Law and Human Behavior
2013, Vol. 37, No. 3, 197–207

0147-7307/13/$12.00 DOI: 10.1037/lhb0000027

The Expression and Interpretation of Uncertain Forensic Science Evidence: Verbal Equivalence, Evidence Strength, and the Weak Evidence Effect

Kristy A. Martire, Richard I. Kemp, Ian Watkins, Malindi A. Sayle, and Ben R. Newell
University of New South Wales

Standards published by the Association of Forensic Science Providers (2009, Standards for the formulation of evaluative forensic science expert opinion, *Science & Justice*, Vol. 49, pp. 161–164) encourage forensic scientists to express their conclusions in the form of a likelihood ratio (LR), in which the value of the evidence is conveyed verbally or numerically. In this article, we report two experiments (using undergraduates and Mechanical Turk recruits) designed to investigate how much decision makers change their beliefs when presented with evidence in the form of verbal or numeric LRs. In Experiment 1 (N = 494), participants read a summary of a larceny trial containing inculpatory expert testimony in which evidence strength (low, moderate, high) and presentation method (verbal, numerical) varied. In Experiment 2 (N = 411), participants read the same larceny trial, this time including either exculpatory or inculpatory expert evidence that varied in strength (low, high) and presentation method (verbal, numerical). Both studies found a reasonable degree of correspondence in observed belief change resulting from verbal and numeric formats. However, belief change was considerably smaller than Bayesian calculations would predict. In addition, participants presented with evidence weakly supporting guilt tended to "invert" the evidence, thereby counterintuitively reducing their belief in the guilt of the accused. This "weak evidence effect" was most apparent in the verbal presentation conditions of both experiments, but only when the evidence was inculpatory. These findings raise questions about the interpretability of LRs by jurors and appear to support an expectancy-based account of the weak evidence effect.

Keywords: forensic science, evidence strength, interpretation, juror decision making

Forensic scientists can never compare the sample (e.g., fingerprint, toolmark, shoeprint) they have with every potential contributing source—as would be required to individualize a sample to a source. Moreover, it is possible for two different sources to leave indistinguishable markings (or, conversely, for the same source to result in two distinguishable markings; Koehler & Saks, 2010). For both of these reasons, expert opinions must have a statistical basis and therefore reflect some degree of uncertainty (National Academies of Science, 2009; Thornton & Peterson, 2007).

Growing acceptance of the value of a probability framework as a coherent logical foundation for forming opinions in the forensic sciences (Aitken et al., 2011) can be seen in documents recommending that evaluative opinions be formulated in line with probability theory and Bayesian principles of belief updating (Berger, 2010). In particular, the Association of Forensic Science Providers (AFSP; 2009) published a set of standards (including a scale of numerical and verbal expressions; see Table 1) indicating opinions should be "based upon the estimation of a likelihood ratio" (p. 161) reflecting the ratio of two probabilities (making no reference to the possibility of error): (a) the likelihood of obtaining a piece of evidence given a proposition broadly consistent with the prosecutions' case; compared with (b) the likelihood of obtaining a piece of evidence given an alternative (defense) proposition. For example, following these guidelines, a forensic scientist could express the results of their analysis of a shoe print in the following manner: "In my opinion the correspondence between the footwear mark at the crime scene and the shoe of the accused is 4.5 times more likely when the shoe has made the mark [Proposition 1], than when the shoe has not made the mark [Proposition 2]."

To date, uptake of these standards has varied considerably across jurisdictions and disciplines. For example, although the use of LRs is now standard practice for cartridge-case and bullet comparisons in the Netherlands (Stoel, 2012), and is actively being explored in other disciplines and jurisdictions (see Morrison, 2011; Neumann, Evett, & Skerret, 2012), in the United States, the use of LRs for any non-DNA comparison is rare.

Even so, in 2011, in the wake of critical attention from the U.K. Court of Appeal in the case of *R v T* (2010), the available standards were extended to address forms of expression (Aitken et al., 2011). In the resultant position statement, 31 leading stakeholders, largely, although not exclusively, from European institutions and organizations, reaffirmed that LRs are the most appropriate foundation for assisting the court. This group also asserted that "a verbal scale based on the notion of the LR is the most appropriate basis for communication of an evaluative expert opinion to the

Kristy A. Martire, Richard I. Kemp, Ian Watkins, Malindi A. Sayle, and Ben R. Newell, School of Psychology, University of New South Wales, Sydney, Australia.

This article was facilitated in part by funding from the Australian Research Council to Richard I. Kemp (Linkage Project LP110100448) and Ben R. Newell (Discovery Grant DP110100797; Future Fellowship FT110100151). Sincerest thanks to Jon Berengut for his kind assistance in the preparation of this article.

Correspondence concerning this article should be addressed to Kristy A. Martire, School of Psychology, University of New South Wales, Sydney, NSW, Australia, 2052. E-mail: k.martire@unsw.edu.au

Table 1
Standards for Numerical and Verbal Expression of Likelihood Ratios (Association of Forensic Science Providers, 2009)

Recommended likelihood ratio terminology	
Numerical expression	Verbal expression (support)
> 1–10	Weak or limited
10–100	Moderate
100–1,000	Moderately strong
1,000–10,000	Strong
10,000–1,000,000	Very strong
> 1,000,000	Extremely strong

court" (see Aitken et al., 2011, Point 5). Importantly, this debate about the best way to present the results of complex forensic science analyses in court has occurred seemingly without reference to the body of empirical evidence amassed by psychologists relating to reasoning under uncertainty.

Psychological evidence suggests that people often have difficulty understanding probabilities and statistics (Gigerenzer & Edwards, 2003) and that they tend to confuse likelihoods and posterior beliefs (as in the prosecutor and defense-attorney fallacies; Koehler, 1996; Thompson, 1989; Thompson & Schumann, 1987) in a way that may sometimes lead to overvaluing of, and at other times undervaluing of, evidence relative to normative (i.e., Bayesian) models (Faigman & Baglioni, 1988; Goodman, 1992; Kaye & Koehler, 1991; Smith, Bull, & Holliday, 2011; Smith, Penrod, Otto, & Park, 1996). The recommendation that forensic scientists should use verbal equivalents may also be problematic, given the psychological evidence indicating that the meaning attributed to a single word can vary from person to person and from context to context (e.g., Brun & Teigen, 1988; Budescu, Broomell, & Por, 2009; Budescu, Por, & Broomell, 2012; Wallsten & Budescu, 1995).

Three key psychological experiments speak particularly to the (non)equivalence of verbal and numerical expressions in the forensic science arena, and to the question of how the decision maker will interpret evidence presented in the form of LRs. The first, by Nance and Morris (2005), investigated various forms for quantifying DNA random-match probabilities (RMPs) and laboratory error rates. Judges and jurors were asked to rate the probability of the defendant's guilt, given a RMP presented in one of three formats: a frequency (e.g., a 1 in 40,000 chance of a coincidental match), a LR (e.g., 40,000 times more likely to match if the accused is the source of the crime scene sample than if he is not), or a chart mapping hypothetical prior and posterior probabilities for the indicated LR. The RMP presentation format was found to significantly influence rated guilt probability, with post hoc comparisons indicating a significant difference between the frequency format (which produced the lowest estimates of guilt probability) and the chart format (which provided the highest estimates), with the LR falling between the two. The authors concluded that the use of LRs appeared to "move juror assessments of evidence in the direction of Bayesian norms" (p. 429), and consequently supported their use in courts over frequentistic expressions of the probative value of DNA analyses.

