50%
51%
52%
53%
54%
55%
56%
57%
58%
59%
60%
61%
62%
63%
64%
65%
66%
67%
68%
69%
70%
71%
72%
73%
74%
75%
76%
77%
78%
79%
how SURE are YOU ?

Legal Standards by the Numbers

Quantifying Burdens of Proof or a Search for Fool's Gold?

by

Richard Seltzer
Russell F. Canan
Molly Cannon
Heidi Hansberry

Just after midnight on a warm summer night, a Caucasian woman was walking alone on the streets of Washington, D.C. All of a sudden, three young men she had never seen before, all African-American, surrounded her, pushed her down, and said, "Shut up, bitch." They menaced her with a rifle-shaped piece of wood, stole her purse, and were gone as soon as they had come upon her. *In re As.H.*, 851 A.2d 456, 457 (D.C. 2004). The woman saw the young men because the street was well-lit but could only provide generalized descriptions of their race and clothing. Later that same evening, however, when the police conducted a show-up by having her view suspects, she was certain the show-up suspects were not the robbers. About a month later, she was shown photos and identified four suspects. She was "very certain" about two of the suspects, which she expressed meant a "seven or eight" on a scale of one to ten. *Id.* at 458.

Five months after the crime, at the juvenile delinquency trial for one of the accused robbers, the woman testified that the juvenile on trial was one of the robbers she had identified from the photos. When asked about her level of certainty, the woman repeated her indication from the photo array, "At the time, on a scale of one to [ten], I said that I was seven or eight." *Id.* at 458. Based on this eyewitness testimony alone — no other witnesses, no fruits of the crime connected to him, no forensic evidence such as DNA or fingerprints tying him to the offense — the juvenile was found guilty of robbery.

On appeal, the District of Columbia Court of Appeals split two to one, with the majority ordering that the guilty finding be reversed based on insufficient evidence — a rare result on appeal. The majority regarded the "seven or eight" certainty level as legally insufficient where eyewitness testimony was the only evidence against the accused. The court based its decision in part on a study by Rita Simon and Linda Mahan (1971) which showed that judges quantified beyond a reasonable doubt higher than 70 to 80 percent.[1] In the cited study, questionnaires were sent to judges asking them to

STUDY OVERVIEW

THIS RESEARCH IS A FOLLOW UP and continuation of the research conducted by Simon & Mahan (1970–71) on translating legal standards into numerical responses. One hundred and twenty-four judges participated in this study. There were four major findings:

1. The results for beyond a reasonable doubt aligned very closely with the judges' perceptions from previous studies.
2. There was no significant relationship between a judge's quantification and factors such as whether the judge was elected or appointed, sat on a criminal or civil docket, or length of tenure.
3. There was a wide variety in the quantification of standards for probable cause, substantial probability, and reasonable articulable suspicion.
4. Judges' quantification of probable cause and preponderance of the evidence was nearly identical, a result with significant legal implications given the standards' differing legal definitions.

The third and fourth findings are new to the literature and may have serious legal consequences.

quantify the beyond a reasonable doubt standard as a percentage. Those judges who responded split roughly into thirds. One-third reported beyond a reasonable doubt to be at 100 percent certainty. One-third reported it at 90 or 95 percent. The final third reported it at 80 percent. The court stated, "very few judges, if any, would have regarded an 80 percent probability as sufficient to prove guilt beyond a reasonable doubt, and . . . all of them would have considered a 70 percent probability as altogether inadequate." *Id.* The dissenting judge dismissed the majority's reliance on the judicial study, noting the inexact nature of the use of a one-to-ten scale and decrying, "[T}he entire effort to quantify the standard of proof beyond a reasonable doubt is a search for fool's gold" because "[a} factfinder's evaluation of credibility and intensity of belief should not be overridden by such inexact and even trivial differences of quantification." *Id.* at 463–64.

The American legal system depends on standards regarding the burden of proof to facilitate outcomes that accurately balance society's interests with an assessment of risk. Judges and juries use these standards to make decisions such as granting bail, assessing the validity of stops or arrests by the police, issuing arrest and search warrants, determining guilt in criminal trials and liability in civil trials, adjudicating child custody disputes, and terminating parental rights. These standards are uniformly expressed verbally rather than numerically. Since an important underlying goal of the legal system is uniform application of the law by decision-makers, both judges and juries, these standards should mean the same thing to different people across time, type of case, and courtroom. Testing the meaning — and the consistency of meaning — of the standards is difficult due to the fact that they are expressed verbally. Translating legal standards into a numerical scale offers an opportunity to test the meaning of the verbally expressed standards in a quantifiable and reproducible manner. This article reviews other attempts to translate probability statements into numerical scales and reports the results of a survey of judges quantifying six legal standards.

STUDY OVERVIEW

The research reported in this article was done as a follow up and continuation of the research conducted by Simon & Mahan and relied upon in *In re As.H*. In this study, participating state- and federal-court trial judges throughout the country translated six legal standards into numerical responses on a 0-to-100 percent scale. The study has four main conclusions, each of which is summarized below.

The first major finding is that the results for beyond a reasonable doubt align very closely with the judges' perceptions from the Simon studies from the late 1960s and early 1970s, which may indicate stability over time.

The second major finding is that there was no significant relationship between a judge's quantification and factors such as whether the judge was elected or appointed, sat on a criminal or civil docket, or length of tenure.

