Surprise vs. Probability as a Metric for Proof

Matthew Ginther[*] *& Edward K. Cheng*[**]

I. INTRODUCTION

In this Symposium issue celebrating his career, Professor Michael Risinger in *Leveraging Surprise* proposes using "the fundamental emotion of surprise" as a way of measuring belief for purposes of legal proof.[1] More specifically, Professor Risinger argues that we should not conceive of the burden of proof in terms of probabilities such as 51%, 95%, or even "beyond a reasonable doubt."[2] Rather, the legal system should reference the threshold using "words of estimative surprise"[3]—asking jurors how surprised they would be if the fact in question were *not* true. Toward this goal (and being averse to cardinality), he suggests categories such as "mildly surprised, surprised, quite surprised, greatly surprised, astonished, shocked, etc."[4]

We find Professor Risinger's proposal intriguing. After all, one can imagine important theoretical reasons why surprise might generate different results from probability. To the extent that the surprise formulation is

[*] Law Clerk to the Honorable Christian J. Moran, United States Court of Federal Claims, Office of Special Masters.

[**] FedEx Research Professor (2017–18) and Professor of Law, Vanderbilt Law School. Thanks to Dale Nance for helpful comments, as well as Michael Risinger and participants at the Symposium on Experts, Inference and Innocence at Seton Hall Law School in October 2017. All recruitment and experimental procedures were approved by the Vanderbilt Institutional Review Board (IRB # 081408).

[1] D. Michael Risinger, *Leveraging Surprise: What Standards of Proof Imply that We Want from Jurors, and What We Should Say to Them to Get It*, 48 SETON HALL L. REV. 965 (2018).

[2] David H. Kaye, *Clarifying the Burden of Persuasion: What Bayesian Decision Rules Do and Do Not Do*, 3 INT'L J. EVIDENCE & PROOF 1 (1999).

[3] Risinger, *supra* note 1.

[4] Risinger, *supra* note 1.

unfamiliar, it might cause jurors to think holistically ("System 1") as opposed to attempting to use rules (often misremembered or misapplied) about probability ("System 2").[5] Surprise might be easier to approach qualitatively, unlike probability, which cries out for quantitative calculation and invokes the fear of numbers for some. Surprise is also notably framed in the negative ("How surprised would you be if the fact were *not* true?") compared to its probability counterpart ("What is the probability that the fact is true?").

Being empiricists, we thus could not help but put Professor Risinger's worthy proposal to the test, if only in a preliminary way. Just what might conceptualizing evidence under a framework of "surprise" look like and get us in practice? Here, we report on a simple experiment where potential jurors were recruited to evaluate evidence using both a surprise framework and a framework using probabilistic language. Using the experiment, we set out to answer two questions: First, does using degrees of "surprise" actually produce intelligible and consistent results among laypersons? And second, does using degrees of "surprise" to evaluate evidence produce results that provide a material benefit, either empirically or normatively, when compared to frameworks based on probability?

II. DESIGN AND METHODS

For our study, we used an online platform, which allowed us to recruit a large sample of individuals to evaluate scenarios under both the proposed surprise and the traditional probability frameworks. Because it was likely that the differences between probabilistic and surprise-based evidence evaluation interact with the weight of the evidence, we wrote scenarios where the weight of the evidence could be set to different levels.

To ensure that the subjects' responses were not influenced by their answers to other questions, we used a between-subjects experimental design: Each subject evaluated only a single scenario and provided a response using either a probabilistic or a surprise framework, but not both.

Seven scenarios were written for the purpose of this study. Each involved the same basic fact pattern in which "Bob" was charged with the murder of his co-worker and friend "James." However, the scenarios varied with regard to the weight of the evidence against "Bob." The scenarios were written so as to impress upon the reader different levels of culpability, ranging from virtually certain innocence ("Evidence Level 1") to virtually certain guilt ("Evidence Level 7"). The seven scenarios are reproduced in Table A below.

[5] See DANIEL KAHNEMAN, THINKING, FAST AND SLOW (2011) for a comprehensive discussion of this dual-process framework of human psychology.