The second study, by McQuiston-Surrett and Saks (2008), examined undergraduate psychology students' rating of the strength of a set of standardized verbal labels proposed for use by the American Board of Forensic Odontology (ABFO). The ratings provided by these mock jurors were roughly the opposite of what the ABFO intended. For example, participants attributed the greatest certainty to testimony of a "match" (86 on a 100-point scale), but this was the phrase that the ABFO reserved for the lowest level of certainty among the four options. Conversely, the phrase the ABFO designated for the strongest expression of certainty, "reasonable scientific certainty," was rated as second most uncertain by participants (70.7/100).

More recently, de Keijser and Elffers (2012) used realistic technical forensic reports to examine how well judges and lawyers (jurists) and experts in the Netherlands understood evaluative expert opinions expressed using the scale recommended by the Netherlands Forensic Institute (NFI; Berger, 2010). The results indicated that although experts (members of the NFI) showed a greater ability to identify correct interpretations of the opinions compared with jurists, both groups commonly made errors interpreting the LRs and had little insight into their limited understanding of the report conclusions.

These studies all suggest that several questions still remain regarding the interpretability of verbal expressions of LRs as formulated under the new standards proposed by the AFSP. Moreover, given that values on the scale range from "weak or limited" to "extremely strong" support, it is also unclear how the degree of numerical and verbal correspondence varies with evidence strength across the scale (see Table 1). Accordingly, the current study was designed to measure change in belief regarding the likely guilt or innocence of the defendant, given verbal or numerical expert evaluative opinions of various strengths. To this end, we measured belief prior to hearing a forensic scientist's opinion (prior belief), and after hearing the expert's opinion (posterior belief). This approach allows us to calculate belief change in response to the evidence, which can be used (a) to assess the equivalence of the verbal and numerical scales over various levels of evidentiary strength, and (b) to compare the observed change from prior belief with posterior belief with the change that was intended by the expert.

It is predicted that evidence attributed greater strength by the expert will result in significantly greater belief change (prior to posterior) than evidence attributed lesser strength. It is also anticipated that opinions presented in numerical format will significantly differ in impact from opinions expressed in "equivalent" verbal form, consistent with the findings of McQuiston-Surrett and Saks (2008), as described previously. Furthermore, there is some reason to anticipate these main effects may be moderated by what is known as a "weak evidence effect" (Fernbach, Darlow, & Sloman, 2011).

The weak evidence effect describes a situation in which a piece of evidence, which weakly supports belief in a given hypothesis (as is the case in which an expert produces an evaluative opinion with a small LR), has the counterintuitive effect of increasing belief in the alternative hypothesis (Lopes, 1987). These directional errors (or "boomerang effects"; Petty & Cacioppo, 1996) have been observed in many contexts, including the evaluation of public policy initiatives (Fernbach et al., 2011), argumentation (McKenzie, Lee, & Chen, 2002), and mock-juror decision making (Smith et al., 1996), and has been explained by various mechanisms from informational neglect (Fernbach et al., 2011), to

expectancy-based decision making, in which a downward revision results from comparing the actual evidence strength against your high expectations of how strong the evidence would or should be (McKenzie et al., 2002). Although it is not yet known if forensic science expert evidence, posed in the form of verbal and numerical LRs, is also susceptible to this counterintuitive style of interpretation, aspects of the recommended expression for low-strength evidence ("weak or limited support") would seem to make this outcome likely. More precisely, although the suffix "support" indicates the evidentiary value is in favor of the proposition rather than against it, the term "weak" may reasonably be interpreted as the opposite of "strong" evidence. Consistent with a weak evidence effect, this could cause people to increase their belief in the alternative proposition rather than weakly increasing their belief in the supported proposition.

Experiment 1

Method

Design. A 3 (evidential strength: low, moderate, high) × 2 (presentation method: verbal, numerical) between-subjects factorial design was employed.

Participants and data screening. Participants were 75 undergraduate psychology students participating for course credit and 545 workers from an online self-enlisted workforce (Mechanical Turk; Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler, & Ipeirotis, 2010), who were compensated US 50¢ for their time. Of the 620 participants completing the experiment, 131 (21.1%) were excluded, based on three predefined criteria,[1] resulting in a final sample of $N = 494$ (Numerical $n_{low} = 87$, $n_{mid} = 73$, $n_{high} = 73$; Verbal $n_{low} = 81$, $n_{mid} = 91$, $n_{high} = 89$). Of this sample, 13.8% resided in Australia, 55.7% in the United States, 20.9% in India, and 9.7% were based in other countries. Three-quarters (75.5%) of the sample reported being native English speakers. On average, the remaining 121 participants had been speaking English for 17.44 years (range 3 to 55, $SD = 8.39$). Females accounted for 54.5% of the sample and the mean age was 31.28 years (range 16 to 81, $SD = 12.36$). Almost 81% (80.8%) of the sample indicated that, to the best of their knowledge, they were eligible for jury duty in their country of residence.

Materials and procedure.

The minimal trial. All participants were presented with a one-page vignette of a hypothetical larceny trial including only the requirements for conviction, the case facts, and the inculpatory testimony of an expert.

Participants were first asked to imagine that they were a juror in a trial and were informed that in order to return a guilty verdict, they must be satisfied beyond a reasonable doubt that (a) the accused person, (b) took and carried away, (c) the property of another, (d) with the intent to permanently deprive the owner of the property, and (e) that taking was without the owner's consent. They were then presented with the following case facts: (a) $300 was stolen from the victim, (b) the accused was arrested with $328 in his possession for which he did not account, (c) the accused was arrested in the vicinity of the theft, (d) the accused did not have an alibi, and (e) at the time of his arrest, the accused was wearing clothes similar to those described by a witness.

Expert evidence. Each participant also read the testimony (loosely based on the evidence provided in *R v T*) of an experienced expert forensic science analyst who compared footwear marks found at the scene of the crime with the shoe the accused was wearing at the time of arrest. As a result of his analysis, the expert stated,

> When assessing the significance of any similarity or differences between a shoe and a mark resulting from an analysis, the likelihood of obtaining that similarity or difference is considered against two alternative propositions: (1) the shoe has made the mark; (2) the shoe has not made the mark.

Participants were then provided one of six possible versions of the expert's opinion, based on their randomly allocated condition (see Table 2). These values and verbal labels were derived from those proposed by the AFSP. For example, someone in the low evidential strength condition next read, "In my opinion the correspondence between the footwear mark at the crime scene and the shoe of the accused [is 4.5 times more likely] (numerical) or [offers weak or limited support] (verbal) when proposition 1 is correct than when proposition 2 is correct."

Participant responses and procedure. After consenting to participate, accessing the experimental materials online, and reading the details of the minimal trial and case facts, participants were asked whether, based on the information presented, they currently believed the accused was more likely to be guilty or not guilty. If the participant indicated a preference for guilt, they were then asked to complete the following sentence using a number greater than 1: "Based on the available evidence I believe that it is ____ times more likely that the accused is *guilty* than *not guilty*" (emphasis added). If the participant expressed an original preference for the "not guilty" option, they were given the same question with the order of the italicized terms reversed. The response to this question was taken as the participant's prior belief in the accused's guilt. Participants were then presented with one of the six types of expert evidence before being asked the same two questions again, providing a posterior-belief value. They were also asked to complete a series of demographic questions and the Subjective Numeracy Scale (SNS; Fagerlin et al., 2007), which assesses numerical fluency. The entire procedure took approximately 15 min to complete.