The third major finding is that there was a wide variety in the quantification of standards for probable cause, substantial probability, and reasonable articulable suspicion.

RICHARD SELTZER is a professor in the Department of Political Science at Howard University.

RUSSELL F. CANAN is a judge on the Superior Court of the District of Columbia. He is the former presiding judge of the Criminal Division and currently the chair of a committee that studies wrongful convictions and proposes reforms to the criminal justice system.

MOLLY CANNON formerly practiced criminal law including court-appointed defense of indigent clients.

HEIDI L. HANSBERRY, Esq., received bachelor's degrees from Yale University and a law degree from Northwestern University School of Law and served as a law clerk to Judge Russell F. Canan.

The fourth major finding is that judges' quantification of probable cause and preponderance of the evidence was nearly identical, a result with significant legal implications given the standards' differing legal definitions.

Although there is debate in the legal academy and the courts as to the utility and place of quantification, cases such as *In re As.H.* demonstrate that quantification has real-world relevance, and, at the very least, quantification is a vehicle for judges to evaluate their judicial decision-making processes.

LITERATURE REVIEW

SOCIAL SCIENCE RESEARCH ON PROBABILITY STATEMENTS

There has been only limited social science inquiry on translating legal, verbal probability statements into numeric estimates. The research is more abundant outside of the legal arena. There are many studies on translating individual verbal probability terms such as "likely," "never," or "often" into actual percentages. Budescu and Wallsten (1995) note that interpretation of verbal probability statements is affected by three factors: context (e.g., "it is likely to snow" is interpreted differently in Minnesota than South Carolina), who makes the statement (e.g., "you will soon recover from your illness" has greater credibility if said by your doctor compared to your neighbor), and who hears it (e.g., "the prison sentence is very light" might be understood very differently by a judge compared to a defendant).[2] They also note that although people prefer receiving numeric over verbal statements, there is no consensus in the literature suggesting the former is more accurate than the latter.

These quantification issues are far from abstract in the medical profession. Physicians often make a prediction when they discuss possible treatments with patients and their families. If a doctor says the likelihood of having a bad reaction from a medication is rare, the patient should know whether the doctor means under a one percent probability, under a five percent probability, or some other figure. There is considerable debate about whether it is better for doctors to use probability statements such as rare, remote, etc., or to use an actual percentage, such as five percent. Some analysts believe that people prefer to communicate without the use of probabilities because it seems more intuitive and natural.[3]

> "The American legal system depends on standards regarding the burden of proof to facilitate outcomes that accurately balance society's interests with an assessment of risk. . . . Since an important underlying goal of the legal system is uniform application of the law by decision-makers, both judges and juries, these standards should mean the same thing to different people across time, type of case, and courtroom.

Another issue is that many people have both very low health literacy[4] and difficulty understanding statistics.[5] Bruine de Bruin, Fischhoff, Millstein, and Halpern-Felsher (2000) argue however that numeric probabilities sharpen the reasoning process of people with such limitations.[6]

On the other hand, verbal statements are not linked to any widely accepted standard (i.e., does rare mean five percent?), and there is often considerable variance in how people interpret these terms.[7] These inconsistencies are higher among people with less education.[8] Some analysts argue that professionals, such as physicians and engineers, exhibit greater consistency than lay people in the interpretation of verbal probabilities.[9] Even so, they argue there is a need for further elaboration of what the probabilities represent in order to minimize the inconsistencies.

The Intergovernmental Panel on Climate Change ("IPCC") addressed the issue concerning inconsistent usage of verbal probability statements by having its authors communicate probabilities using seven common scales. Budescu, Por, and Broomell (2012) had 556 respondents from a Knowledge Network ("KN") survey assign probabilities to eight probabilistic terms involving climate change.[10] They found a low-level of correspondence with the IPCC guidelines. For example, the IPCC stated the phrase "very likely" was to be used when referring to probabilities greater than 90 percent. However, in contrast, the KN respondents reported between 65 and 75 percent on this scale. The authors did not criticize the IPCC for attempting to establish standards to communicate uncertainty. They merely recognized there is no perfect method for achieving this goal. They note from their experiments with the IPCC scales that risk communication is most accurate when one uses a combination of probabilistic terms as well as numeric expressions.

One of the more relevant articles for our analysis examined how medical probability statements were used in court. Merz, Druzdzel, and Mazur (2011) looked at 55 court opinions concerning informed consent between 1951 and 1989 where a verbal probability statement was used by a physician at a civil trial on liability and included a quantitative estimate of the probability.[11] They found large variation in some of the probability terms. Even so, they argued jurors benefit from using numeric figures and visual aids designed to help them understand the terms. They also note that since patients have different levels of literacy, varying methods should be employed to communicate to a diverse group of patients.

Clearly, there is debate on the efficacy of trying to convert probability statements into numeric estimates. In the context of this article, does this type of translation help jurors, judges, and others understand issues surrounding burden of proof?