TABLE A: THE SCENARIOS

	Scenario
Evidence Level 1	Bob has been charged with the murder of his co-worker and friend, James. A security camera shows both of them leaving a sporting event together shortly before James's death. Bob denied the charges from the beginning, reporting that they had gone their separate ways soon after. The murder weapon, a bloody knife, was found stuck in James's body. DNA testing determined that the blood on the knife was a mixture of James and another male, but was not a match to Bob.
Evidence Level 2	Bob has been charged with the murder of his co-worker and friend, James. James was found dead of a gunshot wound in his office and investigators received an anonymous call reporting that Bob and James had recently been feuding. Police found a gun in Bob's car, but later determined that Bob had a permit to carry the gun, and the bullet that killed James was not fired from it. Nonetheless, witnesses can corroborate that Bob and James had been feuding.
Evidence Level 3	Bob has been charged with the murder of his co-worker and friend, James. James was found strangled to death in his own home. Bob is the last person known to have visited James the night before James was murdered. There was no physical evidence found at the scene of the crime and neighbors reported not hearing any noise or activity in James' house after Bob left.
Evidence Level 4	Bob has been charged with the murder of his co-worker and friend, James. A witness reports that Bob and James had a loud argument at a local bar close in time to James's approximate time of death.
Evidence Level 5	Bob has been charged with the murder of his co-worker and friend, James. James was last seen accompanying Bob on a camping trip. Bob reports that the trip was uneventful and that at the end of the trip James drove home. However, no trace of James or his vehicle has ever been found.

	Bob has faced some recent financial troubles, and had disagreed with James on whether to sell a jointly owned apartment building. Upon James's death, Bob became the sole owner of the apartment building.
Evidence Level 6	Bob has been charged with the murder of his co-worker and friend, James. A witness reports that on the night of the murder, Bob and James had a loud argument at a local bar over an affair that James had been having with Bob's wife. Another witness states that she heard the sound of gunshots minutes after the time a 911 call reporting the fight was placed.
Evidence Level 7	Bob has been charged with the murder of his co-worker and friend, James. A witness reports that on the night of the murder, Bob and James had a loud argument at a local bar over an affair that James had been having with Bob's wife. Video surveillance outside the bar shows Bob shooting James. Bob was apprehended fleeing from the scene by an off-duty officer, who found the murder weapon in Bob's waistband.

After evaluating the scenario, subjects were asked two independent questions. First, subjects were asked to provide a numerical evaluation of the strength of the evidence. This first question was framed either in terms of the subjects' estimated probability or the subjects' experience of surprise. Because probability and surprise can pertain to both the defendant's guilt and innocence, we framed the question both ways for each. This resulted in four possible questions that the subject could be asked to evaluate. Which question was presented to any given subject was determined randomly. The four possible questions are presented in Table B.

TABLE B: THE MEASURES

Measure	***Prompt***
p(Guilty)	Based on what is known, what do you believe is the probability (expressed as a percentage) that Bob murdered James?

p(Not Guilty)	Based on what is known, what do you believe is the probability (expressed as a percentage) that Bob **DID NOT** murder James?
surprise(Guilty)	Based on what is known, how surprised would you be to find out that Bob murdered James?
surprise(Not Guilty)	Based on what is known, how surprised would you be to find out that Bob **DID NOT** murder James?

Subjects provided the response to this numerical question on a 101-point Likert scale. We understand that Professor Risinger's proposal expressed a desire to eliminate all "unwarranted cardinality" from representations of evidence,[6] and thus use of this numerical Likert scale may be inconsistent with his vision. For the purposes of this study, however, we wanted to be able to make meaningful comparisons between the probability framework and the surprise framework. We therefore used the 0-100 scale as a matter of convenience. Future research may want to consider different approaches. For example, one might transform Professor Risinger's categories (e.g., moderately surprised, astonished, shocked, etc.) into a numerical space for comparison purposes, though the specifics of such a transform would likely be subjective and contested.

Second, all subjects were asked to make a legal determination of guilt on the basis of the scenario. Specifically, subjects were asked: "Based on what is known, do you think that Bob is guilty of murder?" We included this question so that we could evaluate how the subjects' evaluation of the evidence—under either the probabilistic or the surprise framework—related to their legal determinations.

