IS REASONABLE DOUBT SELF-DEFINING?

Lawrence T. White*
Michael D. Cicchini**

Many courts believe that reasonable doubt is self-defining and, therefore, do not explain the concept to their juries. The empirical evidence, however, suggests otherwise. Controlled studies demonstrate that mock jurors do not distinguish between reasonable doubt, clear and convincing evidence, or even preponderance of evidence standards when reaching their verdicts.

This Article presents our empirical study in which we sought to (1) conduct a more powerful test by remedying the methodological weaknesses of earlier studies, and (2) determine whether, instead of following their burden of proof instruction, mock jurors use a simple heuristic or rule of thumb regarding the quantum of evidence necessary to convict.

Our first finding is consistent with previous findings: we found no significant differences in conviction rates between groups that received different burden of proof instructions. Second, the data also revealed what we call "the 60/65 rule." That is, nearly all study participants either (a) said that less than 60% of the evidence favored the State and voted not guilty, or (b) said that more than 65% of the evidence favored the State and voted guilty.

These findings demonstrate that reasonable doubt is not self-defining. Not only do mock jurors in multiple studies fail to distinguish between reasonable doubt and the two lower, civil burdens of proof, but they are also willing to convict criminal defendants on a quantum of evidence (approximately 65%) that is much lower than what judges expect and the Constitution requires.

Given these findings, we recommend that courts use a relative, comparison-based definition of reasonable doubt to properly convey to jurors the high burden the government must satisfy before depriving a person of life, liberty, or property.

---

* Ph.D., University of California, Santa Cruz (1984); M.A., *with distinction*, California State University at Fresno (1979); B.A., *with honors*, Whittier College (1975). Dr. White is Professor of Psychology at Beloit College in Wisconsin. He also directs the College's Law & Justice Program. Both authors thank Peter White for his technical assistance with the Mechanical Turk online research platform and Mallory Gribble for her assistance with the pilot study that preceded this study and article.

** J.D., *summa cum laude*, Marquette University Law School (1999); C.P.A., University of Illinois Board of Examiners (1997); M.B.A., Marquette University Graduate School (1994); B.S., University of Wisconsin—Parkside (1990). Michael Cicchini is a criminal defense lawyer in Kenosha, Wisconsin. He is also the author or coauthor of four books and twenty law review articles on criminal law and procedure.

TABLE OF CONTENTS

INTRODUCTION

The Constitution protects a defendant from criminal conviction unless the government can prove guilt beyond a reasonable doubt.[1] When instructing juries on this burden of proof, many courts subscribe to one of two philosophies. Some courts go to great lengths to explain the concept of reasonable doubt to the jury.[2] At the other end of the spectrum, many courts believe that reasonable doubt is already self-defining, and therefore do little, if anything, to further explain the concept.[3]

But is the concept of reasonable doubt truly *self-defining*? Or do juries require a definition to fully understand and appreciate this high burden of proof? The existing research demonstrates that jurors fail to distinguish between it and the two lower, civil burdens of proof. That is, in controlled studies, the different standards of proof do not produce different verdict patterns.[4] Further, other studies demonstrate that jurors interpret reasonable doubt to require a relatively low quantum of evidence in the government's favor—somewhere between 63% and 68%—in order to convict.[5]

Earlier studies that tested the impact of different burdens of proof on jury decision-making were limited by small and unrepresentative samples.[6] We therefore decided to improve upon these studies with our own empirical test.[7] Further, if juror decision-making is not influenced by different burdens of proof (as previous studies have demonstrated), we sought to determine if jurors instead use a simple heuristic, or rule of thumb, regarding the strength of the government's evidence that is needed to convict.[8]

[1] Part I.
[2] Id.
[3] Id.
[4] Part II.
[5] Id.
[6] Id.
[7] Part III.A.
[8] Id.

To test our hypotheses, we recruited 495 jury-eligible adults in 45 states and randomly assigned them to read one of four case summaries.[9] These included a battery case with strong evidence of guilt, a battery case with weak evidence of guilt, a trespassing case with strong evidence of guilt, and a trespassing case with weak evidence of guilt.[10] Participants were then randomly assigned to one of three groups, each of which received a different burden of proof instruction: preponderance of evidence, clear and convincing evidence, or proof beyond a reasonable doubt.[11]

As in earlier studies, we found that mock jurors' verdicts were not influenced by the different burden of proof instructions, and this held true for all four case summaries.[12] But jurors did not vote haphazardly. Rather, they were highly sensitive to evidence strength when rendering a verdict. In fact, the vast majority of jurors followed a simple heuristic that we call the 60/65 rule: they either (1) said that *less* than 60% of the evidence favored the state and voted *not* guilty, or (2) said that *more* than 65% of the evidence favored the state and voted *guilty*.[13]

Based on our findings and those of earlier researchers, there is now strong empirical support for a conclusion that reasonable doubt is *not* self-defining—i.e., jurors fail to distinguish between it and the lower burdens of proof and instead will convict defendants on a relatively small quantum of evidence.[14] Therefore, instructions should carefully define reasonable doubt for jurors, and we recommend doing so by using a comparative framework.[15]

Specifically, by comparing and contrasting proof beyond a reasonable doubt with the two lower burdens of proof, jury instructions can provide the necessary context for jurors to understand the high burden the government must satisfy before it may convict a defendant of a crime.[16] We also suggest further research to empirically test our recommended approach to defining reasonable doubt.[17]

I. REASONABLE DOUBT: TO DEFINE OR NOT TO DEFINE?

Before a jury may convict a defendant of a crime, the Constitution requires the government to prove its case beyond a reasonable doubt.[18] Yet, despite this

[9] Part III.B.
[10] Part III.C.
[11] Id.
[12] Part III.D.
[13] Id.
[14] Part IV.
[15] Part V.
[16] Id.
[17] Part VI.
[18] In re Winship, 397 U.S. 358, 364 (1970) ("the Due Process Clause protects the accused against conviction except upon proof beyond a reasonable doubt."). Although this standard was not explicitly adopted until 1970, the Court implicitly recognized it much earlier. *See, e.g.*, Davis v. United States, 160 U.S. 469, 488 (1895) ("How, then, upon principle, or consistently with humanity, can a verdict of guilty be properly returned if the jury entertain a reasonable doubt as to the existence of a fact which is essential to guilt—namely, the capacity in law of the accused to commit that crime?"); Miles v. United States, 103 U.S. 304, 312 (1881) ("The evidence upon which a jury is justified in returning a verdict of guilty must be sufficient to produce a conviction of guilt, to the exclusion of all reasonable doubt.").

(theoretically) uniform standard across states and federal circuits, there are nearly as many jury instructions on reasonable doubt as there are jurisdictions.[19]

When instructing their juries, courts often subscribe to one of two divergent philosophies. A New Hampshire court described one philosophical approach toward reasonable-doubt instructions as follows: "[T]his court feels strongly that a jury must be given some assistance in understanding the concept. . . . [T]he definition of reasonable doubt is perhaps the most important aspect of the closing instruction to a jury in a criminal trial."[20] Courts in this camp often employ lengthy instructions on the burden of proof and presumption of innocence. For example, Wisconsin uses a rambling, 284-word jury instruction;[21] Massachusetts' instruction tallies 285 words;[22] and Alaska's weighs in at an effusive 329 words.[23]

At the other extreme, the Seventh Circuit Federal Court of Appeals believes that the term reasonable doubt "is self-defining, that there is no equivalent phrase more easily understood . . . that the better practice is not to attempt the definition, and that any effort at further elucidation tends to misleading refinements."[24] Similarly, the First Circuit Federal Court of Appeals has warned that defining reasonable doubt "is unnecessary, could confuse the jury, and provides fertile grounds for objections."[25] This approach tends to create shorter instructions. Illinois, for example, tells jurors that "[t]he State has the burden of proving the guilt of the defendant beyond a reasonable doubt" and offers no elaboration of the term; its entire instruction on the presumption of innocence and burden of proof is only 88 words.[26]

It is true, as some courts have warned, that attempts to elucidate "reasonable doubt" often add nothing of value, which is the functional equivalent of not defining it at all. And other attempts have created confusion—or even worse. For example, after discussing reasonable doubt, Wisconsin's pattern instruction admonishes jurors "not to search for doubt" but instead "to search for the truth."[27] The impact of this curious closing mandate has twice been empirically tested. Not only did the language create confusion,[28] it actually lowered the burden of proof, increased conviction rates,[29] and was the functional equivalent of giving no reasonable doubt instruction whatsoever.[30]

---

[19] *See* Victor v. Nebraska, 511 U.S. 1 (1994) (giving courts tremendous leeway when instructing jurors on the government's burden of proof). For examples of the various definitions of reasonable doubt, see Miller W. Shealy, Jr., *A Reasonable Doubt About "Reasonable Doubt"*, 65 OKLA. L. REV. 225 (2013); Hon. Richard E. Welch III, *"Give Me That Old Time Religion": The Persistence of the Webster Reasonable Doubt Instruction and the Need to Abandon It*, 48 NEW ENGLAND L. REV. 31 (2013).