BURDEN OF PROOF

Turning to the legal arena, social science research regarding the effect of different burden of proof instructions on deliberations has been mixed.[12] Kagehiro and Stanton (1985) and MacCoun and Kerr (1988) found, using students in mock criminal trials, that modifying the legal definition of the burden of proof had no effect on the verdict. However,

▶

when Kagehiro and Stanton (1985) used different quantified burdens of proof statements (such as 51 percent, 71 percent, or 91 percent), it had the expected effect: Higher standards led to greater acquittals. Similarly, Kerr, et al (1976), using a sample of 645 students, found that varying the definition of the reasonable doubt had the expected effect.[13] They also found, however, that the students had a substantial lack of understanding of the basic concept, as exhibited in the large variability of the scores, which resulted in greater hung juries.

Horowitz and Kirkpatrick (1996) used 480 jury-eligible adults from the community who were assigned to six-person juries.[14] They first noted that the respondents quantified the prosecution's burden of proof at 61 percent, which is below that of other studies. Nevertheless, the study found different reasonable doubt instructions influenced jurors and their deliberations. Horowitz and Kirkpatrick were particularly concerned that jurors were convicting defendants when the evidence was weak and the reasonable doubt instruction favored acquitting the defendants. They hypothesized that the reasonable doubt instructions they used might be ineffective, or over time the threshold of proof found acceptable by jurors had decreased.

Rita Simon (1969) and Simon and Mahan (1971) conducted empirical studies in the late 1960s and early 1970s with students, jurors, and judges where they were asked to assign probabilities to "burden of proof" statements.[15] Their study of judges was a mail survey of 1,200 state and federal court judges (33 percent responded). The studies also surveyed 69 jurors, who served in Champaign County Court (Illinois) and participated in mock jury deliberations, and 88 students taking sociology classes. The authors argued that there were no noteworthy differences among the three groups on "beyond reasonable doubt." Judges reported a mean of 8.9 out of 10, jurors reported 7.9, and students reported 8.9. They also found what they argued to be a considerable difference among the three groups on "preponderance of the evidence," where the reported scores were 6.1, 7.3, and 7.3, respectively.[16] In other findings, they determined that elected judges had more stringent standards than appointed judges. In questions concerning 14 types of crimes defendants could be charged with (murder, forgery, etc.), they asked judges what "the probability that the defendant committed the act would have to be before you declared him guilty?" They found only slight differences among the various crimes, and the results were consistent with the overall 8.9 level of certainty mentioned above.

McCauliff (1982) conducted a similar study of 195 federal judges.[17] She sent questionnaires to all federal district and circuit court judges as well as justices of the United States Supreme Court. She asked judges to scale nine "burden of proof" questions. Unlike Simon and Mahan, who used scales ranging from 0 to 10, McCauliff used scales ranging from 0 to 100. On reasonable doubt, the average was 90.3 percent, which was very close to Simon's 8.9 (89 percent in McCauliff's scale). She concluded that the current verbal standards were confusing and ambiguous because judges apply different burdens of proof to similar issues and, moreover, interpret the same burden of proof in very different manners. She also noted that although the application of percentages is perhaps not appropriate for use in jury instructions, the results are informative for decision-makers, as it might promote uniformity.

> “Saunders argues that the reasonable doubt standard — and other similarly vague standards — must be quantified to conform to the requirements of equal protection. He argues that two different juries viewing the same set of facts could deliver different verdicts simply because they interpret standards of proof differently. He proposes the use of quantifying standards to reduce this problem.

Offering a counterpoint, Saunders (2005) argues that the reasonable doubt standard — and other similarly vague standards — must be quantified to conform to the requirements of equal protection.[18] He argues that two different juries viewing the same set of facts could deliver different verdicts simply because they interpret standards of proof differently. He proposes the use of quantifying standards to reduce this problem. Saunders provided probability training to respondents and then conducted a survey of 130 college-educated business professionals that revealed a substantial range of responses. However, although the responses ranged from 50 to 99.9 percent, it appears that the mean and standard deviation* was similar to what we and other researchers have found.[19]

THE PRESENT STUDY AND DISCUSSION

BACKGROUND

The present study surveyed judges on the following six standards: 1) reasonable articulable suspicion; 2) probable cause; 3) preponderance of the evidence; 4) substantial probability; 5) clear and convincing evidence; and 6) beyond a reasonable doubt. The standards range from requiring a relatively small amount of evidence to enough evidence to be almost certain. There is a direct correlation between the amount of evidence required and the consequences for error. Essentially, the greater the potential cost of error, the higher the requirement for the burden of proof. For example, in an ordinary civil case where only money is at stake, the standard of proof is preponderance of the evidence. However, as the risks associated with a wrong result increase, so, too, does the required amount of evidence. Thus, when a person's liberty is at stake, the higher standard of proof beyond a

* *Standard deviations tell how much the typical data point differs from the mean.*

reasonable doubt is required to minimize the risk of erroneous convictions. These six standards arise in different contexts, as further described below. Generally speaking, laws — including the United States Constitution, case law from the United States Supreme Court and various lower courts, and statutory law — impose limits on actions that government actors can take, and courts enforce these limits using these various evidentiary standards.

Reasonable Articulable Suspicion

Reasonable articulable suspicion emerged as a standard after *Terry v. Ohio*, 392 U.S. 1 (1968), in which the Supreme Court upheld a police officer's limited search and frisk of a person based on less than probable cause. Reasonable articulable suspicion is commonly used in trial court hearings regarding motions to suppress evidence in a criminal case. The court evaluates in hindsight, based on the facts known at the time, whether there was sufficient requisite suspicion for the officer to conduct a brief, investigatory stop of a person or frisk them for weapons.

