Forensic Science International 240 (2014) 61–68

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect

K.A. Martire *, R.I. Kemp, M. Sayle, B.R. Newell

School of Psychology, The University of New South Wales, Sydney 2052, NSW, Australia

ARTICLE INFO

Article history:
Received 8 October 2013
Received in revised form 28 March 2014
Accepted 1 April 2014
Available online 14 April 2014

Keywords:
Evidence interpretation
Expert evaluative opinions
Communicating uncertainty
Likelihood ratios

ABSTRACT

Likelihood ratios are increasingly being adopted to convey expert evaluative opinions to courts. In the absence of appropriate databases, many of these likelihood ratios will include verbal rather than numerical estimates of the support offered by the analysis. However evidence suggests that verbal formulations of uncertainty are a less effective form of communication than equivalent numerical formulations. Moreover, when evidence strength is low a misinterpretation of the valence of the evidence – a "weak evidence effect" – has been found. We report the results of an experiment involving $N = 404$ (student and online) participants who read a brief summary of a burglary trial containing expert testimony. The expert evidence was varied across conditions in terms of evidence strength (low or high) and presentation method (numerical, verbal, table or visual scale). Results suggest that of these presentation methods, numerical expressions produce belief-change and implicit likelihood ratios which were most commensurate with those intended by the expert and most resistant to the weak evidence effect. These findings raise questions about the extent to which low strength verbal evaluative opinions can be effectively communicated to decision makers at trial.

In many jurisdictions the practice of communicating forensic science expert opinions to courts is undergoing substantial change. Where it was once accepted practice that an expert would testify to categorical individualization [1,2], increasingly expressions reflecting the uncertain nature of forensic analyses are being demanded [3], recommended [4–7], and explored (e.g., [8–17], for detailed consideration).

The use of a likelihood ratio to express the subjectivity and uncertainty associated with forensic science evidence has been embraced by sectors of the forensic science community. In a position statement signed by 31 stakeholders and agencies these scientists declared likelihood ratios to be "the most appropriate foundation for assisting the court in establishing the weight that should be assigned . . ." [6]. The likelihood ratio (LR) is a statement which conveys the probability of the observations given each of the stated propositions or hypotheses (H). For example the likelihood ratio communicates the probability of obtaining the observed similarities between a fingerprint from a known origin and the fingerprint of questioned origin under the hypothesis that the two samples have the same origin (H_1) versus under the hypothesis that they have different origins (H_2) [18].

Critically, however, the signatories to the above mentioned statement appeared to suggest that the preferred form of expression for the likelihood ratio statement is verbal rather than numerical. This is a position supported by the Standards proposed by the Association of Forensic Science Providers [7] who proposed a scale for the translation of numerical likelihood ratios into verbal formats (see Table 1).

Accordingly, taking the approach recommended by Aitken et al. [6] it is preferred, for example, that the expert state: "In my opinion the correspondence between the fingerprint found at the crime scene and the fingerprint taken from the accused offers strong support if the two fingerprints originated from the same person than if the two fingerprints originated from different people"; Rather than: ". . . the correspondence between the fingerprint found at the crime scene and the fingerprint taken from the accused is 5500 times more likely if the two fingerprints originated from the same person . . ."; (see Table 1).

In addition to the perception that verbal expressions of the likelihood ratios are "the most appropriate basis for communication of an evaluative expert opinion to the court . . ." [6], it is also the case that the quantitative data necessary to compute a numerical likelihood ratio are unavailable in many domains of forensic science [12]. This means that, irrespective of their actual or perceived appropriateness, verbal expressions of uncertainty are likely to be observed with increasing frequency in the forensic sciences, at least until issues regarding data availability are resolved.

* Corresponding author. Tel.: +61 2 9385 8563; fax: +61 2 9385 3641.

http://dx.doi.org/10.1016/j.forsciint.2014.04.005
0379-0738/© 2014 Elsevier Ireland Ltd. All rights reserved.

It is also important to consider the intended audience for these expressions of uncertainty [19] and how their interpretations might be influenced by verbal and numerical expressions. Several avenues of research suggests that people often have difficulties understanding probabilities and statistics, and tend to produce suboptimal translations between verbal and numerical expressions of uncertainty. In particular, evidence suggests decision-makers tend to, but don't always [20] under-value probabilistic evidence compared with normative estimates (e.g., [8,10,11,13,21–23]). Furthermore, it is widely acknowledged that different people will understand the same verbal probability expression differently [24–29] leading to conclusions that verbal labels create an "illusion of communication" [30]. Consequently, it is not appropriate to simply assume a particular probability phrase will automatically and reliably result in a specific desired interpretation [9,24].

Despite these concerns it has been suggested that verbal expressions of uncertainty and evaluative labels can also be beneficial, remediating some of the misinterpretations associated with numerical probabilities [4,25,28,31–33]. In an attempt to resolve this ambiguity surrounding the relative "appropriateness" of verbal and numerical expressions, particularly in the context of likelihood ratios, Martire et al. [34] compared the amount of belief change resulting from expert forensic science opinions, expressed as verbal or numerical likelihood ratios of varying strength (low, moderate and high). To do this they used the labels and numerical equivalents recommended by AFSP for evidence offering "weak or limited", "moderately strong" and "very strong" support (see Table 1). Across two web-based studies involving 905 participants Martire et al. measured the extent to which participants' belief in the guilt or innocence of the accused changed after being presented with the testimony of an expert shoe impression examiner. The testimony varied in strength as described above, but always indicated that the likelihood of the observed similarity between the shoe print from the crime scene and the shoe print of the accused was more likely if the two prints shared a common origin (H_1) than if they had different origins (H_2) (i.e., was evidence in support of the prosecution case).

