Running Head: JUROR DECISION MAKING.

Title: Threshold Utilisation in Juror Decision Making.

Lee J. Curley[1*], Rory MacLean[1], Jennifer Murray[1], Andrew C. Pollock[2] and Phyllis Laybourn[1]

1. School of Applied Sciences, Edinburgh Napier University, Sighthill Campus, Sighthill Court, Edinburgh, EH10 4BN, Scotland.
2. Consultant, Irvine, Pladda Avenue, Broomlands, Irvine, Ayrshire, KA11 1DR, Scotland.

*Requests for reprints should be addressed to Lee Curley, School of Applied Sciences, Edinburgh Napier University, Sighthill Campus, Sighthill Court, Edinburgh, EH10 4BN, Scotland (E-mail: Lee.Curley@napier.ac.uk).

Word count (exc. figures/tables/references): 7510

Word count (inc. figures/tables/references): 9760

Abstract

The current research aimed to identify whether a model of juror decision making (i.e. the threshold model) that encompasses both rational and intuitive decision making exists. Sixty participants were selected who would be 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, not guilty or not proven (a verdict type specific to Scotland) counter. Participants were asked after 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 were best described using a flexible model (i.e., a diffusion model) that allowed for information integration. Future research should examine whether or not the diffusion model can explain cognitive fallacies, such as confirmation bias, commonly studied in decision science.

Keywords: Jurors, Decision Making, Stopping Rule, Information Integration, Heuristics, cue utilisation, Law, Psychology, Court Room, Not Proven Verdict.

Threshold Utilisation in Juror Decision Making

Jurors make potentially life changing decisions, with relatively little training and potentially no previous experience. Jurors are given the responsibility of delivering justice to both the state and victim (Roberts & Murray, 2014**;** Wenzel, Okimoto, Feather, & Platow, 2008). This raises the question: how do jurors make decisions despite their lack of legal experience? The current research aimed to identify the process through which legal laypersons reach decisions when they are presented with legal information. The present study specifically focusses on the Scottish legal system.

Much of the legal and decision making research has focussed on the American legal system and American jurors (Dhami & Ayton, 2001). These American focussed research investigations cannot easily be generalised to Scottish jurors or the Scottish legal system. This is because the Scottish legal system has very unique aspects, such as the not proven verdict (Duff, 1999). The not proven verdict is an extra acquittal verdict, which 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 that task difficulty was the only variable that consistently forecasted not proven verdicts. In addition, Hope, Greene, Memon, Gavisk, and Houston (2008) focussed on how a third verdict of not proven would affect conviction rates (i.e., outcome analysis) when the strength of evidence varied. It was found that in cases considered ‘moderate’, the availability of the not proven verdict reduced conviction rates, and that if a not proven verdict was available the number of not guilty verdicts was also reduced. The research by Smithson et al. (2007) and Hope et al. (2008) focussed on decision outcomes, however, and did not explore the cognitive processes that allowed for certain decisions to be reached. To better understand these outcomes, more research is needed that investigates the processes behind them.

One contemporary area of decision making research which puts the focus on the process rather than outcomes and that can apply to the court room 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 the Take The Best heuristic, where an option is chosen based on the first cue that allows a binary set of responses to be discriminated, was as accurate as multiple regression analysis when making decisions in regards 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 adapted it into a more global theory that could explain decision making data more adequately. Their experiment compared a rational decision making strategy (called RAT) with the TTB (where information is searched in the memory of the decision maker from the most valid to the least valid cue; Gigernzer & Goldstein, 1996) model. Both strategies were found to have been used, but not exclusively, across 52.5% of participants in the sample, which means 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). Further, a unified threshold model (which combined RAT and TTB) accounted for more of the decision making behaviour than any single model could do alone (Lee & Cummins, 2004). For instance, the unified threshold model accounted for 84.5% (Compared to RAT= 64% and TTB = 36%) 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) 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 models could potentially explain juror decision making. Each of these threshold models 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 threshold models, the current study concentrates on only two broad categories of threshold decision making models, which are mentioned above.

The models mentioned above are mathematical models, and Ratcliff and Smith (2004) should be consulted for a more mathematical description of the models. However, the current piece of research will be one of the first pieces of research that tries to map the models to juror decision making data and that tries to differentiate the models using inferential statistics and dependent measures: such as likelihood of guilt data, to assess if thresholds are reached through information integration; and, evidence rating data, which will investigate if thresholds 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 thresholds (Gold & Shadlen, 2007; Potter, 2011; Smith & Ratcliff, 2004). For instance, for jurors, the two counters could relate to verdict decisions (Walters, 2007). One counter could be for a guilty verdict, and another counter could be for a not guilty verdict (Smith & Ratcliff, 2004). In these counters, the same mechanism, as described by Lee and Cummins (2004), occurs. The threshold 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 in the guilty counter is reached first, then the juror will give a guilty verdict, whereas if the counter in the not guilty verdict is reached first, then the legal layperson will give a not guilty verdict. Further, in race models, where evidence races towards a threshold, there may be more than two counts, which suggests that Scottish jurors could count not proven evidence until a threshold is reached (Ratcliff & Smith, 2004). Similarly, some count 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).

Furthermore, count data will measured in the current study, by asking participants to count information/evidence into one of three counters. Further, participants will be able to place information into either their guilty counter, their not guilty counter or their not proven counter. Previous research has provided support for the counter model in the court room (Thomas & Hogue, 1970). However, the current study will be the first investigation to establish if count information allows thresholds to be reached, thus allowing verdicts to be given, using inferential statistics.

