts. The not proven verdicts have likelihood ratings that are significantly higher than those of the not guilty verdicts ($p<.001$).

For the variable of decision points, the Greenhouse–Geisser row was consulted as Mauchley's test of sphericity is significant ($p<.001$). There is a significant main effect of decision point, $F(1.6, 243.5)=122.4$, $p<.001$, $\eta_p^2=.44$, large effect size. Bonferroni post hoc tests revealed that the prior point ($M=44.5$) is significantly lower than both the threshold point ($M=65.4$, $p<.001$) and the last point ($M=59.7$, $p<.001$), and the threshold point is significantly higher than the last point ($p<.001$).

There is a significant interaction between the decision points and the verdicts given, $F(4, 306)=33.0$, $p<.001$, $\eta_p^2=.30$, large effect size. See Figure 1 for a visual representation of the interaction.

Simple main effects show that for the guilty verdicts the prior point ($M=47.2$) is significantly lower than the threshold point ($M=83.9$, $p<.001$) and the last point ($M=81.8$, $p<.001$). The threshold point and the last point are not significantly different from one another ($p=.17$). Looking at the not proven verdicts, it is evident that the prior point ($M=44.9$) is significantly lower than the threshold point ($M=62.6$, $p<.001$) and the last point ($M=58.0$, $p<.001$), and that the threshold point is also significantly higher than the last point ($p<.01$). For the not guilty verdicts, the prior point ($M=41.5$) is significantly lower than the threshold point ($M=49.7$, $p<.01$) but does not significantly differ from

the last point ($M=39.2$, $p=.45$). However, the last point is significantly lower than the threshold point ($p<.001$).

When investigating the prior point, the guilty verdicts ($M=47.2$) do not differ significantly from the not proven ($M=44.9$, $p=.41$) and not guilty ($M=41.5$, $p=.07$) verdicts in relation to likelihood of guilt. Likewise, the not proven and not guilty verdicts do not significantly differ from one another in relation to the likelihood of guilt at the prior point ($p=.27$). For the threshold point, the guilty verdicts ($M=83.9$) have a significantly higher likelihood of guilt than the not proven ($M=62.6$, $p<.001$) and not guilty ($M=49.7$, $p<.001$) verdicts. The not proven and not guilty verdicts also significantly differ from one another in relation to the likelihood of guilt measured at the threshold point ($p<.001$), with the not proven verdicts having a higher likelihood of guilt. In relation to the last point, the guilty verdicts ($M=81.8$) have a significantly higher likelihood of guilt in comparison to the not proven ($M=58.0$, $p<.001$) and not guilty ($M=39.2$, $p<.001$) verdicts. Finally, the last point is significantly higher for the not proven verdicts in comparison to the not guilty verdicts ($p<.001$).

Cue utilisation

Table 2 shows the descriptive statistics for cue utilisation. A one-way between-subjects ANOVA was carried out to compare cue utilisation across the different verdict options. There is a significant difference in the cue utilisation between the three verdict types, $F(2, 537)=173.7$, $p<.001$. Post hoc Tukey's tests highlight that significantly fewer cues were utilised when a guilty decision was given

Table 3. Descriptive statistics for the absolute stopping rule.

Variable	Mean	Median	SD
Last count rated as guilty	4.0	4.0	1.9
Last count rated as not guilty	1.8	1.0	1.7
Last count rated as not proven	3.2	3.0	2.1
Guilty verdict	4.0	4.0	1.7
Not guilty verdict	1.3	1.0	1.3
Not proven verdict	3.6	4.0	2.0

($M = 6.2$) in comparison to when a not proven ($M = 11.4$, $p < .001$) or not guilty ($M = 11.7$, $p < .001$) verdict was given. There are no significant differences in cue utilisation between the not guilty and not proven verdicts ($p = .66$).

Count analysis

Descriptive statistics for the last count rating needed and the verdict given

The descriptive statistics shown in Table 3 relate to the last cue/count ratings needed (which allowed a threshold point to be reached) and the verdicts given by the participants. The data were collected categorically (i.e. guilty, not guilty and not proven). For the current analysis, both the frequency of each verdict type and the frequency of each of the different count types (guilty, not guilty and not proven) needed for a threshold to be reached was counted for every participant. For example, if a participant gave 5 guilty verdicts over the 9 trials, the cumulative score for the number of guilty verdicts given would be 5, and if the participant rated the last cue/count needed as not guilty 4 times, the cumulative score for the last cue/count needed for not guilty verdicts would be 4.