Results

Presentation method and evidence strength. Initial analyses were conducted separately for undergraduate and Mechanical Turk participants. The two groups produced the same pattern of belief-change results and accordingly have been analyzed and reported together henceforth.

Belief-change values were calculated for each participant by subtracting the stated prior belief from the posterior belief. For these purposes "not guilty" beliefs were coded as negative values.

[1] Participants were excluded if they (a) completed the experiment in less than 120 s ($n = 10$); (b) failed the "catch-trial" (Paolacci et al., 2010) by not choosing the "not very good" response when asked, "How good are you at surviving one hour without oxygen?" ($n = 55$) or; (c) belief-change value was classified as an "extreme outlier" falling outside 7 times the interquartile range ($n = 61$; Barbato, Barini, Genta, & Levi, 2011).

Table 2
Evidence Strength and Presentation Method

	Presentation method	
Evidentiary strength	Numerical	Verbal
Low	4.5	Weak or limited support
Moderate	450	Moderately strong support
High	495,000	Very strong support

For example, if a participant began by believing the defendant was 4 times more likely to be not guilty than guilty (prior = −4), and finished believing the defendant was 4 times more likely to be guilty than not guilty (posterior = 4), that person will have a belief-change score of 8 (posterior minus prior).

A 2 × 3 ANCOVA was conducted to examine the impact of *presentation method* and *evidential strength* on belief change while controlling for prior-belief value (i.e., the number reflecting how many times more likely one hypothesis is than the other) and score on the SNS (see Figure 1). Of these, only the prior-belief covariate was significant, $F(1, 486) = 8.16$, $MSE = 73.59$, $p < .005$, partial $\eta^2 = 0.017$, 95% CI [.002,.046]. Adjusting for this resulted in a significant main effect for *presentation method*, such that numerical expressions of evidence resulted in greater adjusted mean belief-change ($M = 1.63$) compared with verbal expressions ($M = 0.31$, $F(1, 486) = 23.51$, $MSE = 212.06$, $p < .0005$, partial $\eta^2 = 0.046$, 95% CI [.017, .087]. There was also a significant main effect of *evidential strength*, $F(2, 486) = 13.28$, $p < .0005$, $MSE = 119.78$, partial $\eta^2 = 0.052$, 95% CI [.019, .092], and a significant interaction effect, $F(2, 486) = 7.75$, $MSE = 69.87$, $p < .0005$, partial $\eta^2 = 0.031$, 95% CI [.006, .065]. Pairwise comparisons showed that numerical and verbal expressions resulted in equal amounts of belief change when the evidence strength was moderate or high. However, when evidence strength was low, numerical expression ($M = 1.35$) resulted in significantly greater belief change than the verbal label ($M = -1.39$). Overall, the belief change observed in the low-strength numerical condition was in the direction intended by the expert giving the evidence (i.e., toward guilt); however, the overall belief change observed in the low-strength verbal condition was in the opposite direction to that intended by the expert (i.e., toward innocence).

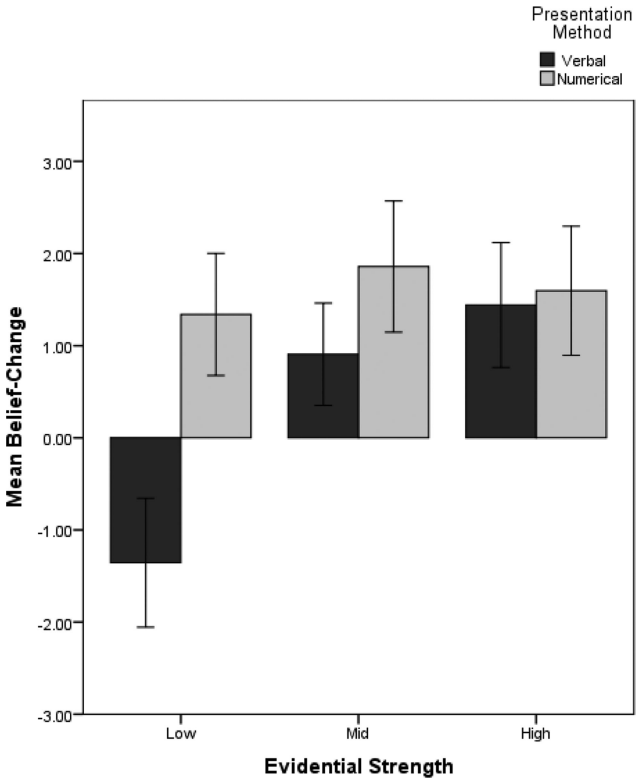

Figure 1. Mean adjusted belief change by presentation method and evidential strength (error bars ± 2 standard errors).

Weak evidence effect. Given that, in all conditions, participants were presented with expert evidence supporting the hypothesis that the accused's shoe had made the mark at the crime scene (and therefore supporting the prosecution case), observed declines in guilt ratings are consistent with a weak evidence effect. That is, inculpatory evidence increased belief in the innocence of the accused. To explore this effect further, the proportion of participants in each condition who revised their belief in the guilt or innocence of the accused downward after reading the experts evidence (i.e., toward "not guilty") was calculated (see Table 3). A majority of those in the low/verbal condition (61.72%) responded in a manner incongruent with the evidence provided by the expert (taking inculpatory evidence to be exculpatory), compared with an average of 12.91% in the remaining conditions. Moreover, a sizable number of those revising their guilty beliefs toward innocence in the low/verbal condition actually crossed the guilty/not guilty threshold when moving from priors to posteriors ($n = 19$; 23.46%). That is, they considered the defendant more likely to be guilty than not guilty before reading the experts' testimony, and considered the defendant more likely to be not guilty than guilty after reading incriminating evidence from the expert.

In order to obtain a clearer picture of the variance captured in the mean belief-change values, particularly compared with Bayesian "norms," box plots were also constructed depicting observed belief change compared with the belief change that would be expected by applying Bayes's theorem to participants' stated priors (*Bayesian belief-change* = *prior-belief* x *LR*; see Figure 2). In all conditions, the overwhelming majority of participants responded more conservatively to the evidence than would have been predicted by the application of Bayes's theorem, and, in the case of the high-strength evidence conditions, the difference between the observed and predicted responses was especially large, amounting to several orders of magnitude.

Table 3
Percent Moving Toward Innocence After Hearing the Expert Evidence by Presentation Method and Evidence Strength

	Percent moving towards innocence (number of participants changing belief preference from guilty to not guilty)	
Evidence strength [LR provided]	Numerical presentation	Verbal presentation
Low [4.5]	12.6% (4)	61.7% (19)
Moderate [450]	8.2% (1)	14.2% (3)
High [495,000]	13.6% (3)	15.7% (2)

Note. LR = likelihood ratio.