[20] State v. Aubert, 421 A.2d 124, 127 (N.H. 1980).

[21] WIS. CRIM. JURY INSTRUCTIONS NO. 140 (2017).

[22] MASS. CRIM. JURY INSTRUCTIONS NO. 2.180 (2015).

[23] ALASKA CRIM. JURY INSTRUCTIONS NO. 1.06 (2012).

[24] United States v. Lawson, 507 F.2d 433, 443 (7th Cir. 1974).

[25] United States v. v. Vavlitis, 9 F.3d 206, 212 (1st Cir. 1993); *see also* United States v. Reives, 15 F.3d 42, 46 (4th Cir. 1994) (discussing the policy against attempting to define reasonable doubt).

[26] ILL. CRIM. JURY INSTRUCTIONS NO. 2.03 (2017). This instruction is based on a long line of Illinois cases holding that "neither the trial court nor counsel should define reasonable doubt for the jury." People v. Downs, 69 N.E.3d 784, 788 (2015) (citing several cases). Several other states, including Texas, also leave the term undefined. *See* Timothy J. Ting, *It's Time to Define "Beyond a Reasonable Doubt"*, 106 ILL. BAR J. 24 (2018).

[27] WIS. CRIM. JURY INSTRUCTIONS NO. 140 (2017).

[28] Mock jurors who received this closing mandate were nearly twice as likely to indicate that their instruction allowed them to convict the defendant even if they had a reasonable doubt about guilt. Michael

But is it true that reasonable doubt is *self-defining* and therefore requires little or no explanation? That is, do juries intuitively understand the meaning of the term without a legally correct definition from the trial judge? If the self-defining hypothesis is correct, then a short "beyond a reasonable doubt" (BRD) instruction—one with minimal or no elaboration—would give defendants more protection than the lower, civil burdens of proof, i.e., preponderance of the evidence (POE) and clear and convincing evidence (CCE). Fortunately, this hypothesis is empirically testable.[31]

II. EARLIER WORK ON STANDARDS OF PROOF

We know of three published studies that tested the effect of different burdens of proof on jurors' verdicts. We discuss the studies here as their methodologies and findings influenced our own study design and conclusions.

In 1973, two British researchers recruited 833 London residents to serve as mock jurors.[32] Participants heard one of two cases (theft or rape) and received one of three instructions: POE, CCE (described as "sure and certain"), or BRD.[33] In the theft case, the different standards did not produce different verdict patterns.[34] In the rape case, the standards again did not operate properly: participants who received a BRD instruction convicted the defendant at a higher rate (32%) than those who received a less stringent CCE instruction (18%).[35]

In 1985, two American researchers conducted two experiments.[36] In the first, 198 undergraduates read a summary of a civil trial in which the plaintiff sued an insurance

---

D. Cicchini & Lawrence T. White, *Testing the Impact of Criminal Jury Instructions on Verdicts: A Conceptual Replication*, 117 COLUM. L. REV. ONLINE 22, 32 (2017).

[29] Michael D. Cicchini & Lawrence T. White, *Truth or Doubt? An Empirical Test of Criminal Jury Instructions*, 50 U. Rich. L. Rev. 1139, 1155 (2016) (finding a statistically significant difference in conviction rates between relevant test groups).

[30] Id. at 1157 (jurors who were given a reasonable doubt instruction and then told "not to search for doubt" but "to search for the truth" convicted at the same rate as those who received no reasonable doubt instruction whatsoever).

[31] There is also anecdotal evidence that jurors do not understand the concept of reasonable doubt, as they often ask their trial judges for clarification or even conduct mid-deliberation internet searches on their smart phones. *See* Bobby Greene, *Reasonable Doubt: Is It Defined by Whatever is at the Top of the Google Search Page?*, 50 J. MARSHALL L. REV. 933, 942-43, 950-53 (2017). As Greene discusses, inquiries from the jury can put the trial judge in a quandary. For example, one trial judge was found to have erred by responding, "It is for the jury to collectively determine what reasonable doubt is." Id. at 942 (quoting *People v. Turman*). Inexplicably, another trial judge in the same state was found to have responded properly by telling the jury, "It is for you to determine." Id. (quoting *People v. Thomas*). *See also* People v. Downs, 69 N.E.3d 784, 789 (2015) (discussing the trial judge's response to jury's request for a definition).

[32] W. R. Cornish & A. P. Sealy, *Juries and the Rules of Evidence*, 1973 CRIM. L. REV. 208, 210 (1973).

[33] Id. at 213-14. The researchers used a very short BRD instruction, offering minimal commentary that a reasonable doubt "is not a fanciful doubt . . ." Id. at 213. They also varied the language of the BRD instructions slightly depending on case type. The instructions totaled either 32 or 35 words, excluding the portion of the instruction on the presumption of innocence. Id.

[34] Id. at 216.

[35] Id. at 217.

[36] Dorothy K. Kagehiro & W. Clark Stanton, *Legal v. Quantified Definitions of Standards of Proof*, 9 L. & HUM. BEHAV. 159 (1985).

company.[37] Participants received one of three instructions—POE, CCE, or BRD—in one of two versions: a standard legal definition[38] or a quantified definition where POE = 51% certainty of guilt, CCE = 71% certainty, and BRD = 91% certainty.[39] Participants rendered verdicts without deliberations.[40] When the standards were defined in words, i.e., standard legal definitions, the instructions had no discernible impact on verdicts.[41] The quantified versions, however, produced a legally-proper pattern of verdicts: the proportion of verdicts favoring the plaintiff decreased as the burden of proof became higher.[42]

In the second experiment, a replication of the first, 220 undergraduates read a summary of a civil trial and received one of three instructions—POE, CCE, or BRD—in one of three versions: a legal definition, a quantified definition, or a combined (legal and quantified) definition.[43] The quantified definitions again produced a legally-proper pattern of verdicts, but the legal definitions and combined definitions did not.[44]

In 1991, an American legal psychologist investigated standards of proof within the context of legal definitions of insanity by conducting two experiments.[45] In the first experiment, 151 undergraduates watched a videotaped reenactment of a trial in which the defendant was charged with killing his daughter and three of her friends.[46] Participants were instructed to apply one of two insanity standards and received one of three burden of proof instructions: POE, CCE, or BRD.[47] The different insanity standards did not affect participants' decisions and, more importantly for our purposes, the different standards of proof had no impact on verdicts.[48]

In the second experiment, a replication of the first, 226 undergraduates watched the videotaped reenactment used in the first study and were assigned to conditions that varied the insanity standard and the burden of proof.[49] The results of this second experiment were identical to the results of the first: different insanity standards and different standards of proof did not produce different verdict patterns.[50]

---

[37] Id. at 163 ("A total of 252 students participated in Experiment 1, but results from 54 . . . were omitted from data analyses" leaving a sample of 198 participants.).

[38] The researchers used a short BRD instruction but offered a brief explanation of the concept. Specifically, they warned that reasonable doubt is "not a mere possible doubt," but rather is one that prevents the jurors from feeling "an abiding conviction, to a moral certainty, of the truth of the plaintiff's case." The BRD instruction totaled 79 words. Id.

[39] Id.

[40] Id. at 164.