Pearson's correlations were conducted to identify whether or not the last piece of evidence needed allowed a decision to be reached. The correlations highlight that significant relationships exist between the last cue/count rating and the verdict given. A significant relationship exists between the last

piece of cue/count evidence needed (when rated as guilty) and the number of guilty verdicts given, $r = .82$, $p < .001$, $r^2 = .67$. There is also a significant relationship between the last piece of cue/count needed (when rated as not guilty) and the corresponding verdict when not guilty verdicts were given, $r = .35$, $p = .007$, $r^2 = .12$. Finally, the number of not proven verdicts given has a significant relationship with the last cue/count needed when rated as not proven, $r = .51$, $p < .001$, $r^2 = .26$.

Discussion

The aim of the current study is to identify which threshold point model of decision-making best describes juror decision-making within the Scottish criminal justice system, where three verdict options are available. The novelty of this study relates to using the verdict that jurors gave as a factor, thus allowing the researchers to investigate the decision processes behind each of the verdicts. This is the first study of its kind to investigate the decision-making processes behind legal verdicts. Additionally, this is the first study that differentiates the not proven verdict from the not guilty verdict in regard to a stopping rule, thus highlighting that Scottish jurors reach not proven verdicts for different reasons than they reach not guilty verdicts.

Diffusion and count models of decision-making were explicitly investigated by looking at final verdict ratings, decision threshold points and both evidence ratings and

likelihood-of-guilt ratings. This investigation provides insight into the cognitive processes of juror decision-making, rather than solely focusing on outcomes like much of the existing literature. It is the first study to empirically compare these models to juror decision-making. It also investigates the unique not proven verdict option that is currently used in the Scottish legal system. Again, this is a novel exploration, as threshold point models are normally investigated using binary outcome options rather than tertiary ones.

The first hypothesis posited that the three decision points (prior, threshold and last) would be significantly different from one another – and this is supported. The general trend over the three verdict types is that the participants started off at the prior point (i.e. without any evidence), moved significantly higher to a threshold point (i.e. a relative stopping rule) and then moved significantly lower to the final point. The fact that the three points are distinct supports the notion that individuals start at a prior point, after which information accrual causes them to reach a threshold point (allowing a response to be given). The notion that information integration (or distortion) takes place even after a decision has been (or can be) made is also supported, allowing for the possibility of a change in the verdict given. The fact that the threshold point and the last point differ also highlights that the threshold point may be a pivotal and independent point in the decision-making process that allows verdicts to be reached. In other words, the confirmation of the hypothesis supports the idea of a diffusion model being applied to legal decision-making.

The second hypothesis is also confirmed, as there is a significant difference in the likelihood ratings across the verdict types, with guilty verdicts having significantly higher mean ratings than not proven and not guilty verdicts. This highlights that the way in which the information was integrated is significantly different across the verdict types. This idea of information integration (represented by the

likelihood-of-guilt values) in mock juror decision-making is supported by previous research. Ostrom, Werner, and Saks (1978) found that jurors initially think that the suspect is innocent and that this belief is averaged with the evidence in a trial to give the final verdict. Estrada-Reynolds et al. (2015) found similar results. The likelihood ratings analysis, therefore, highlights that the evidence from the cases was integrated differently between the three verdict types. This evidence of information integration supports a diffusion (i.e. drift) model of decision-making (Ratcliff & Smith, 2004). The confirmation of this hypothesis also shows that the process behind the not proven verdict is distinct from the process that allows a not guilty verdict to be reached, thus demonstrating that jurors understand the difference between the verdict types and use them differently depending both on how they integrate information and on which threshold point has been reached.

