

How a “Reasonable Doubt” Instruction Affects Decisions of Guilt

Daniel B. Wright
University of Sussex

Melanie Hall
University of Staffordshire

Jurors are instructed to render a guilty verdict if they feel the defendant is guilty beyond a reasonable doubt. The jury is often told that this does not mean an absolute certainty of guilt and that even if it were possible to imagine a scenario in which the defendant is innocent, a guilty verdict may still be appropriate. Here, participants read a case summary. They were either told to say that the defendant was guilty if they believed in guilt beyond a reasonable doubt or were given more detailed instruction stressing that they did not have to be absolutely certain of guilt to give a guilty verdict. In Experiment 1, participants provided “think-aloud” protocols. Content analysis revealed that those who were given this instruction often used the phrase *reasonable doubt* to justify their guilty verdicts by saying that although they were not certain of the defendant’s guilt, their belief exceeded the reasonable doubt threshold. None of the participants in the control group did this. Experiment 2 was designed to test if the instruction affected belief in guilt and the reasonable doubt threshold quantitatively. The instruction affected both people’s belief in guilt and the threshold that they used to define *reasonable doubt*. The implied values for reasonable doubt were 63% for those who received the instruction and 77% for the control group. Implications for jury decision making are discussed.

Before a jury decides the guilt of a defendant, the judge tells them that to return a guilty verdict, they must believe in guilt beyond a reasonable doubt (Shapiro, 1991). According to the U.S. Supreme Court, “the reasonable-doubt standard plays a vital role in the American scheme of criminal procedure. It is a prime instrument for reducing the risk of convictions resting on factual error” (In re Winship, 1970). However, *reasonable doubt* is a difficult concept for legal experts to define and even more difficult to explain to jurors, who are unlikely to be trained in legal terminology. “Beyond a reasonable doubt” is a more stringent belief in guilt than is necessary, for example, in civil cases, where only a preponderance of the evidence is needed in U.S. courts (see Clermont, 2004, for international and historical comparisons). However, it is not absolute certainty. Several surveys (Kramer & Koenig, 1990; Montgomery, 1998) have shown that without an instruction, many potential jurors wrongly believed that reasonable doubt

should be the same as certainty. More worrying is that many legal professionals also hold this view (Zander, 2000).

In many courts the judge defines reasonable doubt for the jury. Kerr and colleagues (1976) showed that some instructions can work well to clarify this difficult concept, but sometimes the definition is so complex that it increases confusion. Consider *Gaines v. Kelly* (2000). *Gaines* had been convicted for setting a grocery store on fire, which killed five people. His counsel argued that the definitions of *reasonable doubt* given at his original trial lowered the prosecution’s burden of proof below that of reasonable doubt. The original trial judge defined the phrase in seven ways. The appeal court agreed that due process had been denied. Several courts discourage judges from attempting to define the term (for example, *Gaines v. Kelly*, 2000). However, others have argued that it is better that a judge, trained in the law, define the phrase rather than leaving the interpretation to naïve jurors (for details see Horowitz, 1997).

Kagehiro (1990; Kagehiro & Stanton, 1985) found that defining reasonable doubt in terms of probabilities improved decision making. In three studies (Kagehiro & Stanton, 1985), participants given quantified definitions of standards

Correspondence should be addressed to Daniel B. Wright, Psychology Department, University of Sussex, Falmer, Brighton, BN1 9QH, UK. E-mail: DanW@sussex.ac.uk

of proof produced more appropriate verdicts than those given nonquantified definitions. However, providing quantities is frowned on in courts and has led to several successful appeals. Creating a good instruction is a difficult task, and it may be impossible to produce a single instruction that is suited for all the different types of criminal cases (Horowitz, 1997). What is clear is that we need to understand how different reasonable doubt instructions may affect jury decision making.

Almost all of the legal discussions about the reasonable doubt instruction center on whether it changes the necessary threshold of belief in guilt to render a guilty verdict (compare the different conclusions from *Cage v. Louisiana*, 1990, and *Victor v. Nebraska*, 1994; see Hemmens, Scarborough, & Del Carmen, 1997, and Kenney, 1995, for critical discussion). Most courts shy away from quantifying reasonable doubt. Some have used William Blackstone's adage (e.g. 18th century English jurist)—that it is better that ten guilty people are set free than one innocent man is imprisoned—to calculate the optimal threshold, but this also requires knowing, among other things, the baseline probability for guilt, which is seldom known (DeKay, 1996). To understand how the baseline affects the threshold if trying to adhere to Blackstone's adage, consider the following example. If 99.9% of defendants were guilty, then to maximize utility using Blackstone's adage jurors should almost always produce guilty verdicts because the probability of a false conviction is low, whereas if only 50% of defendants were guilty, jurors would want to avoid guilty verdicts unless they were fairly certain of guilt.