[41] Id. ("For the legal definitions, the multivariate effect of standard of proof was not significant . . . indicating that the legal definitions of the standards of proof had no effect on the dependent variables").

[42] Id.

[43] Id. at 168.

[44] Id.

[45] James R. P. Ogloff, *A Comparison of Insanity Defense Standards on Juror Decision Making*, 15 L. & HUM. BEHAV. 509 (1991).

[46] Id. at 514.

[47] Id. at 515. The substance of these burden of proof instructions is described as being "standard jury instructions a judge would give in a similar case." Id.

[48] Id. at 516 ("altering the burden and standard of proof do not seem to make a difference in mock jurors' decisions . . .").

[49] Id. at 518.

[50] Id. at 519 ("no significant results were obtained for burden . . . or standard of proof).

To summarize, five experiments in three published studies found little or no evidence that different standards of proof produce different verdict patterns as intended by the courts (and as believed by those judges who contend that reasonable doubt is self-defining and therefore needs little or no explanation). Further, when differences in verdict patterns were observed, they were too small to be statistically significant or did not order themselves properly in terms of how much protection they afforded defendants.[51]

These studies strongly suggest that, instead of interpreting reasonable doubt as the highest burden of proof, jurors apply a lay standard *that differs substantially* from what many courts expect and the Constitution requires. Specifically, when judges are asked to quantify the conviction threshold for reasonable doubt, most set the threshold at 85% or higher.[52] In one study, for example, federal judges throughout the U.S. were surveyed. Of the 171 respondents, 126 (74%) set the threshold at "90% or higher."[53]

Several studies, however, indicate that *jurors* are satisfied with a much lower amount of evidentiary proof in order to convict under the BRD standard. For example, in 1996, researchers divided 480 jury-eligible adults into 80 six-person juries.[54] Each jury observed a reenactment of a murder trial and then received one of five definitions of BRD, all of which were legally permissible.[55] When individual jurors were asked to quantify the BRD standard, they set a mean (average) criterion that ranged from 54% to 70%, depending on the BRD definition they received.[56] None of the instructions caused jurors to do what most courts expect or want them to do: "set the certainty of guilt in the high 80s."[57]

Similarly, in 2007, researchers found that a simple definition of BRD—one merely indicating that proof BRD does not require *certainty* in order to convict—led undergraduate test participants to set the cutoff between a guilty and not-guilty verdict at a 63% chance that the defendant was guilty.[58] In other words, "if someone . . . believed that there was about a 63% chance that the defendant was the culprit, they were as likely to give a guilty verdict as a not guilty verdict."[59] And in 2014, researchers found that

[51] As explained during our discussion of the three studies, *quantified* definitions of the standards of proof did produce a correct verdict pattern, i.e., BRD instructions generated fewer convictions than the civil burdens of proof. However, courts reject the use of quantified definitions. *See* Elisabeth Stoffelmayr & Shari S. Diamond, *The Conflict Between Precision and Flexibility in Explaining "Beyond a Reasonable Doubt"*, 6 PSYCHOL., PUB. POL. & L. 769 (2000).

[52] Lawrence M. Solan, *Refocusing the Burden of Proof in Criminal Cases: Some Doubt About Reasonable Doubt*, 78 TEX. L. REV. 105, 126 (1999) (discussing a poll of federal judges in New York).

[53] Id. (discussing C. M. A. McCauliff, *Burdens of Proof: Degrees of Belief, Quanta of Evidence, or Constitutional Guarantees?*, 35 VAND. L. REV. 1293, 1325 (1982)).

[54] Irwin A. Horowitz & Laird C. Kirkpatrick, *A Concept in Search of a Definition: The Effects of Reasonable Doubt Instructions on Certainty of Guilt Standards and Jury Verdicts*, 20 L. & HUM. BEHAV. 655 (1996).

[55] Id. at 660-61.

[56] Id. at 666.

[57] Id. at 667.

[58] Daniel B. Wright & Melanie Hall, *How a "Reasonable Doubt" Instruction Affects Decisions of Guilt*, 29 BASIC & APPLIED SOCIAL PSYCHOL. 91, 96 (2007).

[59] Id. The control group, or those participants who received no supplemental, descriptive language but were only told that the burden of proof was BRD, set the cutoff at a higher level of 77%. Id.

laypersons, on average, set the subjective probability of guilt needed to convict at a mere 68%.[60]

In sum, the findings of empirical studies using mock jurors are remarkably consistent. Different standards of proof—POE, CCE, and BRD—do not produce different verdict patterns, and mock jurors in criminal cases are satisfied with a level of evidentiary proof that is far lower than what judges expect and what is legally required by the BRD standard.[61] In other words, the evidence thus far points to the conclusion that reasonable doubt is not self-defining, but rather must be explained to the jury.

III. THE STUDY

In our study,[62] we sought to remedy the methodological limitations of the three earlier studies (five experiments) that examined mock jurors' verdicts as a function of the burden of proof instruction they received. Taken as a group, the earlier experiments fared poorly in terms of external validity and statistical conclusion validity.[63]

In all five experiments, participants were unrepresentative of jury-eligible adults in the U.S. Participants were either undergraduates at universities or residents of a single city (London) outside the U.S. Participants responded to a fact pattern that included a particular amount of evidence against the defendant, thereby limiting the generalizability of any conclusions. Finally, four of the five experiments relied upon small samples—fewer than 75 participants per condition. As a result, those studies were underpowered statistically. In other words, different standards of proof may, in fact, operate as intended, but earlier studies could not detect weak to moderate effects because of small sample sizes.

In addition to remedying the limitations of earlier studies, we also sought to investigate, within a single controlled study, the possibility that jurors ignore legal standards of proof and instead use a legally improper decision rule when deciding whether to convict. We formally state here these two related hypotheses.

---

[60] Svein Magnussen, et al., *The Probability of Guilt in Criminal Cases: Are People Aware of Being 'Beyond Reasonable Doubt'?*, 28 APPLIED COGNITIVE PSYCHOL. 196, 199 (2014).

[61] One theoretical framework, the "narrative theory of trial," offers an explanation for this phenomenon: "the side that wins—even in a criminal case—is the side that tells the story that best fits with the evidence presented." Keith A. Findley, *Reducing Error in the Criminal Justice System*, 48 SETON HALL L. REV. 1, 8 (2018). More specifically, "if conviction of a crime fits the facts better than acquittal, it is extremely difficult to overcome the desire to match the facts with the better of the two models, *even if the [state's] case is not very strong*." Solan, *supra* note 52, at 108-09 (emphasis added). Some trial courts exacerbate this problem by specifically instructing jurors to compare the state's theory of guilt with the defendant's theory of innocence. Such instructions require the jury to balance two competing theories, which "suggests that a *preponderance of the evidence standard* is relevant, when it is not." United States v. Kahn, 821 F.2d 90, 93 (2d Cir. 1987) (emphasis added).

[62] Our study materials and procedures were approved by the Beloit College Institutional Review Board.

[63] External validity refers to the extent to which a study's findings can be generalized or applied to other persons and other settings. Statistical conclusion validity refers, in part, to a study's ability to identify statistical relationships that are weak but real. *See* Thomas D. Cook & Donald T. Campbell, QUASI-EXPERIMENTATION: DESIGN & ANALYSIS ISSUES FOR FIELD SETTINGS 37 (1979). For the application of these concepts to research on criminal jury instructions and the burden of proof, see Michael D. Cicchini & Lawrence T. White, *Educating Judges and Lawyers in Behavioral Research: A Case Study*, 53 GONZAGA L. REV. 109 (2017-18).

A. *Hypotheses*

First, do the three standards of proof—POE, CCE, and BRD—provide defendants with different degrees of protection on a sliding scale, with BRD providing more protection than the two civil standards? The courts assume they do,[64] but empirical studies suggest they do not.[65]

Second, do jurors use a lay heuristic—a simple rule of thumb that is independent of legal standards of proof—about how much evidence of guilt is needed to convict? Individuals often rely on heuristics, or mental shortcuts, which can lead to erroneous judgments.[66] If such a rule exists, several studies suggest the verdict threshold for most jurors is somewhere between 63% and 68% probability of guilt.[67]

B. *Participants*

We recruited 500 adults via Amazon's Mechanical Turk (MTurk) research platform.[68] We paid each person $1.00 to participate in an online study of juror decision-making. We eliminated the responses of 25 participants: fifteen because they completed the study in fewer than three minutes, ten because they failed to answer key questions. We replaced these participants with new participants to maintain our desired sample size. After the data collection was concluded, we eliminated five additional participants: four because they were not U.S. citizens, one because he completed the study a second time.