The third hypothesis is also supported, as there is a significant interaction between the two independent variables (verdict type and decision points). The results show that the threshold points (i.e. the relative stopping rule) are different across the verdict types, which suggests that the reaching of a threshold point informed the responses given (Wells, 1992). Additionally, because the last points are significantly different across the verdict types, this shows that once the threshold points were reached they were not crossed back over, implying that the appropriate verdicts were given. For example, drift caused the guilty (or not guilty) threshold point to be reached, and because the threshold point was not crossed again in the opposite direction, the corresponding guilty (or not guilty) verdict was then given. Not proven verdicts can be said to fall between the guilty and not guilty threshold points, and are consequently in the limbo region of not proven. Therefore, it can be said that when a not proven verdict is given, innocence is doubted but guilt cannot be confirmed. The acceptance of the third hypothesis

further suggests that, for each of the verdict types, evidence was accumulated until a threshold point was reached, and the evidence acquired after this point did not affect the response that was given (or cause a change in the verdict). However, it must be pointed out that by asking participants to state the last piece of evidence needed to make a decision, only the final threshold point reached is being examined; participants could have crossed several threshold points throughout a trial, but only the final threshold point that is crossed determines the verdict that is given.

The prior points for each of the verdict types are similar, suggesting that jurors in a three-verdict system are not affected by pre-trial biases. An alternative explanation is that all the participants were biased towards the defence, both because they all started with similar likelihoods of guilt and because the not guilty threshold point is above the prior points for all verdicts. Furthermore, it is likely that all the participants shared the initial belief of 'innocent until proven guilty', as the point at which a not guilty verdict could be given was above all of the prior points, regardless of verdict type.

The threshold points are significantly different from each other, suggesting that information was integrated together, causing likelihood ratings to drift across a threshold point and therefore allowing a response to be given (Ratcliff & Smith, 2004). The results, therefore, suggest that a third verdict can be incorporated into a relative-stopping-rule model by allowing threshold points to be crossed back over in either direction. In addition, the last points used to rate likelihood of guilt, as rated by the participants across the conditions, were found to all be significantly different from each other across the verdict types, further supporting the idea of distinct decision-making processes across the three verdict types. Furthermore, the interaction analysis highlights that the decision-making processes are distinct across the three verdict types, and that information integration and threshold points allowed verdicts to be reached. The

interaction also highlights that the decision-making process is unique for each of the three verdict types, as elaborated on below.

In cases where guilty verdicts were reached, the decision-making process caused drift from the prior point to the threshold point, thus allowing the guilty verdict to be given. There is no difference between the threshold point and the last point for the guilty verdicts. Furthermore, the decision-making process that was used to reach guilty verdicts was found to be non-compensatory (Gigerenzer & Goldstein, 1996); the combination of available cues could not override the guilty verdict, which was reached using 6.2 pieces of information on average.

Nevertheless, it is important to point out that in the theorised diffusion model drift can continue post-threshold, and this drift can either confirm the leading verdict (that is, the threshold point that was last passed) or disconfirm it by passing back across the current threshold point (Carlsón & Russo, 2001). The results show that the information presented post-threshold does not seem to have had a significant effect on drift when a guilty verdict is returned. However, it is likely that if strong defence evidence was presented near the end of a trial, the drift would be negative, thus causing the threshold point to be crossed back over and resulting in a verdict change. Future research can investigate this by manipulating the order of evidence using strength ratings.

In the cases that result in a not proven verdict, the prior points are located in the middle of the scoring range and the likelihood significantly rises towards the threshold point. The last point however is significantly lower than the threshold point, highlighting that in cases where a not proven verdict is reached, there is initially a similar drift to that seen in the cases that result in a guilty verdict. However, through a collection of cues the drift comes to an end in the decision-making zone of not proven, which is between the guilty and not guilty threshold points. For not proven verdicts, the rise may be enough to cross the not guilty