In total, we recruited 593 individuals to participate in this study in October 2017. All recruitment and experimental procedures were approved by the Vanderbilt Institutional Review Board. For recruiting subjects, we used Amazon's Mechanical Turk service (AMT). AMT is an online marketplace where individuals across the globe can perform various tasks for payment from various providers. Such web-based recruiting techniques have been widely validated (that is, online subjects have been demonstrated

[6] Risinger, *supra* note 1.

to display similar behavior to subjects recruited using conventional means), and the resulting population samples are substantially more representative than the convenience samples (e.g., of college students) typically used in such studies.[7] There is, of course, always the possibility that remote subjects may not be fully attentive to the task at hand. To account for this, we excluded subjects who took an abnormal amount of time to respond.[8] Further, we only used subjects who had an established history of satisfactorily completing tasks on AMT. Ultimately, 469 subjects were included in the final analysis.[9]

Subjects who agreed to participate were directed through AMT to the actual experiment, which was hosted by Qualtrics. Qualtrics is a web-based platform that is regularly used by scholars in many fields for hosting surveys and experiments. At the conclusion of their participation, subjects were debriefed and paid at an approximate rate of $6 per hour.

Before participating, Amazon confirmed that all potential participants were over the age of 18 and were United States (US) citizens or residents by means of a US-based bank account as well as their IP address. All experiments ended with subjects providing some demographic information, which allowed us to confirm that the sample was generally representative of the US jury-eligible population (57% male with a median age of 33).

III. RESULTS

Figure 1 displays the responses provided by subjects for each of the four prompts and each evidence level, along with the median (in green). Figure 1 allows us to compare how subjects estimate evidentiary strength under a surprise framework and a probability framework. The figures, however, do not tell us if one of the frameworks is doing a better job than

[7] *See, e.g.*, Michael Buhrmester, Tracy Kwang & Samuel D. Gosling, *Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?*, 6 PERSP. ON PSYCHOL. SCI. 3 (2011); Jon Sprouse, *A Validation of Amazon Mechanical Turk for the Collection of Acceptability Judgments in Linguistic Theory*, 43 BEHAV. RES. METHODS 155 (2011).

[8] Subjects whose timing was two standard deviations faster or slower than the average participant were excluded. This is a customary screening technique to detect non-compliance with task instructions.

[9] A pilot version of this study was run in August 2017. That version was largely similar in design to the version of the study presented here except that the scenarios in the pilot version were scaled from equipoise of guilt to a high probability of guilt; no scenario contained evidence of innocence. In addition to this, we used different scenario facts and the questions were framed to subjects differently. Subjects that completed the pilot version of this study were blocked from participating in the final version of the study. The design implemented here was changed from the pilot version of the study as a result of feedback provided by colleagues who reviewed the manuscript prior to the Symposium, and we thank them for their suggestions. The major findings observed here were also observed in the pilot study.

the other. To further evaluate this question, we can compare the estimates with how likely subjects were (considered in aggregate) to find guilt on the basis of the evidence. We derived this likelihood measure from the subjects' response to the first question (i.e., "Based on what is known, do you think that Bob is guilty of murder?"). Importantly, every subject answered this question, and it was answered prior to the response to the second question. As a result, we can collapse all responses across the surprise and probability conditions. Figure 2 adds an additional line (in red) representing the proportion of subjects finding the defendant guilty (or not guilty) on the basis of the evidence.