Probable Cause

Since reasonable articulable suspicion was developed as a less stringent standard than probable cause, the two remain linked, and reasonable articulable suspicion is often described as ripening into probable cause upon discovery of further facts. Courts evaluate probable cause, also in hindsight, as a predicate to arrest or search, analyzing whether it was reasonable for police to believe a person committed a crime or that fruits of a crime will be found in a particular place. Probable cause is the standard used by courts to rule on motions to suppress evidence based on a warrantless arrest or search and to assess warrant applications prior to a search or arrest.

In describing reasonable articulable suspicion and probable cause, the Supreme Court has noted that they "are not 'finely-tuned standards'" but "are instead fluid concepts that take their substantive content from the particular contexts in which the standards are being assessed." *Ornelas v. United States*, 517 U.S. 690, 696 (1996) (citations omitted). Indeed, the Supreme Court has explained that:

> "[t]he probable-cause standard is incapable of precise definition or quantification into percentages because it deals with probabilities and depends on the totality of the circumstances. We have stated, however, that '[t]he substance of all the definitions of probable cause is a reasonable ground for belief of guilt,' and that the belief of guilt must be particularized with respect to the person to be searched or seized." *Maryland v. Pringle,* 540 U.S. 366, 371 (2003) (citations omitted).

In addition to being somewhat amorphous, probable cause as a term of art is simply confusing. Highlighting the linguistic irony in the label "probable cause," the District of Columbia Court of Appeals essentially classified probable cause as a misnomer: "We have described probable cause as a 'flexible, common sense standard,' which 'does not demand any showing that the officer's belief that he has witnessed criminal behavior be correct or more likely true than false.' Linguistically, this definition is somewhat perplexing, for it is not easy to discern how cause can be probable if the officer's belief that the defendant committed a crime is not 'more likely true than false.'" *Pope v. United States*, 739 A.2d 819, 828 n. 21 (1999) (citations omitted). This misleading terminology may help to explain judges' apparent conflation of probable cause and preponderance of the evidence, as discussed below.

Substantial Probability

In the criminal context, once a person is arrested, one of the key factors in the decision as to whether to detain that person pretrial is whether there is a substantial probability the person committed a dangerous crime. *See* D.C. Code § 23-1322; *Blunt v. United States*, 322 A.2d 579 (D.C. 1974). The standard has been defined by the District of Columbia Court of Appeals as follows: "[A] substantial probability is a degree of proof meaningfully higher than probable cause, intended in the pre-trial detention statute to be equivalent to the standard required to secure a civil injunction — likelihood of success on the merits. We have cautioned against equating substantial probability with the clear and convincing evidence standard." *Blackson v. United States*, 897 A.2d 187, 196 n. 16 (D.C. 2006) (citations omitted).[20] As the court indicated, in the noncriminal context, the substantial probability standard governs whether a court should issue an injunction, which is an order to compel or stop a person from doing a specific act, such as ceasing publication where there is copyright infringement. The substantial probability standard is also used in deciding whether to close criminal trial proceedings to the public and in evaluating the length of time an incompetent accused may be committed. *Press-Enterprise Co. v. Superior Court*, 478 U.S. 1, 14 (1986); *Florida v. Garrett*, 454 U.S. 1004, 1007 (1981).

The prior three standards are used by judges alone because they are used in the pretrial context. The next three — preponderance of the evidence, clear and convincing evidence, and beyond a reasonable doubt — are standards used at the conclusion of a trial by the fact-finder, either a judge or a jury, to determine the outcome.

Preponderance of the Evidence

Preponderance of the evidence is the standard most commonly used in civil cases and means evaluating whether the issue is more likely than not to have occurred. Commonly referred to as anything over 50 percent, some courts have also referenced 50.1 percent as the least amount of evidence required to sustain a verdict under the preponderance of the evidence standard. (*Brown v. Greene,* 577 F.3d 107, 109 n.2 (2d Cir. 2009).)

Clear and Convincing Evidence

Clear and convincing evidence is an intermediate standard of proof that requires more evidence than preponderance of the evidence but less than beyond a reasonable doubt. It is used where the risk related to a wrong decision is substantial. For example, many state fraud cases, medical decisions that affect an incompetent patient, and termination of parental rights all require clear and convincing evidence.[21] In criminal cases, the clear and convincing evidence standard is used in the pretrial detention context. *Pope v. United States*, 739 A.2d 819, 825 (D.C. App. 1999).

▶

TABLE 1. COMPARISONS WITH PREVIOUS STUDIES

	Current Study		Simon		McCauliff	
	Mean	SD	Mean	SD	Mean	SD
Beyond Reasonable Doubt	90.1	7.0	88.9	10.7	90.8	6.8
Reasonable Articulable Suspicion	42.1	22.5			*30.9	15.1
Probable Cause	49.7	16.6			**45.5	12.8
Substantial Probability	55.3	17.6				
Preponderance of the Evidence	54.4	5.4	61.3	11.2	56.0	10.5
Clear and Convincing	73.4	10.6				

** McCauliff used the phrase "reasonable suspicion"*

*** McCauliff used the phrase "probable cause to believe"*

Beyond a Reasonable Doubt

Finally, beyond a reasonable doubt is the standard used by judges and juries in criminal trials to determine whether the government has met its burden to prove the accused guilty of a crime. There is no uniform standard for reasonable doubt, though many jurisdictions' instructions are similar (see Corwin, 2001, and Dumas, 2002). For example, the instruction used in the District of Columbia reads, in part:

> Reasonable doubt, as the name implies, is a doubt based on reason — a doubt for which you have a reason based upon the evidence or lack of evidence in the case. If, after careful, honest, and impartial consideration of all the evidence, you cannot say that you are firmly convinced of the defendant's guilt, then you have a reasonable doubt. Reasonable doubt is the kind of doubt that would cause a reasonable person, after careful and thoughtful reflection, to hesitate to act in the graver or more important matters in life. However, it is not an imaginary doubt, nor a doubt based on speculation or guesswork; it is a doubt based on reason. The government is not required to prove guilt beyond all doubt, or to a mathematical or scientific certainty.[22]

The Maryland instruction states, in part:

> Proof beyond a reasonable doubt requires such proof as would convince you of the truth of a fact to the extent that you would be willing to act upon such belief without reservation in an important matter in your own business or personal affairs.[23]

These instructions are examples of how courts articulate the reasonable doubt standard.[24]

METHODOLOGY

Between December 2007 and April 2012, 124 judges filled out a survey in which they were asked to translate six legal standards into probability statements using five-point increments (see Questionnaire on page XX). In order to make comparisons, we tried to make our questions similar to that of Simon and Mahan.[25] The judges constituted five different groups of predominantly state-court trial judges at various training conferences throughout the country sponsored by the National Judicial College. The use of group self-administered surveys resulted in a response rate approaching 100 percent.

In our survey we asked judges how long they had served on the bench, whether they were appointed or elected, whether they served in state or federal court, and whether their case load was primarily criminal or civil. In the analysis below, we examine whether these factors influenced the judges' probability statements.

RESULTS

The average number of years on the bench for the 124 judges was 11.7 (SD=8.0). Three judges served on U.S. District Court, and the rest served on state trial courts. Thirty-nine percent of the judges were appointed, and the remainder were either elected or faced an election after their initial appointment. Forty percent of the judges' caseloads were primarily criminal. The remainder were civil (4.8 percent), family (4.8 percent), or mixed (47.6 percent).

The means and standard deviations for the six probability statements are in Table 1 (left). The means, from lowest to highest are as follows: reasonable articulable suspicion (42.1 percent), probable cause (49.7 percent), preponderance of the evidence (54.4 percent), substantial probability (55.3 percent), clear and convincing evidence (73.4 percent), and beyond a reasonable doubt (90.1 percent). The means and standard deviations for two questions from the study conducted by Simon (1969) and four questions from McCauliff (1982) are also noted in Table 1.[26] The percentage breakdown for these six questions is in Table 2 (next page).

Comparison with Previous Studies

As discussed above, the Simon & Mahan study and the McCauliff study examined four of the six burdens of proof assessed in our survey.

Beyond a Reasonable Doubt: The results for beyond a reasonable doubt were essentially identical across the three studies, averaging around 90 percent. These differences were not statistically significant.[27]

Preponderance of the Evidence: There was no statistically significant difference between our study and McCauliff's on preponderance of the evidence (54.4 v 56.0; z=1.79). However, Simon's results were significantly higher (54.4 v 61.3; z=11.0).

Reasonable Articulable Suspicion & Probable Cause: There were statistically significant differences between our study and that of McCauliff on reasonable articulable suspicion (42.1 v 30.9; z=4.89) and probable cause (49.7 v 45.5; z=2.4).

Although some of these differences in probability statements across the three studies were statistically significant, the results were remarkably stable given the 30- to 40-year passage of time and the different population of judges. Indeed, across all three studies, the only difference greater than 10 percent was for preponderance of the evidence, where our study had the lowest score (54.4 percent) and

Simon's had the highest (61.3 percent) with McCauliff in between (56.0 percent) (54.4-61.3=6.9; or 12.6 percent).

FACTORS AFFECTING THE JUDGES' RESPONSES

We used three factors to try to predict the responses of the judges. T-tests were used to test statistical significance.

Length of Tenure

The factor length of tenure separated judges into two categories: Judges who had served for less than ten years and those who had served longer. Tenure affected only one of the scales. Judges who had served less than ten years were more likely to give a lower score for substantial probability (51.4 v 59.4; p=.02).

Appointed or Elected

Whether judges were appointed or elected (elected included elected after being appointed) had no impact on the six scales. This result contrasts with Simon (1969), who found that elected judges gave a lower score on the reasonable doubt scale (they were more likely to convict) than judges who were appointed. Simon believed this trend was consistent with the notion that elected judges are more responsive to public opinion than appointed judges.

Type of Cases

There was only one difference between judges who heard primarily criminal cases and other judges. Judges who heard primarily criminal cases had somewhat lower scores for the clear and convincing standard (70.7 v 75.1; p=.03).

In essence, our ability to predict the scores of the judges given the three variables discussed above was limited at best.[28]

QUANTIFICATION AS USEFUL TOOL OR A "SEARCH FOR FOOL'S GOLD"?