The second theoretical approach relates to relative stopping rules. Ratcliff & Smith (2004) outline that in relative stopping rules information integration occurs until one alternative is favoured over another, symbolised by a threshold. Then the appropriate decision is made. In other words, if the integration of evidence causes a certain amount of the information to favour a guilty verdict, by reaching the guilty boundary/threshold, then a guilty response will be given. Therefore, the thresholds 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 thresholds 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 count models (Ratcliff & Smith, 2004). Further, the starting points in the diffusion model can vary depending on how early the decision making process has started, which may fit in well with biased jurors, and jurors who may be affected by pre-trial information (Laming, 1968; Ratcliff & Smith, 2004; Smith & Ratcliff, 2004). This skewed starting (or prior) point may cause the prior point to be closer to one threshold 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 the prior information.

Research on this approach has suggested that the further away the threshold is from the starting point, the slower and more accurate the response will be (Ratcliff & Smith, 2004; Smith & Ratcliff, 2004). This suggests that jurors with higher thresholds (who use more information) are more accurate than jurors with lower thresholds (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, Philiastides, & Sajda, 2009; Ratcliff & Rouder, 1998; Yeung & Summerfield, 2012).

The current study will investigate if information integration (using likelihood of guilt ratings) allows thresholds to be reached. Previous research has utilized similar methods when studying information integration (see Estrada-Reynolds et al., 2015). Further, the research will investigate if different likelihoods of guilt are associated with different thresholds, thus highlighting that once a decision maker's drift reaches a particular threshold, does a specific verdict choice become more appropriate. That is, if a guilty (not guilty) threshold is reached through information integration a guilty (not guilty) verdict is given, unless said threshold is re-surpassed leading to a verdict reversal (Potter, 2011). This has been confirmed in previous research through mathematical modelling (Ratcliff and Smith, 2004). The current research will be the first of its kind to establish if 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, the current research will add an additional point, called the last point. The reasoning for this last likelihood point is to investigate if drift continues/changes in jurors once they have reached a threshold, as unlike previous investigations jurors have no control over information search.

The literature has highlighted three main issues. First, that the threshold model may be a more suitable model for legal decisions in comparison to heuristics and more rational models. Second, that the Scottish legal system, with its three verdict options, is unique and needs proper scientific enquiry. And, finally, that threshold models of decision making may help to explain juror decision making in the unique three verdict system. Therefore, the current study will assess using likelihood data (to symbolise information integration) and evidence ratings, which highlight which verdict (guilty, not guilty and not proven) that particular piece of evidence favours (i.e., to represent count data), in an attempt to find out which model, either the count model or diffusion model, best describes the decision making processes of jurors.

It is expected that in the current study that likelihood of guilt ratings will differ across verdict types. The reasoning for this is because likelihoods will symbolize information integration (or drift) in the current study, and previous research has showed that drift differs depending on the outcome that is reached (Ratcliff & Smith, 2004). Further, it is believed that the three likelihoods points (prior, threshold and last) will differ in the current study. This is because previous research has shown that drift allows decision makers to move from their prior point to a point where a decision can be made (i.e., their threshold point). Therefore, in the current study, it is expected that if the diffusion model is explaining the decision making data that the threshold point will differ from the prior point, thus allowing a decision to be made. Similarly, it is expected that the last point will also differ from the prior point, otherwise a verdict associated with their prior point should be given; that is, the not guilty verdict, as such a verdict may be favored initially at the prior point because of the "innocent until proven guilty" belief.

It is, also, believed that a significant interaction will exist between the verdict given and the likelihood points. This is because it is expected that both different drifts and thresholds (i.e., guilty and not guilty threshold) will lead to different legal decisions, which can be confirmed by previous research by Ratcliff and Smith (2004). Contrary, if the count model explains decision making data the best, it would be expected that the last piece of count data (evidence ratings) needed would have a strong and significant relationship with the verdict that is given. This is because the last piece of count information needed should push decision makers beyond their threshold, thus allowing a response to be made. Therefore, you would expect the response made to correspond to the counter that the piece of evidence was placed in. For example, it would be expected that when the last piece of evidence needed was rated as guilty (not guilty), then a guilty (not guilty) verdict should be given. This expectation can be justified by previous research by Thomas and Hogue (1970), which proposed that jurors could make decisions using independent counters and thresholds.

In summary, drift, or information integration, will be measured by asking participants to state the likelihood of guilt associated with the suspect after each piece of evidence (in a similar manner to Estrada-Reynolds et al., 20015). In addition, count evidence will be measured by asking individuals to supply individual evidence ratings (guilty, not guilty and not proven) after each piece of evidence, thus symbolizing the counter in which specific pieces of evidence have been placed. Further, the last piece of evidence needed to make a decision will symbolize the threshold being reached and will have a corresponding evidence rating (count data) and likelihood (drift), which will allow the researchers to assess if a threshold was reached through information integration or through using a collection of individual pieces of evidence placed in a specific counters. The current research consequently aims to investigate which threshold model of decision making best applies to the Scottish juror system, where three verdicts are available. It is hypothesised that:

H1. There will be significant differences between each of the three (diffusion) points studied (prior, threshold and last) in relation to likelihoods.