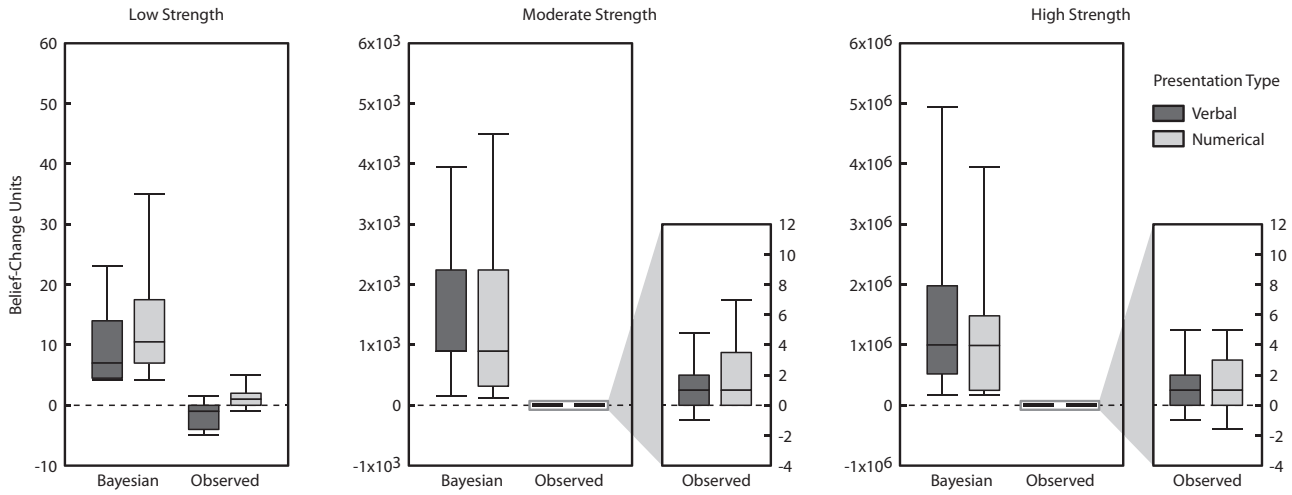

Figure 2. The central 80% of the distribution of observed and Bayesian belief change in the low-evidence (left panel), moderate-evidence (center panel; observed belief-change magnified inset), and high-evidence (right panel; Observed belief-change magnified inset) strength conditions. Each box includes the central 50% of belief-change values, and the solid lines in the boxes mark the medians. Belief change equals stated posterior belief minus stated prior belief.

As another gauge of the degree of correspondence between the expert's intention and the jurors' interpretation of the evidence, the data were used to calculate implicit LRs (ILRs), accounting for each participant's change from prior to posterior belief. This value could then be compared with the LR provided by the expert.

The LR is produced by dividing the posterior-belief odds by the prior-belief odds. However, some adjustment was necessary to allow for the fact that some participants were expressing belief that favored guilt, whereas others favored innocence. To accommodate for this, where belief favored innocence over guilt, the reciprocal of the belief value was calculated, so a participant who thought that the suspect was 2 times more likely to be guilty than innocent was given an odds value 2.0, whereas a participant who thought he was 2 times more likely to be innocent than guilty was given an odds of 0.5 (1/2). Both the prior and the posterior odds were adjusted in this way before calculating the ILR for each participant, using the formula *ILR = posterior odds/prior odds.* For example, if a participant began by believing the defendant was 2 times more likely to be *not guilty* than guilty, they would have a prior of 0.5. If, after hearing the expert evidence, they thought the defendant was 2 times more likely to be *guilty* than not guilty, their posterior odds would be 2. The ILR, which should achieve a change from a prior of 0.5 to a posterior of 2, is 2/0.5 = 4.

The median ILR values are reported in Table 4. These data are consistent with undervaluing the expert's testimony in all conditions. Where the expert attributed values of 4.5, 450, or 495,000 to the evidence, participant's ILRs were roughly 3.7 (low strength), 300 (moderate strength), and 353,571(high strength) *times* smaller than intended by the expert. For example, in the high-strength numerical evidence condition, participants interpreted a statement that explicitly included the LR of 495,000 by using an ILR of 1.4.

Discussion

Examination of the belief change induced by forensic science evidence of varying strengths revealed that, although most belief change was in line with the experts intended interpretation, participants were only weakly sensitive to large differences in evidential strength and underestimated its value compared with Bayesian calculations. This could be due to a misuse of the evidence or, alternatively, may reflect perceptions of the relevance of the evidence to the guilt of the accused; these potential explanations will be considered in more detail in the general discussion.

These data also suggest that verbal and numerical mechanisms proposed by forensic scientists (e.g., AFSP; Aitken et al., 2011) do not have an equivalent impact on participant-jurors' beliefs. Specifically, presenting participants with low-strength expert evidence in a verbal format resulted in an average response that was opposite to the direction intended by the expert giving the evidence, with 61.72% of respondents in this condition treating evidence supporting the defendants' guilt, as though it favored innocence. That is, as hypothesized, a majority of participants in the low-strength/verbal evidence condition treated information that should have further *implicated* the defendant (i.e., inculpatory evidence) as though it actually strengthened the case for the defendants' innocence (i.e., exculpating the defendant). This same inversion, although present, was much less pronounced in the low-strength numerical condition, affecting only 12.64% of participants in the condition. This suggests that numerical expressions may be less

Table 4
Median Implicit Likelihood Ratio by Presentation Method and Evidence Strength

Evidence strength [LR provided]	Median ILR (range)	
	Numerical presentation	Verbal presentation
Low [4.5]	1.2 (0.1–25.0)	0.8 (0.1–35.0)
Moderate [450]	1.5 (0.2–20.0)	1.1 (0.1–6.0)
High [495,000]	1.5 (0.2–25.0)	1.3 (0.1–20.0)

Note. ILR = implicit likelihood ratio; LR = likelihood ratio.

susceptible to this type of misinterpretation, at least when compared with the form of words proposed by the AFSP.

This reversal of intended direction of the evidence appears to be an important and novel demonstration of the weak evidence effect (Fernbach et al., 2011). The possible mechanisms behind this effect will be discussed later, but at this time we note the practical implications of this effect. Although it may appear problematic for decision makers to be treating inculpatory evidence as though it were exculpatory, as observed here, such an outcome is likely more desirable than if the opposite were also true. The criminal justice system is designed to be asymmetric, giving greater emphasis to the protection of the rights of the accused than those of the State (via procedural safeguards, such as a presumption of innocence and the reasonable doubt verdict thresholds, among others). Accordingly, a tendency to discount weak, although incriminating, prosecution evidence is consistent with this intentional bias, and although concerning from a logical decision-making point of view, its implications for the criminal justice system may be less critical. If, however, the opposite were also true—that weak exculpatory evidence was taken to support the guilt of the accused—this would be a much more serious concern, even more so if the newly recommended verbal methods of communication made such an outcome more likely than traditional numerical expressions.

In order to explore this possibility, we conducted a second experiment, using modified materials from Experiment 1, to present participants with either high or low-strength *exculpatory* evidence or low-strength *inculpatory* evidence (thereby replicating part of Experiment 1) in both verbal and numerical formats.