Three main effects emerged across the two studies: (1) A broad sensitivity to evidence strength was observed such that expert opinion evidence of greater strength resulted in significantly more belief change than did lower strength expert evidence; (2) a tendency to underweight the evidence compared to Bayesian norms; specifically, participants did not update their initial beliefs to the extent that would be predicted through the application of Bayes theorem; and (3) a weak evidence effect was observed for low strength verbal evidence when brought by the prosecution. That is, where participants were presented verbal evidence which "weakly" supported the prosecution's version of the case, rather than increase their belief in the guilt of the accused by a small amount as would be appropriate given the additional incriminating evidence, the majority of participants elected instead to decrease their belief in the guilt of the accused. This effect was not statistically significant where the evidence was presented numerically.

The weak evidence or "boomerang" [35] effect describes a situation where weak evidence supporting a proposition, in this case H_1, is wrongly interpreted as evidence supporting the alternate proposition H_2 [36]. In practice this meant that the expert's opinion which should have supported the prosecution's case was interpreted as supporting the defense case by a clear majority of participants in the low strength verbal conditions. Although not previously unknown [10,36–38], and to some extent context dependent [34], this inversion of the valence of the opinions of forensic scientists is somewhat concerning. Specifically, these weak evidence effects are of concern not only because they inaccurately reflect the valence of the expert's opinion, but also because of the stated belief that verbal expressions of evidence should be used by forensic science experts because they provide the most appropriate basis for communication [6]. Overall then, the observed undervaluing and weak evidence effects beg the question, if verbal expressions are not the most appropriate basis for communication, what possible alternative formulation might be?

Budescu and colleagues [25,28] suggest that presenting both verbal expressions and numerical values, in the context of the complete range of possible options (i.e., a table of values and expressions) can improve the interpretation of verbal expressions of uncertainty. In their 2009 study, Budescu and colleagues asked participants to read 13 sentences containing probabilistic terms from the 2007 Intergovernmental Panel on Climate Change (IPCC) report and provide a best estimate of the probability intended by the authors [25]. Participants were allocated to one of four conditions: (1) the control group were provided no instruction regarding the interpretation of the probabilistic term; (2) the translation group were provided a drop-down table including all verbal labels and their numerical equivalents (e.g., >99%); the verbal–numerical group (which was further split into two conditions) were provided with the table including either (3) a broad or (4) a narrow range of numerical values to accompany the verbal expressions which was presented alongside each sentence.

The researchers found that although consistency with the IPCC conversion table was generally low, consistency was significantly higher in the translation than control condition leading the authors to recommend the use of "both verbal terms and numerical values to communicate uncertainties" [25]. This conclusion was confirmed in a follow up study using a nationally representative US sample of 556 participants, which found that verbal–numerical scales (including numerical ranges) increased the differentiation between terms, the internal consistency of each term, and the correspondence with the IPCC report's intended message [28]. These results suggest that a dual form of expression not only provides more information, but also caters to a broad and heterogeneous group of decision-makers. What remains unclear however, is the extent to which the provision of similar verbal–numerical tables will improve interpretations of the more complex form of expression associated with likelihood ratios, specifically, where decision makers are explicitly asked to consider the likelihood of the observations under two competing hypotheses.

Research by de Keijser and Ellfers [12] begins to address this question by explicitly considering the interpretation of likelihood ratios in a forensic science context. Participants were 332 judges and justices, defense lawyers and employees of the Dutch Forensic Institute (NFI) who were presented with likelihood ratios reflecting expert evaluative opinions using either verbal probability statements or visual scales. Decision makers were presented with the expert's conclusion regarding the likelihood of the observations given two scenarios (S) corresponding with the two hypotheses underpinning likelihood ratio (e.g., S1: the tape used to restrain the victim originates from the roll of tape that was seized from the suspect's residence; S2: the tape used to restrain the victim originates from a random other roll of tape). In the visual conclusion condition the expert's opinion was indicated with an 'X' intersecting a horizontal line labeled from "Very strong in favour of Scenario 2" on the left, to "Very strong in favour of Scenario 1" on the right, with a "Neutral" point in the middle. Analyses showed that understanding of likelihood ratios was generally poor and that using visual scales as a "cosmetic attempt" to improve understanding neither improved or impaired participant performance.

It is, however, unclear how these visual scales and numerical–visual dual expressions of likelihood ratios might affect belief-change – rather than comprehension as in de Keijser and Elffers

[12], or correspondence as in Budescu et al. [25,28] – when compared with numerical expressions alone. Additionally, it is not known whether such visual scales and tabular expressions are equally susceptible to weak evidence effects.

In this study we examine differences in the extent of belief-change resulting from four methods of communicating likelihood ratios to lay decision-makers: separate (1) verbal and (2) numerical expressions as in Martire et al. [34]; (3) a dual expression tabular format representing the verbal–numerical translation table produced by the AFSP [7] (see Table 1); and (4) a visual scale depicting the strength of the evidence as a location on a line ranging between the two alternate propositions as per [12]. Each of these presentation methods were tested using a low strength and a high strength evidence formulation. Considering first the issue of belief updating relative to evidence strength (i.e., as measured by belief-change and the calculation of implied likelihood ratios), it is predicted that: (i) The verbal–numerical table condition will provide the highest correspondence of the four presentation modalities based on the findings of Martire et al. [34], which demonstrate closer correspondence as a result of numerical rather than verbal expressions of evidence strength. Similarly, the findings of Budescu et al. [25,28] also place the correspondence resulting from verbal–numerical scales (including numerical ranges) as superior to that of verbal expressions alone; (ii) the correspondence observed in the numerical only condition will be poorer than the verbal–numerical table condition but better than the verbal condition, again reflecting the combined findings of Budescu et al. [25,28] and Martire et al. [34], which place expressions utilizing numerical components as more readily interpretable than verbal only expressions; (iii) given this evidence, the verbal only and visual scale expressions are predicted to provide the poorest correspondence of the four expression types, yet are not predicted to differ from each other based on the absence of differences noted by de Keijser and Elffers [12].

In this experimental context the presence or absence of significant weak evidence effects may also be considered an indicator of the correspondence between expert intentions and lay interpretations of evidence, with a weak evidence effect serving to reduce correspondence. Although various mechanisms have been proposed to account for the weak evidence effects and have found varying support in different contexts, none of the available explanations apply well to scenarios where decision makers are presented with equivalent information in different formats as is proposed in this study.