H2. There will be significant differences between each of the three verdicts types (guilty, not guilty and not proven) in terms of likelihoods.

H3. There will be a significant interaction between verdict type and diffusion points in relation to likelihood of guilt.

H4. There will be a significant association between the ratings (count given) of the last cue needed and the verdict given (for each of the different verdict types).

Method

Design

A mixed factorial quasi-experimental design was used, in which verdict (guilty, not guilty, not proven) was the between-subjects factor, and likelihood point (prior, threshold, last) was the within-subjects factor. These three likelihood levels were applied because it was important to see if they differed across and between the different verdict types that were given. Furthermore, the prior point was the likelihood point given after the context had been provided on a case, but before evidence had been shown. The prior point was situated at this point, so that it would show if participants had pre-trial verdict preferences before they had seen any relevant information. The threshold point was the likelihood point which was associated with the stopping rule. The last point was the likelihood point which was related to the likelihood of a guilt rating given to the last piece of evidence shown in a case.

The first dependent variable utilised in the current research was the likelihood of guilty rating, which ranged from 1 to 100: 1 symbolised definitely not guilty, 50 represented not proven, and 100 represented that the suspect was guilty (Price & Dahl, 2014; Thomas & Hogue, 1976; Windschitl, Scherer, Smith, Rose, 2013). This dependent variable highlighted whether or not information integration was occurring and allowed small changes to be tracked, similar to traditional drift models that have used logarithmic odds (Bitzer, Park, Blankenburg, & Kiebel, 2014; Lee & Cummins, 2004; Vandekerckhove, & Tuerlinckx, 2008).

Participants were also asked to rate each of the pieces of evidence as either guilty, not guilty, or not proven. This was to mirror the count model where evidence is counted separately. Participants were also asked to state the last piece of evidence they needed to make a decision at the end of each vignette. This symbolised when a threshold was reached. This information was also used to measure the dependent measure of cue utilisation (how much evidence was used by the mock jurors). Participants also stated, after each vignette, which verdict they thought was the most appropriate.

Participants

Sixty participants (31 females; 21 students; 39 non-students with a range of professional and manual employment) were recruited via opportunity sampling. Post-hoc power analysis was ran using the software G*power (Faul, Erdfelder, Lang, & Buchner, 2007). The sample size mentioned above was used, the alpha level utilized was $p < .05$, and a large effect size of $F = 0.59$ was chosen because of the uniqueness of the research. The analysis highlighted that the Analysis of Variance (ANOVA) for the diffusion analysis was adequately powered, with an actual power of .98.

Recruitment posters were placed within 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 other eligible individuals. Participants were only eligible to take part in the current study if they were juror eligible (Scottish Court Service, 2015). The mean age of participants was 26.8 years (SD = 9.6 years).

Materials

Standardised information sheets, consent forms and debriefing sheets were employed with all participants.

Legal inventory/demographics questionnaire

This inventory identified prospective participants who could not take part in a real life jury. The demographics questionnaire collected details such as age and gender.

Vignettes

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

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

Each stance (prosecution vs defence) presented between 5-9 pieces of evidence; see counterbalancing in Appendix 1 for more details. This varying number was chosen for two reasons: 1) so that memory constraints did not affect which cues were remembered within each of the stances when making a final verdict, which links with the Miller (1956) 7+-2 rule; 2) the number varied for generalisability purposes, since, some cases in real life trials last longer and present more information than other trials.

Procedure

Participants read an information sheet and completed a consent form. They were told that their data would be used for court appeals (a mild deception); so that participants would think that, their responses had real world consequences. They then completed the legal inventory and if eligible to take part, were given standardised instructions for the study on-screen from within the experiment-software package being used. The procedure for each vignette was identical and was as follows. Participants were provided with an opening statement to provide context to the case. Participants were then asked to give a prior likelihood of guilt, ranging from 1-100. Participants were reminded how to rate the evidence (either guilty, not guilty or not proven) and which buttons to press (G, N or P, respectively) to do this prior to evidence being presented. Next, participants were shown the pieces of evidence, with the prosecution evidence presented first, followed by defence evidence (Englich, Mussweiler, & Strack, 2005; Justice Education Society of BC, 2016). Cross-examinations were not included to avoid problems that could arise by the materials becoming too complex. 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, again ranging from 1-100; which symbolised information integration. Once each piece of evidence had been read and evaluated, participants were asked to give a final verdict: guilty, a not guilty or a not proven verdict. Then participants were asked to identify the last piece of evidence that they needed to make their decision. This response symbolised their threshold being reached, and allowed cue utilisation to be quantified. This process was repeated for each of the nine vignettes. Once all nine vignettes had been completed, participants read through the debrief sheet.

Results

Descriptive Statistics for Verdict Types

Fifty-nine participants 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 were therefore, relatively, similar. Data for each of the points (prior, threshold and last) was averaged within the verdict types given by the participants. For example, if an individual gave four not proven verdicts, the data (prior, threshold and last likelihood points) that were analysed was an average gathered from the four not proven verdict given. Likewise, if the same participant gave two guilty verdicts and three not guilty verdicts, each of the likelihood points would be 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.

Table 1 about here

As shown in Table 1, the mean increases from the prior point to the threshold point for guilty verdicts, 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 and 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 last point. The standard deviation shows a similar pattern. The standard deviation is higher in 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 changes throughout the three points, increasing between the prior and threshold point and then decreasing between the last point and threshold point. The standard deviation decreases through each of the three points.