threshold point, but may not be sufficient to cross the guilty threshold point, leaving the juror's verdict in a not proven limbo. Conversely, jurors who give a not proven verdict may initially cross the guilty threshold point and thus favour a guilty verdict, meaning that for them to finally give a not proven verdict, the evidence presented after they have crossed the guilty threshold point must have caused them to once again cross the guilty threshold point into the not proven zone. Therefore, a not proven threshold point does not exist, and rather, the not proven zone is reached through the jurors crossing either the guilty and/or not guilty threshold point(s). The reason, then, that the point at which a not proven verdict could be reached (or the not proven threshold point as referred to in the analysis) differed significantly from the guilty and not guilty threshold point was because jurors could reach the not proven zone through either crossing the guilty or not guilty threshold point. In other words, some jurors reaching the not proven zone through the guilty threshold point would have had high likelihood of guilt ratings and some jurors reaching the not proven zone through the not guilty threshold point would have had low likelihood of guilt ratings. This then caused the average likelihood of guilt rating for the point at which a not proven verdict could be reached (or the not proven threshold point as referred to in the analysis) to be significantly different from both the not guilty threshold point and the guilty threshold point. If the not proven threshold point had been the same as the not guilty threshold point, then it would suggest that jurors could not cross back over the guilty threshold point once it had been crossed. Furthermore, the results surrounding the not proven verdict suggest that the not proven verdict does not have a specific threshold point and is reached through either the guilty or not guilty threshold point being crossed, and that a not proven verdict can be reached through varying thresholds and information integration.

Had more information had been provided in the not proven cases, one of three outcomes would have resulted:

1. Little drift occurs and a not proven verdict is still given.
2. The drift results in an increase until the guilty threshold point is reached, resulting in a guilty verdict being given.
3. The drift results in a decrease until the not guilty threshold point is reached, resulting in a not guilty verdict being given.

In summary, the drift is constant but the last threshold point crossed determines the verdict that is given. For example, for one juror, the guilty threshold point could be crossed due to one piece of information and then a guilty verdict is given because the rest of the information does not cause a negative drift that results in the guilty threshold point being crossed back over. However, another juror could start off in the not guilty zone, and then the drift could cause a steady increase that after five pieces of evidence results in reaching the guilty threshold point. Then, one piece of information could cause the likelihood to pass back across the guilty threshold point, and if the rest of the evidence does not cause further significant drift, a not proven verdict is rendered.

The decision-making process for cases resulting in not guilty verdicts starts with a middling prior point, which then increases to significantly above the Not Guilty threshold point. Therefore, the Not Guilty threshold point symbolizes when the evidence has caused the juror to have a negative drift, causing them to cross the Not Guilty threshold point from the Not Proven zone. However, the likelihood has significantly dropped at the last point, resulting in the not guilty verdict. This, alongside the fact that on average the not guilty verdicts were reached using 11.7 pieces of information, indicates that initially the participants started off in the not guilty zone, after

which the likelihood increased with the presentation of the evidence. The drift caused the not guilty threshold point to be crossed and the not proven zone to be entered – but after this initial rise, the threshold point is crossed back over as more evidence is presented, causing a shift from the not proven zone back down into the not guilty zone. The likelihood ratings of the participants who returned not guilty verdicts continue to decrease after crossing back over the not guilty threshold point, showing that they became more convinced in the innocence of the defendant after the final crossing.

In addition, the fact that cue utilisation differs between the verdict types supports the notion of threshold point decision-making and information integration. This is because guilty verdicts are reached using fewer cues than not proven and not guilty verdicts, which suggests that the weight of the information that is integrated and the frugality of the threshold points may determine which verdict is reached. The results also show that the more compensatory the process is, the more likely it is that the decision drift will favour a not guilty or not proven verdict rather than a guilty verdict. Interestingly, guilty verdicts were reached after 6.2 pieces of information on average, even though the smallest number of pieces of information shown in a case is 10. This highlights that the participants were making decisions before all the evidence was shown, supporting the view that once a threshold point was reached a decision could be made. Furthermore, the results show that through allowing both threshold points and information integration to vary, both rational and heuristic processing can be mirrored.

In summary, the diffusion threshold model of decision making is a good metaphor to explain juror decision processes in a three-verdict system. Within this adapted model, jurors have three zones, each with a respective verdict type: guilty, not guilty and not proven. These zones are entered through reaching one of two threshold points: guilty and not guilty. If a guilty threshold point is crossed from the not

proven zone, the juror will give a guilty verdict. If the not guilty threshold point is crossed, from the not proven zone, or it is not crossed from the not guilty zone, a not guilty verdict will be given. Finally, if the juror reaches the not proven zone from either the guilty or not guilty threshold point, a not proven verdict will be given. Furthermore, each of the three verdict types are reached through two varying threshold points, information integration and three zones that reflect the verdict options, according to this adapted diffusion threshold model.