The neglect account predicts weak evidence effects where decision makers fail to consider other information relevant to the propositions being evaluated [36]. In the current experimental paradigm however "other" case information is held constant across evidence presentation types, therefore no differences in the presence or prevalence of weak evidence effect would be predicted.

Similarly the averaging [37] and expectancy violation [39] accounts predict weak evidence effects as a result of mistakenly combining (averaging) our prior beliefs, or through a violation of our initial (high) expectations of evidence strength with the addition of (low strength) evidence—causing a final belief that is lower than would be predicted from a normative belief-updating perspective. Yet as with the neglect account above, the different evidentiary presentation styles are not expected to systematically vary with regard to participants' prior beliefs or their expectations regarding the evidence strength; therefore, it is not possible to derive predictions regarding the presence or absence of weak evidence effects in the current study from these theories.

One factor not considered in these explanations of the weak evidence effect, which may differ across the proposed evidence presentation methods, is the ease with which decision makers are able to locate the evidence they have been presented in regard to the propositions they are evaluating. That is, it is possible that weak evidence effects emerge as a result of the decision maker mistaking evidence in favor of one proposition as evidence in favor of the alternative proposition because the form of expression made it difficult to correctly identify which proposition the evidence actually supports.

Consider the verbal expression for low strength evidence proposed by the Association of Forensic Science Providers [7] "weak or limited support". It is possible that weak evidence effects may be observed in response to this formulation because participants mistakenly interpret the phrase "weak support" as being the opposite of (as opposed to less than) "strong support" for the proposition. This may in turn decrease their belief in the proposition producing a downward revision despite the provision of additional supportive evidence.

Importantly, this ease of location account of the weak evidence effect goes some way to helping derive predictions for the current study; more precisely, the weak evidence effect will be less prevalent where participants are presented stimuli that facilitate the location of the evidence with regard to the propositions being evaluated. Understood in this way we would predict that weak evidence effects will be the strongest or the most prevalent in the verbal only condition where participants have no information regarding the range of possible values of evidence strength and where the expression "weak" is used in a somewhat ambiguous manner.

Conversely, we do not anticipate significant weak evidence effects in the numerical, tabular or scale conditions as each of these forms of presenting the evidence facilitate the location of the additional evidence with regard to the two propositions being evaluated. In the numerical condition this is facilitated by using the phrase "times more likely" to indicate a positive multiplier; in the tabular condition it is achieved by pairing the ambiguous verbal expression with the above mentioned positive multiplier; and finally in the visual scale condition it is done by using a mark on a line to locate the evidence with regard to a neutral mid-point and the two alternate propositions.

1. Method

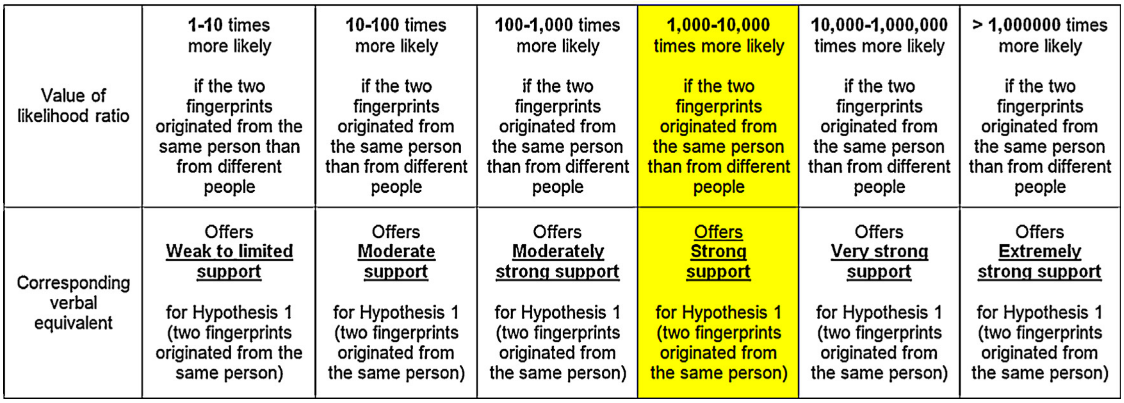
1.1. Design

A 2 (evidential strength: low or high) × 4 (presentation method: numerical, verbal, table or visual scale) between-subjects factorial design was employed.

1.2. Participants and data screening

In order to secure approximately 50 participants per condition, a total sample of 477 residents of the United States were recruited from an online self-enlisted workforce [40,41] and were compensated US50¢ for their time. Of these participants, 73 (15.3%) were excluded based on three predefined criteria[1] resulting in a final sample of $N = 404$, (low strength $n_{\text{numerical}} = 54$, $n_{\text{verbal}} = 53$, $n_{\text{table}} = 53$, $n_{\text{scale}} = 52$; high strength $n_{\text{numerical}} = 44$, $n_{\text{verbal}} = 51$, $n_{\text{table}} = 41$, $n_{\text{scale}} = 56$). The majority (88.1%) of the sample reported a tertiary level of education (at least commenced). Males accounted for 57.9% of the sample and the mean age was 30.09 years (range 18–74, SD = 10.98). Almost the entire sample (91.6%)

[1] Participants were excluded if they: (1) completed the experiment in less than 120 s ($n = 2$); (2) failed the "catch-trial" ($n = 21$) [41]; (3) belief change value was classified as an 'extreme outlier' falling outside 7 times the interquartile range ($n = 50$) [42].