As discussed previously, when surveys ask what the reasonable doubt should be, many respondents give very high estimates. Several psychology studies have found that when people actually make guilty judgments, the reasonable doubt threshold can be lower, between 60% and 90% certainty in guilt (Horowitz, 1997). Values as low as 60% worry many, because this is much lower than most people expect and believe is appropriate (Arkes & Mellers, 2002). This range of observed probability estimates is large and depends on several factors, including the way it is measured. Some researchers (for example, Horowitz & Kirkpatrick, 1996) asked directly what the necessary probability of guilt is for rendering a guilty verdict. Hastie (1993) argued that less direct methods are better and avoid the chance that people report how they should, rather than would, behave. Here we ask participants whether they think the defendant is guilty, and then have them make a judgment about their certainty of guilt. We do not directly ask them to quantify reasonable doubt, but we are able to estimate this across the sample by examining the relationship between belief-in-guilt ratings and verdicts.

It is possible to use people's verdicts and their ratings for belief in guilt to estimate the threshold for reasonable doubt. The method uses a measure from biostatistics called LD50, which stands for the median lethal dose 50%. In biostatistics, this is the dose of a drug for which 50% of the sample would be affected (often meaning from which 50% would die,

hence the morbid name). Here we are interested in the level of belief in guilt where half of the sample would render a guilty verdict. To show how this is calculated, we use a logistic regression framework in which we try to estimate the probability of a guilty verdict ($p[GV_i]$) from the belief in guilt (GB_i):

$$\ln(p[GV_i]/(1-p[GV_i])) = \beta_0 + \beta_1 GB_i,$$

where $GV_i \sim \text{Bernoulli}(p[GV_i])$, which means the probability of a guilty verdict is like a coin toss in which the probability of landing heads (or guilty) is $p[GV_i]$. LD50 is estimated by $-\beta_0/\beta_1$. The asymptotic standard errors are calculated using the `dose.p` function in Venables and Ripley (1999, p. 221).

One limitation of this approach is that participants are asked both to render a guilty verdict and to make a belief guilt judgment. Someone may use their belief rating to justify their verdict. We opted against having one group of participants provide the belief-in-guilt ratings and a different group provide the verdicts, because this would make it difficult to investigate the relationship between these variables.

Rather than focusing only on whether a reasonable doubt instruction can affect the threshold, we are also interested in whether a reasonable doubt instruction may affect other aspects of the decision-making process. Many judges' instructions state how the juror can have some "possible or imaginary doubt" of the guilt, but that a guilty verdict would still be appropriate. For example, in *Gaines v. Kelly* (2000), one definition the original trial judge gave was "reasonable doubt is not mere speculative or imaginary doubt such as anybody might conjure up about anything under the sun." Hence, even if a person can imagine a situation in which the defendant is not guilty, the juror should still give a guilty verdict if the evidence implies guilt beyond a reasonable doubt. One question is whether this aspect of the instruction could do more than only affect the threshold. By inviting jurors to entertain the possibility that the defendant is innocent, this may increase their belief in innocence. This is because the act of imagining an event can increase a person's belief that the event is true (Garry & Polaschek, 2000; Koehler, 1991). Could the instructions used to explain reasonable doubt affect jurors' belief in guilt?

We had two main aims when constructing the instruction. First, we wanted to mention the possibility of imagining the defendant not being guilty, because this has occurred in some instructions and the cognitive psychology literature suggests that it could affect the belief in guilt. Second, we wanted the instruction to be brief. One reason for this is that a long instruction included with a brief trial description could signal to participants that they should focus too much on the instruction. Further, if the instruction was long, it would be difficult to identify which aspects of it affected people's decisions.

There are several models for juror decision making, and several detailed reviews have been published (Devine, Clayton, Dunford, Seying, & Pryce, 2000; Winters & Greene,

2007). One of the most influential models is the story model (Pennington & Hastie, 1992), in which individual pieces of information are used to build up a story. In countries with adversarial judicial systems, such as the United Kingdom and United States, the typical criminal case involves the defense and prosecution each presenting its own story (Pennington & Hastie, 1992; Wagenaar, 1996). In most cases, the defendant is not the culprit for the defense's story, but is for the prosecution's story. The juror decides which story is more believable. If the prosecution's story exceeds some reasonable doubt threshold, then the juror would be predicted to render a guilty verdict. An instruction that invites the juror to consider a defense story may increase the mental activation of this story, thus making it a more viable alternative than if no instruction is given.