Our final sample consisted of 495 jury-eligible adults who hailed from 45 states and the District of Columbia. Two hundred forty-eight participants (50%) identified as female, 244 (49%) as male, and three (< 1%) did not identify as female or male. In terms of ethnicity, 79% identified as White, 7% as African-American, 6% as Asian-American, 4% as Hispanic or Latin American, 2% as mixed race, 1% as Native American or American Indian, and 1% as other.

Participants' ages ranged from 18 to 76; the median age was 33 years. In terms of educational background, 12% of participants had completed high school, 36% some college, 43% a four-year college degree, and 9% a post-graduate degree. Eighty-three participants (17%) had prior jury experience.

[64] *See* Addington v. Texas, 441 U.S. 418 (1979).
[65] *See* Part II.
[66] *See* Brian H. Bornstein & Edie Greene, *Jury Decision Making: Implications For and From Psychology*, 20 CURRENT DIRECTIONS IN PSYCHOL. SCI. 63 (2011) ("jurors sometimes rely on cognitive heuristics when making complicated judgments").
[67] *See* Part II.
[68] *See* Winter Mason & Siddharth Suri, *Conducting Behavioral Research on Amazon's Mechanical Turk*, 44 BEHAV. RES. 1, 1–2 (2012) (MTurk has many advantages, including "easy access to a large, stable, and diverse subject pool, the low cost of doing experiments, and faster iteration between developing theory and executing experiments."); Michael D. Buhrmester, et al., *An Evaluation of Amazon's Mechanical Turk, Its Rapid Rise, and Its Effective Use*, 13 PERSPECTIVES ON PSYCHOLOGICAL SCI. 149, 149 (2017) ("thousands of social scientists from seemingly every field have conducted research using the platform.").

C. *Study Design*

Our study materials used two different criminal cases—battery and trespassing—to increase generalizability. We also constructed strong and weak versions of each case because different standards of proof may produce different verdict patterns only when evidence of guilt is moderate or weak.[69]

We randomly assigned participants to read a case summary that contained one of four fact patterns: a battery case with strong evidence of guilt (1,419 words), a battery case with weak evidence of guilt (1,386 words), a trespassing case with strong evidence of guilt (1,150 words), or a trespassing case with weak evidence of guilt (1,281 words). In all four cases, participants were instructed on the elements of the charged crime.[70]

In the strong-evidence version of the battery case, three witnesses testified. A police officer testified that, on the day of the incident, he observed bruises on the alleged victim's arm and that she signed a complaint against her husband. However, the alleged victim testified (contrary to her earlier complaint) that she and her husband had argued, but her husband did not harm her physically. An expert witness then testified that it is common for an abused spouse to recant a truthful accusation of domestic violence in order to protect the abuser from criminal prosecution. The defendant (the alleged victim's husband) did not testify.

In the weak-evidence version of the battery case, the same three witnesses testified. However, this time the police officer admitted that he could not remember the size of the bruise ("it could have been very small"); the alleged victim admitted that, although she did sign a complaint, she may have exaggerated the incident; and the expert witness admitted that there are reasons why a person might make a false police report against a spouse.

In the strong-evidence version of the trespassing case, two witnesses testified. One of the victims[71] testified that a man with a handgun entered his apartment and demanded money but left abruptly without taking anything. The victim further testified that the police located a suspect about five blocks away and that he, the victim, went to the suspect's location and identified him as the perpetrator. A police officer testified that the suspect did not have a handgun and the police were unable to locate a handgun in the neighborhood. The defendant did not testify.

In the weak-evidence version of the trespassing case, the police officer testified that, even though it took the police eight minutes to arrive, the suspect was found standing on a corner a mere one block (rather than five blocks) away from the victim's apartment. And while the victim positively identified the suspect as the perpetrator (just as he did in the strong-evidence version, above) the officer admitted that a second victim—the first victim's roommate—was unable to pick the defendant out of a lineup the following day.

---

[69] Horowitz & Kirkpatrick, *supra* note 54, found that one version of BRD—an instruction that told juries to convict only if they were "firmly convinced" of the defendant's guilt—produced fewer guilty verdicts than other versions, but only when the case against the defendant was weak. When evidence of guilt was strong, juries convicted most of the time, regardless of how BRD was defined.

[70] All study materials and a complete data file of participants' responses are available at https://osf.io/xm5jr/ — a site hosted by the Open Science Foundation.

[71] We use the word "victim" here because in this fact pattern a crime was committed, and the case turned on the issue of identification.

After reading one of the four trial summaries, participants were randomly assigned to read a jury instruction on the burden of proof to be applied when reaching their verdict. Some participants (n = 99) received the following POE instruction.

> *Preponderance of Evidence*: The State has the burden of proving the defendant's guilt by a preponderance of evidence. Preponderance of evidence means that it is *more probable than not* that the defendant is guilty. If you are so persuaded, you should find the defendant guilty. However, if you are not so persuaded, then you must find the defendant not guilty.[72]

Some participants (n = 100) received the following CCE instruction.

> *Clear and Convincing Evidence*: The State has the burden of proving the defendant's guilt by clear and convincing evidence. This is a higher burden of proof than "more probable than not." Clear and convincing evidence must persuade you that it is *highly probable* that the defendant is guilty. If you are so persuaded, you should find the defendant guilty. However, if you are not so persuaded, then you must find the defendant not guilty.[73]

Some participants (n = 296) received a BRD instruction.[74]

> *Beyond a Reasonable Doubt*: The State has the burden of proving the defendant's guilt beyond a reasonable doubt. If, after carefully considering all of the evidence, you are convinced beyond a reasonable doubt, then you should find the defendant guilty. However, if you have a reasonable doubt, you must find the defendant not guilty.[75]

After reading their particular burden of proof instruction, participants rendered a verdict (guilty or not guilty) and indicated how certain they were about the correctness of their verdict on a 10-point scale. They also indicated what percent of the evidence *they believed* favored the State (the prosecution).[76]

---

[72] This instruction is based on SEVENTH CIR. CIVIL JURY INSTRUCTIONS NO. 1.27 (2015).

[73] This instruction is based on SEVENTH CIR. CIVIL JURY INSTRUCTIONS NO. 1.28 (2015).

[74] Previous studies used short BRD instructions but often included *some* descriptive language. To investigate the possibility that such language—short of a full definition or substantial explanation—would impact verdicts, we constructed three different BRD instructions. The first, reproduced above, leaves BRD completely undefined. The second equated being convinced beyond a reasonable doubt with having "a firm belief in the truth of the charge." (This language is comparable to that tested in an earlier controlled study. *See* Kagehiro & Stanton, *supra* note 36, at 163.) The third reminded jurors to find the defendant not guilty if they have a reasonable doubt, "even if you think that the charge is probably true." The three instructions totaled 50, 63, and 60 words respectively. Our three BRD instructions did not produce statistically significant differences in verdict patterns and had no impact on participants' responses to other questions, so we combined all participants who received a BRD instruction into a single group (n = 296).

[75] SEVENTH CIR. CRIM. JURY INSTRUCTIONS NO. 1.03 (2012).

[76] Not only is this test question consistent with the methodologies employed in the published research, but assigning a numeric, strength-of-evidence value also makes intuitive sense. For example, when seeking

Participants then answered an attention-check question about the number of people who testified as a witness in the case, a question about the meaning of the judge's burden of proof instruction, and a set of questions about their demographic characteristics.

The results of an earlier, face-to-face pilot study led us to believe participants could complete the study thoughtfully in nine to ten minutes, depending on condition.[77] The median time used by MTurk participants to complete the study was 11.5 minutes.