FIGURE 1: MEAN EVIDENCE STRENGTH UNDER SURPRISE AND PROBABILISTIC FRAMEWORKS

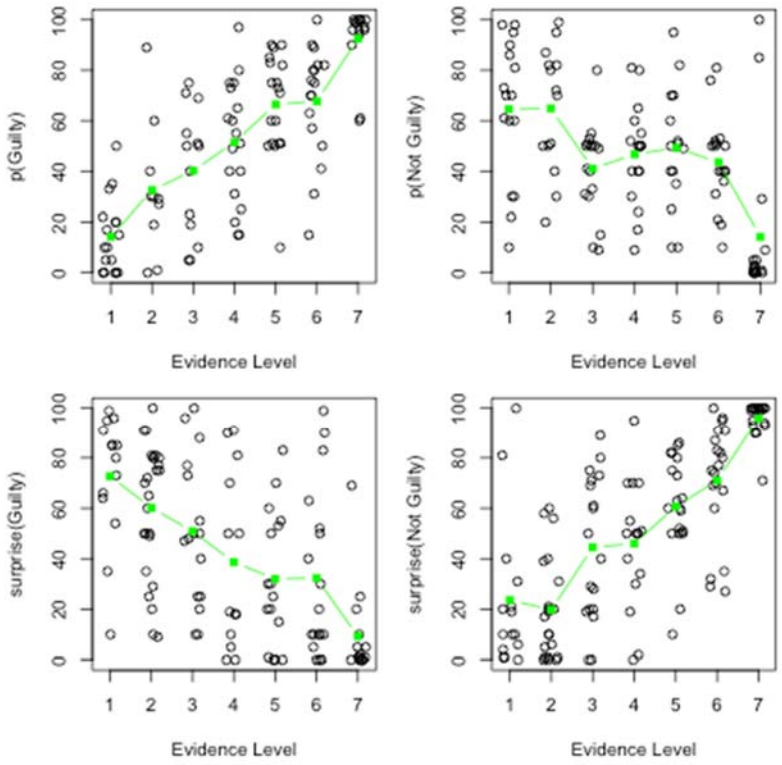


FIGURE 2: MEDIAN EVIDENCE STRENGTH UNDER SURPRISE AND PROBABILISTIC FRAMEWORKS COMPARED TO GUILT DETERMINATIONS

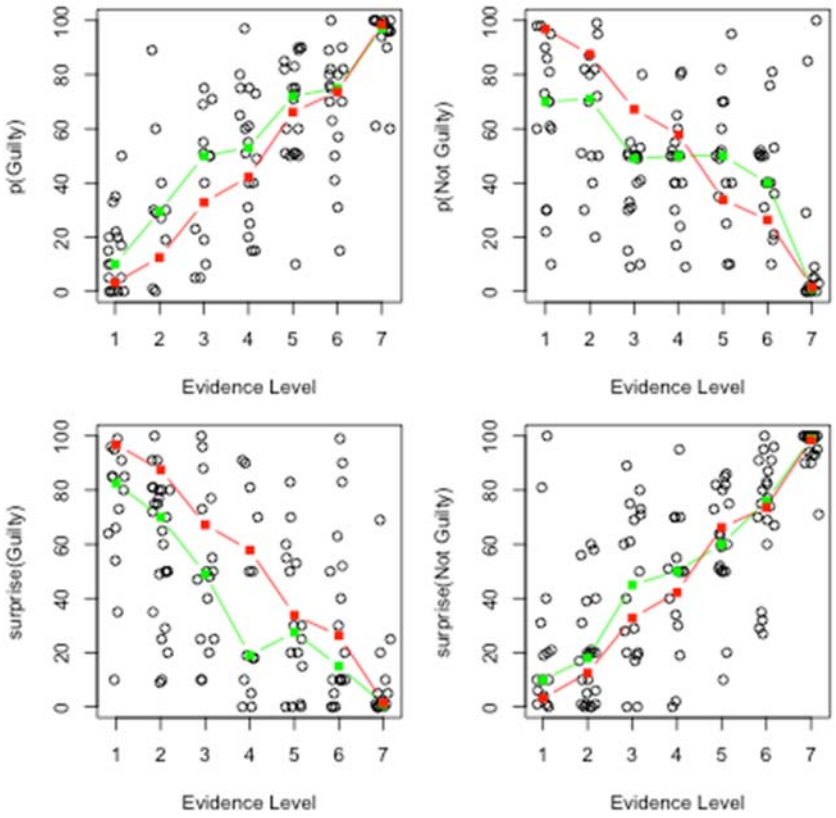
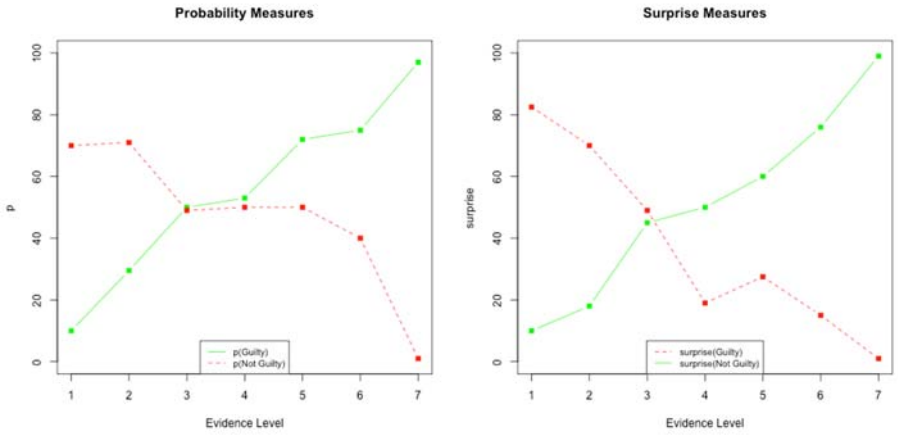


Figure 2 suggests that the median of the *surprise(Not Guilty)* metric tracks the actual conviction rate as well as, if not better than, the median of the *p(Guilty)* metric. The inverse metrics *p(Not Guilty)* and *surprise(Guilty)* tracked the actual acquittal rate less well,[10] though the way in which they deviated differed.