Two studies were conducted. The aim of the first study was to explore the justifications people give for their verdicts. For this we asked participants to think aloud during their decision making and to provide retrospective justification. This allowed us to make qualitative inferences about the decision process. The second study examined the effects of the instruction quantitatively, using a larger sample. We examined whether the instruction affects the belief in guilt and, using logistic regression and the LD50 statistic, we estimated the reasonable doubt threshold for those given the instruction versus those not given the instruction.

EXPERIMENT 1

The aim of this experiment was to determine any qualitative differences in the decision-making processes for people given a particular reasonable doubt instruction versus those not given this instruction. Participants were asked to think aloud while making their decision and were also asked to retrospect about their decisions. Although people are not aware of all of the cognitive strategies that they use (Nisbett & Wilson, 1977), this procedure often provides useful insight into decision-making processes (Ericsson & Simon, 1980).

Method

Twenty-six undergraduates from the University of Bristol volunteered. They were either paid £3 or given a half-hour course credit. Participants read about the rape case of Nancy Von Roper (real names were not used) described in Loftus and Ketcham (1991). The case involved eyewitness misidentification and police corruption, leading to the false conviction of Tom Hoyle.¹ The summary that participants read (available on <http://www.sussex.ac.uk/Users/danw/roper>.

¹This is a distressing case. A reporter was able to find the real rapist. Tom Hoyle sued the police. After having to cross several hurdles, his case was successful, but he died before it was decided (for details see Loftus & Ketcham, 1991).

htm) was selected so that approximately half of the participants would give a guilty verdict.

Participants were told to decide whether the defendant was guilty. Half ($n = 13$) of the participants were told that they should render a guilty verdict only if they believed that the defendant was guilty beyond a reasonable doubt, without any further clarification. Participants in the instruction condition ($n = 13$) were also told this, but had an additional instruction:

You do not have to be certain of the defendant's guilt. You may be able to imagine a scenario in which the defendant is not guilty, but still believe the defendant is guilty "beyond a reasonable doubt."

We devised this instruction so that it was very short and briefly mentioned the possibility that a guilty verdict could still be made even if it were conceivable that the defendant was not guilty. Most instructions used in court are much longer; however, given the overall length of the summary, a brief instruction was needed, otherwise it would have stood out. This is, of course, only one possible reasonable doubt instruction. Although we constructed it based in part on existing instructions, others could also be used.

Participants were asked to rate their belief that the defendant was the culprit using a 0 to 100 probability scale. Participants were told to think aloud as they made their decisions. Finally, the participants were asked to write a description of how they reached their decision.

Results

The purpose of this study was to shed light on the decision processes, not to estimate the proportions of guilty verdicts or the mean ratings of guilt. These quantitative measures are reported for completeness, but the power is too low to realistically expect significant effects for them. The means for the belief-in-guilt comparison were not significantly different (control mean 71%, instruction mean 62%, $t[24] = 1.10$, $p = .28$, $d = 1.56sd$),² nor were the verdicts (control 38% guilty, instruction 54% guilty, $\chi^2(1) = 0.62$, $p = .43$, odds ratio = 1.87). Belief in guilt was a significant predictor of rendering a guilty verdict ($\chi^2(1) = 21.50$, $p < .001$, odds ratio = 1.17 [belief is on a 0–100 scale, which is why this value is near 1]), but no other effects approached significance. The estimate of reasonable doubt (LD50, from Venables & Ripley, 1999) is 71% with a 95% confidence interval from 62% to 79% for the entire sample. The estimate for the control group alone is 74% and for the instruction group is 70%. Given the sample size, the difference between these is not statistically significant.

The protocols were content analyzed in three ways: word counts, units of evidence, and use of the phrase *reasonable*

²The data for belief are negatively skewed (-0.31), but given the small sample size, this asymmetry is not statistically significant ($SE = 0.46$). Given the results of the second study it is worth reporting that the t test on this variable squared yields nearly identical results: $t(24) = 1.09$, $p = .29$.

doubt. First, word counts were calculated for the think-aloud and retrospective written protocols, broken down by verdict and condition. As expected, on average more words were spoken (mean $\pm 95\%$ confidence interval: 186 ± 41) than written (61 ± 10). Further analyses revealed no other significant effects on word counts.