D. *Findings*

Jurors who received a BRD instruction (n = 296) convicted at the rate of 43.6%; those who received a CCE instruction (n = 100) convicted at the nearly identical rate of 43.0%; and those who receive a POE instruction (n = 99) convicted at the rate of only 37.4%. Just as in prior studies, these conviction rates were not consistent with the different burdens of proof, i.e., the POE instruction offered defendants more protection (a slightly lower conviction rate) than the BRD instruction, and jurors did not distinguish at all between the CCE and BRD burdens of proof.

Further, these differences in conviction rates were *not* statistically significant.[78] Expressed mathematically: $\chi^2(2) = 1.20$, $p = .55$. The p-value indicates it is more likely than not—in fact, a 55% probability—that the observed differences in conviction rates were attributable to pure chance. Further, the p-value is *far* above the conventional threshold of statistical significance in the social sciences, which is typically defined as $p = .05$.[79]

---

clarification of the concept of proof beyond a reasonable doubt, one jury asked the trial judge, "What is your definition of reasonable doubt? 80%, 70%, 60%?" People v. Downs, 69 N.E.3d 784, 786 (2015). Additionally, our wording of the question was designed to elicit jurors' confidence level in the defendant's guilt based on their evaluation of the evidence in their particular case summary. Other researchers have worded the question differently when testing this hypothesis. *See* Part II. We worded the question in terms of how much evidence favored the state to invoke the commonly used weight-of-evidence or scales-of-justice analogy. For example, one journalist used such an analogy in explaining the POE standard as follows: "If . . . *50.1% of the evidence supports a claim* but 49.9% does not, that 50.1% is still enough to tip the scale, to prove the claim." Alan Abrahamson, *Tragedy at Sea Pits What-Ifs Against Legal Proof*, L.A. TIMES (Aug. 26, 1991) (emphasis added), at http://articles.latimes.com/1991-08-26/local/me-820_1_circuit-court/2.

[77] We conducted a pilot study with 40 participants to determine the time needed to complete the study and to confirm that (1) participants could easily understand the instructions, case summaries, and questions; (2) the battery and trespassing cases would produce a mix of guilty and not guilty verdicts; and (3) the strong-evidence versions of the cases would produce more guilty verdicts than the weak-evidence versions.

[78] For each participant, we combined their dichotomous verdict (guilty or not guilty) with their level of certainty (1-10) about the correctness of their verdict choice to create a new variable called Scaled Verdict. Values of Scaled Verdict ranged from 1 to 20, with 1 being very certain that the defendant was not guilty and 20 being very certain that the defendant was guilty. Dichotomous verdicts and Scaled Verdicts were highly correlated, $r = .95$, $p < .001$. (The statistic r is used to measure the degree to which two variables are mathematically related to each other. Values of r can range from 0 to 1.00.) Therefore, for purposes of simplicity we discuss only dichotomous verdicts in the text; information on Scaled Verdicts is relegated to the footnotes.

[79] In addition to the p-value, the other statistic provided in the text, above, is known as the chi-square statistic ($\chi 2$). It measures the degree to which observed frequencies differ from expected frequencies. If burden of proof instructions have no impact on juror verdicts, then we expect the percentage of guilty votes

Neither did the different burden of proof instructions have any impact on verdicts when the data were analyzed by case type and evidence strength, i.e., the strong-evidence version of battery, the weak-evidence version of battery, the strong-evidence version of trespass, or the weak-evidence version of trespass.[80]

However, as we expected, the case summaries that presented stronger evidence of the defendant's guilt produced a higher proportion of guilty verdicts. That is, jurors who received a strong-evidence case (n = 249), either battery or trespass, convicted at the rate of 55.8%. Those who received a weak-evidence case (n = 246), again either battery or trespass, convicted at the much lower rate of 28.5%. These results are statistically significant. Expressed mathematically: $\chi^2(1) = 38.00$, $p < .001$. The p-value indicates that we are more than 99% certain [1-p] that the difference in conviction rates between strong- and weak-evidence groups is a real difference that did *not* occur by chance.

After choosing a verdict, participants estimated the percent of evidence in the case that they believed favored the State, i.e., the amount of evidence they believed indicated the defendant was probably guilty. As expected, jurors assigned to one of the strong-evidence cases offered higher estimates (63.0%) than jurors assigned to a weak-evidence case (46.1%).[81] Participants' estimates of the strength of the State's case were strong predictors of their verdicts.[82]

When participants believed that less than 60% of the evidence favored the State, 93% of participants (233 of 250) voted not guilty. When participants believed that more than 65% of the evidence favored the State, 86% of participants (178 of 206) voted guilty. The tipping point—that is, the point at which the majority of participants, for the first time, voted guilty instead of not guilty—occurred somewhere between 60% and 65%. This is illustrated in the following table.[83]

---

in a given case to be essentially the same, regardless of the instruction received. This is, in fact, what we observed. *See* Arthur Aron & Elaine N. Aron, STATISTICS FOR PSYCHOLOGY 509-11 (3d ed. 2003).

[80] In the strong-evidence versions of the battery and trespassing cases, the different burden of proof instructions had no impact on Scaled Verdicts, $F(2, 245) = 0.50$, $p = .61$. Even in the weak-evidence versions of the battery and trespassing cases, the different instructions had no significant impact on Scaled Verdicts, $F(2, 242) = 1.60$, $p = .20$.

[81] The difference expressed mathematically is $t(490) = 6.61$, $p < .001$.

[82] This was true for both dichotomous verdicts ($r = .72$, $p < .001$) and Scaled Verdicts ($r = .75$, $p < .001$). As indicated earlier, dichotomous verdicts (guilty or not guilty) and Scaled Verdicts were highly correlated.

[83] Three participants failed to answer the question regarding the percentage of evidence favoring the state, leaving us with 492 respondents instead of 495.

Amount of Evidence that Favors the State	*Votes for Guilty*	*Votes for Not Guilty*	*Total Votes*	*Percent Guilty*	*Percent Not Guilty*
Less than 50%	12	202	214	6%	94%
50% - 55%	5	31	36	14%	86%
60%	9	18	27	33%	67%
65%	5	4	9	56%	44%
70% - 75%	43	14	57	75%	25%
More than 75%	135	14	149	91%	9%
Total	209	283	492		

For each participant, we then determined if they used what we call "the 60/65 rule." A participant used the rule if (a) they said that *less* than 60% of the evidence favored the State and voted *not* guilty, or (b) they said that *more* than 65% of the evidence favored the State and voted *guilty*. According to these criteria, 84% of all participants used the rule, 9% did not use the rule, and 7% could not be categorized (because they estimated that 60% or 65% of the evidence favored the State). Of those participants who could be categorized, 90% (411 of 456) followed the 60/65 rule. Most importantly, the different burden of proof instructions had *no effect* on whether a participant used the 60/65 rule. This null effect can be expressed mathematically: $\chi^2(4) = 4.06$, $p = .40$, $V = .09$.[84]

Seventy-seven percent of participants correctly answered the attention-check question about the number of witnesses who testified.[85] Eighty-four percent (247 of 293) of the participants who received a BRD instruction correctly understood that, if they had a reasonable doubt about the defendant's guilt, they must not convict. Only 42% (83 of 199) of the participants who received a POE or CCE instruction correctly understood that, if they were convinced the charge was true, they could convict the defendant even if they had a reasonable doubt about guilt. Put another way, 74% of all participants reported that the standard of proof in the case before them was BRD, even though only 60% of participants (296 out of 495) actually received a BRD instruction.[86]

[84] Additionally, we found that participants' use of the 60/65 rule was not associated with gender, $V = .06$, $p = .66$, and was not associated with prior jury experience, $V = .08$, $p = .43$. The statistic V is a correlation coefficient that measures the degree to which two variables are mathematically related to each other. Like its statistical sibling r, values of V can range from 0 to 1.00. The statistic V is used when both variables are nominal (categorical) variables, e.g., a participant's prior jury experience (yes or no) and a participant's use of the 60/65 rule (yes or no).

[85] Our attention-check question was more difficult than the questions used in some MTurk studies (e.g., which of the following best describes the shape of a ball?). Also, the wording of the question may have confused some participants. Two participants contacted us off-line to say they were confused by the word "witness" in the question; they pointed out that "witness" could refer to a person who testified at trial or to a person who observed the incident or evidence of it.