Value of likelihood ratio	**1-10** times more likely if the two fingerprints originated from the same person than from different people	**10-100** times more likely if the two fingerprints originated from the same person than from different people	**100-1,000** times more likely if the two fingerprints originated from the same person than from different people	**1,000-10,000** times more likely if the two fingerprints originated from the same person than from different people	**10,000-1,000,000** times more likely if the two fingerprints originated from the same person than from different people	**> 1,000000** times more likely if the two fingerprints originated from the same person than from different people
Corresponding verbal equivalent	Offers **Weak to limited support** for Hypothesis 1 (two fingerprints originated from the same person)	Offers **Moderate support** for Hypothesis 1 (two fingerprints originated from the same person)	Offers **Moderately strong support** for Hypothesis 1 (two fingerprints originated from the same person)	Offers **Strong support** for Hypothesis 1 (two fingerprints originated from the same person)	Offers **Very strong support** for Hypothesis 1 (two fingerprints originated from the same person)	Offers **Extremely strong support** for Hypothesis 1 (two fingerprints originated from the same person)

Fig. 1. Example of table presentation method for high strength evidence.

indicated that, to the best of their knowledge, they were eligible for jury duty.

1.3. Materials and procedure

1.3.1. The minimal trial

All participants were presented with a brief written description of a hypothetical burglary trial. Participants were asked to imagine that they were a juror in a trial and were presented with the following case facts: (a) a house in a wealthy suburb was broken into while its occupants were at work; (b) the perpetrator entered the house by breaking a window; (c) the perpetrator unsuccessfully attempted to open a safe inside the house; (d) the accused was questioned in the vicinity of the house shortly after the crime; (e) at the time of his arrest the accused was wearing clothes similar to those described by a witness; (f) the accused was under significant financial stress at the time of the offense; and (g) the accused did not have an alibi for the time of the burglary. They were also informed that the accused denied ever having been to the house in question and plead not guilty. Participants were then informed that in order to return a guilty verdict they must be satisfied beyond a reasonable doubt that (1) the accused person; (2) broke into and entered the premises; (3) with the intent to steal property.

1.3.2. Expert evidence

Each participant also read the testimony of an experienced expert forensic science analyst who compared a fingerprint found on the safe at the crime scene with the fingerprint of the accused. As a result of his analysis the expert stated: "when assessing the significance of any similarity or differences between two fingerprints, the likelihood of obtaining that similarity or difference is considered against two alternative propositions: (Hypothesis 1 (H_1)) the two fingerprints originated from the same person; (Hypothesis 2 (H_2)) the two fingerprints did not originate from the same person". Participants were then provided one of eight possible versions of the expert's opinion based on their allocated condition (evidence strength: low or high by evidence presentation method: numerical, verbal, table or visual scale).

In all conditions the expert testimony was derived from the Standards proposed by the Association of Forensic Science Providers [7]; see Table 1. For example, a participant in the numerical presentation method condition next read: "In my opinion the correspondence between the fingerprint found on the safe from the crime scene and the fingerprint taken from the accused [is **5.5 times more likely**; (low strength)] or [is **5500 times more likely**; (high strength)] if the two fingerprints originated from the same person (Hypothesis 1) than if the two fingerprints originated from different people (Hypothesis 2)" [emphasis added]. In the verbal condition the text in bold was replaced with the words "offers weak or limited support" in the low strength evidence condition, or "offers strong support" in the high strength condition.

Those receiving the numerical–verbal table of evidence read that "the correspondence between the fingerprint found on the safe from the crime scene and the fingerprint taken from the accused is as highlighted in the table below" and were presented with Fig. 1 where the relevant evidence strength cells (low or high) were highlighted in yellow. As in Budescu et al. [25,28], both the verbal probability phrase and a range of numerical values were presented. Furthermore, to convey a likelihood ratio, as opposed to a simple likelihood statement, we also included the two alternative propositions considered by the expert when formulating their opinion in a form of words matching each expression type.

Those presented with a visual scale depicting the evidence read that the "correspondence ... is as marked with regard to the two hypotheses using the red 'X" on the line below" and were presented with either Panel A (low strength) or Panel B (high strength) in Fig. 2 based on the scale used by de Keijser and Ellfers [12].

The visual representation was constructed such that it encompassed values from −10,000 (in favor of Hypothesis 2) to +10,000 (in favor of Hypothesis 1), with a neutral point in the middle offering no support for either hypothesis. The indicator for low strength evidence was placed as close as possible to the neutral point without intersecting it to reflect the likelihood ratio of 5.5. The indicator for high strength evidence was placed fractionally past the midpoint of the positive side of the scale to reflect the likelihood ratio of 5500.

1.3.3. Participant responses and procedure

After accessing the experimental materials online, consenting to participate, 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 not in original]. If the participant had expressed an original preference for the "not guilty" option they were given the same question with the order of the bold terms reversed. The response to this question was taken as the participant's prior-belief in the guilt of the accused.

Participants were then presented with one of the eight types of expert evidence (according to their random allocation to condition) before being asked the same two questions again; providing a