The story was divided into 33 individual pieces of evidence. We recorded for both the think-aloud and the retrospective written reports whether the person used each of these and whether they used them in support of innocence or guilt. The total numbers of pieces of evidence used were summed for each individual for the think-aloud and retrospective protocols. A mixed $2 \times 2 \times 2 \times 2$ analysis of variance was run with two within-subjects factors (concurrent vs. retrospective protocol and innocence vs. guilt) and two between-subject factors (verdict and condition). There were significant effects for people saying more during the task than for writing afterward (11.60 versus 9.38 pieces of evidence), $F(1,22) = 4.77, p = .04$, partial $\eta^2 = 0.18$, and a large interaction for people describing evidence that was consistent with their verdicts, $F(1,22) = 77.82, p < .001$, partial $\eta^2 = 0.78$. Both these effects were predicted and unsurprising. No other effects were significant.

Next we looked at the use of the phrase *reasonable doubt*. This is a crude measure of whether the participant relied on the instruction to justify their verdict. The transcriptions were searched for this phrase. Everyone who used the phrase in their retrospective written protocols also used it in their concurrent think-alouds. We have therefore classified people simply as using the phrase or not. Table 1 shows the number using this phrase broken down by their verdicts and by their condition. Only people in the instruction condition used the phrase and still rendered a guilty verdict. A logistic regression was run using condition and whether the phrase was used to predict the verdict, and the interaction was statistically significant, $\chi^2(1) = 6.04, p = .01$.

Several of the participants' simply used the phrase in stating their verdict, "not guilty beyond a reasonable doubt," without further clarification of what was meant by the phrase. However, some descriptions were much more enlightening (all descriptions are available at <http://www.sussex.ac.uk/Users/danw/roper.htm>). Five people in the control

condition used the phrase, and all said not guilty. Their descriptions can be summarized by participants 1 and 21: "I wouldn't say that was absolutely spot on that he wasn't guilty" and "Isn't enough evidence to say 100% guilty." These participants were using the reasonable doubt criterion to justify not guilty verdicts on the basis of some doubt. The phrases *absolutely spot on* and *100% guilty* suggest that the threshold is assumed to be high. Ten people in the instruction condition used the phrase *reasonable doubt*. Four of these participants rendered verdicts of not guilty. Three of these four did not expand on what was meant by the phrase. Participant 16 gave an interesting description, noting that if he were on a "real jury" that the criterion is a "slight bit of doubt." This seems similar to the justifications given by those people in the control condition who said not guilty.

Although no one in the control condition who used the phrase *reasonable doubt* gave a guilty verdict, six people in the instruction condition did. There are two main ways in which people could give a guilty verdict. First, they could be absolutely certain of guilt and therefore simply state that this belief exceeds reasonable doubt. The story was designed so that no one should have been absolutely certain about guilt, and therefore no one in this study appeared absolutely certain of guilt. Second, a person could have doubts about guilt, but believe the likelihood that the defendant is guilty exceeds reasonable doubt. All six of the descriptions described uncertainty. However, they all expressed how their belief in guilt was high, exceeding reasonable doubt. Examples include: "very unlikely all are coincidences" (S4), "it's definitely not certain but ..." (S10), and "even though it is just circumstantial it just seems too obvious" (S25).

Discussion

The ways in which mock jurors made decisions about the guilt of a defendant were examined through both think-aloud protocols and retrospective written explanations. Some highly predictable effects were observed, such as longer spoken than written protocols and that people gave evidence consistent with their verdict. The most interesting effects concerned the use of the phrase *reasonable doubt*. In the control condition, people used this phrase only to justify not

TABLE 1
The number of Guilty Verdicts by use of the Phrase *Reasonable Doubt* and Condition

	Condition		Total
	Control (n = 13)	Instruction (n = 13)	
Did not use <i>reasonable doubt</i>	5 of 8 (63%)	1 of 3 (33%)	6 of 11 (55%)
Used <i>reasonable doubt</i>	0 of 5 (0%)	6 of 10 (60%)	6 of 15 (40%)
Total	5 of 13 (38%)	7 of 13 (54%)	12 of 26 (46%)

Note. The significant interaction between use of the phrase and condition is because only people in the instruction condition used the phrase and still rendered a guilty verdict.

guilty verdicts. Everyone who used this phrase in their descriptions made a not guilty verdict. Many argued that they felt the person was probably guilty, but not to the necessary level. Those given the instruction used the phrase in two ways. Some used it to justify not guilty verdicts, like those in the control condition. However, the majority (60%) who used the phrase did so to justify guilty verdicts. They argued that although they were not certain about guilt, they felt there was enough evidence against the defendant that their belief in guilt exceeded reasonable doubt.