There is robust debate about the role of statistics and mathematics in the administration of justice. This debate has centered on the validity and interpretation of evidence, including DNA, eyewitness identification, blood types, and discrimination.[29] Laurence Tribe (1971) wrote a seminal article over 40 years ago critiquing this trend.[30] He argued that mathematics at trial has the potential to help reveal the truth but can overwhelm the other evidence. In addition, he critiqued the use of mathematical models (primarily Bayesian) to directly determine a verdict. This article set off a storm of debate. One critic of Tribe noted how mathematics requires the user to think rigorously and precisely.[31] More recently, however, Tillers characterized this debate as "unproductive and sterile."[32]

There has been some discussion in the judiciary about quantification of proof. Judge Jack B. Weinstein of the Eastern District of New York addressed the issues regarding quantification of burdens of proof and noted, "An attempt to quantify in order to provide some uniformity in application of the rule is justified even though it must be conceded that the percentage chosen is based on public policy favoring enforcement of constitutional rights and somewhat arbitrary."[33]

In *United States v. Fatico*,[34] Judge Weinstein again discussed burdens of proof and the necessary level of certainty, citing the Simon study and others. The court discussed at length historical attempts to quantify the burdens of proof, noting variance in such quantifications. In his opinion, Judge Weinstein included the results of his own survey of the ten judges of the Eastern District of New York, which listed responses for the beyond a reasonable doubt standard as ranging between 76 percent and 95 percent. *Id.* at 410.

With respect to the appropriateness of quantification, Judge Posner commented, "Numerical estimates of probability are helpful in investments, gambling, scientific research, and many other activities but are not likely to be helpful in the setting of jury deliberations. . . . It is one thing to tell jurors to set aside unreasonable doubts, another to tell them to determine whether the probability that the defendant is guilty is more than 75, or 95, or 99 percent." *United States v. Hall*, 854 F.2d 1036, 1044-45 (7th Cir. 1988) (Posner, J., concurring).

As discussed at the beginning of this article, the District of Columbia Court of Appeals dealt with quantification of standards of proof when it reversed a conviction on the grounds of insufficient evidence based on a single eyewitness identification where the witness testified her certainty level was a seven or eight on a scale of one to ten. *In re As.H.*, 851 A.2d 456 (D.C. 2004). In the dissenting opinion, Judge Michael W. Farrell

▶

TABLE 2. PERCENTAGE BREAKDOWN OF SIX SCALES

	RD	AS	PC	SP	PE	CC
0-10%	0	8.3	0	1.8	0	0
11-20%	0	14.2	4.1	0.9	0	0
21-30%	0	18.3	15.4	11.4	0	0.8
31-40%	0	12.5	13.0	9.6	0.8	1.6
41-50%	0	6.7	16.3	10.5	12.9	0
51-60%	0.8	22.5	32.5	28.1	81.4	8.2
61-70%	0	5.0	7.3	21.1	0.8	22.1
71-80%	13.3	9.2	9.8	14.9	4.0	53.3
81-90%	41.7	2.5	0	0.9	0	12.3
91-100%	44.2	0.8	1.6	0.9	0	1.6

RD: Beyond Reasonable Doubt
AS: Reasonable Articulable Suspicion
PC: Probable Cause
SP: Substantial Probability
PE: Preponderance of the Evidence
CC: Clear and Convincing

criticized the reasoning of the majority and its reliance on the quantification studies cited. The use of quantification of burdens of proof has been percolating in academic scholarship for decades without much adoption by the judiciary. It is beyond the scope of this article to enter this debate. Instead, we are looking at how judges view burdens of proof and the extent to which these interpretations might aid those who use burdens of proof in their decision-making, i.e. judges and jurors. The question remains — and we do not address — whether, as Judge Farrell posits, such quantification is "a search for fool's gold."

CONCLUSION

This study presents four major findings: First, with the exception of the question on preponderance of the evidence, our results are remarkably similar to those found in the studies by Simon & Mahan and McCauliff. We realize that caution is called for in comparing the various studies, given differing methodologies and the fact that our sample is best described as a convenience sample. Nevertheless, the evidence suggests that judges' quantifications of burdens of proof have been fairly stable over time.

Second, characteristics of the judges (length of tenure, appointed versus elected, and type of cases) had little impact on how they interpreted the six legal standards.

Third, some of the scales in our survey showed substantial variability. Though the means for the six standards reflected the expected upward numerical progression from lower risk requiring lower certainty, the averages do not reflect the entirety of responses for three of the standards. There was very high variability on reasonable articulable suspicion ($\bar{x}$=42.1, SD=22.5), substantial probability ($\bar{x}$=55.3, SD=17.6), and probable cause ($\bar{x}$=49.7, SD=16.6). These are the three standards used exclusively by judges, and the high standard deviations reveal a lack of uniform application of these legal standards. This has not been previously discussed in the literature and is of concern. Practically speaking, this means that the standards used for bail decisions and Fourth Amendment rulings — including the reasonableness of stops, frisks, searches, and arrests — are not consistent across judges, and the same facts will yield different results based on the ruling judge's quantification of the standard.

Among the standards that are used to determine the outcome of cases, there was more consistency: Clear and convincing evidence ($\bar{x}$=73.4, SD=10.6), preponderance of the evidence ($\bar{x}$=54.4, SD=5.4), and beyond a reasonable doubt ($\bar{x}$=90.1, SD=7.0). Therefore, there appears to be greater uniformity among judges interpreting standards commonly used by juries.