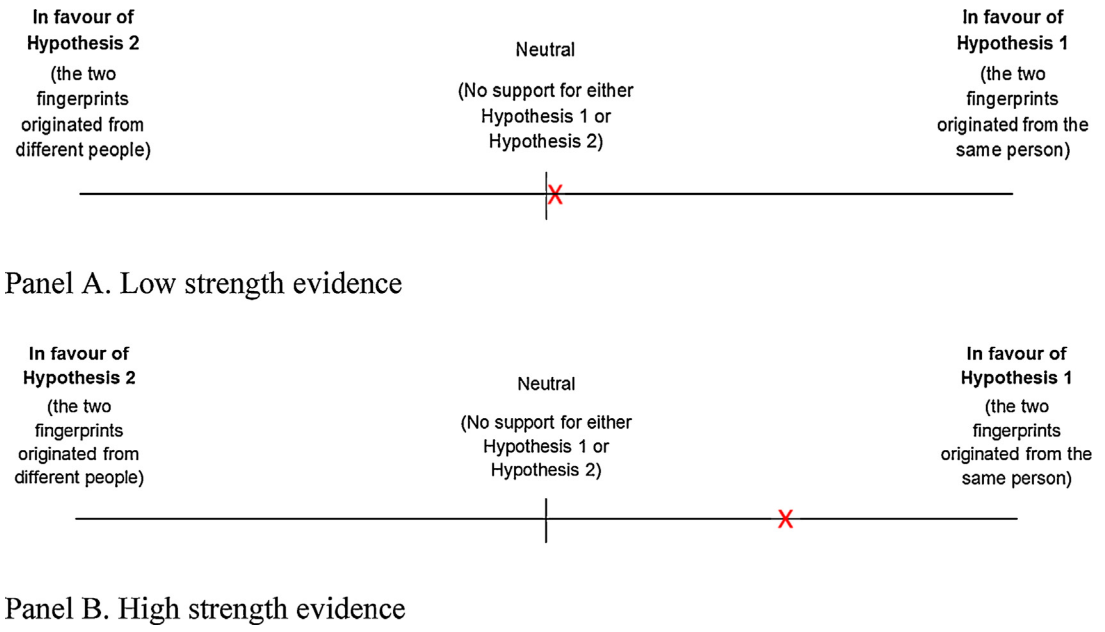

In favour of Hypothesis 2 (the two fingerprints originated from different people)
Neutral (No support for either Hypothesis 1 or Hypothesis 2)
In favour of Hypothesis 1 (the two fingerprints originated from the same person)
X
Panel A. Low strength evidence
In favour of Hypothesis 2 (the two fingerprints originated from different people)
Neutral (No support for either Hypothesis 1 or Hypothesis 2)
In favour of Hypothesis 1 (the two fingerprints originated from the same person)
X
Panel B. High strength evidence

Fig. 2. Visual scales depicting low and high strength evidence. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

posterior-belief value. They were also asked to complete a series of demographic questions and the subjective numeracy scale (SNS: [43]) which assesses numerical fluency. The entire procedure took approximately 15 min to complete.

2. Results

2.1. Belief-change

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. For example if a participant began by believing the defendant was 2 times more likely to be not guilty than guilty (prior = −2), and finished believing the defendant was 2 times more likely to be guilty than not guilty (posterior = 2), that person will have a belief-change score of 4 (posterior minus prior).

A 2 × 4 ANCOVA was conducted to examine the impact of evidence strength and presentation method 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 Fig. 3). Neither of these covariates were significant, leaving a significant main effect for evidential strength such that high strength evidence resulted in greater adjusted mean belief-change ($M = 3.43$, 95% CI: 2.76–4.10) than low strength evidence ($M = -0.08$, 95% CI: −0.72–0.55, $F(1,394) = 55.99$, MSE = 1222.14, $p < 0005$, partial $\eta^2 = 0.124$); and a significant main effect of presentation method ($F(3,394) = 6.52$, $p < 0005$, MSE = 142.26, partial $\eta^2 = 0.047$). There was also a significant interaction effect ($F(3,394) = 4.55$, MSE = 99.29, $p < 005$, partial $\eta^2 = 0.033$). An inspection of cell means suggested that when evidence strength was high the presentation method formats were equivalent. However, when evidence strength was low, numerical expressions resulted in significantly greater belief-change ($M = 2.71$, 95% CI: 1.46–3.96) than was observed in the other conditions ($M_{\text{verbal}} = -2.38$, 95% CI: −3.65 to −1.12; $M_{\text{table}} = -0.25$, 95% CI: −1.51–1.02; $M_{\text{scale}} = -0.42$, 95% CI: −1.69–0.86).

This interpretation is supported by two follow up one-way ANCOVAs each examining the effect of presentation method on belief-change given low or high strength evidence after controlling for SNS score and prior belief values. Where evidence strength was low (and neither covariate was significant) there was a main effect of presentation method ($F(1,203) = 11.67$, MSE = 232.90, $p < 0005$, partial $\eta^2 = 0.145$). Simple comparisons showed the numerical presentation differed significantly from the verbal, table and visual scale presentation formats, causing greater belief-change. By contrast, where evidence strength was high (and in the absence of significant covariates), presentation method was not significant ($F(1,186) = 0.80$, MSE = 19.15, $p = 0.50$, partial $\eta^2 = 0.013$).

2.2. Correspondence between provided and implicit likelihood ratios

As another gage of the degree of correspondence between the expert's intention and the jurors' interpretation of the evidence,

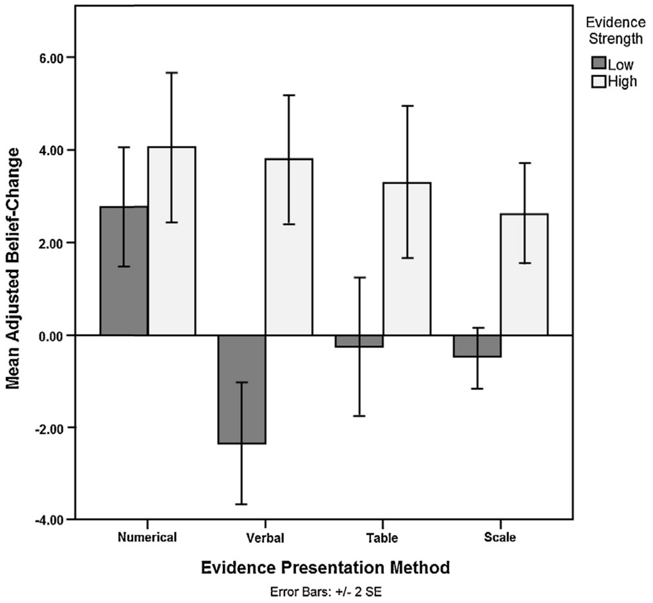


Fig. 3. Mean adjusted belief-change by evidence presentation and strength.

Table 1
Standards for numerical and verbal expression of likelihood ratios.

Recommended likelihood ratio terminology	
Numerical expression	Verbal expression (support)
>1–10	Weak or limited
10–100	Moderate
100–1000	Moderately strong
1000–10,000	Strong
10,000–1000,000	Very strong
>1000,000	Extremely strong

the data were used to calculate the likelihood ratio that would account for each participant's change from their prior to their posterior belief. This "implicit" likelihood ratio value was then compared to the likelihood ratio provided in the evidence given by the expert.