Significant associations were found in support of the fourth hypothesis, which investigates whether or not individual counters, or evidence types, can allow decision-makers to reach a threshold point and thus return the corresponding verdict. The association between guilty verdicts and the last guilty cue/counter needed is significant. The direction is positive and the correlation coefficient is strong. Likewise, the association between not guilty verdicts and the last not guilty cue/counter needed is also significant and positive, but with a moderate correlation coefficient. Further, the relationship between not proven verdicts and the last not proven cue/count needed is significant and positive, with a moderate correlation coefficient. This provides some support that information was placed in one of three counters – guilty, not guilty or not proven – and that the first threshold point that was reached using these counters informed the decision that was ultimately made.

However, although the significance levels suggest that these results are below a 1% likelihood of occurring by chance, the correlation coefficients indicate that the ratings for the last cue/counter needed do not always allow a threshold point to be reached. Furthermore, because the correlation coefficients are not 1, this implies that simply counting information does not allow a threshold point to be reached, and therefore that the decision-making process is more likely to be integrative. In other words, because count data alone do not always allow threshold points to be reached, information

integration may better explain the process. For example, some pieces of evidence that are counted as guilty may be considered weak, therefore causing a drift away from the guilty threshold point (or resulting in no significant change). In summary, the correlations hold some support for the count model, but also cast doubt on how well this model may fit a mock or real-life juror setting.

One potential limitation of this study is that there could be a lack of ecological validity, as the study is quasi-experimental. This artificial setting – with its lack of real-life cues – may have affected the decision-making process, meaning that what is observed in this study could differ from what occurs in real-life trials (Simon, 1956; Wiener et al., 2011). However, previous research comparing the culpability ratings given by participants who read transcripts versus those who view eyewitness testimony on video has found that it is the information, rather than the medium, that is important to informing participants' beliefs about the culpability of defendants (Pezdek, Avila-Mora, & Sperry, 2010). In addition, cognitive processes such as decision-making happen in a similar manner irrespective of whether they are produced naturally or artificially (Watt & Quinn, 2008). Nevertheless, future research may want to try to increase the ecological validity of mock juror experiments through the use of audiovisual stimuli in order to achieve closer alignment to real-life trials.

A confounding effect may have been introduced by asking participants to rate the evidence as guilty, not guilty or not proven and then to also rate the likelihood of the guilt of the suspect from 1 to 100. The initial evidence rating may have anchored the likelihood-of-guilt ratings – but there is no other way to collect both independent evidence ratings and likelihood-of-guilt ratings. Even if the likelihood-of-guilt ratings was anchored by the evidence ratings (i.e. the count data), this does not take away from the fact that the diffusion model can explain the decision-making data better than the count model. Future research may want to

repeat the current experiment using a between-subjects method where one group gives likelihood-of-guilt ratings and one group gives evidence ratings (count data), and then compare which model explains each group's decision-making data in the most satisfactory manner.

This study is the first to demonstrate that the diffusion model is a good fit for the decision-making process that jurors use, and that this process may involve integrating information in a way that allows a threshold point to be reached, which then allows a verdict to be given if the threshold points is not passed back across said threshold points. Furthermore, this study highlights that the diffusion model can explain how jurors reach each of the three verdict types used in Scotland through incorporating two threshold points, information integration and three separate zones; each zone symbolizes one of the three verdicts. The overarching findings of the study support the fit of the diffusion model to the Scottish criminal justice system.

Acknowledgements

This research was carried out in part-completion of the first author's PhD at Edinburgh Napier University, Scotland.

Ethical standards

Declaration of conflicts of interest

Lee J. Curley has declared no conflicts of interest.

Rory MacLean has declared no conflicts of interest.

Jennifer Murray has declared no conflicts of interest.

Andrew C. Pollock has declared no conflicts of interest.

Phyllis Laybourn has declared no conflicts of interest.

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of Edinburgh Napier University's Research and Integrity Committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent

Informed consent was obtained from all individual participants included in the study.