Experiment 2

Method

Design. The principal design of Experiment 2 was a 2 (evidence: exculpatory high strength, exculpatory low strength) × 2 (presentation method: verbal, numerical), with both factors manipulated between subjects. In addition, we repeated two conditions from Experiment 1 (inculpatory low-strength verbal and inculpatory low-strength numerical) in an attempt to replicate the previously observed weak evidence effect with inculpatory evidence.

Participants. Participants were 139 undergraduate psychology students participating for course credit and 399 Mechanical Turk workers who were compensated US 50¢ for their time. Data collected online was screened using the same three criteria employed in Experiment 1, resulting in the following removals: 4 on the basis of completion time; 104 failed the "catch trial"; and 19 based on the Inter Quartile Range (IQR). Overall, of the 538 participants completing the experiment, 127 (23.6%) were excluded based on these criteria, resulting in a final sample of 411 (Numerical $n_{\text{exculp/high}} = 66$, $n_{\text{exculp/low}} = 67$, $n_{\text{inculp/low}} = 66$; Verbal $n_{\text{exculp/high}} = 74$, $n_{\text{exculp/low}} = 71$, $n_{\text{inculp/low}} = 67$). Of the remaining participants, 33.8% resided in Australia, 20.9% in the United States, 30.2% in India, and 15.1% were in other countries. Almost 60% (59.9%) of the sample reported being native English speakers. The remaining 165 participants had been speaking English for an average of 18.59 years (range 5 to 52, $SD = 10.16$). Males accounted for 52.1% of the sample, and the mean age was 26.89 years (range 17 to 66, $SD = 10.26$). Almost half (49.6%) of the sample indicated that, to the best of their knowledge, they were eligible for jury duty in their country of residence.

Materials and procedure. The materials and procedure for Experiment 2 were identical to Experiment 1 except where the direction of the expert evidence was reversed. In the *inculpatory* (replication) conditions, as in Experiment 1, the expert testified against two alternative propositions: "(1) the shoe has made the mark; (2) the shoe has not made the mark." For the *exculpatory* conditions, the order of these propositions was reversed such that the expert testified that the two alternatives propositions were "(1) the shoe has not made the mark; and (2) the shoe has made the mark." Irrespective of whether the expert provided inculpatory or exculpatory evidence, the opinion statement was the same as for Experiment 1: "In my opinion the correspondence between the footwear mark at the crime scene and the shoe of the accused is [4.5 times more likely/offers weak or limited support] when proposition 1 is correct than when proposition 2 is correct," with the precise wording of each of the propositions remaining on screen throughout.

Results

Initial analyses were conducted separately for undergraduate and Mechanical Turk participants. The two groups produced the same pattern of belief-change results and accordingly have been analyzed and reported together.

Inculpatory (replication) conditions. A one-way ANCOVA was conducted using the belief-change values obtained from the inculpatory conditions (low-strength verbal and numerical, $N = 133$) to establish whether the weak evidence effect could be replicated in Experiment 2 (see Figure 3). Once again, the prior-belief covariate was significant, $F(1, 130) = 18.16$, $MSE = 150.58$, $p < .0005$, partial $\eta^2 = 0.123$, 95% CI [.036, .230]. Adjusting for this resulted in a significant main effect for *presentation method*, such that numerical expressions of evidence resulted in greater adjusted mean belief change ($M = 1.54$) compared with verbal expressions ($M = -2.54$, $F[1, 130] = 66.71$, $MSE = 553.30$, $p < .0005$, partial $\eta^2 = 0.339$, 95% CI [.212, .448]. These results are consistent with an overall weak evidence effect in the inculpatory low-strength verbal evidence condition, but not the inculpatory low-strength numerical evidence condition.

Exculpatory expert evidence. A 2 × 2 ANCOVA was conducted to examine the impact of *evidence strength* (high or low) and *presentation method* on belief change (controlling for priors) in the exculpatory expert conditions ($N = 278$; see Figure 2). In this case, only the main effect of *evidence strength* was significant, with high-strength evidence resulting in greater belief-change (toward innocence, $M = -2.43$) than the low-strength evidence ($M = -1.47$, $F[1, 273] = 3.99$, $MSE = 78.02$, $p < .05$, partial $\eta^2 = 0.014$, 95% CI [.000, .054]. On average, belief-change scores in each of the low-strength exculpatory conditions were in the direction intended by the expert (negative). This is not consistent with a weak evidence effect, in which positive belief change (i.e., movement toward guilt) in response to weak exculpatory evidence would have been expected.

The direction of individual belief-change decisions was examined in order to establish what proportion of individuals in each condition revised their belief-change in a manner that was incon-

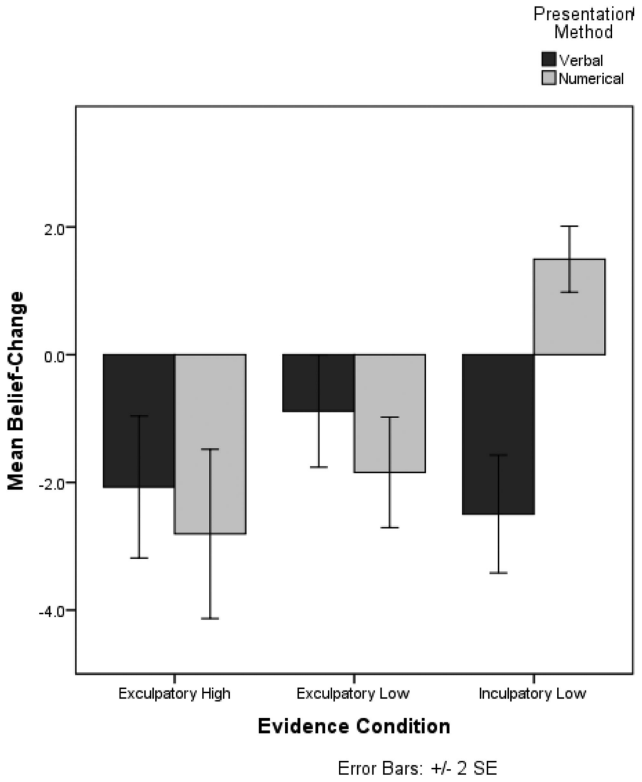

Figure 3. Mean adjusted belief change from prior to posterior by presentation method and evidence condition (error bars ± 2 standard errors).

gruent with the expert evidence (see Table 5). As in Experiment 1, a majority of those in the inculpatory low/verbal condition (67.16%) responded in the opposite direction to that intended by the expert, compared with 30.99% in the exculpatory low verbal condition and around 12% to 20% in the remaining conditions. As in Experiment 1, a substantial proportion of those making incongruous changes in the inculpatory low-strength verbal condition changed their preference from one of "guilty" to "not guilty" (38.81%). In contrast, only 18.18% of participants in the exculpatory low-strength/verbal condition changed from "not guilty" to "guilty." However, none of the participants in the exculpatory low-strength numerical condition made incongruous changes of this sort.

