

Running head: MEASURING REASONABLE DOUBT

Dhami, M. K. (2008). On measuring quantitative interpretations of reasonable doubt. *Journal of Experimental Psychology: Applied*, 14, 353-363.

On Measuring Quantitative Interpretations of Reasonable Doubt

Mandeep K. Dhami

University of Cambridge, UK

Send correspondence to:

Mandeep K. Dhami
Institute of Criminology
University of Cambridge
Sidgwick Avenue
Cambridge
CB3 9DT
England, UK
E-mail: mkd25@cam.ac.uk
Tel: + 44 (0)1223 335385
Fax: + 44 (0)1223 335356

Acknowledgements

The author would like to thank Hal Arkes and Tom Wallsten for their helpful comments on the present research.

On Measuring Quantitative Interpretations of Reasonable Doubt

Abstract

Beyond reasonable doubt represents a probability value that acts as the criterion for conviction in criminal trials. We introduce the membership function (MF) method as a new tool for measuring quantitative interpretations of reasonable doubt. Experiment 1 demonstrated that three different methods (i.e., direct rating, decision theory-based, and MF) provided significantly different and uncorrelated interpretations of reasonable doubt, although all methods predicted verdicts equally well, and showed inter-individual variability in interpretations. In Experiment 2 only the direct rating method demonstrated a significant effect of judicial instructions on reasonable doubt. In both Experiments, the MF method showed intra-individual variability in interpretations of reasonable doubt. The methods may be capturing different aspects of the concept of reasonable doubt. These findings have implications for the validity of past research findings on reasonable doubt, and for the utility of triangulation of methods in future research.

Keywords: Standards of proof, beyond reasonable doubt, membership function, linguistic probabilities, juror decision making

In criminal trials, it is difficult if not impossible to determine what actually occurred, and so decision makers rely on a standard of proof that specifies the degree of belief in guilt required, in terms of probability, before conviction. In this way, the standard acts as a criterion for making verdicts. In many adversarial systems such as the Anglo-North American, beyond reasonable doubt is the standard of proof used in criminal trials (see Shapiro, 1991). This represents a principle of due process (Newman, 1993) which is designed to minimize false convictions by setting a stringent criterion (high probability) for conviction, albeit, at the expense of possible false acquittals (see Arkes & Mellors, 2001; and DeKay, 1996). It is generally agreed that this is the most stringent standard in law, and should represent a value above .90 (see Newman, 1983; United States v. Fatico, 1978; which does not imply the standard should be quantified).

Studies measuring judges and mock jurors' quantitative interpretations of reasonable doubt have found that interpretations are sometimes below and sometimes around .90, and there are inter-individual differences in interpretations (for a review see Hastie, 1993). However, researchers have used different methods. Dane (1985, p. 143) demonstrated that different methods yielded different interpretations, and that some methods provided interpretations that better matched "recreated" verdicts with actual verdicts (or had greater validity). In this paper, we compare two commonly used methods against a new method which has not been previously used to measure quantitative interpretations of reasonable doubt. The membership function method was developed to measure people's quantitative interpretations of linguistic probabilities in general such as "very likely" (Karelitz, Budescu, & Wallsten, 2000). Two experiments examine the effect of method on measurement of quantitative interpretations of reasonable doubt. Such research can have implications for law-makers, legal professionals, and researchers. Determining how methods can impact research findings can help law-makers evaluate the reliability and validity of past findings on

quantifying reasonable doubt. It can also help future researchers choose the most effective and efficient method in research on standards of proof. Ultimately, a better understanding of how people interpret standards of proof can help legal professionals appreciate how they are used by jurors.

Measuring Quantitative Interpretations of Reasonable Doubt

In measuring quantitative interpretations of reasonable doubt it is assumed that, given the evidence, people assess the probability that the defendant committed the crime as charged. If the probability of commission is greater than the interpretation of the standard then an individual will convict (find the defendant guilty), otherwise the individual will acquit. The standard of proof represents the balance placed between a defendant's rights to freedom and society's rights to be protected from crime. As a principle of due process, reasonable doubt puts the balance in favor of the defendant. Researchers have measured peoples' quantitative interpretations of reasonable doubt, particularly, as evidence indicates that interpretations affect verdicts (e.g., Horowitz & Kirkpatrick, 1996; Kagehiro & Stanton, 1985; Kerr, Atkin, Stasser, Meek, Holt, & Davis, 1976).

McCauliff (1982) reported that the value of reasonable doubt ranged from 50% to 100% with a mean of 90% across judges. (Researchers often use percentages instead of probabilities). Hastie, Penrod, and Pennington (1983) found that on average members of the public interpreted reasonable doubt as .92. Martin and Schum (1987) demonstrated that students interpreted the standard as from .91 to .99 across offence types. However, other studies have found that mean interpretations of reasonable doubt are somewhat below .90 for judges (e.g., Simon, 1969; Simon & Mahan, 1971) and mock jurors (students; e.g., Horowitz & Kirkpatrick, 1996; Kagehiro, 1990; Kassin & Wrightsman, 1979; Kerr et al., 1976; Nagel, 1979; Nagel & Neef, 1979; Simon, 1970; Simon & Mahan, 1971). The above body of

research also reveals wide inter-individual differences across peoples' interpretations of reasonable doubt.

Effect of Judicial Instructions

One potential source of the variability in interpretations of reasonable doubt lies in its imprecision. Jurors are uncertain what the standard means (e.g., Young, Cameron, & Tinsley, 2001). Some jurisdictions have developed and use judicial instructions that aim to define the standard for jurors (see Hemmens, Scarborough, & del Carmen, 1997). For example, the Federal Judicial Center (1987, Instruction number 21) in the US which is responsible for research and education in the federal courts, has defined reasonable doubt as "...proof that leaves you firmly convinced of the defendant's guilt..." In principle, there are four possible scenarios that may arise when the judge instructs jurors: reasonable doubt is left undefined (i.e., "the defendant is presumed innocent unless the prosecution has proved guilt beyond a reasonable doubt"), a qualitative definition is used (e.g., see above definition by Federal Judicial Center, 1987), a quantitative definition is used (e.g., "the defendant is presumed innocent unless the evidence against him has at least 90% probability of truth"), or a combined qualitative-quantitative definition is used. Researchers have studied the effects of these scenarios on interpretations of reasonable doubt (for a review see Horowitz, 1997).

Kerr et al. (1976) found that interpretations of reasonable doubt were significantly higher under a "stringent" qualitative definition (i.e., where doubt must be substantial) than under no definition or a "lax" qualitative definition (i.e., where any doubt qualified as reasonable doubt). Nagel (1979) reported that interpretations were lower under qualitative than quantitative instructions. Kagehiro and Stanton (1985) similarly demonstrated that quantitative instructions had the intended effect on reducing guilty verdicts by increasing the standard whereas qualitative instructions did not. Later, Kagehiro (1990) reported that whereas combined qualitative followed by quantitative instructions did not have the intended

effect on verdicts, combined quantitative followed by qualitative instructions did (see also Kagehiro & Stanton, 1985). Horowitz and Kirkpatrick (1996) revealed that five different qualitative instructions had differential effects on interpretations of the standard. Others have also found effects of judicial instructions on reasonable doubt (e.g., Koch & Devine, 1999; Montgomery, 1998).