The implicit likelihood ratio was computed 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 which favored guilt while 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 while 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 likelihood ratio which would need to be used to move each participant from their prior-odds to their posterior-odds, using the formula likelihood ratio = 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 likelihood ratio which would achieve a change from a prior of 0.5 to a posterior of 2 is 2/0.5 = 4.

The implicit likelihood ratios necessary to account for the change from prior- to posterior-odds were calculated for each participant. The median values are reported in Table 2 below. These data suggest a considerable undervaluing of the experts' testimony in all conditions, with the undervaluing reaching massive proportions in the case of the high strength condition. Across the low and high strength conditions the expert attributed values of either 5.5 or 5500, participants adjusted their prior-beliefs using a median likelihood ratios that were between 2.51 and 11 times smaller than intended in the low strength conditions and between 1000 and 2000 times smaller than intended in the high strength conditions (Table 2).

Table 2
Median implied likelihood ratio by presentation method and evidence strength.

Evidence presentation method	Median implied likelihood ratio adopted by participants (range)	
	Low strength	High strength
Numerical	2.19 (0.13–15.15)	4.54 (0.00–50.00)
Verbal	0.50 (0.00–2.50)	4.61 (0.00–100.00)
Table	1.41 (0.00–10.00)	2.90 (0.00–10.00)
Visual scale	1.00 (0.00–4.00)	2.74 (0.00–25.00)

Table 3
Percent of participants updating their belief toward innocence after hearing the expert evidence.

Evidence presentation method	Percent moving toward innocence	
	Low strength	High strength
Numerical	12.96	11.36
Verbal	64.15	1.96
Table	32.08	12.2
Visual scale	38.46	16.07

2.3. Weak evidence effect

Returning now to the belief-change scores, further investigation allows for the examination of the weak evidence effect. The average 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 average belief-change observed in the low strength verbal, table and visual scale presentation conditions was in the opposite direction to that intended by the expert (i.e., the presentation of the evidence increased belief in innocence—see Fig. 3).

To assess whether the average belief-change in the low evidence strength, verbal, table and visual scale conditions differed significantly from zero one sample *t*-tests were conducted for each of these three conditions. The results indicated that an overall weak evidence effect was found in the verbal condition ($M_{\text{diff}} = -2.34$, $t_{52} = -3.58$, $p < 0.005$) but not in either the table ($M_{\text{diff}} = -0.26$, $t_{52} = -0.35$, $p = 0.725$) or visual scale conditions ($M_{\text{diff}} = -0.50$, $t_{51} = -1.52$, $p = 0.135$).

The weak evidence effect was also examined by calculating the proportion of participants in each condition who made a downward revision of their belief in the guilt of the accused after reading the expert's evidence, see Table 3. A majority of those in the low strength verbal condition (64.15%) demonstrated a weak evidence effect, compared to 32.08% in the table, 38.46% in the visual scale and only 12.96% in the numerical conditions. In the high evidence strength conditions the average percentage of those demonstrating a weak evidence effect was only 10.40% (across presentation conditions).

Significantly, the weak evidence effect was not due to participants initially believing in the innocence of the accused and sticking to that belief despite the evidence of the expert. About half of the participants in the low strength verbal condition ($n = 17$) initially favored guilt but then shifted to a not-guilty decision after hearing the evidence.

3. Discussion

Likelihood ratios are increasingly being embraced as the most appropriate means for communicating the uncertainty associated with expert opinions in the forensic sciences [4,6,7]. The manner in which these expressions are interpreted by decision-makers is, however, of vital importance [19] and not well understood by either forensic science policy makers or psychologists. Previous examinations of lay comprehension of probabilistic statements have identified generalized undervaluing of evidence as well as idiosyncratic interpretations of verbal expressions of uncertainty [13,23,34,36].

Our results suggest that numerical expressions, as opposed to verbal, dual verbal–numerical (table) expression and visual (scale) methods, produce belief-change and implicit likelihood ratios most commensurate with the intentions of the expert. This method of presentation was also the most resistant to weak evidence effects. This pattern of results is broadly consistent with studies which have raised concerns regarding the potential "illusion of communication" [30] associated with verbal expressions as compared to numerical formats [9,24], and provides preliminary insights into the relative performance of dual expressions and visual presentations as compared to numerical expressions of likelihood ratios.

Despite these results supporting previous expressions of concern regarding verbal expressions of uncertainty, our specific prediction that the verbal–numerical presentation method would provide the greatest correspondence with expert intentions was not borne out.

Closer inspection reveals that performance in the verbal, table and scale conditions was undermined by weak evidence effects. For the low strength verbal condition this is a clear replication of the results in Martire et al. ([34], also showing directional errors of around 64%) and consistent with the ease of location account. In contrast ease of location cannot account for the substantial proportion of those in the table and visual scale conditions who also made directional errors (32.08% and 38.46%, respectively); nor the errors (12.96%) observed in the numerical condition.

Overall then, given the contextual frameworks available in the numerical, table and visual scale conditions, it is difficult to argue that the proposed location error accounts entirely for the weak evidence effects observed in this study, although the inability to contextualize the information in the verbal condition may have contributed to the much larger weak evidence effect observed there. The fact that all participants successfully complied with quality control requirements (see footnote 1) makes it unlikely that inattention is the sole underlying cause, although, again we cannot completely rule out some contribution from a failure to engage with the material as might be suggested by the roughly 10% of participants who also made directional errors in response to 'strong' evidence.

In all likelihood elevated rates of directional errors arise via a combination of different evaluation mechanisms—and identifying the precise manner in which these mechanisms interact awaits future research. Notwithstanding the complete specification of these contributory mechanisms, these data suggest clearly that numerical expressions of evidence, particularly low strength evidence are the most commensurate with the intentions of the expert and are the most resilient to weak evidence effects. For this reason there appears to be evidence-based wisdom in using numerical expressions to express uncertainty whenever possible.