Funding

This work was supported by the McGlashan Charitable Trust; and the East Lothian Educational Trust.

References

- Ashill, N. J., & Yavas, U. (2006). Vignette development: An exposition and illustration. *Innovative Marketing*, 2(1), 28–36. Available from: https://businessperspectives.org/component/option,com_journals/task,journal/id,5/Itemid,74/
- Bergert, F. B., & Nosofsky, R. M. (2007). A response-time approach to comparing generalized rational and take-the-best models of decision-making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(1), 107–129. doi: 10.1037/0278-7393.33.1.107.
- Bitzer, S., Park, H., Blankenburg, F., & Kiebel, S. J. (2014). Perceptual decision-making: Drift-diffusion model is equivalent to a Bayesian model. *Frontiers in Human Neuroscience*, 8, 1–17. doi: 10.3389/fnhum.2014.00102.
- Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision-making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113 (4), 700–765. doi: 10.1037/0033-295X.113.4.700
- Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. *Cognitive Psychology*, 57(3), 153–178. doi:10.1016/j.cogpsych.2007.12.002.
- Carlson, K. A., & Russo, J. E. (2001). Biased interpretation of evidence by mock jurors. *Journal of Experimental Psychology: Applied*, 7(2), 91–103. doi: 10.1037//1076-898X.7.2.91.
- Cedrus Corporation. (2014). *Superlab 5 [computer software]*. San Pedro: Cedrus Corporation.
- Dhami, M. K., & Ayton, P. (2001). Bailing and jailing the fast and frugal way. *Journal of Behavioral Decision-making*, 14(2), 141–168. doi:10.1002/bdm.371.
- Duff, P. (1999). The Scottish criminal jury: A very peculiar institution. *Law and Contemporary Problems*, 62(2), 173–201. Available at: <http://www.jstor.org/>
- Englich, B., Mussweiler, T., & Strack, F. (2005). The last word in court—a hidden disadvantage for the defense. *Law and Human Behavior*, 29(6), 705–722. doi: 10.1007/s10979-005-8380-7.
- Estrada-Reynolds, V., Gray, J., & Nuñez, N. (2015). Information integration theory, juror bias, and sentence recommendations captured over time in a capital trial. *Applied Cognitive Psychology*, 29(7), 713–722. doi: 10.1002/acp.3155.
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. doi:10.3758/BF03193146
- Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4), 650–669. doi: 10.1037/0033-295X.103.4.650
- Gigerenzer, G., & Goldstein, D. G. (1999). *Betting on one good reason: Take the best and its relatives. Simple Heuristics that Make Us Smart*. New York: Oxford University Press.
- Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision-making. *Annual Review of Neuroscience*, 30, 535–574. doi: 10.1146/annurev.neuro.29.051605.113038.
- Grace-Martin, K. (2017). *When unequal sample sizes are and are not a problem in ANOVA*. Retrieved from: <https://www.theanalysisfactor.com/when-unequal-sample-sizes-are-and-are-not-a-problem-in-anova/>
- Heverly, M. A., Fitt, D. X., & Newman, F. L. (1984). Constructing case vignettes for evaluating clinical judgment: An empirical model. *Evaluation and Program Planning*, 7, 45–55. doi:10.1016/0149-7189(84)90024-7.
- Hope, L., Greene, E., Memon, A., Gavisk, M., & Houston, K. (2008). A third verdict