This effect suggests that the instruction may cause participants to lower their threshold for reasonable doubt. Although the effects were in the predict directions, with a lower estimated threshold for the instruction group, the purpose of this study was to examine the content of the justifications rather than make quantitative comparisons of the verdicts. The second experiment used quantitative methods and a larger sample to investigate whether the threshold and belief in guilt are affected by the instruction.

EXPERIMENT 2

The first study showed that many people in the instruction condition (6 of the 13) said that they felt the defendant may be guilty, but not beyond a reasonable doubt. No one in the control condition provided this justification. This suggests that the instruction may lower the reasonable doubt threshold. Further, ratings of belief in guilt were higher in the instruction condition. The difference was nonsignificant, but with a small sample, and therefore it is important to examine this effect with a larger sample.

Thus, we are looking for two possible effects of the reasonable doubt instruction. The first is lowering the threshold for rendering a guilty verdict away from absolute certainty. Second, the instruction invites the juror to contemplate a story in which the defendant is not guilty. Assuming that people create and compare different stories when deciding the guilt of a defendant (Pennington & Hastie, 1992), if the instruction strengthens the activation of the defense story, this may decrease the belief that the defendant is the culprit. If there were no other effects, this would decrease the likelihood of a guilty verdict. If both of these effects are in operation, then they may cancel each other out with respect to the actual verdict. However, their effects may still be apparent. The threshold hypothesis predicts that, controlling for participants' belief of guilt, those given the additional instructions should be more likely to give guilty verdicts.

Method

One hundred seventy-two undergraduates from the University of Bristol volunteered. None had taken part in Experiment 1. They read the same case description as in the first study. Participants had to make their verdict: guilty or not

guilty. Half ($n = 86$) of the people were told that they should render a guilty verdict only if they believed that the defendant was the culprit "beyond a reasonable doubt," without any further clarification. The instruction condition participants ($n = 86$) were also told this, but in addition had the reasonable doubt instruction used in the first study. Finally, all participants were asked to rate their belief that the defendant was in fact the culprit. This was done on a 0 to 100 probability scale.

Results

Prior to testing our predictions, descriptive statistics were performed. Forty-three percent of the people gave a guilty verdict (compared with 46% in the first study). The mean probability that the defendant was the culprit was 61% (compared with 67% in the first study). The distribution of belief scores was negatively skewed (skewness = -0.80 , $SE = 0.19$). This variable was squared, resulting in a more symmetrical distribution (skewness = -0.08 , $SE = 0.19$) with a mean of .4245 (the square root of .4245 is .65).³ The transformed variable will be used for statistical inference purposes, though we back-transform the squared variable for the graph.

The first question is: Does the additional instruction affect the verdicts? In the control condition, 37% said that the defendant was guilty, compared with 49% in the instruction condition. This is nonsignificant, $\chi^2(1) = 2.37$, $p = .12$, odds ratio = 1.61. From a psychological standpoint, the lack of significance suggests either that any effects are relatively small or that the threshold and change-in-belief effects may both occur but may cancel each other out with respect of the verdict.

Next, we looked at the squared probability of the belief that the defendant was the culprit across the two conditions. A t test showed, as predicted, that people in the instruction condition gave lower probabilities, $t(170) = 2.06$, $p = .04$ (two tailed). The effect size is approximately 0.30 of a standard deviation (between a small and medium effect size using Cohen's [1988] classification). When transformed back into the unsquared probabilities, the mean for the instruction group was 62%, and the mean for the control group was 68%.

Finally, we ran a series of logistic regressions to predict whether the participants would reach a guilty verdict. As predicted, inclusion of the squared probability variable had a significant, $\chi^2(1) = 81.43$, $p < .001$, and large effect ($R_L^2 = .35$).⁴ The main question was whether condition (instruction vs. control) was a significant predictor of rendering a guilty verdict after accounting for belief in guilt. It was. People in the instruction condition were more likely to make a guilty

³In general, transforming a variable, finding its mean, and then back-transforming the mean will not result in the original mean.