QUESTIONNAIRE SENT TO JUDGES

In every jurisdiction in the United States, the *burden of proof* necessary to convict a defendant in a criminal trial is that the defendant's guilt must be established *beyond a reasonable doubt*. We would like you to do the following: Translate the phrase beyond a reasonable doubt into a statement of probability.

What level of certainty must you have to find a criminal defendant guilty *Beyond a Reasonable Doubt*?

— — — — — — — — — — — — — — — — — — — —
100% 90 85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0%

What level of certainty must you have to find *Reasonable Articulable Suspicion* in a Terry analysis?

— — — — — — — — — — — — — — — — — — — —
100% 90 85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0%

What level of certainty must you have to find *Probable Cause* to arrest at a motions hearing?

— — — — — — — — — — — — — — — — — — — —
100% 90 85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0%

What level of certainty must you have to find *Substantial Probability* that defendant committed an offense in a bail hearing?

— — — — — — — — — — — — — — — — — — — —
100% 90 85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0%

What level of certainty must you have to find *Preponderance of the Evidence* at a civil trial?

— — — — — — — — — — — — — — — — — — — —
100% 90 85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0%

What level of certainty must you have to find that evidence is *Clear and Convincing*?

— — — — — — — — — — — — — — — — — — — —
100% 90 85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0%

The fourth significant finding in this study is the similarity of the judges' quantifications of preponderance of the evidence and probable cause (54.4 percent for preponderance of the evidence and 49.7 percent for probable cause). The legal difference between these two standards is significant. Preponderance of the evidence requires a finding of more likely than not, whereas probable cause is a lower standard that requires reasonable grounds to believe. The judges' quantifications seem to conflate these two standards, which may have serious legal consequences.

Judges' variability in quantifications of reasonable articulable suspicion, substantial probability, and probable cause, and their mistaken conflation of probable cause and preponderance of the evidence, indicate a lack of consensus for standards that ought to be understood and applied in a consistent manner. This problem is important to highlight to ensure the correct thresholds are being employed at the proper times. Thus, while some may critique the use of quantifications as a search for fool's gold, this study bolsters the work by Simon & Mahan and McCauliff and further establishes that the quantification of legal concepts provides unique insight into judicial decision-making.

[1] Rita James Simon & Linda Mahan, *Quantifying Burdens of Proof: A View from the Bench, the Jury, and the Classroom*, 5 L. Soc'y Rev. 319 (1971).

[2] David V. Budescu & Thomas S. Wallsten, *Processing Linguistic Probabilities: General Principles and Empirical Evidence*, in Decision Making from a Cognitive Perspective 275–318 (Jerome Busemeyer, Reid Hastie and Douglas L. Medin ed., 1995).

[3] Danielle Timmermans & Juriaan Oudhoff, *Different Formats for the Communication of Risk: Verbal, Numerical and Graphical Formats*, in Wiley Encyclopedia of Operations Research & Management Science (2011); Thomas S. Wallsten et al., *Preferences and Reasons for Communicating Probabilistic Information in Verbal or Numerical Terms*, 31 Bull. Psychonomic Soc'y 135, 135–38 (1993).

[4] Mandana Vahabi, *The Impact of Health Communication on Health-Related Decision Making: A Review of Evidence*, 107 Health Educ. 27 (2007).

[5] Tim M. Gale et al., *Do Mental Health Professionals Really Understand Probability? Implications for Risk Assessment and Evidence-Based Practice*, 12 J. Mental Health 417 (2003); Daniel Kahneman, Maps of Bounded Rationality: A Perspective on Intuitive Judgment and Choice, Nobel Prize Lecture (Dec. 8, 2002).

[6] Wandi Bruine de Bruin et al., V*erbal and Numerical Expressions of Probability: "It's a Fifty-Fifty Chance"*, 81 Organizational Behav. & Hum. Decision Processes 115 (2000).

[7] *See generally* Ruth Beyth-Marom, *How Probable is Probable? A Numerical Translation of Verbal Probability Expressions*, 1 J. Forecasting 257 (1982); Budescu & Wallsten, *supra* note 2; Valerie A. Clarke et al., *Ratings of Orally Presented Verbal Expressions of Probability by a Heterogeneous Sample*, 22 J. of Applied Soc. Psychol. 638 (1992); Michael Theil, *The Role of Translations of Verbal into Numerical Probability Expressions in Risk Management: A Meta-Analysis*, 5 J. Risk Res. 177 (2002); Dennis J. Mazur & Jon F. Merz, *Patients' Interpretations of Verbal Expressions of Probability: Implications for Securing Informed Consent to Medical Interventions*, 12 Behav. Sci. & L. 417 (1994); Dennis J. Mazur & D.H. Hickman, *Patient Interpretations of Terms Connoting Low Probabilities When Communicating About Surgical Risk*, 8 Theoretical Surgery 143 (1993); Geoffrey. D Bryant & Geoffrey R. Norman, *Expressions of Probability: Words and Numbers*, 302 New Eng. J. Med. 411 (1980).

[8] One of the more interesting criticisms on the use of numeric probabilities comes from Dr. Richard Feynman, the Nobel Prize-winning physicist who criticized the final report of the Rogers Commission on the Challenger disaster, of which he was a member. He notes that NASA took engineers' subjective judgments of risk and attempted to quantify them. By so doing, they ignored historical data and divorced the analysis from any scientific basis. *See* Eliot Marshall, *Feynman Issues his Own Shuttle Report, Attacking NASA's Risk Estimates*, 232 Science 1596 (1986). This type of assessment resulted in projections of risk that were absurdly low.