[10] To formally assess "how well" the estimations tracked the guilt determinations, we performed a correlation analysis between subjects' estimations and their decisions of guilt for each of the metrics. This analysis allows us to calculate the strength of the relationship between the two responses. The r^2 values were as follows: *p(Guilty)*, 0.63; *p(Not Guilty)*, 0.25; *surprise(Guilty)*, 0.32; *surprise(Not Guilty)*, 0.52. All of the metrics showed significant correlations (*p*s < 0.05). However, the strength of the correlations for *p(Guilty)* and *surprise(Not Guilty)* were significantly higher than for the other two metrics (*Z*s > 2.09; *p*s < 0.05).

Another way of assessing metrics is to evaluate their internal coherence. For example, *p(Guilty)* and *p(Not Guilty)* should be complements, as should *surprise(Guilty)* and *surprise(Not Guilty)*. Figure 3 plots the two probability metrics together, and the two innocence metrics together. Theoretically, the curves of the median estimates should form an "X," which is the case for the surprise metrics. The probability metrics, however, do not, as the *p(Not Guilty)* curve exhibits some flattening for non-extreme evidence levels.

FIGURE 3: EXAMINATION OF INTERNAL COHERENCE



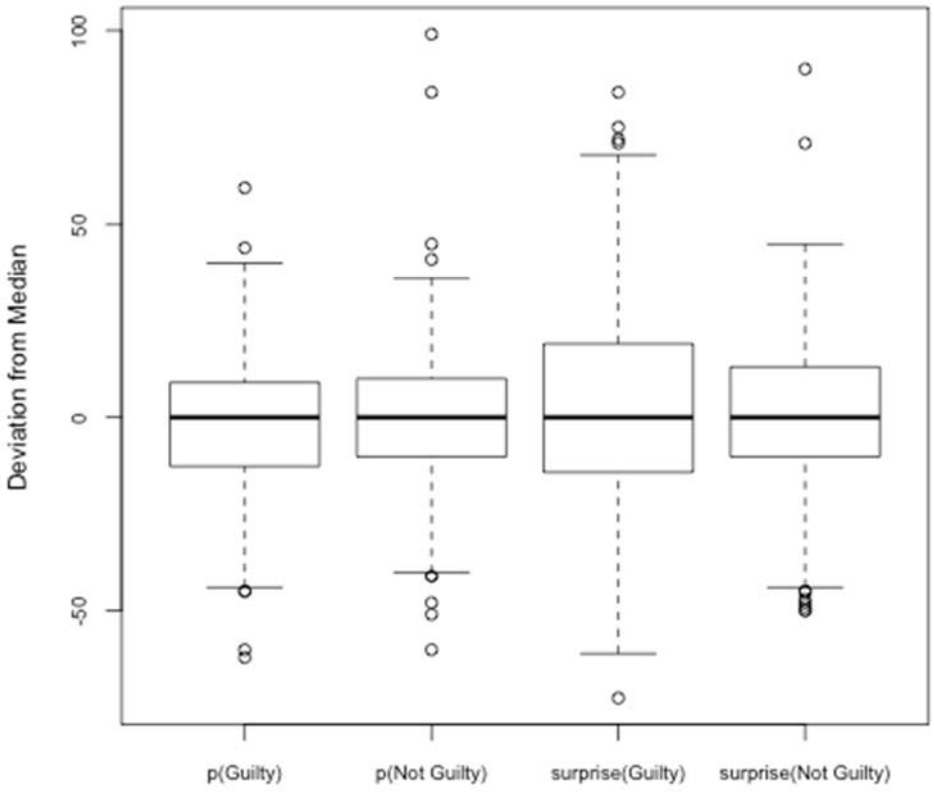
One final possibility for assessing the four metrics is to consider their level of dispersion. Even if we cannot agree on a "ground truth" by which to evaluate the accuracy of the metrics, we might agree that getting a tight distribution of estimates from participants is preferable. To examine how consistent subjects are when making probability or surprise evaluations of evidence, we looked at the results in the following way: We first centered[11] the data for each probe and evidence level. After centering, the data from each evidence level for a given prompt was combined. Figure 4 plots the histograms of the deviations from the mean for each metric. Visual examination suggests that the spreads are all similar. A statistical analysis indicates that the variance in subjects' responses for p(Guilty), p(Not Guilty), and surprise(Not Guilty) were statistically similar.[12] Subjects'