⁴ R_L^2 is the proportion of total χ^2 accounted for by the model. It can range from 0 to 1 and like R^2 from linear regression can be interpreted as a measure of fit of the model (Field, 2000).

verdict conditional on belief in guilt. This is statistically significant, $\chi^2(1) = 15.35$, $p < .001$, and represents a modest increase in the model's fit ($R_L^2 = .41$). Inclusion of the interaction between squared belief and group in the model improved the fit of the model, $\chi^2(1) = 5.13$, $p = .02$, though modestly ($R_L^2 = .43$). If logistic regressions are run separately for the two conditions the fitted models are:

$$\text{control: } \ln(p[\text{GV}_i]/(1-p[\text{GV}_i])) = -4.11 + 7.00 \text{ SQGB}_i,$$

$$\text{instruction: } \ln(p[\text{GV}_i]/(1-p[\text{GV}_i])) = -5.67 + 14.09 \text{ SQGB}_i,$$

where SQGB_i is the square of the belief score in percentages (so it can range from 0–1). Figure 1 shows these two models. The reasonable doubt threshold is shown by the dashed line. It is estimated by the negative of intercept of the model divided by the slope ($-\beta_0/\beta_1$), then square rooted (standard errors estimated from Venables and Ripley, 1999, p. 221). For the control group reasonable doubt is estimated at 63% (95% CI from 59% to 67%). For those given the instruction it is 77% (95% CI from 71% to 82%). So, if a person believes that there is about 75% likelihood that the defendant is the culprit, the prediction is that about half the time those in the control group will render a guilty verdict compared with about 90% of those in the instruction group. As can be seen from Figure 1, the two groups are basically indistinguishable until belief in guilt reaches about 50%.

A complementary way of interpreting interactions in logistic regressions is to focus on how steep each of the curves is. The steeper the curve, the stronger the relationship between the belief in guilt and the probability of rendering a guilty verdict. Using the language of item-response modeling, this is similar to a two-parameter item-response model where there is higher discriminability for the instruction group. The difference in discriminability is measured by the interaction. According to Figure 1, there is more agreement on the response function between belief in guilt and whether the verdict should be guilty or not guilty for people in the instruction group than for those in the control condition (see Saris, 1988, for discussion of response functions; see Wright, Gaskell, & O'Muirheartaigh, 1994, for examples).

Discussion

A simple and brief instruction, providing a definition of *reasonable doubt*, produced two effects. First, by inviting the participants to consider the possibility that the defendant may not be the culprit, participants in the instruction condition gave lower ratings for belief in guilt than did participants in the control group. Second, the instruction lowered the threshold of belief in guilt necessary for giving a guilty verdict. In this study, if someone given the additional instruction believed that there was about a 63% chance that the defendant was the culprit, they were as likely to give a guilty ver-

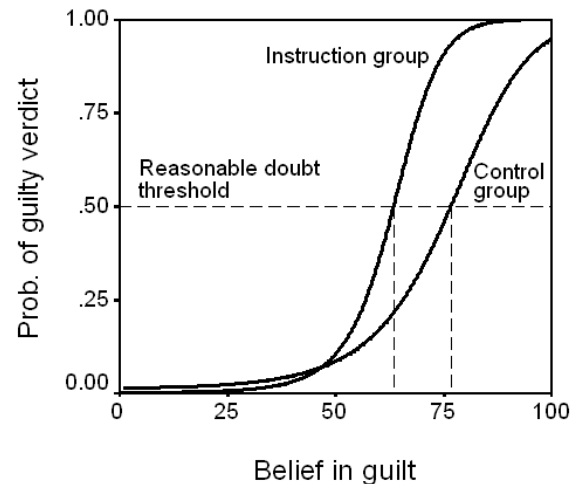


FIGURE 1 Comparing the group given the instruction with the control group, with the belief in guilt. The model includes main effects for condition and belief in guilt, as well as their interaction. The horizontal dashed line shows the reasonable doubt threshold. The vertical lines show the estimated level for reasonable doubt for the group given the instruction (63%) and the control group (77%).

dict as a not guilty verdict. Thus, for the instruction participants this was the level of reasonable doubt. For the control group, reasonable doubt appears to be approximately 77%.

That both of these effects appear to be present means that although the instruction lowered the threshold for a guilty verdict, because belief in guilt was also lowered, there was no significant effect on the number of guilty verdicts. However, as the wording of the instruction could be made to stress either of the two mechanisms, it is likely that some instructions would increase the probability of guilty verdicts, whereas others would decrease this probability.