[86] Participants' understanding of their burden of proof instructions is discussed in Part VI.

IV. DISCUSSION OF THE FINDINGS

Our study's two major findings—(a) different burden of proof instructions did not produce different verdict patterns, and (b) nearly all jurors used a simple 60/65 rule of thumb when deciding whether to convict—are discussed below.

A. *Different Burdens of Proof*

The first finding of our study—that different burdens of proof had no impact on juror verdicts—is consistent with the results of studies conducted decades ago. As we discussed earlier, those studies were limited by weak external validity (i.e., generalizability) and weak statistical conclusion validity. However, we were able to successfully remedy those problems with our study.

Specifically, our study participants were jury-eligible adults between the ages of 18 and 76 who hailed from 45 states. They responded to four different criminal cases, each with its own fact pattern and evidence strength. We also used a large sample (N = 495) to produce a high-powered test.[87] And, once again, the different standards of proof—POE, CCE, and BRD—had no discernible impact on mock jurors' verdicts. We observed the same null effect across four criminal cases with different fact patterns and varying evidence strength. In short, the BRD instruction did not offer any greater protection than the two civil burdens of proof, and this held true regardless of the charged crime and regardless of the strength of the State's case.

To our knowledge, no researcher has been able to demonstrate that the concept of BRD, particularly when left undefined or only minimally defined, provides defendants with more protection than the civil burdens of proof.[88] To the contrary, multiple researchers have now observed that burden of proof instructions have no effect on verdicts, regardless of the characteristics of the participants used, the charges against the

[87] We used a 2x2x3 between-subjects study design, i.e., twelve cells. We did not predict a three-way interaction nor did we test for one. Rather, we tested for a main effect: whether different standards of proof produced different verdict patterns. Because we compared three different instructions, we had three test groups (not twelve). The two additional independent variables (strength of evidence and type of crime) were included to assess the generalizability of our finding, i.e., to see if the null effect would be observed in different kinds of cases. The estimated power (likelihood) for detecting a small effect in a three-group experiment with an average of 165 participants per group is .98, i.e., there would be a 98% chance of detecting a small effect, assuming one exists. (When participants are unevenly distributed across test conditions, statistical power is lowered slightly.) We did not detect an effect; we can therefore be confident that one does not exist. *See* STATISTICAL DECISION TREE, P*ower Calculator: Calculation for a One-Way Independent ANOVA*, at https://www.anzmtg.org/stats/PowerCalculator/PowerANOVA (last visited May 19, 2018).

[88] We are aware of one study where variations in the laxity or stringency of a BRD definition affected individual and mock jury verdicts in the expected way. *See* Norbert L. Kerr et al., *Guilt Beyond a Reasonable Doubt: Effects of Concept Definition and Assigned Decision Rule on the Judgments of Mock Jurors*, 34 J. PERSONALITY & SOCIAL PSYCHOL. 282 (1976) (testing [1] an instruction leaving BRD undefined, [2] a lax-criterion BRD instruction telling the jury that "[a] reasonable doubt is not just a possible doubt, not a capricious or trivial doubt," and [3] a stringent-criterion BRD instruction warning the jury that, in order to convict, "the prosecution must have convinced you to a moral certainty, with absolute and positive proof, that the defendant is guilty.").

defendant, the specific facts and strength of evidence in the case under consideration, and the decade in which the study was conducted.

B. *The 60/65 Rule*

Although the mock jurors in our study were not sensitive to variations in the burden of proof instructions, they were highly sensitive to variations in the strength of the evidence. When evidence against the defendant was strong, participants generally voted to convict, regardless of the burden of proof; when evidence against the defendant was weak, participants generally voted to acquit, regardless of the burden of proof. Juror sensitivity to evidence strength was observed in both cases (battery and trespassing) and across all three burdens of proof. A "strength of evidence" effect has also been observed in other studies of jury decision-making.[89]

Put another way, although mock jurors' verdict choices were not influenced by their burden of proof instruction, neither did they choose their verdicts in a haphazard, unpredictable fashion. In fact, we found strong evidence that jurors used a simple rule, unencumbered by legal standards of proof, about how much evidence of guilt is needed to convict. We call this the 60/65 rule. Participants who used this rule—and 90% of them in our study did—voted to acquit when they believed that less than 60% of the evidence favored the State but voted to convict when they believed that more than 65% of the evidence favored the State.

The existence of a 60/65 rule—or something very similar to it—is substantiated by human judgment studies that have found the tipping point from acquittal to conviction to be somewhere between 63% and 68% probability of guilt.[90] Jury researcher Irwin Horowitz was apparently correct when he wrote, more than twenty years ago, that "[r]ather than having to move jurors from 0% to 90% certainty, all prosecutors need do is move the needle on the scale from 50% to perhaps 65% certainty."[91]

A great deal of research now demonstrates that reasonable doubt is not self-defining. If the BRD standard is to offer defendants greater protection than lower burdens of proof, and if it is to require more than a mere "65% certainty" in jurors' minds before they convict, then the concept of reasonable doubt must be properly defined for the jury.

V. HOW TO DEFINE REASONABLE DOUBT

Trial judges are responsible for ensuring that juries understand the law.[92] But when the concept of reasonable doubt is left undefined (or, worse yet, is defined

[89] For a review, see Dennis J. Devine et al., *Jury Decision Making: 45 Years of Deliberating Groups*, 7 PSYCHOL., PUB. POL. & L. 622 (2001).

[90] *See* Part II.

[91] Irwin A. Horowitz, *Reasonable Doubt Instructions: Commonsense Justice and Standard of Proof*, 3 PSYCHOL., PUB. POL. & L. 285, 294 (1997).

[92] *See* State v. Neumann, 832 N.W.2d 560, 584 (Wis. 2013) ("A circuit court must, however, exercise its discretion in order to fully and fairly inform the jury of the rules of law applicable to the case and to assist the jury in making a reasonable analysis of the evidence."). This is true even when jury instruction committees have issued jurisdiction-wide pattern instructions. *See Preface*, PATTERN CRIM. JURY INSTRUCTIONS FOR THE DISTRICT COURTS OF THE FIRST CIRCUIT, DRAFTING COMMITTEE (Nov. 1997)

improperly) defendants are at risk of being convicted with a level of proof that is significantly lower than what is expected by the courts and required by the Constitution.

What can be done to prompt jurors in criminal cases to apply the concept of reasonable doubt in a legally appropriate manner? As a New Hampshire court stated, "a jury must be given some assistance in understanding the concept."[93] We believe this "assistance" must be substantial, as the brief yet clear instructions used in burden of proof studies have failed to produce legally-proper verdict patterns.

First, burden of proof jury instructions should begin, as most already do, by explaining the presumption of innocence. This part of the instruction should also include, as Hawaii's does, a warning such as this: "The presumption of innocence is not a mere slogan but an essential part of the law that is binding upon you."[94]

Second, because jurors are more likely to follow instructions when trial judges explain the rationale behind them,[95] judges should identify proof beyond a reasonable doubt as the applicable standard and also explain *why* it is such an important concept in criminal law: (1) it protects us from loss of "life, liberty, and property" at the hand of the government, and (2) nearly as important, it helps to ensure the community's confidence in the criminal justice system.[96] Put another way, "It is critical that the moral force of the criminal law not be diluted by a standard of proof that leaves people in doubt whether innocent men are being condemned."[97] This principle should be communicated to the jury in plain English.

Third, with regard to the reasonable doubt definition itself, some researchers have advised trial judges to instruct jurors in a way that places the relevant burden of proof within its legal context.[98] There is empirical support for this recommendation. For example, when study participants received a set of instructions that included all three standards of proof—POE, CCE, and BRD—from a single jurisdiction, they were able to compare and contrast the standards.[99] Based on participants' rating of each standard in terms of how difficult it would be for a plaintiff to win a civil case, the researchers

---

(although "the pattern instructions . . . will be helpful in crafting a jury charge in a particular case, it bears emphasis that no district judge is required to use the pattern instructions."); United States v. Gonzalez-Balderas, 11 F.3d 1218, 1223 (5th Cir. 1994) (stating that, with regard to the mandate "to seek the truth," "although the sentence is taken from the Fifth Circuit Pattern Jury Instructions, trial courts, in an abundance of caution, may wish to delete it from their instructions.").