- option: Exploring the impact of the not proven verdict on mock juror decision-making. *Law and Human Behavior*, 32(3), 241–252. doi: 10.1007/s10979-007-9106-8.
- Justice Education Society of BC. (2016). How a criminal trial works. Supreme Court BC: Online help guide. Retrieved the 15 of January, 2016, from: <http://www.supremecourtbc.ca/criminal/how-a-criminal-trial-works>.
- Laming, D. R. J. (1968). *Information theory of choice reaction time*. New York: Wiley.
- Lee, M. D., & Cummins, T. D. (2004). Evidence accumulation in decision-making: Unifying the 'take the best' and the 'rational' models. *Psychonomic Bulletin & Review*, 11(2), 343–352. doi: 10.3758/BF03196581
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97. doi: 10.1037/h0043158
- Murray, J., & Thomson, M. E. (2010). Applying decision-making theory to clinical judgements of violence risk assessment. *Europe's Journal of Psychology [online]*, 2010(2), 150–171, available: <http://ejop.psychopen.eu/index.php/ejop>.
- Newell, B. R., & Lee, M. D. (2009). Learning to adapt evidence thresholds in decision-making. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 473–478). Austin, TX: Cognitive Science Society. Available from: <http://csjarchive.cogsci.rpi.edu/proceedings/2009/papers/85/paper85.pdf>.
- Ostrom, T., Werner, C., Saks, M. (1978). Integration theory analysis of jurors' presumptions of guilt or innocence. *Journal of Personality and Social Psychology*, 36(4), 436–450. doi: 10.1037/0022-3514.36.4.436.
- Pezdek, K., Avila-Mora, E., & Sperry, K. (2010). Does trial presentation medium matter in jury simulation research? Evaluating the effectiveness of eyewitness expert testimony. *Applied Cognitive Psychology*, 24(5), 673–690. doi: 10.1002/acp.1578.
- Potter, K. W. (2011). *When You are Confident that You are Wrong: Response Reversals and the Expanded Poisson Race Model* (Doctoral dissertation, The Ohio State University). Available from: https://etd.ohio-link.edu/pg_10?0::NO:10:P10_ACCESSION_NUM:osu1321454142.
- Price, H. L., & Dahl, L. C. (2014). Order and strength matter for evaluation of alibi and eyewitness evidence. *Applied Cognitive Psychology*, 28(2), 143–150. doi: 10.1002/acp.2983.
- Ratcliff, R., Philastides, G., & Sajda, P. (2009). Quality of the evidence for perceptual decision-making is indexed by trial-to-trial variability of the EEG. *Proceedings of the National Academy of Sciences*, 106(16), 6539–6544. doi:10.1073/pnas.0812589106
- Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, 9(5), 347–356. doi: 10.1111/1467-9280.00067.
- Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, 111(2), 267–333. doi: 10.1037/0033-295X.111.2.333.
- Roberts, S. C., & Murray, J. (2013). Applying the revenge system to the criminal justice system and jury decision-making. *Behavioral and Brain Sciences*, 36(01), 34–35. doi: 10.1017/S0140525X12000581
- Rouder, J. N. (2001). Testing evidence accrual models by manipulating stimulus onset. *Journal of Mathematical Psychology*, 45, 334–354. doi: 10.1006/jmps.2000.1319.
- Sangero, B., & Halpert, M. (2007). Why a conviction should not be based on a single piece of evidence: A proposal for reform. *Jurimetrics*, 48(43), 43–94. Available at: <https://www.ssrn.com/en/>
- Scottish Court Service. (2015). Jury Service: In the High Court and Sheriff Court. 1–16.
- Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63(2), 1–11. doi: 10.1037/h0042769.
- Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, 27(3), 161–168. doi: 10.1016/j.tins.2004.01.006.
- Smithson, M., Deady, S., & Gracik, L. (2007). Guilty, not guilty, or ...? Multiple options in jury verdict choices. *Journal of Behavioral Decision-Making*, 20(5), 481–498. doi: 10.1002/bdm.572.
- Thomas, E. A., & Hogue, A. (1976). Apparent weight of evidence, decision criteria, and confidence ratings in juror decision-making. *Psychological Review*, 83(6), 442–465. doi: 10.1037/0033-295X.83.6.442
- Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods*, 40(1), 61–72. doi: 10.3758/BRM.40.1.61
- Walters, G. D. (2007). Using Poisson class regression to analyze count data in correctional and forensic psychology a relatively old solution to a relatively new problem.