Investigating Verdict Choices across Likelihood Points

A two-way mixed-groups Analysis of Variance (ANOVA), with verdict given (guilty, not guilty, not proven) as the between-subjects factor and likelihood point (prior, threshold, last) as the within-subjects factor was carried out. There was 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 highlighted that guilty verdicts had significantly higher likelihood ratings (*M*=71) than not guilty verdicts (*M*=43.4) and not proven verdicts (*M*=55.2; *p*=.001, *p*<.001, respectively). Not proven verdicts had likelihood ratings that were significantly (*p*<.001) higher than the likelihood ratings in not guilty verdicts.

For the variable of likelihood points, the Greenhouse-Geisser row was conulsted as Mauchley's test of sphericity was significant: *p*<.001. There was a significant main effect of likelihood 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 likelihood point (*M*=44.5) was significantly lower than both the threshold likelihood point (*M*=65.4) and the last likelihood (*M*=59.7) point (*p*<.001, *p*<.001, respectively). The threshold likelihood point was significantly (*p*<.001) higher than the last likelihood point.

There was a significant interaction between the likelihood points and the verdicts given: *F*(4,306)=33; *p*<.001, $\eta_p{}^2$=.30, which is a large effect size. See Figure 1 for visual representation of interaction.

Figure 1 about here

Simple main effects showed that for guilty verdicts the prior point (*M*=47.2) was significantly lower than the threshold point (*M*=83.9) and last point (*M*=81.8; *p*<.001, *p*<.001, respectively). The threshold point and last point were not significantly different from one another (*p*=.17). When investigating not proven verdicts, it was evident that the prior point (*M*=44.9) was significantly lower than the threshold point (*M*=62.6) and the last point (*M*=58; *p*<.001, *p*<.001, respectively). The threshold point was also significantly higher than the last point (P<.01). For not guilty verdicts, the prior point (*M*=41.5) was significantly (*p*<.01) lower than the threshold point (*M*=49.7). The prior point did not significantly (*p*=.45) differ from the last point (*M*=39.2). However, the last point was significantly (*p*<.001) lower than the threshold point.

When investigating the prior point, guilty verdicts (*M*=47.2) did not differ significantly from not proven (*M*=44.9) or not guilty (*M*=41.5) verdicts (*p*=.41, *p*=.07, respectively) in relation to likelihood of guilt. Likewise, not proven and not guilty verdicts did not significantly (*p*=.27) differ from one another in relation to the likelihood of guilt at the prior point. For the threshold point, guilty verdicts (*M*=83.9) had a significantly higher likelihood of guilt than not proven (*M*=62.6) and not guilty (*M*=49.7) verdicts (*p*<.001, *p*<.001, respectively). Not proven and not guilty verdicts significantly differed from one another in relation to the likelihood of guilt measured at the threshold point (*p*<.001), with not proven verdicts having a higher likelihood of guilt. In relation to the last point, guilty verdicts (*M*=81.8) had a significantly (*p*<.001) higher likelihood of guilt in comparison to both not proven (*M*=58) and not guilty (*M*=39.2) verdicts. Finally, the last point was significantly higher in not proven verdicts in comparison to not guilty verdicts (*p*<.001).

Cue Utilisation

Table 2 shows the descriptive statistics for Cue utilisation.

Table 2 about here

A one-way between subjects ANOVA was carried out to compare cue utilisation across the different verdict options. There was a significant difference in the cue utilisation between the three verdicts: *F*(2, 537)=173.7; *p*<.001. Post hoc Tukey’s tests highlighted that significantly fewer cues were utilised when a guilty decision was given (*M*=6.2) in comparison to when a not proven (*M*=11.4) and not guilty verdict (*M*=11.7) was given (*p*<.001, *p*<.001, respectively). There were no significant differences (*p* = .66) in cue utilisation between not guilty and not proven verdicts.

Count Analysis

Descriptive Statistics for Last Count Rating needed and Verdict Given

The descriptive statistics shown in Table 3 relate to the last cue/count ratings needed (which allowed a threshold to be reached), and the verdict given by the participants.

Table 3 about here

The data was collected using categorical data (i.e. guilty, not guilty and not proven). For the current analysis, the frequency of each verdict type was counted for each of the participants. Likewise, the frequency for 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 five guilty verdicts over the nine trials, their cumulative score for the number of guilty verdicts given would be five, and if they rated the last cue/count they needed as not guilty four times, their cumulative score for the last cue/count they needed for not guilty verdicts would be four.

Pearson's correlations were conducted to identify if the last piece of evidence needed allowed a decision to be reached. The correlations highlighted that significant relationships existed between the last count rating and the verdict which was given. A significant relationship existed between the last piece of count evidence needed (when rated as guilty) and the number of guilty verdicts given: r=.82, p<.001, r^2=.67. There was a significant relationship between the last piece of evidence 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 had 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 was to identify which threshold model of decision making best describes juror decision making within the Scottish criminal justice system, where three verdict options are available. This was the first study of its kind that investigated the decision making processes behind legal verdicts. Additionally, this was the first study that has differentiated the not proven verdict from the not guilty verdict in regards to a stopping rule, thus highlighting that Scottish jurors reach not proven verdicts for different reasons to why they reach not guilty verdicts.