[93] State v. Aubert, 421 A.2d 124, 127 (N.H. 1980).

[94] HAW. CRIM. JURY INSTRUCTIONS NO. 3.02 (2014). Such language is important, as research has shown that jurors may simply be glossing over the presumption of innocence. That is, "while the law ostensibly creates a presumption of innocence, it is widely recognized, both as a matter of theory and empirics, that prosecutors are actually aided by a presumption of guilt, at least once the first bits of evidence are introduced . . ." Findley, *supra* note 61, at 20 (citing Michael J. Saks & D. Michael Risinger, *Baserates, the Presumption of Guilt, Admissibility Rulings, and Erroneous Convictions*, 2003 MICH. ST. L. REV. 1051, 1062 (2003)).

[95] *See* Nancy Steblay et al., *The Impact on Juror Verdicts of Judicial Instruction to Disregard Inadmissible Evidence: A Meta-Analysis*, 30 L. & HUM. BEHAV. 469 (2006).

[96] In re Winship, 397 U.S. 358, 362 (1970).

[97] Id. at 364.

[98] *See* Stoffelmayr & Diamond, *supra* note 51, at 776 ("By providing this explicit contrast with a less stringent standard of proof, the definition encourages jurors to adopt an appropriately high threshold for conviction. It could be strengthened even further by adding an additional contrast with clear and convincing").

[99] Kagehiro and Stanton, *supra* note 36, at 172.

concluded that the different burdens of proof "might affect verdicts as intended by the law, *if they were presented in comparative context.*"[100]

We strongly agree with these recommendations. Specifically, trial judges should instruct juries on all three standards of proof and explicitly state that POE is the lowest standard, CCE is higher, and BRD is higher still. In criminal cases, defining "proof beyond a reasonable doubt" by comparing it to other, lower standards would provide the necessary context for jurors to appreciate this high standard. To illustrate, the language below is designed to produce legally-proper verdict patterns.

> Some civil cases use the preponderance of evidence standard. In those cases, it is only necessary to prove that something is probably true, or more likely true than not. But this is a criminal case, and the State's proof must be more powerful than that.
>
> Other civil cases use the clear and convincing evidence standard. In those cases, it is necessary to prove that the truth of something is highly probable. But this is a criminal case, and the State's proof must also be more powerful than that.
>
> In criminal cases such as this, you can convict the defendant only if the State's proof satisfies you beyond a reasonable doubt that the defendant is guilty. If it does not, you must find the defendant not guilty even if you think the charge is probably true, and even if you think it is highly probable that the charge is true.[101]

Fourth, the instruction should conclude by conveying the Supreme Court's mandate that, to convict a defendant under this high burden of proof, the jury must "reach a subjective state of *near certitude* of the guilt of the accused[.]"[102] One way to convey this is to end the instruction as North Carolina does: "Proof beyond a reasonable doubt is proof that fully satisfies or entirely convinces you of the defendant's guilt."[103]

In sum, criminal defendants in many states and federal jurisdictions are not adequately protected from conviction because jurors do not apply the BRD standard in the manner that judges expect and the Constitution requires. The reason for this state of affairs is clear: The legal concept of proof beyond a reasonable doubt is *not* self-defining.

---

[100] Id. (emphasis added).

[101] Michael D. Cicchini, *Instructing Jurors on Reasonable Doubt: It's All Relative*, 8 CALIF. L. REV. ONLINE, 72, 85 (2017). Some jurisdictions already instruct jurors on a comparative basis. *See, e.g.,* MASS. CRIM. JURY INSTRUCTIONS No. 2.180 (2015) ("It is not enough to establish a probability, even a strong probability, that the defendant is more likely to be guilty than not guilty. That is not enough."); ARIZ. CRIM. JURY INSTRUCTIONS No. 5b(1) (2015) ("In civil cases, it is only necessary to prove that a fact is more likely true than not or that its truth is highly probable. In criminal cases such as this, the State's proof must be more powerful than that."). Our recommended instruction draws heavily from Arizona's instruction, but more forcefully distinguishes the BRD standard from the two civil burdens of proof. Conversely, other jurisdictions *also* instruct jurors on a comparative basis, but do more harm than good in the process. For example, stating that BRD is a higher standard than "mere suspicion of guilt" says nothing of value, is grossly misleading, and could even lower the burden of proof. *See* Michael D. Cicchini, *Roger Federer, Michael Cicchini, and Pennsylvania's Burden of Proof*, THE LEGAL WATCHDOG (June 17, 2017), at http://thelegalwatchdog.blogspot.com/2017/06/roger-federer-michael-cicchini-and.html.

[102] Jackson v. Virginia, 443 U.S. 307, 315 (1979) (emphasis added).

[103] N.C. CRIM. JURY INSTRUCTIONS NO. 101.10 (2008).

To remedy this problem, criminal courts should instruct jurors on BRD in a way that includes four components. First, judges should explain the presumption of innocence; second, judges should identify BRD as the applicable burden of proof and explain its importance to the jury; third, judges should present the BRD standard within its legal context by comparing and contrasting the three burdens of proof; and fourth, judges should conclude by stressing the Supreme Court's "near certitude" requirement for conviction.

VI. STUDY LIMITATIONS AND FURTHER TESTING

All empirical studies are flawed in the sense that methodological decisions designed to solve one problem often exacerbate another. Possible limitations of our study include the following.

First, mock jurors in our study were jury-eligible adults who responded to realistic fact patterns, but they did not participate in an actual trial. We chose to test the impact of different instructions in an artificial setting because we needed a high level of experimental control. We were able to systematically vary instructions among participants and hold other variables constant in a way that could never be achieved in actual trials. Real jurors in actual trials, however, may give more attention to a burden of proof instruction and make a greater effort to apply the standard in a legally-proper manner. However, this would first require that jurors *understand* the burden of proof. In one study, most real-life jurors who had completed jury duty were still confused about the meaning of BRD.[104]

Second, the mock jurors in our study read a short trial summary; they did not watch a lengthy videotaped reenactment of a trial. We chose to use abbreviated summaries for two reasons. To begin, we wished to give the burden of proof instruction every opportunity to have an impact on jurors' verdicts by minimizing the amount of additional information that might wash out a standard of proof effect. Further, mock jurors typically do not react differently to abbreviated and more elaborate case summaries.[105] Nevertheless, researchers may wish to employ more realistic simulations when they examine the impact of judicial instructions on verdict patterns.

Third, the mock jurors in our study did not deliberate with others before choosing a verdict. It seems reasonable to believe that jurors who deliberate will come to understand the judge's instruction as they deliberate and then apply the standard when choosing a verdict. On the other hand, in a study discussed earlier, the mock jurors

[104] Geoffrey P. Kramer and Dorean M. Koenig, *Do Jurors Understand Criminal Jury Instructions? Analyzing the Results of the Michigan Juror Comprehension Project*, 23 U. MICH. J. L. REFORM 401 (1990). Similarly, with regard to the closely related concept of the presumption of innocence, about one-third of real-life Wyoming jurors "believed that the reasonable doubt standard does cause a shift in the burden of proof from the government to the defendant, despite instructions by the court explaining the presumption of innocence." Solan, *supra* note 52, at 120 (discussing Bradley Saxton, *How Well Do Jurors Understand Jury Instructions? A Field Test Using Real Juries and Real Trials in Wyoming*, 33 LAND & WATER L. REV. 59 (1998)).