4. Limitations

The direct comparison of implied likelihood ratios with those likelihood ratios provided by the expert in their evidence continues to pose a challenge. As discussed in Martire et al. [34] participants in this experiment were ultimately evaluating propositions regarding the guilt of the defendant, yet the evidence the expert presented related to the likelihood of observed similarities between the crime scene and suspect print under the defense (different sources) and prosecution (same source) hypotheses. This disconnect between the evidence presented and the propositions being evaluated by the decision makers (i.e., non-aligned hypotheses), although potentially pervasive in real testimony [44], could reasonably lead to a disconnect between the strength of (fingerprint) evidence provided, and the amount of (guilt) belief-updating observed among participants. That is, evidence that the defendant's fingerprint was found at the crime scene is not necessarily relevant to determining whether he is guilty of burglary in this case. Accordingly, it might be reasonable to expect smaller changes in guilt belief than would be the case if the evidence also pertained directly to the guilt of the defendant.

We attempted to address this issue in these experimental materials by purposely making the fingerprint evidence highly incriminating (i.e., very relevant to the issue of guilt). We did this by: (a) placing the questioned fingerprint in an incriminating location (on the safe); (b) by selecting a crime (burglary) where the perpetrators presence at the crime scene was inculpatory (i.e., having entered the premises); and (c) by having the defendant deny ever having been at the crime scene, thereby removing any innocent explanation for the presence of his fingerprint at the location. Thus, the crime scenario was constructed such that if the expert's evidence made a juror believe that the fingerprint found at the crime scene belongs to the suspect, then she should also increase her belief in the likelihood that the suspect is guilty of the crime burglary. That is, under the assumption that participants had a clear understanding of the implications of the minimal case, implicit likelihood ratios should approach Bayesian normative predictions. Yet despite this, the evidence continues to suggest a substantial underweighting of the evidence.

From a pragmatic standpoint it is also possible to argue that it is suboptimal to exclude the responses of participants on the basis that they are statistical outliers (falling outside 7 times the interquartile range) given that in the real world this fact would have no bearing on juror eligibility. We willingly acknowledge this inconsistency as a limitation to our approach, however we note that in this study the extreme variability in participant belief-change values (which are ultimately free to take on any positive or negative value) made a trade-off between statistical interpretability and participant representativeness unavoidable. Only by reducing the enormous variability introduced by a small proportion of the overall sample (12.3%) was it possible to statistically analyze and describe the broader trends reflected in the responses of the remaining 87%. Although it is not clear how to best optimize this trade-off, future research examining the extent to which extreme belief-change by individuals in a group (jury) might impact upon the belief-change of others in the group would undoubtedly be fruitful.

5. Conclusion

The claim made in the 2011 position statement [6] that verbal formulations of likelihood ratios are the most appropriate basis for communicating with decision-makers looks increasingly questionable, as does the broader ideal of the Bayesian juror. Verbal expressions for low strength evidence, at least those suggested by the AFSP [7], reliably generate weak evidence effects in our studies, indicating that decision makers invert the valence of the expert evidence when presented in this manner. Moreover, in high strength evidence conditions, verbal expressions are no more likely than numerical, dual verbal–numerical or visual methods to produce interpretations of expert evidence which closely correspond with intended meanings.

Thus, where an expert must render an opinion regarding low strength evidence, our results suggest numerical, rather than verbal evidence will produce the highest correspondence between expert intentions and decision-maker interpretations. Furthermore, given the high likelihood of miscommunication associated with low strength verbal expressions of uncertainty, in instances where numerical values for low strength evidence cannot be provided, it would seem appropriate to question whether expert opinions in the form of verbal likelihood ratios should be offered at all. This is especially true in situations where jurors are not presented with a contextual framework in which to interpret that evidence. However, before making such a drastic recommendation, it would be valuable to conduct further research focusing on alternative forms of expression to those suggested by the AFSP [7]. In particular, future work could examine the potential for juror education to remedy directional errors associated with low strength evidence implicating the accused.

Acknowledgments

This article was facilitated in part by funding from the Australian Research Council to Kristy A. Martire (Discovery Early Career Researcher Award DE140100183), Richard I. Kemp (Linkage Project LP110100448) and Ben R. Newell (Future Fellowship FT110100151).

References

[1] K. Inman, N. Rudin, Principles and Practices of Criminalistics: The Profession of Forensic Science, CRC Press, Boca Raton, FL, 2001.

[2] J.W. Osterburg, The evaluation of physical evidence in criminalistics: subjective or objective process? J. Crim. Law, Criminol. Police Sci. 60 (1969) 97–101.

[3] Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, 2009.

[4] I. Evett, Towards a uniform framework for reporting opinions in forensic science casework, Sci. Justice 38 (1998) 198–202.

[5] C.E.H. Berger, Criminalistics is reasoning backwards, Ned. Juristenblad 85 (2010) 784–789.

[6] C. Aitken, C.E.H. Berger, J.S. Buckleton, C. Champod, J. Curran, A. Dawid, et al., Expressing evaluative opinions: a position statement, Sci. Justice 51 (2011) 1–2.

[7] Association of Forensic Science Providers, Standards for the formulation of evaluative forensic science expert opinion, Sci. Justice 3 (2009) 161–164.

[8] D. McQuiston-Surrett, M.J. Saks, The testimony of forensic identification science: what expert witnesses say and what factfinders hear, Law Hum. Behav. 33 (2009) 436–453.

[9] D. McQuiston-Surrett, M.J. Saks, Communicating opinion evidence in the forensic identification sciences: accuracy and impact, Hastings Law J. 59 (2008) 1159.

[10] B.C. Smith, S.D. Penrod, A.L. Otto, R.C. Park, Jurors' use of probabilistic evidence, Law Hum. Behav. 20 (1996) 49–82.

[11] L.L. Smith, R. Bull, R. Holliday, Understanding juror perceptions of forensic evidence: investigating the impact of case context on perceptions of forensic evidence strength, J. Forensic Sci. 56 (2011) 409–414.

[12] J. de Keijser, H. Elffers, Understanding of forensic expert reports by judges, defense lawyers and forensic professionals, Psychol. Crime Law 18 (2012) 191–207.