Diffusion and count models of decision making were explicitly investigated, by looking at final verdict ratings, decision thresholds, and both evidence ratings and likelihood of guilt ratings. This investigation has provided greater insight into the cognitive processes of juror decision making, rather than a sole focus on outcomes, as has been much of the focus in the existing literature. It is the first piece of research to empirically compare these models to juror decision making. It also investigated the ‘unique’ Not Proven verdict option, currently used within the Scottish legal system. Again, this is a novel exploration, as threshold models are normally investigated using binary outcome options rather than tertiary.

The first hypothesis posited that the three likelihood points (i.e. prior, threshold and last) would be significantly different from one another; this was supported. The general trend over the three verdicts was to start off at a prior point (i.e. without any evidence) move significantly higher to a threshold likelihood (i.e. a relative stopping rule), and then move significantly lower to a final point. The fact that the three points are distinct supports the notion that individuals start at a prior point, then information accrual causes the them to reach a threshold point (allowing a response to be given) and that information integration (or distortion) continues once a decision has/can be made. This may allow for the possibility of verdict reversals. The fact that the threshold and last point differ also highlighted that the threshold point may be a pivotal and independent point in the decision making process, which 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 was also confirmed: there was a significant difference in likelihood ratings across verdicts, with guilty verdicts having significantly higher mean ratings than not proven verdicts, and not guilty verdicts. This highlighted that the way that the information was integrated was significantly different across the verdicts. This idea of information integration (represented by the likelihood of guilty values) in mock juror decision making is supported by previous research. Ostrom, Werner and Saks (1978) found that jurors initially thought that the suspect was innocent and that this belief was averaged, with the evidence in a trial, to give the final verdict. Estrada-Reynolds, Gray and Nunez (2015) found similar results. The likelihood rating analysis, therefore, highlighted that the evidence from the cases was integrated differently between the 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. Therefore, showing that jurors understand the differences between the verdicts, and use them differently depending on how they integrate information and which threshold has been reached.

The third hypothesis was also accepted: there was a significant interaction between the two independent variables (i.e. verdict and likelihood points). Because the results showed that the likelihood thresholds (i.e. relative stopping rule) were different across the verdicts, this proposes that the reaching of a likelihood boundary informed the responses given (Wells, 1992). Additionally, because the last points were significantly different across the verdicts, thresholds were not re-surpassed, implying that appropriate responses were made. For example, drift caused the guilty (not guilty) threshold to be reached, and because the threshold was not re-surpassed a guilty (not guilty) was then given. Not proven verdicts can be said to be verdicts that fall in-between guilty and not guilty thresholds, and are consequently in the limbo boundary of not proven. Therefore, it can be said that the not proven verdict is a verdict where innocence is doubted but guilt cannot be confirmed. The acceptance of the third hypothesis further suggests that in each of the verdicts, evidence accumulated until a threshold was reached, and that evidence acquired after this point did not affect the response that was given (or cause a verdict reversal). Although, it must be mentioned that by asking participants to state the last piece of evidence, the researchers were only looking at the last threshold reached. Furthermore, participants could have crossed several thresholds throughout a trial, but it is the last threshold crossed that determines the verdict that is given.

The starting points for each of the verdicts were similar, suggesting that jurors in a three verdict system are not affected by pre-trial biases, and that the jurors believed initially that neither of the Anglo-American verdicts had been proven yet. An alternative explanation is that all of the participants are biased towards the defence. This is because all of the participants start with similar likelihoods of guilt and because the not guilty threshold point is above all of the different verdicts prior points. Furthermore, it is more likely that all of the jurors had a similar initial belief in “innocence 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 were significantly different from each other, suggesting that information is integrated together, causing likelihood ratings to drift to a boundary, which then allowed a response to be made (Ratcliff & Smith, 2004). The results, therefore, suggest that a third verdict can be incorporated into a relative stopping rule model by allowing thresholds/boundaries to be re-surpassed. In addition, the last points all varied from each another across the verdicts. This further supported the idea of distinct decision making processes across the three verdict types. Furthermore, the interaction analysis highlighted that the decision making processes were distinct across the three verdict types, and that information integration and thresholds allowed verdicts to be reached.

The interaction also highlighted that the decision making process for each of the verdicts varied. For example, with guilty verdicts, the decision making process drifted from the prior point to the threshold point, thus allowing a guilty verdict to be given. There was no difference between the threshold point and last point is guilty verdicts. Furthermore, the decision making process of guilty verdicts was found to be non-compensatory (Gigerenzer & Goldstein, 1996): the combination of available cues could not over-ride the guilty verdict made using, on average, 6.2 pieces of information.

Nevertheless, it is important to point out that in the theorised diffusion model that drift can continue post-threshold and the drift can either confirm the leading verdict (that is the threshold that was last passed) or disconfirm the current threshold. The current research has shown that the information presented post-threshold in guilty verdicts seemed not to have a significant effect on drift. However, it is likely that if strong defence evidence was placed near the end of a trial, that the drift would be negative, thus causing the threshold to be re-surpassed and allowing a verdict reversal. Future research will investigate this by manipulating the order of evidence using strength ratings.