The main statistical procedure for this situation is a logistic regression. A significant interaction was found; the instruction increased the discriminability with respect to belief in guilt. The instruction group was more consistent on rendering verdicts for different levels of belief of guilt than was the control group. One purpose of reasonable doubt instructions should be to clarify the meaning of a complex legal phrase, so that everyone has a similar definition. From this standpoint, the interaction suggests that some reasonable doubt instruction should be used. However, it is worth noting that several different models could have been fit to these data. The choice of these can affect how the curves are interpreted. Logistic regression is perhaps the most popular technique and has good statistical properties, but we encourage further research on the exact form of the response function between belief of guilt and the probability of a guilty verdict. In essence, this response function shows how people use the phrase *reasonable doubt*, so this response function should be at the heart of any quantitative investigation of the meaning of the phrase.

GENERAL DISCUSSION

The phrase *reasonable doubt* lies at the heart of the criminal justice system. There is much discussion about what it means. Given that experts disagree, it is likely that naïve jurors will be confused about this phrase. This linguistic variability is important. Pretend that you are a defendant in a criminal trial, like Tom Hoyle. Would you like to be tried by a jury who feels that being 60% sure of guilt is adequate for conviction, or one that has a higher threshold? Similarly, if you are a rape victim, like Nancy Von Roper, would you be interested in the threshold a jury was using? There are questions about the threshold that should be addressed by society, through our elected officials and appointed judges. Perhaps academics in moral philosophy and law can also provide information to help society make tough decisions about whether there is a single appropriate belief in guilt for convicting somebody and whether this belief should be expressed with a number. As psychologists, our role is showing how the reasonable doubt criterion is used by potential jurors, rather than how it should be used. Legal experts dispute whether judge's should define the phrase *reasonable doubt* to jurors and, if so, what the instruction should be. Here we created a very brief instruction based on several instructions that have been used in trials and explored the effects of this instruction on decision making.

Our studies have some limitations that are worth addressing. First, undergraduate students were used. This was done largely because of their availability. We encourage further research on samples more representative of the general population. However, having a relatively homogeneous sample decreases within-group variation and therefore increases the power for comparing between the control and instruction groups. Thus, for these studies this was an advantage. Second, we used only one instruction and only one case. In a sense, therefore, these are case studies (Wells & Windschitl, 1999; Wright, 1998). We are therefore cautious about generalizing to other instructions and other types of cases. In particular, the instruction we used produced two effects, a lowering of belief in guilt and a lowering of the reasonable doubt threshold. Other instructions may stress these aspects to different degrees. Instructions provided by judges are usually much longer than the instruction that we used. A short instruction was necessary to identify the causes of the effects we observed (Wright, in press), but if judges decide to use longer instructions it is necessary to examine the potential influences of all parts of these instructions.

These studies showed that an instruction can have important effects on juror decision making. In Experiment 1 we showed that many participants used the instruction to justify rendering a guilty verdict when they were not certain about the guilt of the defendant. The instruction allows participants the opportunity both to express uncertainty and to deliver a guilty verdict. In Experiment 2 we showed that this instruction affected people's belief in guilt and their reasonable

doubt thresholds. The instruction lowered people's belief in guilt and their threshold for reasonable doubt. Figure 1 shows a significant interaction. This can be interpreted as the participants in the instruction group having more similar definitions of reasonable doubt than did participants in the control condition. From a psychological perspective it makes sense that an instruction used to define a phrase should lower the variability in how the phrase is interpreted. However, in practice some definitions, particularly when phrased in legalese, can increase confusion (see *Victor v. Nebraska*, 1994). It is desirable to have jurors using similar criteria and therefore important to have little variation in definitions among individual jurors. Thus, on the basis of the significant interaction and desire to decrease variability, these data suggest that a brief and clear instruction should be used.

In summary, the instructions often given to jurors to help explain reasonable doubt are complex. Often their purpose is to explain to jurors that reasonable doubt does not mean absolute certainty. Many legal authorities (see *Victor v. Nebraska*, 1994) have noted that the phrasing is both outdated and includes legal jargon, both of which are likely to confuse the jurors. They criticize phrases such as *moral certainty* that are supposed to convey the appropriate meaning, but are more likely to be misinterpreted than to be helpful. It is critical that we understand how people use the phrase *reasonable doubt* and how people map a belief in guilt onto a verdict. In these studies we showed that a relatively simple and brief instruction affected people's belief in guilt, their reasonable doubt threshold, and the way in which they justify their verdicts.

