Reasonable Doubt Survey

Jason Ralston Stanton Hudja

April 2024

1 Introduction

The standard of evidence known as "proof beyond a reasonable doubt" (hereafter, "reasonable doubt") is common to many legal systems. For example, the reasonable doubt standard is used in the United States, the United Kingdom, Canada, and Japan. In all of these countries, jurors under this standard are instructed to convict a defendant if and only if the defendant is guilty beyond a reasonable doubt. The prevalence of this standard, in conjunction with its position in the criminal justice system, makes the reasonable doubt standard important to understand. In this survey, we highlight some of the experimental research on reasonable doubt.

Researchers have focused on a few questions in regard to reasonable doubt. One of the two questions that we will focus on in this survey is how do jurors respond to the reasonable doubt standard. This question is inspired by the vague definitions that are provided to jurors by courts. Generally, these definitions tend to describe the reasonable doubt standard as something that would be clear to a reasonable person. In addition to the vague definitions that are provided to jurors, there is no standard accepted definition of reasonable doubt. This allows for vast heterogeneity in how states define reasonable doubt. Some states appear to focus more on reducing Type I error (i.e., convicting an innocent person), while other states appear to focus more on reducing Type II error (i.e., acquitting a guilty person). This heterogeneity makes it even more imperative that researchers uncover how jurors respond to the reasonable doubt standard.

The second question that we will focus on in this survey is whether jurors' behavior under the reasonable doubt standard differs from their preferences towards trial outcomes. While it is important to understand how jurors behave under the reasonable doubt standard, it is also important to understand whether their preferences over trial outcomes are inconsistent with their behavior. If these preferences differ from their behavior, then it means that jurors are being constrained by the reasonable doubt standard. Additionally, it allows for the possibility that jurors' behavior may differ from their interpretation of reasonable doubt. For example, jurors may deviate from their interpretation of reasonable doubt due to their preferences for Type I and Type II error.

2 Experiments

Researchers have used experimental methods to address many important questions regarding reasonable doubt. In this section, we will address the experimental literature on two of these questions. First, we will address the question of how jurors behave under the reasonable doubt standard. Second, we will address whether individuals' preferences towards trial outcomes are inconsistent with their behavior under reasonable doubt. In the following subsections, we examine evidence from several experimental studies to summarize where the literature stands on these issues.

2.1 Behavior Under the Standard

Legal scholars have long been interested in understanding how jurors make decisions under the reasonable doubt standard. Theoretically, under the reasonable doubt standard, researchers have considered jurors as having a threshold probability that they use to determine whether to convict a defendant. Jurors are modeled as using the evidence presented at trial to formulate a subjective probability that the defendant is guilty. Jurors vote to convict if this subjective probability meets or exceeds the reasonable doubt threshold; otherwise, they acquit. This modeling of juror behavior has inspired researchers to try to uncover individuals' reasonable doubt threshold and whether this threshold can be influenced.

Early experiments utilized mock juries to determine whether their decisions would differ under different juror instructions (Davis et al., 1975; Kerr and Atkin, 1976; Kagehiro, 1990; Horowitz and Kirkpatrick, 1996; Koch and Devine, 1999). In these experiments, participants read a hypothetical case and were later asked to render judgements as individuals and as groups. Subjects are divided into different groups and each group is given a different prompt for reasonable doubt (or some other standard) before they made their verdict. One early paper, Kerr and Atkin (1976), found that different definitions of reasonable doubt changed both individual verdicts and mock jury verdicts. They found that instructions that used language related to reducing Type I error produced more "not guilty" verdicts than instructions that highlighted careful deliberation and instructions that left reasonable doubt undefined. While Kerr and Atkin (1976) found differences based on reasonable doubt definitions, Kagehiro and Stanton (1985) found no difference in conviction rates between three different standards of proof: preponderance of evidence, "clear and convincing evidence", and reasonable doubt. A few papers by Michael Cicchini and Lawrence White (Cicchini and White, 2016; Cicchini and White, 2017; White and Cicchini, 2019) studied whether the definition of reasonable doubt or different standards of proof affected juror behavior on MTurk. These papers found that while definitions changed behavior, different standards of proof did not. Additionally, a few studies (Kagehiro and Stanton, 1985; Horowitz and Kilpatrick, 1996) defined reasonable doubt as a percentage of guilt and showed that presenting participants with definitions with higher percentages decreased conviction rates.