The decision making process of not proven verdicts was unique in relation to the other three verdicts. For instance, not proven verdicts started off with their prior likelihood near the middle of the scoring system, then significantly rose to the threshold point. The likelihood score at the last point was significantly lower than that at the threshold point. This highlighted that the drift of not proven verdicts initially rose in a similar fashion to guilty verdicts. However, through a collection of cues the drift reaches a decision making zone of not proven, which is in-between the guilty and not guilty thresholds. Further, for not proven verdicts, the rise may be enough to escape the not guilty boundary, but may not be sufficient enough to cross the guilty threshold, leaving the juror’s verdict in a not proven limbo.

Nevertheless, if more information had been provided, the drift could have went in one of three ways:

1) The drift may not have changed much, and a not proven verdict would still have been given;
2) The drift may have increased until a guilty threshold was reached, thus allowing a guilty verdict to be given,
3) A decrease in drift may have occurred until the not guilty threshold was surpassed, allowing a not guilty verdict to be given.

In summary, drift is constant, but the last thresholds crossed determines the verdict that is given. For example, one juror could cross the guilty thresholds using one piece of information and then give a guilty verdict because the rest of the information does not cause the guilty threshold to be re-surpassed. However, another juror could start of in the not guilty boundary, and then their drift could increase after five pieces of information until they reach the guilty boundary. Then, once piece of information could cause them to drop below the guilty boundary, and if the rest of the information did not cause a change in the decision making drift, the juror would then give a not proven verdict.

The decision making process of not guilty verdicts was distinct from the other two verdicts. The analysis highlighted that not guilty verdicts started with a middling prior likelihood, and then increased, significantly, to the threshold point. Then the likelihood value significantly dropped to the last point. This, alongside the fact that, on average, not guilty verdicts used 11.7 pieces of information, indicates that initially the decision maker may have started off at a point of innocence, and then increased their likelihood ratings. Moreover, the threshold point would then symbolise when the jurors drift re-surpassed the not guilty threshold, after an initial rise, from the not proven area into the not guilty zone. Jurors who gave not guilty verdicts continued to decrease their likelihood ratings after crossing the not guilty threshold, showing that they became more convinced in the innocence of the suspect post-threshold.

In addition, the fact that cue utilisation differed between the verdict types supports the notion of threshold decision making and information integration. This is because some verdicts (guilty) were reached using fewer cues in comparison to other verdict types (not guilty and not proven), which suggests that the weight of the information that is integrated and the frugalness of a threshold may determine the verdict that is reached. The results also show that the more compensatory the process is, the more likely the decision drift will favour acquittal verdicts in comparison to conviction verdicts. Interestingly, guilty verdicts were reached after 6.2 pieces of information, even though the smallest number of pieces of information shown in a case was 10 pieces of evidence. This, therefore, highlighted that individuals were making decisions before all the evidence was shown, which supports the view that once a threshold was reached a decision could be made. Furthermore, the current research has shown that through allowing both thresholds and information integration to vary both rational and heuristic processing can be mirrored.

Significant associations were found, which supported the fourth hypothesis. This hypothesis related to investigating if individual counters, or evidence types, could allow decision makers to reach a threshold and thus give an associating verdict. The association between guilty verdicts and the last guilty cue/counter needed was significant. The direction was positive and the correlation coefficient was strong. Likewise, the association between the last not guilty cue/count needed and not guilty verdicts was significant, positive and the coefficient was modest. Further, the relationship between the last not proven cue/count needed and not proven verdicts was significant, positive and the correlation coefficient was moderate. Thus providing some support that information was placed in one of three counters; guilty, not guilty and not proven; and that the thresholds that was reached first in each of the counters then informed the decision made.

Despite the significant associations, the results indicated that count information might not be enough to allow thresholds to be reached. For instance, although the significance levels suggest that these results are below a 1% likelihood of occurring by chance, the correlation coefficients propose that the last cue/count ratings needed do not always allow a threshold to be reached. Further, because the correlation coefficients are not one, it proposes that more than just count information is needed to reach thresholds. Consequently, this implies that simply counting information does not allow a threshold to be reached; and, therefore that the decision making process is more integrative. In other words, count data does not always allow thresholds to be reached on their own, and information integration may explain why a boundary was reached more adequately. For example, evidence counted as guilty may be weak and may actually drift a juror away from a guilty threshold response (or keep the decision making process constant). In summary, the correlations do hold some support for the count model, but also cast doubt on how realistic the model may be in a juror setting.

One potential limitation of this study is that there could be a lack of ecological validity, as the study was quasi-experimental. This artificial setting, with the lack of real life cues, may have affected the decision making process, and made what was observed in the current experiment different from what may have occurred in a real life trial (Simon, 1956; Wiener et al., 2011). However, previous research comparing culpability ratings from participants who used transcripts versus participants who viewed eyewitnesses on a video camera found that it was the information, not the meduim, that was important in regards to the jurors beliefs of the culpabability of the suspects (Pezdek, Avila-Mora, & Sperry, 2010). In addition, cognitive processes, such as decision making, happen in a similar manner irrespective of whether or not the cognitve processes are produced naturally or artifically (Watt & Quinn, 2008). Nevertheless, future research may want to try and increase the ecological validity of mock juror experiments through audio/visual stimuli, to more closely align to processes that would occur in a real life trial.

A confounding effect may have been introduced by asking participants to rate the evidence as guilty, not guilty, or not proven. and then rate the liklihood of guilt of the suspect from 1-100. The initial evidence rating may have anchored the liklihood of guilt ratings. There was no other way to collect both independent evidence ratings and liklihood of guilt data, however. Even if the liklihood of guilt evidence was anchored by the evidence ratings (i.e. count data), it does not take away from the fact that the diffusion model could explain the juror decision making data more sufficently than the count model could. Future research may want to repeat the current experimnet in a betweeen subjects method, where one group gives liklihood of guilt ratings and one group gives evidence ratings (count data), and then compare which model explains their own groups decision amking data in the most satisfactory manner.