Methodological Considerations

Importantly, different studies have employed different methods to measure quantitative interpretations of reasonable doubt. For example, Hastie et al. (1993), Kagehiro (1990), and McCauliff (1982) all used a direct rating method, whereas Nagel (1979) used an indirect method. Research on the effects of judicial instructions has mostly employed the direct rating method (Horowitz & Kirkpatrick, 1996; Kagehiro, 1990; Kerr et al., 1976; Montgomery, 1998).

Hastie (1993) classified methods as direct and indirect. Two commonly used types of methods are the direct rating method (also called minimum probability or self-report method), and the decision theory-based method (also called statistical decision theory method), which is classified as an indirect method. There are several versions of these two types of methods.

Typically, the direct rating methods asks participants what would be the *minimum probability* that the defendant committed the crime as charged that they would require before deciding to convict. Responses may be provided on a 0 to 1 probability scale (or some variant such as a 0 to 100 percent scale).

By contrast, the decision theory-based methods assume that a decision to convict should be made if the *utility* of conviction is greater than that of acquittal (Kaplan, 1968). Participants are asked to assign values to the four possible outcomes of a trial (i.e., acquitting

the innocent, convicting the guilty, convicting the innocent, and acquitting the guilty). Fried, Kaplan, and Klein's (1975) formula is typical (for a critique see Arkes and Mellers [2002]):

$$\text{Decision Criterion} = \frac{\text{Utility (Acquit Innocent)} - \text{Utility (Convict Innocent)}}{\text{Utility (Acquit Innocent)} - \text{Utility (Convict Innocent)} + \text{Utility (Convict Guilty)} - \text{Utility (Acquit Guilty)}}$$

In a review of studies using these methods, Hastie (1993) concluded that direct methods such as the direct rating methods produced higher (or more stringent) interpretations of reasonable doubt than did indirect methods such as the decision theory-based methods. Indeed, direct rating methods tended to provide interpretations of approximately .80 or above whereas decision theory-based methods yielded interpretations from .50 to .55. These differences were also apparent within studies using more than one method (e.g., Nagel, 1979; see also Connolly's [1987] discussion of the difference in standards obtained via direct probability ratings versus establishing utilities). Thus, direct rating methods appear to emphasize an individual's reluctance to make a false conviction, whereas decision theory-based methods appear to highlight the undesirability of a false acquittal.

Dane (1985) empirically compared different methods including the commonly used self-report (or direct rating) and statistical decision theory-based methods. Mock jurors were presented with three versions of a fictitious trial involving aggravated assault. They read the trial transcript along with the judge's instructions on the standard of proof, completed a questionnaire pre-deliberation that obtained information for measuring reasonable doubt using the different methods, deliberated in juries to reach a unanimous verdict, and completed the questionnaire again post-deliberation. It was found that the mean pre-deliberation (and post-deliberation in parentheses) measures of reasonable doubt were: .66 (.70) using the self-report method and .52 (.52) using the statistical decision theory-based method.

In an attempt to examine the validity of each method, Dane (1985) converted the interpretations of reasonable doubt measured by each method into verdicts by comparing participants' probability of commission ratings to their interpretations of reasonable doubt as elicited by each method, and assuming guilty verdicts if the commission values were greater than the reasonable doubt values (otherwise assuming not guilty verdicts). Thus, Dane (1985) *recreated* verdicts, and these were then compared to participants' actual verdicts to see how well they matched. This approach is different from the one of using regression models to predict verdicts from measures of reasonable doubt used in the present study. According to Dane (1985), the methods accurately recreated the following percentages of verdicts at pre-deliberation (and post-deliberation in brackets): 77% (73%) for self-report method and 82% (85%) for statistical decision theory-based method. Therefore, both methods provided verdicts that were congruent with actual verdicts above chance level, although the statistical decision theory-based method proved somewhat better than the self-report method.

However, to date, no published studies have used measures of reasonable doubt to directly predict verdicts, and so the *predictive validity* of the methods remains unknown. This makes it difficult to determine which method ought to be used in future research, and which method is capable of capturing people's "true" interpretations of reasonable doubt. Similarly, to date, there is no published research on the intra-individual variability in interpretations of reasonable doubt, which like inter-individual variability, may have adverse implications for due process requirements and the accuracy of verdicts. Finally, if different methods do yield different interpretations of reasonable doubt, it would be useful to know to what extent the effects of judicial instructions are attributable to researchers' choice of direct-rating methods.

The present paper aims to extend Dane's (1985) analysis by comparing the predictive validity of different methods, examining the extent of both inter- and intra-individual variability in interpretations of reasonable doubt indicated by different methods, and

determining whether different methods yield different findings on the effects of judicial instructions on reasonable doubt. In particular, two commonly used methods (i.e., direct rating and decision theory-based) will be compared to a new method that has not been previously used in this context. This is the membership function (MF) method (Karelitz et al., 2000) which can for an individual simultaneously capture both point and variability in interpretations, thus providing measures of inter- and intra-individual variability. Next, we provide a brief review of this method and the theory that underlies its use.

Theory of Linguistic Probabilities and the Membership Function Method

Wallsten and Budescu's (1995) theory of linguistic probabilities states that phrases such as "very likely" can be represented as fuzzy subsets of the probability interval (see also Budescu & Wallsten, 1995). Thus, these qualitative probability phrases can be interpreted quantitatively. According to the theory, the "membership function" (MF) of a phrase has a peak and spread of probabilities. The MF, $\mu_w(p)$, for a phrase w evaluated at probability p equals 0 if the individual considers p *not at all* described by w , the function equals 1 if p is considered to be *absolutely* described by w , and it equals a number between 0 and 1 if p is considered to be described by w to *some degree*. The peak (or point) interpretation of a phrase is the probability value that is rated the highest on the *not at all* to *absolutely* scale for its ability to substitute for the phrase. The spread (or range) of interpretation of a phrase refers to all those probability values either side of the peak which have a greater than *not at all* rating, indicating that these values can substitute for the phrase but only to *some degree*. Comparing peaks across people shows inter-individual variability whereas the spread within people demonstrates intra-individual variability.

The multi-stimuli MF method that is used to elicit people's quantitative interpretations of probability phrases was developed and validated by Karelitz et al. (2000). Briefly, participants are asked to rate the extent to which specific probability values from 0 to

1 (in .1 intervals) “substitute” for a phrase (e.g., very likely). Ratings are made on scales from 0 representing *not at all* to 100 for *absolutely* (in 10-point intervals). The responses can be joined-up to provide a pictorial representation of the peak and spread of the interpretation of a phrase. Figure 1 shows a hypothetical example of an individual’s interpretation of “very likely.” As illustrated, the MF for this phrase peaks at .6 with a spread of 8 points from .1 to .9. This phrase has one peak, but MFs can be bi/multi-modal or plateau, and this example is of a phrase representing a single probability whereas reasonable doubt represents a range of probabilities beyond the peak. The MF method can be classified as a direct method.