While some studies tried to uncover whether juror decisions depend on juror instructions, other studies directly solicited jurors' thresholds under reasonable doubt. In two studies by Mandeep Dhami (Dhami, 2008; Dhami et al., 2015), participants acted as mock jurors and read a trial summary. Participants are told that reasonable doubt "[...] 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". They were informed to only convict if "[...] the prosecution has proved guilt beyond a reasonable doubt". However, along with rendering a verdict, participants were asked a question regarding the maximum probability that the defendant is innocent that would make the participant hesitant to convict.[1] These two papers found evidence that reasonable doubt thresholds are not significantly affected by different definitions of reasonable doubt. Across the two studies, they estimated that jurors' thresholds under reasonable doubt were between 85% and 89%.

While these previous studies had purely hypothetical decisions, there is recent research that tries to explore the effects of reasonable doubt using a *consequential experiment*. Throughout the rest of the survey, we will use the term consequential experiment to refer to an experiment where a juror's decisions affect another participant.[2] For example, Aimone et al. (2023) explore the effects of different reasonable doubt definitions on juror behavior in a consequential experiment. The experiment consisted of two stages. In the first stage, two participants were paired and each played a real effort task to earn \$50. One of the two paired subjects was given the opportunity to take a fixed amount of the other subject's earned \$50. In the second stage, different participants were asked to act as individual jurors and were informed on the details of the first stage. It was explained to jurors that the decisions that they made could affect a defendant from the first stage. This is a large departure from previous studies as jurors now understood that their actions had consequences (and were not just hypothetical decisions). In this experiment, a defendant from the first stage being found guilty would cost that defendant \$40.

In Aimone et al. (2023), jurors explicitly chose their own reasonable doubt threshold. It was explained to jurors that the defendant under consideration would have an undisclosed probability (Y) of being guilty, which was to be uniformly drawn from the integers between 0 and 100 (inclusive). Jurors had to specify their guilt threshold such that a realization of Y below (at or above) the threshold would result in a judgement of "not guilty" ("guilty"). Before jurors made their choice of guilt threshold, they were presented with a real definition of reasonable doubt from a state in the United States. The definition that was provided varied between subjects. These different definitions were

[1] This question comes from Dhami (2008). Dhami (2015) uses the question "How much proof would make you hesitate to find someone guilty of a crime?".

[2] Consequential experiments are not incentivized in the way that we generally incentivize decisions in experimental economics. However, these studies provide jurors with an incentive to thoroughly consider their decisions as their decisions affect another participant. This difference in incentivization stems from the fact that jurors do not face monetary consequences from their decisions, but their decisions have consequences for someone else.

selected to vary how much they appeared to emphasize reducing Type I error or Type II error. The study found no significant differences in subjects' guilt thresholds across the different definitions. Additionally, the average guilt threshold was around 72%.

Overall, the papers that focus on jurors' reasonable doubt thresholds provide somewhat conflicting takeaways. Early studies on whether reasonable doubt definitions affect behavior differ in their conclusion. Additionally, studies that elicit reasonable doubt thresholds tend to result in different magnitudes. Our view from surveying the literature is that reasonable doubt definitions may matter, but this effect is likely to be small. Additionally, we view the research that elicits reasonable doubt thresholds as inconclusive as these elicitations may be sensitive to the method used. As we will mention later, more consequential experiments can help us better address the questions from this subsection.

2.2 Preferences and Other Considerations

While many studies elicit an individual's reasonable doubt threshold, it is unclear if this threshold is solely driven by a juror's interpretation of reasonable doubt. In the studies discussed in the previous subsection, we summarized research that tried to uncover how reasonable doubt definitions might influence a juror's decision to vote to convict. In this subsection, we instead summarize research that explores whether jurors' preferences towards trial outcomes are inconsistent with their behavior under the reasonable doubt standard.