- Criminal Justice and Behavior*, 34(12), 1659–1674. doi: [10.1177/0093854807307030](https://doi.org/10.1177/0093854807307030).
- Watt, R., & Quinn, S. (2008). It depends on what you do in the laboratory. *British Journal of Psychology*, 99, 351–354. doi: [10.1348/000712607X267976](https://doi.org/10.1348/000712607X267976).
- Wells, G. L. (1992). Naked statistical evidence of liability: Is subjective probability enough? *Journal of Personality and Social Psychology*, 6(5), 739–752. doi: [10.1037/0022-3514.62.5.739](https://doi.org/10.1037/0022-3514.62.5.739).
- Wenzel, M., Okimoto, T. G., Feather, N. T., & Platow, M. J. (2008). Retributive and restorative justice. *Law and Human Behavior*, 32, 375–389. doi: [10.1007/s10979-007-9116-6](https://doi.org/10.1007/s10979-007-9116-6).
- Wiener, R., Krauss, D., & Lieberman, J. (2011). Mock jury research: Where do we go from here? *Behavioural Sciences and Law*, 29, 467–479. doi: [10.1002/bsl.989](https://doi.org/10.1002/bsl.989).
- Windschitl, P., Scherer, A., Smith, A. & Rose, J. (2013). Why so confident? The influence of outcome desirability on selective exposure and likelihood judgment. *Organizational Behavior and Human Decision Processes*, 120(5), 73–86. doi: [10.1016/j.obhdp.2012.10.002](https://doi.org/10.1016/j.obhdp.2012.10.002).
- Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions Society*, 367, 1310–1321. doi: [10.1098/rstb.2011.0416](https://doi.org/10.1098/rstb.2011.0416).

Appendix 1

The details included in the nine vignettes across the three verdict types handed down in the real-life trials are presented below.

Factor	Guilty	Not guilty	Not proven
Familiarity with victim	2 familiar 1 unfamiliar	1 familiar 2 unfamiliar	2 familiar 1 unfamiliar
Vulnerability of victim	2 vulnerable 1 not vulnerable	1 vulnerable 2 not vulnerable	2 vulnerable 1 not vulnerable
Crime details	1. Housebreaking, stabbing 2. Argument with multiple injuries 3. Body not found	1. Head injury, strangled 2. Neck injury, asphyxiation 3. Self-defence, stabbing	1. Throat cut 2. Body not found 3. Multiple injuries, blow to the head
Victim's age (years)	51, 16, 33	19, 44, 26	18, 43, 26
Victim's gender	2 females 1 male	1 female 2 males	3 females 0 males
Accused's age (years)	20, 22, 39	33, 49, 23	33, 21, 33
Accused's gender	3 males	3 males	3 males
Weapon used	2 yes, 1 unknown	2 yes, 1 no	2 yes, 1 unknown
No. of words in opening statement	123, 110, 112 $M = 115$	129, 104, 114 $M = 116$	125, 105, 122 $M = 117$

Appendix 2

The cue types and number per vignette (case) across the three verdict types handed down in the real-life trials are presented below.

	Guilty	Not guilty	Not proven
Number of cues	Case 1: prosecution = 7 defence = 5 Case 3: prosecution = 9 defence = 9 Case 6: prosecution = 9 defence = 6	Case 2: prosecution = 7 defence = 7 Case 4: prosecution = 5 defence = 6 Case 5: prosecution = 5 defence = 5	Case 7: prosecution = 6 defence = 5 Case 8: prosecution = 9 defence = 7 Case 9: prosecution = 7 defence = 6
Severity rating (<i>M</i>)	Case 1 = 4.69 Case 3 = 4.13 Case 6 = 3.56	Case 2 = 4.25 Case 4 = 3.88 Case 5 = 3.94	Case 7 = 4.13 Case 8 = 3.38 Case 9 = 1.94
Familiarity rating (<i>M</i>)	Case 1 = 2.06 Case 3 = 2.00 Case 6 = 1.94	Case 2 = 2.56 Case 4 = 1.69 Case 5 = 1.94	Case 7 = 1.69 Case 8 = 1.50 Case 9 = 1.94
Realism rating (<i>M</i>)	Case 1 = 3.81 Case 3 = 3.50 Case 6 = 3.44	Case 2 = 4.00 Case 4 = 3.50 Case 5 = 4.19	Case 7 = 3.56 Case 8 = 3.75 Case 9 = 4.19

Note: The severity, familiarity and realism scores were gathered from the pilot. Each of the scores mentioned could be ranked from 1 to 5, where 1 indicates a low score and 5 indicates a high score.