INSERT FIGURE 1 HERE

It has been shown that the MF method can reliably and validly measure peoples’ quantitative interpretations of probability phrases (e.g., Dhimi & Wallsten, 2005; Karelitz et al., 2000). Past research has found that people have broad interpretations for most phrases in their linguistic probability lexicons (e.g., Wallsten, Budescu, & Zwick, 1993), that interpretations of the same phrase differ across people (e.g., Budescu & Wallsten, 1990), and that interpretations may be affected by the context in which a phrase is used (e.g., Cohen & Wallsten, 1992; Shapiro & Wallsten, 1994 and Tsao & Wallsten, 1994 both unpublished studies cited in Wallsten & Budescu, 1995). The findings on linguistic probabilities using the MF method thus appear to mirror those of research on reasonable doubt using other methods.

Experiment 1

Aims and Hypotheses

The main aims of Experiment 1 were to: (1) examine the effect of method (i.e., direct rating, decision theory-based, and MF) on measurement of quantitative interpretations of reasonable doubt, (2) determine the validity of these methods in terms of their ability to predict verdicts, and (3) measure inter- and intra-individual variability in interpretations of reasonable doubt.

Based on past research on reasonable doubt it was hypothesized that direct methods such as the direct rating and MF methods would produce higher mean interpretations of reasonable doubt than the indirect, decision theory-based method. Thus, the two direct methods were hypothesized to indicate that people aim to minimize false convictions whereas the indirect method was hypothesized to show that people aim to reduce false acquittals. Second, based on Dane's (1985) findings, it was hypothesized that direct and indirect methods would differ in their ability to predict verdicts, and that indirect methods would be more predictively valid (although note that the methods in Dane's study were used to *recreate* rather than *predict* verdicts). Finally, based on past research on reasonable doubt and linguistic probabilities it was hypothesized that there would be wide inter- and intra-individual variability in interpretations of reasonable doubt (at least when undefined).

Method

Participants. Forty-nine jury eligible students at the University of Victoria, Canada, volunteered to participate. They received two percentage points towards a course. Their mean age was 23.60 years, and 81.63% were female. Two participants reported having been called for jury service.

Design. A within-subjects design was employed with three levels of method (i.e., direct rating, decision theory-based, and MF). The order of presentation of methods was randomized across participants.

Measures and procedure. Data were collected in small groups and the Experiment took approximately 45 minutes to complete. Participants first read a 5-page summary of a real criminal trial in the US (California v. Suzanne Johnson, 1998). This was chosen partly because Canadian students would likely be unaware of the case, and partly because the original trial culminated in a hung jury, which would likely lead to more variance in verdicts to be predicted in the present study. However, the original charge was murder and a pilot

study indicated that few participants would convict for this charge, but may consider conviction for a lesser charge such as manslaughter. Thus, for the present purposes, Suzanne Johnson, a 53 year-old day care provider, was tried for the manslaughter (rather than murder) of a 6 month-old baby girl in her care. The trial summary included details of the charge and penalty, background to the case, the victim's injuries, the prosecution case, the defense case, and the defendant's statements. (A copy of the trial summary is available from the author.)

Next, participants read the judge's typical legal instructions on the presumption of innocence, burden of proof, and standard of proof. Regarding the latter, the instructions included the following statement "The defendant is presumed innocent unless the prosecution has proved guilt beyond a reasonable doubt." Thus, the standard is left undefined.

Following this, participants completed a three part "Pre-deliberation Questionnaire" based on Dane's (1985). In Part 1, participants rated the probability that the defendant committed the crime as charged, their confidence in their verdict, how well they understood the judge's instructions, how confident they were that they obeyed the judge's instructions, how difficult it was for the prosecution to meet the required standard of proof, the strength of the prosecution case against the defendant, and the strength of the defense case for the defendant. Probabilities were measured on 0-1 scales with .05-point intervals, confidence ratings were made on 0-100 scales with 10-point intervals, and the other ratings were made on 0-10 scales with 1-point intervals. Participants also made a verdict (i.e., not guilty or guilty).

Part 2 of the questionnaire measured participants' interpretations of reasonable doubt using each of the three methods. In the direct rating method, participants were asked "what is the minimum probability for the defendant having committed the crime that he/she is charged with which you would require before you convict?" They provided responses by circling a

point a 0% to 100% scale with 5-point intervals. Responses were converted to a probability scale for analysis.

In the decision theory-based method, participants were asked to separately rate the value they attached to the four possible outcomes of a criminal trial: convicting a person who is innocent, acquitting a person who is guilty, convicting a person who is guilty, and acquitting a person who is innocent. Responses were measured on 21-point scales with 1-point intervals.

In the MF method, participants were presented with the phrase beyond reasonable doubt along with 11 21-point scales labeled *not at all* to *absolutely* at each end. Each scale corresponds to one of 11 percentage values ordered from 0% to 100% (in 10% intervals; these values were converted to probabilities for analysis). These values were presented in the same order across participants. Participants were asked to respond to the following question: “To what extent, from *not at all* to *absolutely*, would each of these percentage values (e.g., 0%, 10%, etc) substitute for the phrase ‘beyond reasonable doubt?’ Make sure you circle a point on each scale.”

Finally, Part 3 collected data on participants’ demographic details including their gender, age, jury experience, and their prior knowledge of the case.

Results

Data from three participants were excluded from analysis because they reported being aware of the Johnson case. It is worth noting that 11 of the 49 participants said they found the instructions to the MF task confusing such that they used the scale in the reverse direction, and so their responses were also excluded from analyses involving the MF method (this problem was remedied in Experiment 2).

Effect of method on measures of reasonable doubt. The direct rating and the decision theory-based methods both yield point interpretations of reasonable doubt, whereas the MF

method yields a peak and spread of interpretation. For present purposes, the peak of the MF (the probability value that absolutely substitutes for the phrase) was used in data analysis unless otherwise stated.

There were no significant correlations across interpretations of reasonable doubt provided by the three methods: $r = -.18$ for the direct rating and decision theory-based methods, $r = .11$ for the direct rating and MF methods, and $r = -.08$ for the decision theory-based and MF methods, $ps > .05$. Scatter plots (not shown here) ruled out the possibility of non-linear relationships existing. Although other factors such as a restricted range would reduce the magnitude of a correlation, this was not considered relevant as reasonable doubt is not expected to span the entire probability interval, and standard deviations indicated that one measure was not restricted compared to others.

A within-subjects ANOVA revealed a significant main effect of method on reasonable doubt, $F[2, 66] = 177.99, p = .001$, partial $\eta^2 = .84$. The MF method (peak) yielded the highest mean interpretations ($M = .96, SD = .10$), followed by the direct rating method ($M = .85, SD = .12$), and the decision theory-based method ($M = .53, SD = .09$). Further analyses using paired-sample t -tests indicated that mean interpretations of reasonable doubt measured by the three methods were significantly different from each other, $t[43] = 17.01, p = .001$, Cohen's $d = 2.31$ for direct rating and decision theory-based methods, $t[34] = 4.69, p = .001, d = -1.00$ for direct rating and MF methods, and $t[33] = 17.45, p = .001, d = -4.78$ for decision theory-based and MF methods.