This study is the first to demonstrate that the diffusion model is a good metaphor for the decision making processes that occur within jurors, and that the decision making process of jurors may involve integrating information in a way that allows a threshold to be reached, which then allows a verdict response to be given if the threhsold is not re-surpassed. Further, the current piece of research has highlighted that all three verdict options in use within the

Scottish criminal justice systems have distinct threshold points, thus demonstrating that jurors have a different interpretation of not proven and not guilty verdicts, contratsing with older research on the topic. The overarching findings of the study supports 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**.**

Disclosure statement

The authors report no declarations of interest.

Funding

Funding for my PhD was provided from the McGlashan Charitable Trust, and the East Lothian Educational Trust.

Conference Presentation

The international Association of Forensic Mental Health Services.

Geographical statement

The study was carried out at Edinburgh Napier University, Scotland.

Ethics Declaration

The current study was granted ethical approval by Edinburgh Napier's Research and Integrity committee.

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/

Ask, K., Rebelius, A., & Granhag, P. A. (2008). The "Elasticity" of criminal evidence: A moderator of investigator bias. *Applied Cognitive Psychology*, *22* (1), 1245–1259. doi: 10.1002/acp.1432

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. (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.

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: doi:10.3758/BF03196581

MacCoun, R. J. (1989). Experimental research on jury decision-making. *Science, 244* (4908), 1046-1050. Available from: http://www.jstor.org/

Mehlhorn, K., Ben-Asher, N., Dutt, V., & Gonzalez, C. (2014). Observed variability and values matter: Toward a better understanding of information search and decisions from experience. *Journal of Behavioral Decision Making, 27* (4), 328-339. doi: 10.1002/bdm.1809

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. Austin, TX: Cognitive Science Society. 473-478. Available from: http://csjarchive.cogsci.rpi.edu/proceedings/2009/papers/85/paper85.pdf.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology, 2* (2), 175- 220. doi: 10.1037/1089-2680.2.2.175.

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.

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the story model for juror decision making. *Journal of Personality and Social Psychology, 62* (2), 189-206. doi: 10.1037/0022-3514.62.2.189.

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.ohiolink.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., & 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), 333-267. doi: 10.1037/0033-295X.111.2.333.

Ratcliff, R., Philiastides, 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

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*, 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, D. (2004). A third view of the black box: Cognitive coherence in legal decision making. *The University of Chicago Law Review*, 511-586. Available from: http://www.jstor.org/.

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.

Sommers, S. R., & Ellsworth, P. C. (2000). Race in the courtroom: Perceptions of guilt and dispositional attributions. *Personality and Social Psychology Bulletin*, *26* (11), 1367-1379. doi: 10.1177/0146167200263005

The Analysis Factor (2017). *When unequal sample sizes are and are not a problem in ANOVA*. Retrieved from: http://www.theanalysisfactor.com/

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.

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.

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.

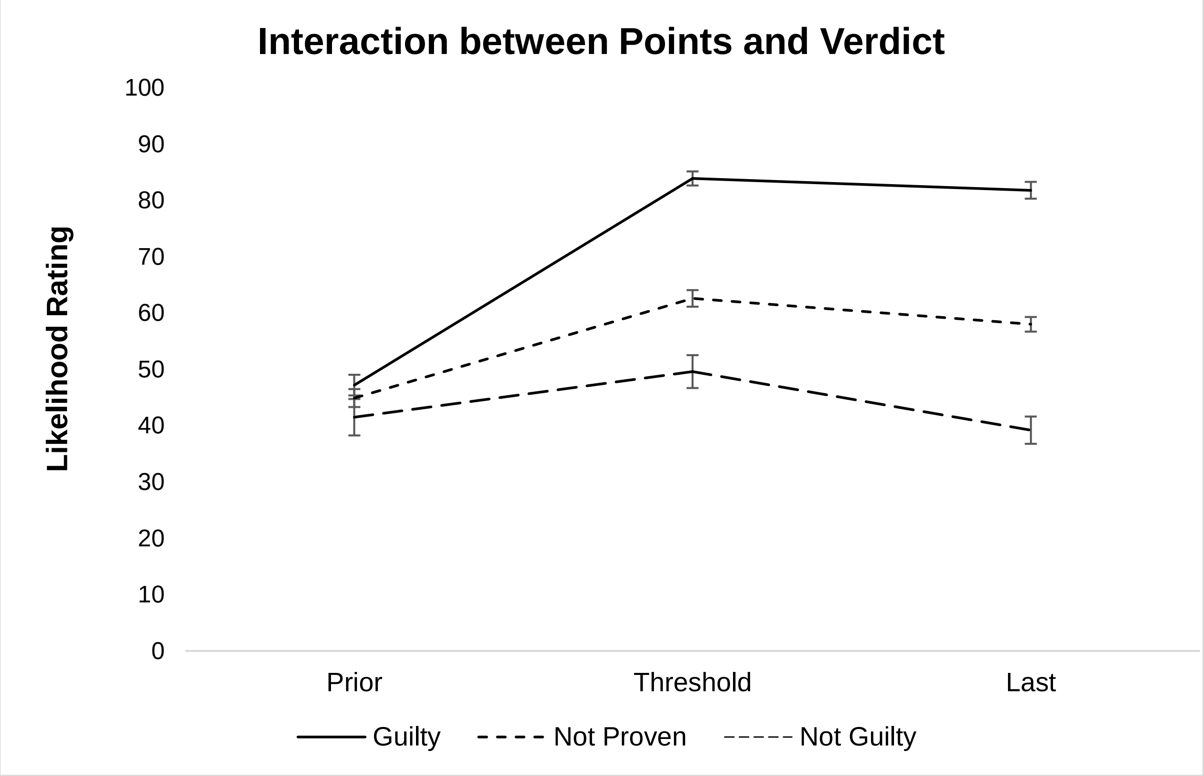
Wenzel, M., Okimoto, T. G., Feather, N. T., & Platow, M. J. (2008). Retributive and restorative justice. *Law and Human Behavior, 32,* 375-389. doi: 10.1007/s10979-007-9116-6.