[11] To center means to subtract the mean of the set from each item in the set. This transformation normalizes for mean, but not for variance. To normalize for both the mean and variance is called a z-transformation.

[12] Brown-Forsythe Test: $F(2,348) = 0.35$, $p = 0.701$.

variance in their responses for surprise(Guilty), however, were greater than the other three.[13]

FIGURE 4: BOXPLOTS OF DEVIATION FROM MEDIAN



[13] Brown-Forsythe Test: $F(3,364) = 4.37$, $p = 0.005$.

IV. DISCUSSION

This exploratory study is meant to be nothing more than prefatory. We hope that it can begin a conversation and provide a starting point for future, more comprehensive studies. Concordantly, we caution the reader to be especially guarded when making conclusions. With those caveats in mind, there are several aspects to the results that excite us for what they may ultimately reveal.

As an initial matter, surprise appears to be a viable metric. This outcome is interesting if for no other reason than the fact that surprise is an unusual way of conceptualizing the strength of evidence. In our experience, individuals are quite familiar with estimating their confidences using a probabilistic framework. By contrast, expressing confidence in a conclusion by indicating a high level of hypothetical surprise over the opposite outcome is a procedure few are familiar with. Nonetheless, this study shows that surprise measures are highly consistent across subjects. In fact, when we compare the amount of variance that subjects display when making responses using a surprise framework we find that it is no different from the variance in responses using a probabilistic framework. In other words, subjects appear to be equally consistent when gauging the strength of evidence using surprise as they are with probabilities.

This finding on consistency partially mitigates Professor Risinger's concern that there may be "significant variability" between individuals when making assessments based on surprise.[14] To be sure, his concern was that individuals might value *word-based* strength categories differently. What we have shown here, however, is that at least with respect to *numerical* presentations, potential jurors seem to share a common "surprise scale," or at least one no worse than for numerical probabilities. An insistence on word-based strength categories may generate additional problems, but at least there is nothing inherent about the concept of surprise that promotes greater inconsistency among subjects.

One possible explanation for the similar variability between the probability and surprise metrics is that subjects are using a similar mental calculus to determine their response. For example, perhaps all subjects basically think in terms of probabilities, whether asked about probability explicitly or asked about level of surprise. This explanation, however, is undercut by the data in Figure 1. While it is true that subjects evaluate evidence similarly (if not identically) under a *p(Guilty)* and a *surprise(Not Guilty)* framework, that is not true for the converse measures: *p(Not Guilty)* and *surprise(Guilty)*. The divergence between *p(Not Guilty)* and *surprise(Guilty)*, which we might reasonably expect to also track each other 1-to-1, is remarkable. For example, at Evidence Levels 4–6, a typical subject

[14] Risinger, *supra* note 1.

reports a probability of innocence of about 50% while a typical subject reports surprise at the defendant's guilt to be about 20 out of 100. As the amount of inculpatory evidence increases, this gap narrows somewhat, but remains substantial until Evidence Level 7.

Why the lack of correspondence between the measures of *p(Not Guilty)* and *surprise(Guilty)* given the high level of correspondence between *p(Guilty)* and *surprise(Not Guilty)*? Our current data reveals no answers. Future studies may want to consider debriefing subjects to understand their thought process when making the determinations. Future studies may also try to design an experiment that associates this divergence with a known psychological phenomenon.

Where should others go from here? To begin, there is no better statistical measure of the robustness of a result than replication.[15] Though not presented here beyond a footnote, we think it is important to note that the results presented here served as a replication of a pilot study, finding substantially the same results. That said, we hope that others will try to replicate these results in other contexts. For example, do the results hold if mock jurors are instructed using pattern jury instructions (or their corollary) rather than using a simple rating scale? What happens if the quantitative scales are replaced by word categories as Professor Risinger proposed? What happens if the vignettes become more factually complex, or draw from the civil context, or involve potential biasing mechanisms like race or gender?