Predicting verdicts. Participants said they had understood the judge's instructions quite well ($M = 8.70, SD = 1.23$), and were highly confident they had obeyed these instructions ($M = 79.89, SD = 16.36$). The verdicts were fairly evenly split across participants: 54.35% rendered a guilty verdict. As would be expected, verdicts (0 = not guilty and 1 = guilty) were significantly negatively correlated with participants' ratings of the

strength of the defense case ($r = -.52, p = .001$) and significantly positively correlated with their ratings of the strength of the prosecution case ($r = .53, p = .001$). Similarly, as expected, the mean rating of the probability that the defendant committed the crime as charged was significantly greater for participants who rendered a guilty verdict ($M = .82, SD = .11$) than those who voted not guilty ($M = .48, SD = .20$), $t[30] = 7.12, p = .001, d = 2.11$. Also as expected, participants who voted not guilty said it was significantly more difficult for the prosecution to meet the required standard of proof ($M = 7.52, SD = 1.50$) than those who voted guilty ($M = 5.84, SD = 2.64$), $t[39] = 2.71, p = .010, d = -0.78$. Participants who voted guilty were significantly more confident in their verdicts ($M = 78.54, SD = 14.02$) than those who voted not guilty ($M = 55.48, SD = 24.13$), $t[31] = 3.85, p = .001, d = 1.17$.

The validity of the three methods in predicting verdicts was examined by computing separate binary logistic regression analyses where the interpretations of reasonable doubt elicited by each method were used to predict participants' verdicts. All three models were statistically significant and predicted verdicts above chance level (i.e., 54%), $\chi^2[1, N = 46] = 4.24, p = .039, \phi = .03$ for direct rating method, $\chi^2[1, N = 44] = 4.65, p = .031, \phi = .33$ for decision theory-based method, and $\chi^2[1, N = 35] = 6.39, p = .011, \phi = .43$ for MF method. Specifically, the methods accurately predicted the following percent of verdicts: 60.87% for direct rating method, 63.64% for decision theory-based method, and 62.86% for MF method.

Finally, a logistic regression model was computed to predict verdicts in which interpretations of reasonable doubt captured by the three methods were entered simultaneously as predictors. It was found that the model accurately predicted 76.47% of verdicts ($\chi^2[3, N = 34] = 4.24, p = .001, \phi = .35$), and only the measures provided by the direct rating and MF methods were significant predictors, $ps < .05$.

Inter- and intra-individual variability in reasonable doubt. The direct rating method showed the greatest inter-individual variability in interpretations of reasonable doubt (i.e.,

range = .75), followed by the decision theory-based method (i.e., range = .56). The MF method showed the least inter-individual variability in interpretations (i.e., range = .40).

For interest, Figure 2 illustrates the average MF of reasonable doubt across participants, showing the peak and spread of interpretations. As mentioned earlier, the MF method is capable of capturing intra-individual differences in interpretations by analyzing the spread/range (rather than peak) of the MF. Across participants, the mean spread of the MF was .52 (*SD* of range = .35), with four participants showing no spread and six indicating that reasonable doubt represented to some degree values along the whole 0-1 probability interval.

INSERT FIGURE 2 HERE

Discussion

Experiment 1 demonstrated that different methods produce different quantitative interpretations of reasonable doubt. As hypothesized, direct methods (i.e., direct rating and MF methods) produced higher mean interpretations of the standard than the indirect, decision theory-based method. The two commonly used methods (i.e., direct rating and decision theory-based methods) on average yielded values of reasonable doubt of .85 and .53, respectively, which is consistent with past research (see Hastie, 1993). By contrast, the MF method, which has not previously been used in this context, yielded an average value of .96. Thus, whereas the decision theory-based method suggests that people aim to reduce false acquittals, the direct rating and MF methods suggest they focus on reducing false convictions.

In support of our prediction, and consistent with past research on reasonable doubt (e.g., McCauliff, 1982) and linguistic probabilities (see Wallsten & Budescu, 1995), all three methods showed wide inter-individual variability in interpretations of reasonable doubt. The greatest variability was indicated by the direct rating method, whereas the MF method indicated the least variability. These findings may reflect real variability that exists or may suggest that the latter method is used more consistently across participants.

Also as hypothesized, and consistent with past research on linguistic probabilities (see Wallsten & Budescu, 1995), the MF method revealed intra-individual variability in interpretations of reasonable doubt as the spread of the MF suggested that most participants took the standard to represent to *some degree* probability values below and above their *absolute* interpretation. Indeed, for most participants reasonable doubt represented values along half of the 0-1 probability interval. This is the first study of intra-individual variability in interpretations of reasonable doubt.

These findings on inter- and intra-individual variability raise concerns about the consistency of juror behavior. Inter-individual variability suggests that different jurors may find it difficult to reach consensus on a verdict (Kalven & Zeisel, 1966), whereas intra-individual variability suggests that the same juror may arrive at a different verdict on the same or similar case on another occasion if the juror uses a value along the spread of the MF as the threshold. However, as Heffer (2006) points out, there is no legal consensus that reasonable doubt should be fixed (see also Stoffelmayr & Diamond, 2000). Nevertheless, it is worth exploring the extent to which such variability is driven by systematic factors rather than being random. Identifying the effect of systematic factors can help to develop strategies to minimize variability in interpretations of reasonable doubt. Experiment 2 provided data to enable initial exploration of this issue by examining the effect of judicial instructions.

Participants were fairly evenly divided in their verdicts on the case. Logistic regression analyses showed that all three methods accurately predicted verdicts above chance level. Furthermore, although all three methods, as hypothesized, differed in their ability to accurately predict verdicts, the difference was negligible, and the decision theory-based method did not do as well in the full model. The *predictive validities* of the direct rating and decision theory-based methods were somewhat lower than the validities reported by Dane (1985) who *recreated* verdicts. Dane (1985) used participants' probability of commission

values, which are necessarily related to actual verdicts, to also assume the verdicts determined by the reasonable doubt standard measured by each method. This increases the likelihood that the verdicts generated by the methods matched actual verdicts and so inflates the estimated validity of each method.

The fact that the three methods yielded uncorrelated interpretations of reasonable doubt suggests that either they contain measurement error or that they may be capturing unique and useful components of the concept of reasonable doubt. A closer examination of each method indicates potential sources of the differences across methods. For instance, the direct rating method asks for the *minimum* probability value that represents reasonable doubt, whereas the MF method (peak) is the probability value that *absolutely* substitutes for the phrase, and the decision theory-based method asks for the *utility* of the outcomes of a criminal trial. The two direct methods may also have in common an ability to draw people's attention to the goal of reducing false convictions by focusing them on the *likelihood* of the event (i.e., person committing the crime). By contrast, the indirect method may draw people's attention to the goal of reducing false acquittals by focusing them on the *value* of the outcome of the trial. A triangulated approach to studying reasonable doubt may be useful.

Before drawing firm conclusions about the effect of method on reasonable doubt, it is worth noting that Experiment 1 demonstrated the effect of method within the context of a manslaughter case. Past research, mostly using the direct rating method, suggests that offence type may influence interpretations of reasonable doubt (Martin & Schum, 1987; Montgomery, 1998; Simon, 1969; 1970; Simon & Mahan, 1971). Thus, the interpretations of reasonable doubt obtained in Experiment 1 may not necessarily generalize to cases involving other offence types. Nevertheless, there is no reason to believe that the effect of method would differ across offence types. Rather, a meaningful analysis may be to compare the effect of method outside of case context. Experiment 2 provided data to enable such analysis.