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.

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.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosphical Transactions Society, 367,* 1310-1321. doi: 10.1098/rstb.2011.0416.

Figure.1


Interaction between Points and Verdict
Likelihood Rating
100
90
80
70
60
50
40
30
20
10
0
Prior
Threshold
Last
Guilty
Not Proven
Not Guilty

Tables

Table 1. Descriptive Statistics for Verdict Type

Table 2. Descriptive statistics of cue utilisation between verdicts

Table 3. Descriptive statistics for the absolute stopping rule

Table 1. Descriptive Statistics for Verdict Type

	Likelihood of Guilt Rating		
	Prior	Threshold	Last
Guilty			
Mean	47.2	83.9	81.8
Median	50.0	85.0	82.5
Standard Deviation	14.1	9.6	11.5
Not Guilty			
Mean	41.5	49.7	39.2
Median	50.0	50.0	40.0
Standard Deviation	20.2	18.2	15.1
Not Proven			
Mean	44.9	62.6	58.0
Median	50.0	63.5	59.0
Standard Deviation	12.1	11.2	9.9

Table 2. Descriptive statistics of cue utilisation between verdicts

Verdict	Mean	Median	Standard Deviation
Guilty	6.2	5	3.4
Not Proven	11.4	11	3.3
Not Guilty	11.7	11	3.0

Table 3. Descriptive statistics for the absolute stopping rule

Variable	Mean	Median	SD
Last Count Rated as Guilty	4.0	4	1.9
Last Count Rated as Not Guilty	1.8	1	1.7
Last Count Rated as Proven	3.2	3	2.1
Guilty Verdict	4.0	4	1.7
Not Guilty Verdict	1.3	1	1.3
Not Proven Verdict	3.6	4	2

Appendices

Appendix 1. Details included in the nine vignettes, across the three verdict types handed down in the real-case trials**.**

	Guilty	Not Guilty	Not proven
Familiar victim	2 familiar 1 not	1 familiar 2 not	2 familiar 1 not
Vulnerable victim	2 vulnerable 1 not	1 vulnerable 2 not	2 vulnerable 1 not
Crime details	1.Housebreak/ stabbing 2.Argument/multiple injuries 3. Body not found.	1.Head injury/strangled 2.Neck injury/affixation 3.self-defence/stabbing	1.Cut throat 2.Body not found 3. Multiple injuries/blow to head.
Victim age (years)	51, 16, 33	19, 44, 26	18, 43, 26
Victim gender	2 female 1 male	1 female 2 males	3 female
Accused age	20, 22, 39.	33, 49, 23.	33, 21, 33.
Accused gender	3 males	3 males	3 males
Weapon used	2 yes, 1 unknown.	2 yes, 1 no.	2 yes, 1 unknown.
N words in opening statement	123, 110, 112. Mean =115	129,104, 114. Mean=115.7	125, 105, 122 Mean= 117.3

Appendix 2. Cue types and number per vignette (case), across the three verdict types handed down in the real-case trials**.**

	Guilty	Not Guilty	Not Proven
Number of cues	Case 1: prosecution 7. Defence 5.	Case 2: prosecution 7. Defence =7.	Case7: prosecution 6. Defence 5.
	Case 3: prosecution 9. Defence 9.	Case 4: prosecution 5. Defence 6.	Case 8: prosecution 9. Defence 7.
	Case 6: prosecution 9 Defence 6	Case 5: prosecution 5. Defence 5.	Case 9: prosecution 7. Defence 6.
Severity rating of case (mean)	Case 1=4.69	Case 2=4.25	Case 7=4.13
	Case 3=4.13	Case 4=3.88	Case 8=3.38
	Case 6=3.56	Case 5=3.94	Case 9=1.94
Familiarity rating of case (mean)	Case 1=2.06	Case 2=2.56	Case 7=1.69
	Case 3=2.00	Case 4=1.69	Case 8=1.50
	Case 6=1.94	Case 5=1.94	Case 9=1.94
Realism rating (mean)	Case 1=3.81	Case 2=4.00	Case 7=3.56
	Case 3=3.50	Case 4=3.50	Case 8=3.75
	Case 6=3.44	Case 5=4.19	Case 9=4.19

The severity, familiarity and realism scores were gathered from the pilot. Each of the scores mentioned could be ranked from one to five; one indicating a low score and five symbolising a high score.