One factor that might influence jurors' decisions is how they feel about various trial outcomes. For example, jurors may be concerned about an outcome that falsely convicts an innocent defendant (i.e., Type I error) or falsely acquits a guilty defendant (i.e., Type II error). Individuals' attitudes towards various trial outcomes may lead them to deviating from their interpretation of reasonable doubt. This could arise if jurors try to minimize the probability of an outcome that they find unsatisfactory (instead of behaving consistently with their interpretation of reasonable doubt).

There is early research that focused on the possibility that jurors' attitudes towards trial outcomes could influence their behavior at trial. This experimental work is based on preference-based models of juror-decision making, which were initially suggested by Kaplan (1968) and Connolly (1987). Some early experimental studies that tested the validity of the preference-based model approach were Nagel et al. (1981) and Dane (1985). These studies used hypothetical trials and asked subjects to assign values to each of the four possible outcomes of a trial: (i) acquitting an innocent person, (ii) acquitting a guilty person (i.e., Type II error), (iii) convicting an innocent person (i.e., Type I error), and (iv) convicting a guilty person. These values were then used to calculate a guilt threshold using a calculation similar to Kaplan (1968) (and Feddersen and Pesendorfer, 1998). In a review of several preference-based experiments, Hastie (1994) found that employing a preference based calculation resulted in juror guilt thresholds between 50% and 55%. These preference-based thresholds are notably lower than those where participants have their thresholds elicited directly. This suggests that there may be a disconnect between

how jurors behave under the reasonable doubt standard and how individuals prefer to make decisions involving guilt.

Similar to the previous subsection, there is recent research that tries to uncover how individuals treat Type I and Type II error in a consequential experiment. Hudja et al. (2024) use an experiment where participants trade-off Type I and Type II error. In this study, participants are placed in decision environments where Type I and Type II error initially occur with certainty. This means there is a 100% chance that an innocent person will be convicted and a 100% chance that a guilty person will be acquitted. Participants are given an endowment of points where each point can either reduce Type I or Type II error by one percentage point. Subjects are placed into a hypothetical and a consequential environment. In the consequential environment, participants make decisions that affect an actual subject from a different study. A common finding in both decision environments is that participants consistently choose to reduce Type I error more than Type II error. However, there are also a decent number of subjects who appear to prefer reducing Type II error or to prefer balancing these errors when they can. These results further suggest that there may be a disconnect between participants' preferences towards trial outcomes and their behavior under the reasonable doubt standard.

3 Future Work

While there is a lot of experimental research on reasonable doubt, there is still more that needs to be done. One important area for future research is to conduct more consequential experiments. It is crucial to understand whether the results obtained under consequential experiments differ from the results obtained from hypothetical decisions. Until recently, almost all reasonable doubt studies have utilized hypothetical decisions. This is not surprising as it is difficult to incentivize juror decisions in an experiment using the standard experimental methodology. Best experimental practices strongly recommend paying subjects for their decisions, but in the real world there are no direct financial consequences for juror decisions. However, recent research has provided a path forward by having juror decisions affect another participant; this type of design provides incentives for a juror without having to pay them for their decisions. It is important for future research to analyze questions of reasonable doubt with consequential experiments as participants can often act carelessly without incentives (Voslinsky and Azar, 2021).

While conducting consequential experiments is a first step, it is also important for future work to add more components of the jury decision-making process into the design. It is important to remember that convictions in criminal cases depend on the decision of a jury. Thus, it is crucial that future work uses consequential experiments to understand how juries determine a verdict under reasonable doubt. Some possible areas of research are determining how jurors deliberate, uncovering the average reasonable doubt threshold of a jury, and understanding how different voting aggregation rules (unanimous vote, majority vote, etc.) affect conviction rates.