Finally, if different methods consistently yield different findings, then it is useful to establish the implications of using each of the three methods for researching specific issues such as the effect of judicial instructions on interpretations of reasonable doubt. As stated earlier, past research has reported that instructions do influence interpretations (e.g., Kagehrio & Stanton, 1985; Nagel, 1979). Interestingly, this body of research has mostly employed the direct rating method, and so it is unclear if the findings would be supported if another method was employed. Therefore, Experiment 2 was designed to examine the robustness of the findings on the effect of judicial instructions on reasonable doubt.

Experiment 2

Aims and Hypotheses

The main aims of Experiment 2 were to: (1) examine the effect of judicial instructions on quantitative interpretations of reasonable doubt as measured by the direct rating, decision theory-based, and the MF methods, (2) explore the extent of inter- and intra-individual variability in interpretations of reasonable doubt across judicial instructions, and (3) replicate the results of Experiment 1 on the effect of method on reasonable doubt outside of case context.

Based on past research on the effect of judicial instructions (Kagehiro & Stanton, 1985; Nagel, 1979), it was predicted that according to the direct rating method quantitative instructions would yield higher interpretations of reasonable doubt than qualitative instructions or when the standard is undefined (there is insufficient past research to make an a priori prediction about the effect of instructions using the decision theory-based and MF methods). Second, although there is a lack of previous research to make a priori predictions about inter- and intra-individual variability in interpretations of reasonable doubt across judicial instructions, it may be appropriate to expect less variability under quantitative than qualitative instructions or when the standard is undefined. This is because quantitative

instructions might better narrow the possible range of interpretations that an individual may have. Finally, it was expected that the results of the effect of method on reasonable doubt reported in Experiment 1 that were obtained in the context of a manslaughter case would be replicated outside of case context in Experiment 2.

Method

Participants. Two-hundred and three jury eligible students at the University of Victoria, Canada, volunteered to participate. They received two percentage points towards a course. Their mean age was 20.12 years, and 75.77% were female. Three participants reported having been called for jury service.

Design. A 3 x 3 mixed factorial design was employed with judicial instructions (i.e., undefined, qualitative, quantitative) as the between-subjects factor and method (i.e., direct rating, decision theory-based, and MF methods) as the within-subjects factor. The order of presentation of methods was randomized across participants.

Measures and procedure. Data were collected in small groups and the Experiment took approximately 30 minutes to complete. Participants first read the judge's instructions. In the undefined condition the instructions were as follows: "The defendant is presumed innocent unless the prosecution has proved guilt beyond a reasonable doubt." The qualitative instructions were: "The defendant is presumed innocent, and the law does not require the prosecution to prove a defendant guilty beyond all possible doubt, and neither is it sufficient to prove that the defendant is probably guilty." Finally, the quantitative instructions were: "The defendant is presumed innocent unless the evidence against the defendant has at least 90% probability of truth."

Participants then completed a questionnaire that measured their interpretations of reasonable doubt using each of the three methods described in Experiment 1. The instructions for the MF task were, however, made clearer to avoid confusion: "Imagine that you had to

use a number instead of the phrase ‘beyond reasonable doubt.’ To what extent does each of the numbers below (e.g., 0%, 10%, etc) substitute for the phrase? If you think the phrase can *not at all* be substituted by the number then circle the left-most point on the scale. If you think the phrase can be *absolutely* substituted by the number then circle the right-most point on the scale. If you think the phrase can be *somewhat* substituted by the number then circle a point somewhere along the scale in-between the left- and right-most points on the scale that represents your thought. Make sure you circle a point on each scale.”

Finally, data were collected on participants’ demographic details including gender, age, and jury experience.

Results

Effect of judicial instructions on reasonable doubt. Of the 203 participants, five did not complete the direct rating task, two did not complete the decision theory task, and 17 did not complete the MF task. The peak of the MF for reasonable doubt was used when analyzing the data from the MF method unless otherwise stated. None of the participants reported being confused with the MF task.

There were no significant correlations among the interpretations of reasonable doubt as measured by the three methods across the judicial instructions (this was also the case for within instructions): $r = -.01$ for the direct rating and decision theory-based methods, $r = .08$ for the direct rating and MF methods, and $r = -.07$ for the decision theory-based and MF methods, $ps > .05$. Scatter plots (not shown here) again ruled out the possibility of non-linear relations. As before, other factors such as restricted range could also not account for the lack of significant correlations.

A mixed ANOVA was computed to examine the effect of judicial instructions and method on interpretations of reasonable doubt. Judicial instruction was the between-subjects factor and method was the within-subjects factor. There was a significant main effect of

method on interpretations of reasonable doubt, $F[2, 346] = 553.58, p = .001, \text{partial } \eta^2 = .76$ (see Table 1). There was no significant main effect of judicial instructions on interpretations, $F[2, 173] = 0.76, p = .469, \text{partial } \eta^2 = .01$ (see Table 1). However, there was a small, but statistically significant interaction effect of judicial instructions by method on interpretations of reasonable doubt, $F[4, 346] = 3.68, p = .006, \text{partial } \eta^2 = .04$ (see Table 1). Further analyses using one-way ANOVAs showed that the interpretations of reasonable doubt measured by the decision theory-based and MF methods did not vary significantly across the three judicial instructions, $ps > .05$. However, although the effect is small, the interpretations measured by the direct rating method significantly increased from the undefined condition to the condition with quantitative instructions, $F[2, 197] = 4.09, p = .018, \text{partial } \eta^2 = .04$.

INSERT TABLE 1 HERE

Inter- and intra-individual variability in reasonable doubt. As Table 2 shows, all three methods showed greatest inter-individual variability in interpretations of reasonable doubt under the qualitative instructions (i.e., range = .95 for direct rating method, range = .68 for decision theory-based method, and range = .50 for MF method). The MF method showed least inter-individual variability in interpretations under the quantitative instructions (range = .20), whereas this was the case for the direct rating and decision theory-based methods in the undefined condition (i.e., range = .75 for direct rating method and range = .62 for decision theory-based method).

INSERT TABLE 2 HERE

With regard to the effect of instructions on intra-individual variability in interpretations, a one-way ANOVA revealed that there were no significant differences in the spread of the MF for reasonable doubt across the different instructions, $F[2, 183] = 1.41, p = .247, \text{partial } \eta^2 = .02$. The mean spread of the MF across instruction conditions was .59 (*SD*

of range = .32), with 8 participants showing no spread and 14 indicating that reasonable doubt represented to some degree values along the whole 0-1 probability interval.