[9] *See* Augustine Kong et al., *How Medical Professionals Evaluate Expressions of Probability*, 315 New Eng. J. Med. 740, 740–45 (1986); Michael A. Nakao & Seymour Axelrod, *Numbers are Better than Words*, 74 Am. J. Med. 1061, 1061–65 (1983).

[10] David V. Budescu, Han-Hui Por, & Stephen B. Broomell, *Effective Communication of Uncertainty in the IPCC*, 113 Climate Change 181 (2012).

[11] Jon F. Merz, Marek J. Druzdzel and Dennis J. Mazur, *Verbal Expressions of Probability in Informed Consent Litigation*, 11 Med. Decision Making 273 (1991).

[12] Dorothy K. Kagehiro & W. Clark Stanton, *Legal vs. Quantified Definitions of Burden of Proof*, 9 L. & Hum. Behav. 159 (1985); Robert J. MacCoun & Norbert L. Kerr, *Asymmetric Influence in Mock Jury Deliberation: Jurors' Bias for Leniency*, 54 J. Personality & Soc. Psychol.

[13] Norbert L. Kerr et al., *Guilt Beyond a Reasonable Doubt: Effect of Concept Definition and Assigned Decision Rule on the Judgments of Mock Jurors*, 34 J. Personality & Soc. Psychol. 282 (1976). Mock jurors were told that reasonable doubt was either (1) essentially any doubt, (2) substantial doubt, or (3) not given any definition. Those given the first standard were least likely to convict.

[14] 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 Verdict*, 20 L. & Hum. Behav. 655 (1996).

[15] Rita Simon, *Judge's Translations of Burdens of Proof into Statements of Probability*, Trial Law. Guide 29 (1969); Simon & Mahan, *supra* note 1.

[16] It could be argued that the difference between 8.9 and 7.9 (judges and students versus jurors on reasonable doubt) is not "little" and is almost the same as the difference on preponderance of the evidence (6.1 and 7.3).

[17] C.M.A. McCauliff, *Burdens of Proof: Degrees of Belief, Quanta of Evidence, or Constitutional Guarantees*, 35 Vanderbilt L. Rev. 1293 (1982).

[18] Harry D. Saunders, Quantifying Reasonable Doubt: A Proposed Solution to an Equal Protection Problem 1–23 (2005).

[19] Saunders did not report the mean or standard deviation, and it is not possible to accurately compute these given the graph provided.

[20] In the District of Columbia, the clear and convincing evidence standard is also used in determining whether to hold a person in jail without bond awaiting trial. D.C. Code § 22-1322.

[21] *See, e.g.*, *Avery v. State Farm Mut. Auto. Ins. Co.*, 216 Ill. 2d 100, 192 (Ill. 2005); *Cruzan v. Mo. Dep't of Health*, 497 U.S. 261, 282–83 (1990); *Santosky v. Kramer*, 455 U.S. 745 (1982).

[22] Criminal Jury Instructions for the District of Columbia, 2.108 (2013).

[23] Maryland Criminal Pattern Jury Instructions, 2:02 (2012).

[24] An example of an instruction used in federal court is as follows: ". . . Proof beyond a reasonable doubt means proof which is so convincing that you would not hesitate to rely and act on it in making the most important decisions in your own lives. If you are convinced that the govern-

▶

ment has proved the defendant guilty beyond a reasonable doubt, say so by returning a guilty verdict. If you are not convinced, say so by returning a not guilty verdict." Pattern Criminal Jury Instructions for the Sixth Circuit, 1.03 (2013), *available at* http://www.ca6.uscourts.gov/internet/crim_jury_insts/pdf/07_Chapter_1.pdf.

[25] For issues on phrasing questions on reasonable doubt, *see* Mandep. K. Dhami, *On Measuring Quantitative Interpretations of Reasonable Doubt*, 14 J. Experimental Psychol.: Applied 353 (2008).

[26] Simon and McCauliff did not include standard deviations. We were able to compute them given their frequency distributions.

[27] Using a two-sample z-test** for comparing two means ($p<.05$).

[28] With 18 separate t-tests**, we would expect approximately one difference to be statistically significant with random data. Having two that are statistically significant with no a-priori hypotheses would suggest that there were no differences.

[29] David McCord, *Primer for the Nonmathematically Inclined on Mathematical Evidence in Criminal Cases: People v. Collins And Beyond*, 47 Wash. & Lee L. Rev. 741 (1990).

[30] Lawrence H. Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 Harv. L. Rev. 1329 (1971).

[31] Peter Tillers, *Symposium: Probability and Inference in the Law of Evidence: Introduction*, 66 B.U. L. Rev. 381, 386 (1986).

[32] Peter Tillers, *Trial by Mathematics—Reconsidered*, 10 L., Probability & Risk 167 (2011).

[33] *United States v. Copeland*, 369 F. Supp. 2d 275, 344 (E.D.N.Y. 2005) (determining on remand that defendant failed to demonstrate a reasonable probability that he would have obtained a waiver of deportation in the absence of constitutional error).

[34] 458 F. Supp. 388, 411 (E.D.N.Y. 1978).

** *T-test and z-scores are used in tests of significance. Tests of significance state the probability that a result is a function of chance.*