One last area for future work is determining how preferences towards trial outcomes

influence reasonable doubt thresholds. Preference-based studies have estimated thresholds that are very different from the reasonable doubt thresholds that we see elicited from subjects. This suggests that subjects' reasonable doubt thresholds are generally different from what we would expect given subjects' preferences for trial outcomes. While this suggests that individuals behave differently in a trial than they would if they were unconstrained, it is still unclear if subjects' preferences towards trial outcomes play a role in their behavior at trial. It is thus important for researchers to determine if reasonable doubt thresholds are solely driven by jurors' interpretation of reasonable doubt or if preferences play a role.

4 References

Aimone, J., Hudja, S., Law, W., North, C., Ralston, J. and Rentschler, L., 2023. An experimental exploration of reasonable doubt. *Journal of Economic Behavior & Organization*, 212, pp. 873-886.

Cicchini, M. and White, L., 2016. Truth or doubt? An empirical test of criminal jury instructions. *University of Richmond Law Review*, 50(4), pp. 1139-1167.

Cicchini, M. and White, L., 2017. Testing the impact of criminal jury instructions on verdicts: A conceptual replication. *Columbia Law Review Online*, 117, pp. 22-35.

Connolly, T., 1987. Decision theory, reasonable doubt, and utility of erroneous acquittals. *Law and Human Behavior*, 11, pp. 101-112.

Dane, F., 1985. In search of reasonable doubt. *Law and Human Behavior*, 9, pp. 141-158.

Davis, J., Kerr, N., Atkin, R., Holt, R. and Meek, D., 1975. The decision processes of 6- and 12-person mock juries assigned unanimous and two-thirds majority rules. *Journal of Personality and Social Psychology*, 32(1), pp. 1-14.

Dhami, M., 2008. On measuring interpretations of reasonable doubt. *Journal of Experimental Psychology: Applied*, 14(4), pp. 353-363.

Dhami, M., Lundrigan, S. and Mueller-Johnson, K., 2015. Instructions on reasonable doubt: Defining the standard of proof and the juror's task. *Psychology, Public Policy, and Law*, 21(2), pp. 169-178.

Feddersen, T. and Pesendorfer, W., 1998. Convicting the innocent: The inferiority of unanimous jury verdicts under stategic voting. *American Political Science Review*, 92(1), pp. 23-35.

Hastie, R., 1994. Inside the juror: The Psychology of Juror Decision making. s.l. Cambridge University Press.

Horowitz, I. and Kirkpatrick, L., 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(6), pp. 655-670.

Hudja, S., Aimone, J., North, C., Ralston, J., and Rentschler, L., 2024. Prosecutorial discretion: a theoretical and experimental investigation. *Unpublished Manuscript.*

Hudja, S., Ralston, J., Wang, S., Aimone, J., Rentshler, L., and North, C., 2024. Attitudes Towards Type 1 and Type 2 Error. *Unpublished Manuscript.*

Kagehiro, D., 1990. Defining standards of proof in jury instructions. *Psychological Science*, 1(3), pp. 194-200.

Kagehiro, D. and Stanton, W., 1985. Legal vs. quantified definitions of standards of proof. *Law and Human Behavior*, 9(2), pp. 159-178.

Kaplan, J., 1968. Decision theory and the factfinding process. *Stanford Law Review*, 20, pp. 1065–1092.

Kerr, N. and Atkin, S., 1976. Guilt beyond a reasonable doubt: Effect of concept definition and assigned decision rule on the judgements of mock jurors. *Journal of Personality and Social Psychology*, 34(2), pp. 282-294.

Koch, C. and Devine, D., 1999. Effects of reasonable doubt definition and inclusion of a lesser charge on jury verdicts. *Law and Human Behavior*, 23(6), pp. 653-674.

Nagel, S., Lamm, D., and Neef, M., 1981. 'Decision Theory and Juror Decision-Making' in Sales, B. (ed.) *The Trial Process: Perspectives in law and psychology.* New York: Plenum Press, pp. 353-386.

Vaslinsky, A. and Azar, O., 2021. Incentives in experimental economics. *Journal of Behavioral and Experimental Economics*, 93:101706.

White, L. and Cicchini, M., 2019. Is Reasonable Doubt Self-Defining? *Villanova Law Review*, 64(1), pp. 1-24.