Replication outside of case context. The data from the undefined condition in Experiment 2 were analyzed to determine the extent to which the results of Experiment 1 were replicated outside of case context. A within-subjects ANOVA revealed a significant main effect of method on reasonable doubt, $F[2, 124] = 205.15, p = .001, \text{partial } \eta^2 = .77$. The MF method (peak) yielded the highest mean interpretation of reasonable doubt ($M = .98, SD = .06$), followed by the direct rating method ($M = .77, SD = .19$), and the decision theory-based method ($M = .53, SD = .10$). Indeed, these statistics are close to those reported in Experiment 1 (i.e., $M = .96$ and $SD = .10$ for MF method, $M = .85$ and $SD = .12$ for direct rating method, and $M = .53$ and $SD = .09$ for decision theory-based method). Similarly, further analyses indicated that mean interpretations of reasonable doubt measured by the three methods were significantly different from each other, $t[66] = 6.02, p = .001, d = 1.43$ for direct rating and decision theory-based methods, $t[63] = 8.90, p = .001, d = -1.49$ for direct rating and MF methods, and $t[64] = 32.13, p = .001, d = -5.08$ for decision theory-based and MF methods.

Discussion

The main effect of method found in the analysis of data from the undefined condition in Experiment 2 replicated the results of Experiment 1 on the effect of method on interpretations of reasonable doubt outside of case context. Indeed, there is no reason to believe that the effect of method should differ across offence types or from a case to no case context, even though interpretations of reasonable doubt may themselves differ. Consistent with Experiment 1, the average values of reasonable doubt obtained outside of case context in Experiment 2 were higher for the direct methods (i.e., direct rating and MF methods) than the indirect, decision theory-based method. Furthermore, as in Experiment 1, there were no

significant correlations among interpretations of reasonable doubt provided by the three methods.

In support of our prediction, Experiment 2 demonstrated that the direct rating method yields on average significantly higher interpretations of reasonable doubt under quantitative instructions than when the standard is left undefined. This is consistent with past research (Kagehiro & Stanton, 1985; Nagel, 1979). To add to this literature, and in contrast to the effect observed using the direct rating method, it was revealed that interpretations of reasonable doubt provided by both the decision theory-based and MF methods did not differ significantly under different judicial instructions. In addition, instructions did not have a significant effect on intra-individual variability in interpretations captured by the MF method. According to all three methods inter-individual variability was greatest under the qualitative instructions. It was only the MF method that revealed the expected finding of least inter-individual variability in interpretations of reasonable doubt under quantitative instructions.

The present findings imply that previous research findings concerning the effect of judicial instructions on reasonable doubt may have been affected by the method used to measure interpretations of reasonable doubt. Indeed, most of the past research on the effect of judicial instructions has used the direct rating method (Horowitz & Kirkpatrick, 1996; Kagehiro, 1990; Kerr et al., 1976; Montgomery, 1998). One explanation for the effect demonstrated by the direct rating method is that instructions may only impact the *minimum* value of reasonable doubt captured by this method. Alternatively, and consistent with the present findings, there are reasons to believe that instructions would not necessarily impact interpretations. Jurors are active information processors and they use commonsense, so they may reject or amend instructions, particularly if they conflict with their sense of justice (e.g., Dann, 1993; Finkel, 1995). Courts are also often ineffective in communicating the standard to

lay jurors (Heffer, 2006). Future research could probe into the issue of the sensitivity or susceptibility of different methods to the effects of judicial instructions on reasonable doubt.

Meanwhile, it may be wise to postpone policy discussions concerning the utility of different judicial instructions until further research is conducted using various methods. Adoption of specific instructions also ought to take into account their impact on intra- and inter-individual variability in interpretations of reasonable doubt. The present findings suggest that qualitative instructions may be problematic since they tend to increase inter-individual variability, which may make it difficult for jurors to agree on a verdict. On the other hand, there has been widespread objection to the use of quantitative instructions (Notes, 1995). It should be noted that the present study used only one example of a qualitative and quantitative instruction and so replication is necessary using other such instructions, as well as instructions that may combine both quantitative and qualitative elements.

General Discussion

The present findings have several implications. First, people interpreted the standard either below the .90 probability generally agreed by judges and law-makers (according to the direct rating and decision theory-based methods), or above the threshold (according the MF method). Effort should be made to bring people's interpretations of reasonable doubt closer to that intended.

Second, all three methods showed wide inter-individual variability in interpretations of reasonable doubt measured in terms of range (which was greater under qualitative instructions), and the MF method also revealed intra-individual variability. This raises concerns about the ability of jurors to agree on a verdict as well as the consistency of individual jurors' verdicts. Future research could further explore the sources of such variability, so that strategies can be designed to minimize variability.

Third, the effect of judicial instructions on interpretations of reasonable doubt were only apparent using the direct rating method, but not the decision theory-based nor MF methods. This questions the reliability and validity of past research findings on judicial instructions. Legal policy-makers and practitioners may wish to postpone use of instructions or at least be cautious of their use. Choice of instructions ought to be informed by knowledge of whether they have the desired impact on peoples' understanding of reasonable doubt, while ensuring that they do not increase variability in interpretations. In fact, the MF method can be used to determine how people interpret particular phrases such as "firmly convinced" that may form part of a qualitative instruction by identifying phrases with little inter- and intra-individual variability (spread) and whose MFs typically peak at .90 probability.

Finally, the interpretations of reasonable doubt captured by the three methods were uncorrelated, and the three methods predicted verdicts equally well. These methods appear to capture different aspects of the concept of reasonable doubt. Future research ought to seek further evidence of the relative and combined predictive validity of the methods. For now, it may be prudent for researchers to use a triangulated approach to studying reasonable doubt. Alternatively, researchers may use a method on grounds of practicality or the type of questions it can help answer (see also Connolly, 1987).

Balancing the Evidence for the Membership Function Method

The present paper introduced the MF method to the study of standards of proof. This method represents a viable addition/alternative to the direct rating and decision theory-based methods. One unique advantage of the MF method is that it captures intra-individual variability in interpretations of a phrase. Another advantage is that the reliability and validity of the MF method has been repeatedly confirmed (e.g., Dhimi & Wallsten, 2005; Karelitz et al., 2000). Indeed, Dhimi and Wallsten (2005) demonstrated that the MF method can capture quantitative interpretations of probability phrases in individuals' unique lexicons and that

these interpretations can predict how individuals use their phrases. A third advantage lies in the fact that the MF method showed the least inter-individual variability in interpretations (measured either in terms of range or standard deviation), implying that it may be used fairly consistently across individuals.

The MF method also has the advantage of direct methods. For instance, the direct methods present participants with a fairly simple task (although it may be argued that people are not necessarily familiar with using probability scales). By contrast, the decision theory-based method presents a relatively more complex task which may seem peculiar (e.g., asking for the utility of convicting the innocent). As an indirect method, the decision theory-based method also requires more assumptions on the part of the researcher in converting utilities to a criterion value and some have commented on the difficulties of establishing a consistent set of utilities (e.g., Connolly, 1987; although Arkes and Mellers [2002] present a formula, that has not been used to date, which avoids asking peculiar questions and which they claim is more representative of what jurors may do if they wish to maximize utility).