[13] D.H. Kaye, J.J. Koehler, Can jurors understand probabilistic evidence? J. R. Stat. Soc. Ser. A (Stat. Soc.) 154 (1991) 75–81.

[14] V.P. Hans, D.H. Kaye, B.M. Dann, E.J. Farley, S. Albertson, Science in the jury box: jurors' comprehension of mitochondrial DNA evidence, Law Hum. Behav. 35 (2011) 60–71.

[15] M. Juanchich, K.H. Teigen, G. Villejoubert, Is guilt 'likely' or 'not certain'?: contrast with previous probabilities determines choice of verbal terms, Acta Psychol. 135 (2010) 267–277.

[16] G. Jackson, Understanding forensic science opinions, in: J. Fraser, R. Williams (Eds.), Handbook of Forensic Science, Willan, Devon, UK, 2009, pp. 419–445.

[17] Special issue: papers on the R v T debate, C. Aitken (Ed.), Law Probability and Risk, 11, 2012, pp. 255–375.

[18] G.S. Morrison, Measuring the validity and reliability of forensic likelihood-ratio systems, Sci. Justice 51 (3) (2011) 91–98.

[19] P. Campbell, Understanding the receivers and the reception of science's uncertain messages, Philos. Trans. R. Soc. Ser. A: Math. Phys. Eng. Sci. 369 (2011) 4891–4912.

[20] W.C. Thompson, S.O. Kaasa, T. Peterson, Do jurors give appropriate weight to forensic identification evidence? J. Empirical Legal Stud. 10 (2013) 359–397.

[21] D.L. Faigman, A. Baglioni, Bayes' theorem in the trial process, Law Hum. Behav. 12 (1988) 1–17.

[22] J. Goodman, Jurors' comprehension and assessment of probabilistic evidence, Am. J. Trial Advoc. 16 (1992) 361.

[23] A. Corner, A.J.L. Harris, U. Hahn, Conservatism in belief revision and participant skepticism, in: Proceedings of the 32nd Annual Conference of the Cognitive Science Society, Cognitive Science Society, Austin, TX, 2010, pp. 1625–1630.

[24] T.S. Wallsten, D.V. Budescu, A review of human linguistic probability processing: general principles and empirical evidence, Knowledge Eng. Rev. 10 (1995) 43–62.

[25] D.V. Budescu, S. Broomell, H.H. Por, Improving communication of uncertainty in the reports of the intergovernmental panel on climate change, Psychol. Sci. 20 (2009) 299–308.

[26] W. Brun, K.H. Teigen, Verbal probabilities: ambiguous, context-dependent, or both, Organ. Behav. Hum. Decis. Processes 41 (1988) 390–404.

[27] N. Shaw, P. Dear, How do parents of babies interpret qualitative expressions of probability, Arch. Dis. Child. 65 (1990) 520–523.

[28] D.V. Budescu, H.H. Por, S. Broomell, Effective communicating of uncertainty in the IPCC reports, Clim. Change 113 (2012) 181–200.

[29] V.H.M. Visschers, R.M. Meertens, W.W.F. Passchier, N.N.K. De Vries, Probability information in risk communication: a review of the research literature, Risk Anal. 29 (2009) 267–287.

[30] D.V. Budescu, T.S. Wallsten, Consistency in interpretation of probabilistic phrases, Organ. Behav. Hum. Decis. Processes 36 (1985) 391–405.

[31] N.F. Dieckmann, E. Peters, R. Gregory, M. Tusler, Making sense of uncertainty: advantages and disadvantages of providing an evaluative structure, J. Risk Res. 15 (2012) 717–735.

[32] C.R. Fox, J.R. Irwin, The role of context in the communication of uncertain beliefs, Basic Appl. Soc. Psychol. 20 (1998) 57–70.

[33] M.C. Politi, P.K.J. Han, N.F. Col, Communicating the uncertainty of harms and benefits of medical interventions, Med. Decis. Making 27 (2007) 681–695.

[34] K.A. Martire, R.I. Kemp, I. Watkins, M.A. Sayle, B.R. Newell, The expression and interpretation of uncertain forensic science evidence: verbal equivalence, evidence strength and the weak evidence effect, Law Hum. Behav. 37 (2013) 197–207.

[35] R.E. Petty, J.T. Cacioppo, Attitudes and Persuasion: Classic and Contemporary Approaches, Westview Press, Boulder, CO, US, 1996.

[36] P.M. Fernbach, A. Darlow, S.A. Sloman, When good evidence goes bad: the weak evidence effect in judgment and decision-making, Cognition 119 (2011) 459–467.

[37] L.L. Lopes, Procedural debiasing, Acta Psychol. 64 (1987) 167–185.

[38] A. Harris, A. Corner, U. Hahn, James is polite and punctual (and useless): a Bayesian formalisation of faint praise, THINK REASONING (2013), http://dx.doi.org/10.1080/13546783.2013.801367.

[39] C.R.M. McKenzie, S.M. Lee, K.K. Chen, When negative evidence increases confidence: change in belief after hearing two sides of a dispute, J. Behav. Decis. Making 15 (2002) 1–18.

[40] M. Buhrmester, T. Kwang, S.D. Gosling, Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data, Perspect. Psychol. Sci. 6 (2011) 3–5.

[41] G. Paolacci, J. Chandler, P.G. Ipeirotis, Running experiments on Amazon Mechanical Turk, JUDGM DECIS MAK 5 (2010) 411–419.

[42] G. Barbato, E. Barini, G. Genta, R. Levi, Features and performance of some outlier detection methods, J. Appl. Stat. 38 (2011) 2133–2149.

[43] A. Fagerlin, B.J. Zikmund-Fisher, P.A. Ubel, A. Jankovic, H.A. Derry, D.M. Smith, Measuring numeracy without a math test: development of the subjective numeracy scale, Med. Decis. Making. 27 (2007) 672–680.

[44] N. Fenton, D. Berger, D. Lagnado, M. Neil, A. Hsu, When 'neutral' evidence still has probative value (with implications from the Barry George Case), Sci. Justice, http://dx.doi.org/10.1016/j.scijus.2013.07.002.