A final possibility is to push further on the thought that surprise encourages a more holistic or System 2 approach to the problem of evidence. Along these lines, one question is whether surprise might protect against common probabilistic fallacies. One of the most ubiquitous probabilistic fallacies in legal contexts is the base rate fallacy. This fallacy concerns the little weight subjects give to base rates when computing likelihood. To formally test whether using measures of surprise might ameliorate the effects of this fallacy in decision-making, we presented 357 subjects[16] with a variant of Daniel Kahneman & Amos Tversky's classic hypothetical involving green and blue cabs. Specifically:

> Following a recent baseball game, a number of cars in the ballpark parking lot were found to have had their windows smashed in.

---

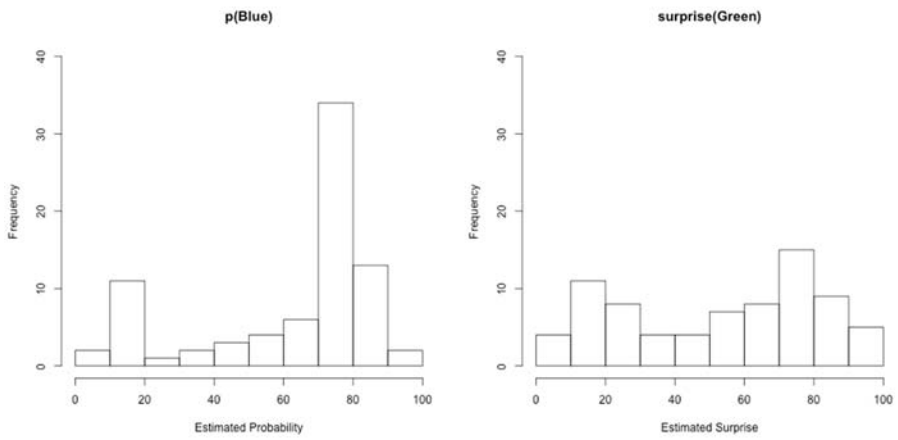
[15] STEVEN J. LUCK, AN INTRODUCTION TO THE EVENT-RELATED POTENTIAL TECHNIQUE 310 (2d ed. 2014) ("Replication does not depend on assumptions about normality, sphericity, or independence. Replication is not distorted by outliers. Replication is a cornerstone of science. Replication is the best statistic.").

[16] In total, 379 subjects were recruited using the same methods and procedures described for the main experiment. *See* Part II, *supra*.

> At the game, 85% of the fans supported the green team and were wearing green. Meanwhile 15% of the fans were supporting the blue team and wearing blue. Only individuals with tickets had access to the parking lot.
>
> A witness, the parking lot attendant, reports that the person who vandalized the cars was wearing blue. However, it was nighttime when the vandalism occurred. The court tested the reliability of the witness under the same circumstances that existed on the night of the crime and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.

The correct answer is 41%. When subjects were asked for the probability that the perpetrator wore blue, most respondents answered 80%, which represents the classic base rate fallacy result. Respondents focused on the accuracy of the witness and did not account for the low base rate of blue team fans. Intriguingly, however, when asked how surprised they would be that the perpetrator wore green, the responses indicated that significantly fewer people were susceptible to the base rate effect. Figure 5 presents histograms of the responses. While it is difficult to formally compare the two distributions since the parameter is different, we note that 80% of the responses estimated a probability of greater than 50%, while only 57% of the responses provided a level of surprise greater than 50 out of 100.

FIGURE 5: HISTOGRAM OF RESPONSES FROM BASE RATE EXPERIMENT



While the dispersion associated with surprise(Green) may be undesirable, these very preliminary results suggest that surprise may indeed cause people to think about evidence differently, and may be a tool that the legal system can harness to address base rate neglect.

V. CONCLUSION

Surprise offers an intriguing framework both for communicating the burden of proof to jurors and for having the jurors think about the evidence in a case. Our preliminary results show that surprise is a viable metric that jurors can apply in meaningful and reasonably consistent ways. Yet, as one might expect, our results raise more questions than they answer. Why is the probability of innocence handled differently by mock jurors, so that its estimate coheres with neither estimates of the probability of innocence nor estimates of surprise? And can surprise offer a useful way to address base rate neglect? Professor Risinger has certainly opened the door to some interesting questions.