However, the MF method also has some of the limitations of direct methods, as well as other potential limitations. For instance, direct methods may be more prone to social desirability response bias. In addition, asking participants the extent to which values along the probably scale at .1 intervals substitute for a phrase inevitably means that the quantitative interpretation of any phrase provided by the MF method is not as fine-grained as it would be if the intervals were smaller (e.g., .05). This may partly explain why the interpretations of reasonable doubt and the intra-individual variability in interpretations were so great. Yet, there is a trade-off between using smaller intervals and the time taken to complete the task. Similarly, presenting probability values in the same ascending order across participants in the MF method may encourage them to respond in a certain way (e.g., increase MF for values as they ascend). However, this order is not inherent to the method and can be changed, although

it might add noise to the data by making the task more difficult. Regardless, it is clear that the MF method is amenable to revisions that enhance the efficiency of data collection while retaining its theoretical properties.

Strengths and Limitations

The present research used methods and procedures that are typical of jury research in general and research on standards of proof in particular. Nevertheless, it could be argued that the external validity of the present findings is limited because they are based on research involving participants who may not be representative of actual jurors, that reasonable doubt was examined in either no case context or in the context of a written summary of a case which diverges from the detailed oral and visual representations available at a real trial, and that reasonable doubt and verdicts were only measured at the pre-deliberation stage. For the reasons outlined below, these apparent limitations are minimized in the present research.

It would be inappropriate to study real juries in real trial situations, for ethical reasons when one is manipulating a variable such as judicial instructions, and for pragmatic reasons when one wants to conduct a robust study using a within-subjects design since the same jury will not try the same case more than once. Regardless, the present findings may be considered representative, reliable, and valid for several reasons. First, the mock (student) jurors in the present research were jury eligible and a few said they had been called for jury service. Some studies have also shown few differences between mock and real jurors (e.g., MacCoun & Kerr, 1988). Second, effort was made to use a detailed summary of a real trial rather than a brief description of a fictitious trial. Additionally, Bornstein (1999) concluded that there was little negative impact on the validity of studies using student mock jurors and written presentations of trial material. Finally, there is some evidence that the pre-deliberation distribution of verdicts influences the post-deliberation verdict (e.g., Kalven & Zeisel, 1966;

Newkirk, 1981; Sandys & Dillehay, 1995), and that interpretations of reasonable doubt differ very little from pre- to post-deliberation (Dane, 1985; Horowitz & Kirkpatrick, 1996).

Concluding Remarks

The present findings highlight the need to re-evaluate the reliability and validity of past research findings on quantifying reasonable doubt and on the effects of judicial instructions on reasonable doubt, since they may have been affected by the method used to measure interpretations. The present findings can also help future researchers choose the appropriate method when studying standards of proof, by emphasizing the advantages and limitations of different methods, and the potential for a triangulated approach. Gaining a better understanding of how people interpret standards of proof such as beyond reasonable doubt can help legal professionals appreciate how it is used by jurors, and how they may be assisted to use it appropriately.

References

- Arkes, H. R., & Mellers, B. A. (2002). Do juries meet our expectations? *Law and Human Behavior, 26*, 625-639.
- Bornstein, B. H. (1999). The ecological validity of jury simulations: Is the jury still out? *Law and Human Behavior, 23*, 75-91.
- Budescu, D. V., & Wallsten, T. S. (1990). Dyadic decisions with numerical and verbal probabilities. *Organizational Behavior & Human Decision Processes, 46*, 240-263.
- Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. In J. Busemeyer, D. L. Medin, & R. Hastie (Eds.), *Decision making from a cognitive perspective* (pp. 275-318). New York: Academic Press.
- Cohen, B. L., & Wallsten, T. S. (1992). The effect of constant outcome value on judgments and decision making given linguistic probabilities. *Journal of Behavioral Decision Making, 5*, 53-72.
- Connolly, T. (1987). Decision theory, reasonable doubt, and utility of erroneous acquittals. *Law and Human Behavior, 11*, 101-112.
- Dane, F. C. (1985). In search of reasonable doubt. *Law and Human Behavior, 9*, 141-158.
- Dann, B. M. (1993). "Learning lessons" and "speaking rights": Creating educated and democratic juries. *Indiana Law Journal, 68*, 1229-1279.
- DeKay, M. (1996). The difference between Blackstone-like error ratios and probabilistic standards of proof. *Law & Social Inquiry, 21*, 95-132.
- Dhimi, M. K. & Wallsten, T. S. (2005). Interpersonal comparison of subjective probabilities. *Memory & Cognition, 33*, 1057-1068.
- Federal Judicial Center (1987). *Pattern criminal jury instructions*. Washington, DC: Federal Judicial Center.

Finkel, N. J. (1995). *Commonsense justice. Jurors' notions of the law*. Cambridge, MA:

Harvard University Press.

Fried, M., Kaplan, K. J., & Klein, K. W. (1975). Juror selection: An analysis of voir dire. In

R. J. Simon (Ed.), *The juror system in America: A critical overview* (pp. 58-64).

Beverly Hills, CA: Sage.

Hastie, R., (1993). (Ed.). *Inside the juror: The psychology of juror decision making* (pp. 84-

115). Cambridge: Cambridge University Press.

Hastie, R., Penrod, S., & Pennington, N. (1983). *Inside the jury*. Cambridge, MA: Harvard

University Press.

Heffer, C. (2006). Beyond 'reasonable doubt': The criminal standard of proof instructions as

communicative act. *The International Journal of Speech, Language and the Law*, 13,

159-188.

Hemmens, C., Scarborough, K. E., & del Carmen, R. V. (1997). Grave doubts about

'reasonable doubt': Confusion in State and Federal courts. *Journal of Criminal*

Justice, 25, 231-254.

Horowitz, I. A. (1997). Reasonable doubt instructions. Commonsense justice and standard of

proof. *Psychology, Public Policy, and Law*, 3, 285-302.

Horowitz, I. A., & Kirkpatrick, (1996). A concept in search of a definition: The effects of

reasonable doubt instructions on certainty of guilt standards and jury verdicts. *Law*

and Human Behavior, 20, 655-670.

Kagehiro, D. K. (1990). Defining the standard of proof in jury instructions. *Psychological*

Science, 1, 194-200.

Kagehiro, D. K., & Stanton, W. C. (1985). Legal vs. quantified definitions of standards of

proof. *Law and Human Behavior*, 9, 159-178.

- Kalven, H., & Zeisel, H. (1966). *The American jury*. Boston, MA: Little, Brown and Company.
- Kaplan, J. (1968). Decision theory and the factfinding process. *Stanford Law Review*, 20, 1065-1092.
- Karelitz, T., Budescu, D. V., & Wallsten, T. S. (2000). *Validation of a new technique for eliciting membership functions of probability phrases*. Poster presented at the Meeting of the Society for Judgment and Decision Making, New Orleans, USA.
- Kassin, S. M., & Wrightsman, L. S. (1979). On the requirements of proof: The timing of judicial instruction and mock juror verdicts. *Journal of Personality and Social Psychology*, 37, 1877-1887.
- Kerr, N. L., Atkin, R. S., Stasser, G., Meek, D., Holt, R. W., & Davis, J. H. (1976). Guilt reasonable doubt: Effects of concept definition and assigned decision rule on the judgments of mock jurors. *Journal of Personality and Social Psychology*, 34, 282-294.
- Koch, C. M., & Devine, D. J. (1999). Effects of reasonable doubt definition and inclusion of a lesser charge on jury verdicts. *Law and Human Behavior*, 23, 653-674.
- Martin, A. W., & Schum, D. A. (1987). Quantifying burdens of proof: A likelihood ratio approach. *Jurimetrics Journal*, 27, 383-402.
- McCauliff, C. M. A. (1982). Burdens of proof: Degrees of belief, quanta of evidence, or constitutional guarantees? *Vanderbilt Law Review*, 35, 1260-1335.
- MacCoun, R. J., & Kerr, N. L. (1988). Asymmetric influence in mock jury deliberation: Jurors' bias for leniency. *Journal of Personality and Social Psychology*, 54, 21-33.
- Montgomery, J. W. (1998). The criminal standard of proof. *New Law Journal*, 582-584.
- Nagel, S. S. (1979). Bringing the values of jurors in line with the law. *Judicature*, 63, 189-195.