Table 5
Percent Moving Opposite to Expert Evidence by Presentation Method and Evidence Strength

	Percent moving incongruous to the evidence (number of participants in condition who changed their original guilty/not guilty preference)	
Evidence condition	Numerical presentation	Verbal presentation
Exculpatory high	18.2 (2)	18.9 (5)
Exculpatory low	20.9 (0)	31.0 (4)
Inculpatory low	12.1 (0)	67.2 (26)

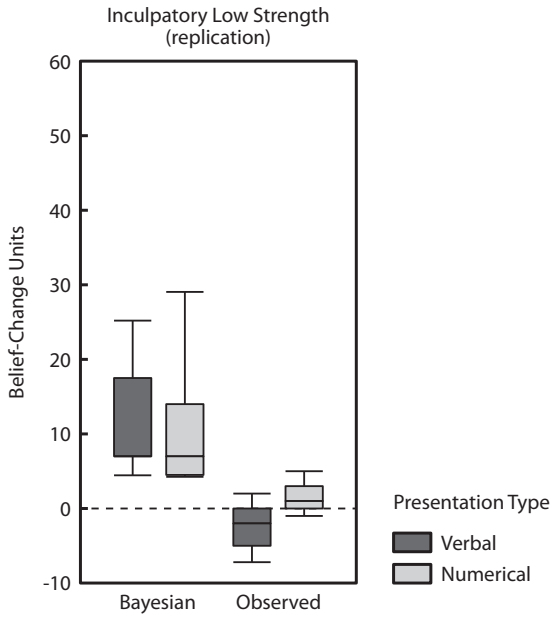

Figure 4. The central 80% of the distribution of observed and Bayesian belief-change in the inculpatory low-strength (replication) condition. Each box includes the central 50% of belief-change values, and the solid lines in the boxes mark the medians. Belief change equals stated posterior belief minus stated prior belief.

Box plots of observed and Bayesian belief-change are presented in Figure 4 (replication condition) and Figure 5 (exculpatory evidence conditions). Again, participant belief change is conservative compared with Bayesian norms, with the low-strength inculpatory verbal evidence resulting in a weak evidence effect for a majority of participants.

The median ILRs are presented in Table 6. The ILRs used in the exculpatory conditions were inverted here in order to permit direct comparisons between the ILR and the provided LR. Consistent with Experiment 1, the data show that the ILRs were markedly more conservative than those provided.

General Discussion

We conducted two experiments investigating the impact of the newly agreed format for the presentation of forensic science evidence in court. Our results suggest that, although in some ways, the numerical scale and verbal equivalents proposed by the AFSP performed as intended, in other ways, they did not and hence are far from ideal.

Verbal-Numerical Equivalence

When tested using the inculpatory evaluative opinion of a forensic scientist, significant differences between verbal and numerical expressions of evidence were revealed only for low-strength evidence. This suggests verbal-numerical equivalence at the mid and high levels of the scale. Similarly, when tested using exculpatory testimony, the only significant difference observed was between high and low-strength evidence—as one would hope—

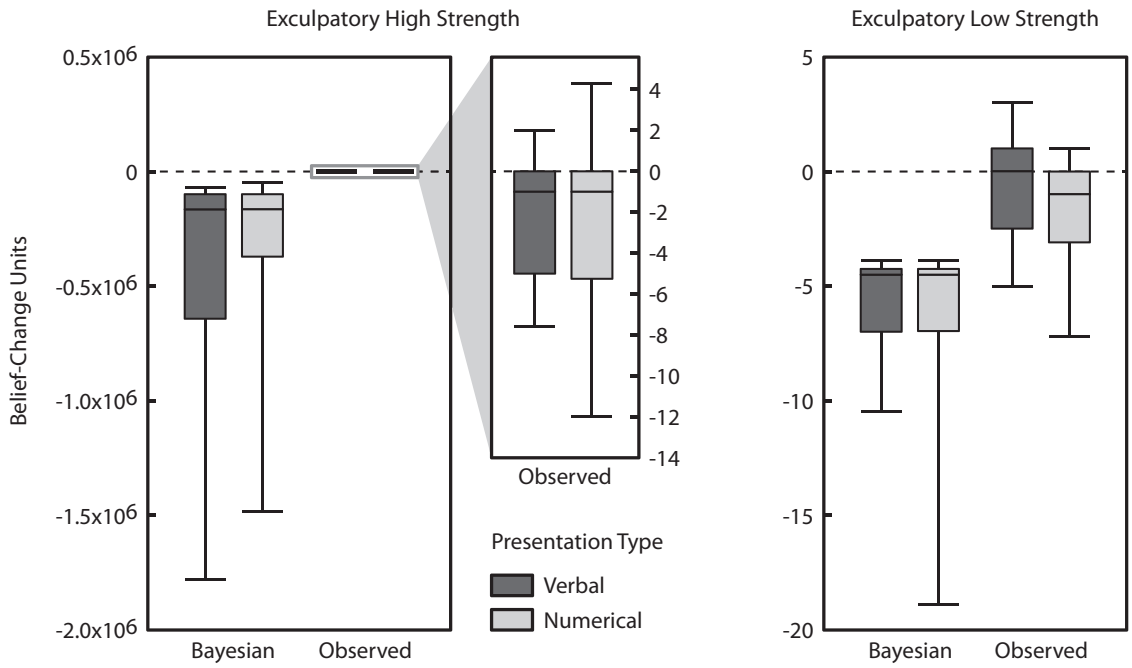

Figure 5. The central 80% of the distribution of observed and Bayesian belief change in the exculpatory high strength (left panel; observed belief change magnified inset) and exculpatory low strength (center panel). Each box includes the central 50% of belief-change values, and the solid lines in the boxes mark the medians. Belief change equals stated posterior belief minus stated prior belief.

and not between verbal and numerical expression. Yet these positive indications belie a much more nuanced set of results.

Correspondence Between Intentions and Interpretations

The significant main effects observed in Experiment 1 were modified by a significant interaction between evidence strength and presentation type. This means that participant belief change from before to after hearing the expert's evaluative opinion did not increase in a simple fashion in response to increasing evidence strength. Specifically, decision makers presented with evidence that one hypothesis was 450 times more likely, given an inculpatory rather than exculpatory account, showed the same amount of belief change as was observed among those presented with evidence that was 495,000 times more likely. Furthermore, although these two evidence strengths differ by three orders of magnitude, the mean difference in belief change from moderate strength to high strength was −0.36 units in the numerical condition and 0.53 units in the verbal condition. Similarly, the difference between implicit and provided LRs show a change in the median odds (in the wrong direction) from moderate to high strength of −0.01 in the numerical condition and 0.02 in the verbal condition.

The observation that the degree of belief change achieved in the moderate compared with high condition was far from commensurate with the actual difference cited in the evidence and intended by the expert suggests that participants were insensitive to the relative weights of the evidence. Further comparison of the Bayesian belief-change distributions and the values assigned by the expert with observed belief change and participant ILRs also reveals a pattern consistent with a conservative application of the probabilistic evidence.

This apparent insensitivity to the intended value of the evidence displayed by participants in Experiment 1 is moderated by the

Table 6
Median Implicit Likelihood Ratio by Presentation Method and Evidence Strength

Evidence condition [LR provided]	Median ILR (range)	
	Numerical presentation	Verbal presentation
Exculpatory high [495,000][a]	1.5 (0.1–80.0)	1.4 (0.2–56.0)
Exculpatory low [4.5][a]	1.2 (0.4–35.0)	1 (0.2–72.0)
Inculpatory low [4.5]	1.3 (0.3–15.0)	0.6 (0.0–6.0)

Note. ILR = implicit likelihood ratio; LR = likelihood ratio.
[a] ILRs were inverted to permit direct comparisons with the provided LRs.

results from Experiment 2. In particular, in Experiment 2, a significant main effect of evidence strength was observed, such that participants engaged in significantly greater belief change when presented with LRs of 495,000 compared with 4.5. However, a comparison of the observed belief change and ILRs against Bayesian belief change and provided LRs again shows an absence of correspondence.