[105] Geoffrey P. Kramer & Norbert L. Kerr, *Laboratory Simulation and Bias in the Study of Juror Behavior: A Methodological Note*, 13 L. & HUM. BEHAV. 89 (1989).

deliberated in a group but still returned verdicts that were not affected by the standard of proof assigned.[106] Other studies have found that deliberations play a minor role, as actual jurors devote less than one percent of their deliberation time to issues related to the burden of proof.[107] Further, the verdict favored by a majority of jurors before deliberation almost always becomes the jury's final verdict.[108]

Fourth, MTurk workers have a financial incentive to complete tasks quickly and may have skimmed the trial summary and judge's instruction. But we believe this scenario is unlikely for two reasons. First, the online participants in our study devoted as much time to the required tasks as did participants in our face-to-face pilot study. Second, a large majority of the online participants correctly answered a difficult attention-check question and were highly sensitive to variations in the strength of evidence against the defendant. Participants could not have distinguished so clearly between the strong- and weak-evidence cases if they had merely skimmed the trial summary. Our positive experience with the participants in our study is not surprising: "evaluations have found that MTurk particpants' attention is equal to or better than undergraduate participants' attention."[109]

Finally, although our findings align very closely with earlier findings that jurors are not sensitive to different burdens of proof, and that jurors will vote guilty even when evidence does not reach a quantitative threshold consistent with proof beyond a reasonable doubt, we are not able to say, precisely, *why* this is so. Put another way, did we obtain null effects because jurors failed to *identify* the proper burden of proof, i.e., POE, CCE, BRD, before rendering a verdict? Did they identify the proper burden but simply *not understand* it? Or did they both identify and understand the burden of proof but simply refuse to follow it?

There is evidence that some of our participants misidentified the burden of proof from the outset. Specifically, we tested the three burdens of proof using *criminal*-case fact patterns. Previous researchers used civil-case fact patterns because "confusion might arise if subjects were asked to apply any standard other than reasonable doubt to a criminal case."[110] In our study, we included a post-verdict question that can be used to evaluate this concern.

As stated earlier, while 84% (247 of 293) of our study participants who received a BRD instruction correctly described their instruction in a post-verdict question, only 42% (83 of 199) of the participants who received a POE or CCE instruction were able to do so. That is, they described their instruction as being BRD. This could partly explain why participants were not sensitive to the different standard of proof instructions: some of them failed to identify the proper standard to begin with.

---

[106] Cornish & Sealy, *supra* note 32.

[107] James R. P. Ogloff, *Judicial Instructions and the Jury: A Comparison of Alternative Strategies*, FINAL REPORT (BRITISH COLUMBIA LAW FOUNDATION, 1988).

[108] Shari S. Diamond, *Illuminations and Shadows from Jury Simulations*, 21 L. & HUM. BEHAV. 561, 564 (1997); Bornstein & Greene, *supra* note 66, at 65 ("In approximately 90% of trials, the position favored by the majority at the beginning of deliberations becomes the jury verdict.").

[109] Buhrmester, et al., *supra* note 68, at 151. Further, juror inattention is a common issue an issue in real-life jury trials as well. *See* State v. Chestnut, 643 S.W.2d 343, 346 (Tenn. Ct. App. 1982) (upholding the denial of defendant's motion for new trial despite evidence of two jurors sleeping through evidentiary portions of the case).

[110] Kagehiro & Stanton, *supra* note 36, at 162.

Yet, when previous researchers tested the effect of three burden of proof instructions using a *civil*-case fact pattern, there was certainly a risk of the reverse happening: confusion might arise because subjects were asked to apply a BRD standard to a mere civil matter. Despite the inherent difficulty of using either a civil or a criminal fact pattern to test three different burdens of proof, both studies—the previous study using a civil case and our study using criminal cases (battery and trespass)—produced verdict patterns that failed to conform to the different standards of proof.

But while some jurors in our study misidentified their burden of proof at the outset, there is very strong evidence that many of them either misunderstood the concept of BRD or simply refused to apply it when reaching their verdicts.[111] More specifically (and as discussed above) we know from responses to the post-verdict question that 74% of all participants reported—some correctly, some incorrectly—that they had received a reasonable doubt instruction.[112] Despite nearly three-fourths identifying BRD as their applicable standard, nearly all participants (90%) followed the 60/65 rule. That is, they either (a) said that less than 60% of the evidence favored the State and voted not guilty, or (b) said that more than 65% of the evidence favored the State and voted guilty.

Therefore, regardless of whether jurors misidentified, misunderstood, or simply refused to apply their burden of proof, the results of this study and of the earlier studies discussed in this Article provide strong evidence that reasonable doubt is *not* self-defining. Rather, it needs to be explicitly and properly defined so that prosecutors must do more than "move the needle on the scale . . . to [merely] 65% certainty"[113] in order to win a criminal conviction.

Given the numerous studies already demonstrating that reasonable doubt is not self-defining, future researchers should test various BRD *definitions*—particularly our comparative, context-based definition discussed earlier—to determine which definitions compel mock jurors to abandon the 60/65 rule and instead demand more evidence of guilt before convicting criminal defendants under the BRD standard of proof.

CONCLUSION

When instructing juries on reasonable doubt, many courts subscribe to one of two philosophical camps. On the one hand, many courts go to great lengths to define, explain, or illustrate the concept of reasonable doubt.[114] On the other hand, many courts believe that the concept of reasonable doubt is already self-defining; therefore, they

[111] Many studies and reviews have found that judicial instructions of all kinds are poorly understood or simply ignored. A comprehensive review of jury decision-making studies published between 1945 and 1999 found that "jurors often do not make decisions in the manner intended by the courts, regardless of how they are instructed." Devine, *supra* note 89, at 699. Judicial instructions concerning the standard of proof are probably not an exception to the general rule. *See* Kramer & Koenig, *supra* note 104.

[112] This can be calculated using the numbers in the earlier paragraph. Of those who received a BRD instruction, 247 properly identified it. Of those who received one of the civil instructions, 116 (199 – 83) nonetheless identified their instruction as BRD. Therefore 363 (247 + 116) of 492 (293 + 199) or 73.8% identified their burden of proof as BRD. (The total number of participants was 495, but three failed to answer the post-verdict question about the burden of proof, leaving 492 respondents to the question.)

[113] Horowitz, *supra* note 91, at 294.

[114] Part I.

believe, jury instructions should not attempt to further define or explain the concept, as any attempts to do so may confuse the jury.[115]

It is true that many jury instructions explaining reasonable doubt create confusion—or, worse yet, lower the burden of proof below the constitutionally-mandated standard.[116] However, the existing empirical evidence suggests that reasonable doubt is not self-defining. Not only do mock jurors in controlled experiments fail to distinguish between reasonable doubt and the lower, civil burdens of proof, but they are also willing to convict defendants based on a quantum of evidence that is much lower than what many judges expect and the Constitution requires.[117]

In this Article, we present our study that tested the impact of different burdens of proof on juror decision-making. We sought to remedy the limitations of earlier studies and determine whether, instead of applying the burden of proof instruction, jurors use a simple rule of thumb about the quantum of evidence needed to find a defendant guilty.[118]

Our findings confirmed those of previous studies: Mock jurors are not sensitive to burden of proof instructions. That is, reasonable doubt instructions provide defendants with no greater protection than the lower, civil burdens of proof.[119] Further, we found that instead of following their burden of proof instruction, jurors used a simple heuristic that we call the 60/65 rule.[120]

Specifically, nearly all mock jurors in our study either (a) said that *less* than 60% of the evidence favored the State and voted *not* guilty, or (b) said that *more* than 65% of the evidence favored the State and voted *guilty*.[121] In other words, the tipping point—the point at which jurors were as likely to convict as to acquit—occurred somewhere between 60% and 65%.

The empirical evidence has repeatedly demonstrated that reasonable doubt is not self-defining; instead, jurors need assistance in understanding and appreciating the high burden of proof that the government must meet when it attempts to deprive a person of life, liberty, and property.[122]

We therefore recommend that instructions on reasonable doubt begin by describing the presumption of innocence.[123] Then, instructions should identify BRD as the applicable burden and explain the *reasons* that our Constitution requires such a high burden of proof before the jury may convict.[124] The heart of any instruction should then define reasonable doubt by relating the BRD standard to the lower POE and CCE standards.[125] Such context will help jurors understand, on a comparative basis, the high level of proof that the government must produce. Finally, instructions should conclude

[115] Id.
[116] Id.
[117] Part II.
[118] Part III.C.
[119] Part III.D.
[120] Id.
[121] Id.
[122] Part IV.
[123] Part V.
[124] Id.
[125] Id.

with language that communicates the Supreme Court's "subjective state of near certitude" requirement for a criminal conviction.[126]

[126] Id.