- Nagel, S. S., & Neef, M. G. (1979). *Decision theory and the legal process*. Lexington, MA: Lexington Books.
- Newkirk, J. M. (1981). The effect of different reasonable doubt definitions and group versus individual deliberation on the behaviour of mock jurors. *Dissertation Abstracts International*, 42, 1673-B-1674-B.
- Newman, J. O. (1993). Beyond “reasonable doubt.” *New York University Law Review*, 68, 979-1002.
- Notes (1995). Reasonable doubt. *Harvard Law Review*, 108, 1716-1733, 1955-1972.
- Sandys, M., & Dillehay, C. (1995). First-ballot votes, pre-deliberation dispositions, and final verdicts in jury trials. *Law and Human Behavior*, 19, 175-195.
- Shapiro, B. (1991). *Beyond reasonable doubt and probable cause: Historical perspectives on the Anglo-American law of evidence*. Berkeley, CA: University of California Press.
- Simon, R. J. (1969). Judges’ translations of burdens of proof into statements of probability. *Trial Lawyer’s Guide*, 13, 103-114.
- Simon, R. J. (1970). Reasonable doubt. *The Journal of Applied Behavioral Science*, 6, 203-209.
- Simon, R. J. & Mahan L. (1971). Quantifying burdens of proof. A view from the bench, the jury, and the classroom. *Law and Society Review*, 319-330.
- Stoffelmayr, E., & Diamond, S. S. (2000). The conflict between precision and flexibility in explaining ‘beyond a reasonable doubt.’ *Psychology, Public Policy and Law*, 6, 769-787.
- United States v Fatico. 458 F.Supp. 388. (E.D.N.Y. 1978).
- Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *Knowledge Engineering Review*, 10, 43-62.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science*, 39, 176-190.

Young, W., Cameron, N., & Tinsley, Y. (2001). *Juries in criminal trials*. New Zealand Law Commission Report No. 69. Wellington: New Zealand Law Commission.

Table 1.

Means and Standard Deviations of Interpretations of Reasonable Doubt by Method and Judicial Instructions

	Method					
	Direct Rating		Decision Theory- Based		Membership Function	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Undefined	.77	.19	.53	.10	.98	.06
Qualitative	.79	.20	.54	.10	.95	.10
Quantitative	.85	.17	.52	.08	.96	.05
Total	.80	.19	.53	.09	.96	.07

Table 2.

Inter-Individual Variability – Ranges of Interpretations of Reasonable Doubt by Method and Judicial Instructions

	Method		
	Direct Rating	Decision Theory-	Membership
		Based	Function
Judicial Instructions	Range	Range	Range
Undefined	.75	.62	.40
Qualitative	.95	.68	.50
Quantitative	.90	.64	.20

Figure Captions

Figure 1. *Example membership function of probability phrase 'very likely'.*

Figure 2. *Mean membership function of reasonable doubt across participants with 95% confidence intervals in Experiment 1.*

Figure 1

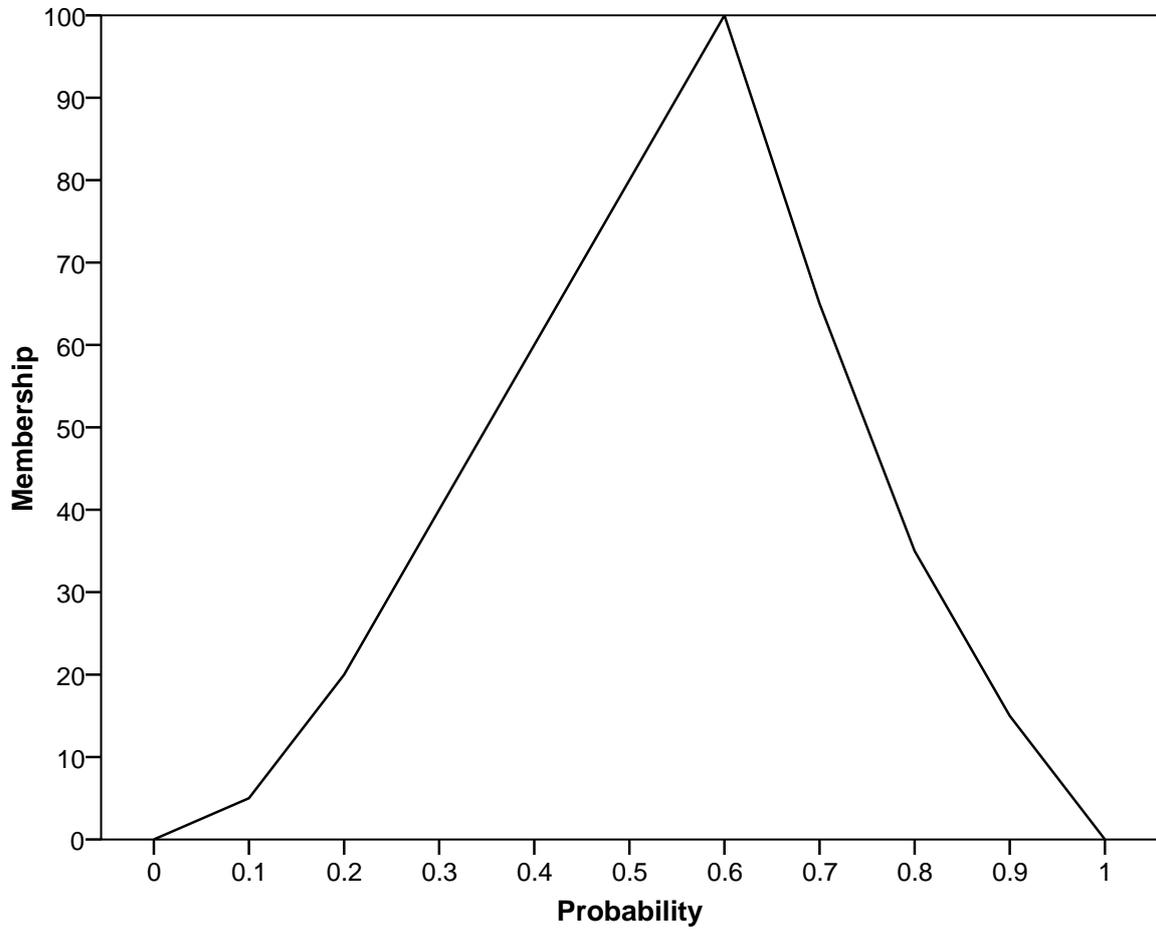


Figure 2