Overall, then, although the verbal and numerical levels of the scale tested here appear to induce similar levels of belief change in the participants (except for the low-strength verbal inculpatory evidence, which we will address later), there is little correspondence between the meaning intended by the expert and the interpretations made by the participants. This finding is consistent with literature demonstrating that decision makers tend to undervalue probabilistic information (Faigman & Baglioni, 1988; Goodman, 1992; Kaye & Koehler, 1991; Smith et al., 1996, 2011).

Alternatively, this pattern of results may reflect an appropriate weighting of the evidence, given its relevance to the actual *guilt* of the accused (Schum & Martin, 1982). Specifically, the presence of the defendants' shoeprint at the crime scene is a piece of circumstantial evidence that has the potential to implicate the defendant in the crime but does not directly speak to whether the defendant actually committed larceny or not. Put another way, having been at a crime scene does not mean that you are the perpetrator of the crime. Our participants may have been sensitive to this distinction, and, accordingly, it may be that the low levels of weight attributed to the evidence by the decision makers demonstrate a sophisticated appreciation for the value of circumstantial evidence within the broader case context, rather than a misapplication of the evidence. Although this alternative explanation does not undermine observed between-groups differences or the observed weak evidence effect, further studies in which the expert testifies regarding evidence with direct implications for the guilt of the accused should be conducted to tease apart these alternative accounts and clarify the cause of the apparent undervaluation.

The Weak Evidence Effect

In both studies, the inversion characterizing a type of weak evidence effect (Fernbach et al., 2011) was most frequently observed in low-strength conditions, particularly for which verbal communication methods were used. These data suggest that verbal rather than numerical methods of expression are more open to this kind of misinterpretation. It is particularly interesting, however, that participants presented with exculpatory rather than inculpatory expert evidence were not affected in the same way. When a significant interaction between evidence strength and presentation method was observed in Experiment 1, this same interaction was not evident in the exculpatory conditions in Experiment 2. That is, participants presented with weak evidence pointing toward innocence interpreted this as increasing the likelihood of innocence, whereas those presented with weak evidence consistent with guilt also interpreted this as increasing the likelihood of innocence.

From a criminal justice perspective, this pattern of results is encouraging in that it appears to demonstrate that individuals making decisions within a criminal justice context (even if experimentally induced) are sensitive to the asymmetry built into the system. Jurors participating in a trial are explicitly required to give more weight to the rights of the accused than to the State by virtue of the presumption of innocence, a range of procedural rules, and burden of proof standards. Similarly, participants in our experiments were asked to imagine they were a juror in a trial, were made aware of the requirements for a larceny conviction, and were advised to use a reasonable-doubt threshold. Thus, within such a context, it may be appropriate to see that weak prosecution evidence (inculpatory) generally does not strengthen belief in guilt (and, in fact, does the opposite), whereas weak defense evidence (exculpatory) is more likely to be used to strengthen belief in the innocence of the accused.

This pattern of results is, however, more concerning if considered from the perspective of the expert. Expert evaluative opinions, whatever their specific formulation, are presented to fact finders in order to help them to reach an "accurate resolution to a dispute in issue" (*United States v Downing*, cited in Cutler & Penrod, 1995, p. 27). Neither the possible undervaluing of the weight of the expert's evidence, nor the incongruent revision of beliefs that we have observed to result from the application of the standards proposed by the AFSP, are consistent with this aim. Accordingly, it would seem appropriate for the AFSP to reconsider at least the use of these verbal equivalents in the presentation of low-strength inculpatory evidence.

Lastly, when considered from a psychological standpoint, the asymmetry of the weak evidence effect observed here may have important theoretical implications. At least three competing mechanisms have been proposed to account for the weak evidence effects observed in various contexts. Stated simply, the *neglect* account suggests that the effect results from a failure to consider other information (e.g., other circumstantial evidence of guilt) that could also impact belief in a proposition (Fernbach et al., 2011). The *averaging* explanation contends that the error is a function of averaging the amount of support offered to the proposition by the new evidence with our prior beliefs in that proposition, thereby reaching a midpoint between the two (Lopes, 1987). The *expectancy violation* mechanism suggests that an incongruent revision occurs when we compare the actual support offered with the proposition by the new evidence, against the amount of support we expected the new evidence would provide. When the actual support offered is smaller than what we expected, we revise down our original belief, despite being given more evidence to support it (McKenzie et al., 2002).

Although each of these theoretical accounts has found support in the literature, only the expectancy violation explanation can account for the results we report here. The fact that weak evidence effects were more prominent in the incriminating rather than the exculpatory conditions cannot be explained by neglect, as the external information available (to be neglected) in both instances was the same. Moreover, given that the inculpatory and exculpatory evidence also provided the *same amount of support* ("weak or limited") for the proposition (although the order of these propositions differed across conditions), the data would only support the averaging explanation if (a) most people in the exculpatory condition began with prior beliefs lower than the amount of support eventually offered by the expert (thereby preventing the midpoint between the two being lower than their priors), and (b) most in the inculpatory condition began with prior beliefs higher than the amount of support offered by the expert (resulting in a midpoint between the two being lower than their priors). To test the validity of this explanation, a comparison of the mean prior beliefs in the

verbal low-strength inculpatory and exculpatory conditions was conducted and revealed no significant difference (M_{inculp} = 2.21, M_{exculp} = 3.19, t = 0.58, df = 136, p = .571, two-tailed), indicating the averaging account is not supported here.

Only the expectancy violation explanation allows for the different expectations one might have regarding exculpatory and inculpatory evidence to affect the belief revision process. Specifically, it is reasonable to believe that decision makers in a forensic context might expect the prosecution to introduce highly incriminating evidence, if they are to secure a conviction. Thus, when presented with evidence offering only "weak or limited support" for an inculpatory proposition, decision makers revise, in a downward direction, their belief in the guilt of the accused, even though they have been presented with additional information supporting the hypothesis, in essence saying, "If this is the best the prosecution can offer, then I am not persuaded the defendant is guilty at all." Conversely, decision makers may have lower expectations regarding exculpatory evidence. In fact, they may properly have no expectations at all, given that they are required to presume the defendant is innocent and that the burden of proof is borne by the prosecution (i.e., the defendant has no obligation to mount a positive defense). Thus, even weak exculpatory evidence has the potential to surpass the decision-maker's expectation, removing any need to revise down their belief in the innocence of the accused in the face of weak evidence.

Limitations and Future Directions

We did not target, select, or analyze only jury-eligible respondents in these experiments. We would argue that an emphasis on jury-eligible respondents was unnecessary, given that we were primarily interested in the impact of various types of expert evidence on decision making. As we have no basis for believing that the individuals in our studies would differ systematically or significantly from a jury-eligible sample in their response to expert evidence, we stand by our approach.