REFERENCES

- Arkes, H. R., & Mellers, B. A. (2002). Do juries meet our expectations? *Law and Human Behavior*, 26, 625–639.
- Cage v. Louisiana, 111 S. Ct. 328 (1990).
- Clermont, K. M. (2004). Standards of proof in Japan and the United States. *Cornell International Law Journal*, 37, 263–284.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- DeKay, M. L. (1996). The difference between Blackstone like error ratios and probabilistic standards of proof. *Law and Society: Journal of the American Bar Foundation*, 21, 95–132.
- Devine, D. J., Clayton, L. D., Dunford, B. B., Seying, R., & Pryce, J. (2000). Jury decision making: 45 years of empirical research on deliberating groups. *Psychology, Public Policy, and Law*, 7, 622–727.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87, 215–257.
- Field, A. P. (2000). *Discovering statistics using SPSS for Windows*. London: Sage.
- Gaines v. Kelly (2000). US 2nd Circuit Court of Appeals. Docket No. 96–2761. Retrieved from <http://laws.findlaw.com/2nd/962761v2.html> (accessed March 14, 2007)
- Garry, M., & Polaschek, D. L. L. (2000). Imagination and memory. *Current Directions in Psychological Science*, 9, 6–10.
- Hastie, R. (1993). Algebraic models of juror decision processes. In R. Hastie (Ed.), *Inside the juror: The psychology of juror decision making* (pp. 81–115). New York: Cambridge University Press.

- 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, L. C. (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.
- Kenney, S. (1995). Fifth Amendment: Upholding the constitutional merit of misleading reasonable doubt jury instructions. *Journal of Criminal Law & Criminology*, 85, 989–1027.
- Kerr, N. L., Atkin, D., Stasser, R., Meek, D., Holt, R., & Davis, J. H. (1976). Guilt beyond a 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.
- Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin*, 110, 499–519.
- Kramer, G., & Koenig, D. (1990). Do jurors understand criminal jury instructions? Analyzing the results of the Michigan juror comprehension projects. *University of Michigan Journal of Law*, 23, 401–438.
- Loftus, E. F., & Ketcham, K. (1991). *Witness for the defense: The accused, the eyewitness, and the expert who put memory on trial*. New York: St. Martin's Press.
- Montgomery, J. W. (1998). The criminal standard of proof. *New Law Journal*, 148, 582.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259.
- Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the story model for juror decision making. *Journal of Personality and Social Psychology*, 62, 189–206.
- Saris, W. E. (1988). *Variation in response functions: A source of measurement error in attitude research*. Sociometric Research Foundation: Amsterdam.
- Shapiro, B. J. (1991). *Reasonable doubt and probable cause*. Berkeley, CA: University of Berkeley Press.
- Venables, W. N., & Ripley, B. D. (1999). *Modern applied statistics with S-Plus (3rd ed.)*. New York: Springer-Verlag.
- Victor v. Nebraska, 114 S. Ct. 1239 (1994).
- Wagenaar, W. A. (1996). Anchored narratives: A theory of judicial reasoning and its consequences. In G. M. Davies, S. Lloyd-Bostock, M. McMurrin, & J. C. Wilson (Eds.), *Psychology, law and criminal justice* (pp. 267–285). Berlin, Germany: de Gruyter.
- Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling in social psychological experimentation. *Personality and Social Psychology Bulletin*, 25, 1115–125.
- In re Winship*, 397 U.S. 358 (1970).
- Winters, R. J., & Greene, E. (2007). Jury decision making. In F. Durso, R. Nickerson, S. Dumais, S. Lewandovsky, & T. J. Perfect (Eds.), *Handbook of Applied Cognition (2nd ed.)*. (pp. 739–761) Chichester, UK: Wiley.
- Wright, D. B. (1998). People, materials, and situations. In J. A. Nunn (Ed.), *Laboratory psychology* (pp. 97–116). Hove, England: Lawrence Erlbaum Associates, Inc.
- Wright, D. B. (2006). Causal and associative hypotheses in psychology: Examples from eyewitness testimony research. *Psychology, Public Policy, and Law*, 12, 190–213.
- Wright, D. B., Gaskell, G. D., & O'Muirheartaigh, C. A. (1994). How much is "Quite a bit"? Mapping absolute values onto vague quantifiers. *Applied Cognitive Psychology*, 8, 479–496.
- Zander, M. (2000). The criminal standard of proof—how sure is sure? *New Law Journal*, 150, 1517.