A second issue worthy of consideration is the recruitment of participants via the online marketplace Mechanical Turk. It is possible that individuals completing our study online may have been less diligent or engaged in the task compared with individuals completing the task under the supervision of the experimenter, and this may threaten the validity of the results. Various steps were taken in these experiments to ensure this was not the case. First, in both experiments, we ran a validation sample of undergraduates alongside those completing online. The patterns of results produced by our student and online samples were the same. Second, we set rigorous predefined exclusion criteria using a recommended "catch test" in addition to a speed-based performance measure, leading to the exclusion of participants whose data was of questionable quality. Third, Experiment 2 involved the repetition and clear replication of two conditions from Experiment 1. As a result, we are confident that the data obtained online speaks to real-world decision-making performance on this task.

In conclusion, it is clear from this research that decision makers vary widely in their responses to uncertain forensic science evidence, revising their beliefs in vastly different ways than those predicted by Bayesian calculations. What remains unclear, however, is what mechanism(s) account for the observed discrepancies and therefore how they can be minimized. Further research examining central, rather than peripheral, evidence, as well as competing accounts for over and undervaluing of evidence, is necessary to clarify our understanding of uncertain evidence and to inform its communication.

References

Aitken, C., Berger, C. E. H., Buckleton, J. S., Champod, C., Curran, J., Dawid, A., . . . Jackson G. (2011). Expressing evaluative opinions: A position statement. *Science & Justice, 51,* 1–2. doi:10.1016/j.scijus.2011.01.002

Association of Forensic Science Providers. (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice, 49,* 161–164. doi:10.1016/j.scijus.2009.07.004

Barbato, G., Barini, E., Genta, G., & Levi, R. (2011). Features and performance of some outlier detection methods. *Journal of Applied Statistics, 38,* 2133–2149. doi:10.1080/02664763.2010.545119

Berger, C. E. H. (2010). Criminalistics is reasoning backwards. *Nederlands Juistenblad, 85,* 784–789.

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes, 41,* 390–404. doi:10.1016/0749-5978(88)90036-2

Budescu, D. V., Broomell, S., & Por, H. H. (2009). Improving communication of uncertainty in the reports of the Intergovernmental Panel on Climate Change. *Psychological Science, 20,* 299–308. doi:10.1111/j.1467-9280.2009.02284.x

Budescu, D. V., Por, H. H., & Broomell, S. (2012). Effective communicating of uncertainty in the IPCC reports. *Climatic Change, 113,* 181–200. doi:10.1007/s10584-011-0330-3

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3–5. doi:10.1177/1745691610393980

Cutler, B. L., & Penrod, S. D. (1995). *Mistaken identification: The eyewitness, psychology, and the law*. Cambridge, UK: Cambridge University Press.

de Keijser, J., & Elffers, H. (2012). Understanding of forensic expert reports by judges, defense lawyers and forensic professionals. *Psychology, Crime & Law, 18,* 191–207. doi:10.1080/10683161003736744

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the subjective numeracy scale. *Medical Decision Making, 27,* 672–680. doi:10.1177/0272989X07304449

Faigman, D. L., & Baglioni, A. (1988). Bayes' theorem in the trial process. *Law and Human Behavior, 12,* 1–17. doi:10.1007/BF01064271

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). When good evidence goes bad: The weak evidence effect in judgment and decision-making. *Cognition, 119,* 459–467. doi:10.1016/j.cognition.2011.01.013

Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: From innumeracy to insight. *British Medical Journal, 27,* 741–744. doi:10.1136/bmj.327.7417.741

Goodman, J. (1992). Jurors' comprehension and assessment of probabilistic evidence. *American Journal of Trial Advocacy, 16,* 361.

Kaye, D. H., & Koehler, J. J. (1991). Can Jurors Understand Probabilistic Evidence? *Journal of the Royal Statistical Society, Series A,* 75–81. Retrieved from: http://www.jstor.org/stable/2982696

Koehler, J. J. (1996). On conveying the probative value of DNA evidence: Frequencies, likelihood ratios, and error rates. *University of Colorado Law Review, 67,* 859–886.

Koehler, J. J., & Saks, M. J. (2010). Individualization claims in forensic science: Still unwarranted (June 1, 2010). *Brooklyn Law Review, 75,* 1187.

Lopes, L. L. (1987). Procedural debiasing. *Acta Psychologica, 64,* 167–185. doi:10.1016/0001-6918(87)90005-9

McKenzie, C. R. M., Lee, S. M., & Chen, K. K. (2002). When negative evidence increases confidence: Change in belief after hearing two sides of a dispute. *Journal of Behavioral Decision Making, 15,* 1–18. doi:10.1002/bdm.400

McQuiston-Surrett, D., & Saks, M. J. (2008). Communicating opinion evidence in the forensic identification sciences: Accuracy and impact. *Hastings Law Journal, 59,* 1159–1190.

Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice, 51,* 91–98. doi:10.1016/j.scijus.2011.03.002

Nance, D. A., & Morris, S. B. (2005). Juror understanding of DNA evidence: An empirical assessment of presentation formats for trace evidence with a relatively small random match probability. *The Journal of Legal Studies, 34,* 395–444.

National Academies of Science. (2009). *Strengthening forensic science in the United States: A path forward*. Washington, D.C.: The National Academies Press.

Neumann, C., Evett, I. W., & Skerret, J. (2012). Quantifying the weight of evidence from a forensic fingerprint comparison: A new paradigm. *Journal of the Royal Statistical Society Series A (Statistics in Society), 175,* 371–415. doi:10.1111/j.1467-985X.2011.01027.x

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5,* 411–419.

Petty, R. E., & Cacioppo, J. T. (1996). *Attitudes and persuasion: Classic and contemporary approaches*. Boulder, CO: Westview Press.

R v T, EWCA Crim 2439 (Court of Appeal–Criminal Division, 2010).

Schum, D. A., & Martin, A. W. (1982). Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review, 17,* 105–151.

Smith, B. C., Penrod, S. D., Otto, A. L., & Park, R. C. (1996). Jurors' use of probabilistic evidence. *Law and Human Behavior, 20,* 49–82. doi:10.1007/BF01499132

Smith, L. L., Bull, R., & Holliday, R. (2011). Understanding juror perceptions of forensic evidence: Investigating the impact of case context on perceptions of forensic evidence strength. *Journal of Forensic Sciences, 56,* 409–414. doi:10.1111/j.1556-4029.2010.01671.x

Stoel, R. (2012). A practitioners view on the application of the likelihood ratio approach in cartridge case and bullet comparison. In the *21st International Symposium on the Forensic Sciences: Convicts to criminalistics*. Retrieved from http://events.cdesign.com.au/ei/viewpdf.esp?id=314&file=//srv3/events/eventwin/docs/pdf/anzfss2012abstract00374.pdf

Thompson, W. C. (1989). Are juries competent to evaluate statistical evidence? *Law and Contemporary Problems, 52,* 9–41.

Thompson, W. C., & Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior, 11,* 167–187. doi:10.1007/BF01044641

Thornton, J., & Peterson, J. (2007). The general assumptions and rationale of forensic identification. In D. L. Faigman, D. Kaye, M. J. Saks, J. Sanders, & E. Cheng (Eds.), *Modern scientific evidence: The law and science of expert testimony* (pp. 157–222). St Paul, MN: Thomson West.

Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *Knowledge Engineering Review, 10,* 43–62. doi:10.1017/S0269888900007256

Received September 17, 2012
Revision received February 3, 2013
Accepted February 4, 2013 ■