Forensic Science International: Synergy 1 (2019) 24–34

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: https://www.journals.elsevier.com/forensic-science-international-synergy/

Juror comprehension of forensic expert testimony: A literature review and gap analysis

Heidi Eldridge

RTI International, 3040 E. Cornwallis Rd., Research Triangle Park, NC, 27709, USA

ARTICLE INFO

Article history:
Received 19 October 2018
Received in revised form
1 March 2019
Accepted 4 March 2019
Available online 9 March 2019

Keywords:
Expert testimony
Juror comprehension
Verbal scale
Likelihood ratio
Strength of evidence
Cognitive psychology

ABSTRACT

Forensic scientists and commentators including academics and statisticians have been embroiled in a debate over the best way to present evidence in the courtroom. Various forms of evidence presentation, both quantitative and qualitative, have been championed, yet amidst the furor over the most "correct" or "accurate" way to present evidence, the perspective of the fact-finder is often lost. Without comprehension, correctness is moot. Unbeknownst to many forensic practitioners, there is a large, though incomplete, body of literature from the cognitive psychology domain that explores the question of what jurors understand when forensic scientists testify. This body of work has begun to test different proposed methods of testimony in an effort to understand which are most effective at communicating the strength of evidence that is intended by the expert. This article is a review of that literature that is intended for the forensic scientist community. Its aim is to educate that community on the findings of completed studies and to identify suggestions for further research that will inform changes in testimony delivery and ensure that any modifications can be implemented with confidence in their effectiveness.

1. Introduction

For years, a growing segment of the forensic science community has been striving to institute reforms in the way conclusion testimony is delivered in a court of law. Entire conferences[1] and special publications of journals[2] have been dedicated to the question of the best way to responsibly present forensic evidence, and the 2009 National Research Council report, *Strengthening Forensic Science in the United States: A Path Forward* [1], was generally critical of the way forensic scientists at the time overstated the strength of evidence[3] without supporting data (see, e.g., pp. 4, 21–22, 52–53, 87–88, 127–182).

The forensic science and statistical communities have been earnestly debating the "best" way to present conclusions, and this debate is still active and energetic today.[4] However, with the best of intentions, this debate seems to frequently neglect to include an important group of commentators—the cognitive psychologists. While the practitioners and academics are embroiled in a scientific debate over the most logically, mathematically, or reliably "correct" or "accurate" way to present forensic evidence, they often forget to stop and consider the consumer of this information—the fact-finder.

While the goal of providing accurate and logically cohesive evidence is a laudable one, if juries cannot understand or appropriately apply the testimony that is given, then forensic science has not been effective, transparent, or ultimately, useful to the trier of fact. Psychologists have conducted a large body of research to examine the question of how laypeople (taken as proxies for potential jurors) understand conclusions presented in many different quantitative and qualitative formats.

This article reviews a large portion of the available literature in juror comprehension of forensic science testimony, offers some insight into the common themes in the literature, and includes

E-mail address: heldridge@rti.org.

[1] For example, the National Institute of Standards and Technology hosted both a Technical Colloquium on the Weight of Evidence (2016 and 2017) and an International Forensic Science Error Management Symposium (2015 and 2017).

[2] For example, the journal Science & Justice dedicated a special issue (issue 56, 2016) to airing the debate over the use of likelihood ratios in presenting forensic science evidence.

[3] A phrase used to describe how much support the findings give to one proposition versus the other (such as the proposition that two impressions originated from the same source versus from different sources).

[4] The recent Impression, Pattern and Trace Evidence Symposium featured several interactive panel discussions centered around this topic. Information and archived recordings can be found at: https://forensiccoe.org/2018-impression-pattern-and-trace-evidence-symposium/.

https://doi.org/10.1016/j.fsisyn.2019.03.001

recommendations for future research. Its intent is to expose the forensic science community to a relevant body of literature with which many may not be familiar, and to provide a snapshot of the current state of the field. This article is a review only and does not present any new research. It would be both premature and disingenuous for this article to suggest solutions that should be implemented, although it does endeavor to point out both helpful conclusions that have been previously reached and areas where further research would be beneficial.

The structure of the article is broken into four sections. First, we will summarize the literature that has explored the pros and cons of several different modes of presenting evidence to fact-finders (Section 2). Second, we will discuss juror perceptions of credibility—both of the testifying expert, and of the evidence type itself (Section 3). Third, we will discuss the problem of how to teach a jury complex science within the constraints of testimony (Section 4). Finally, we will review generalizations that can be gleaned from the reviewed literature and make recommendations for future research (Section 5).

2. Evidence presentation methods

Historically, many forensic science disciplines have reported their results in absolute terms—that this trace *was left* by that source. Recent critical documents (see, e.g. Refs. [1,2]) have denounced this language as overstating the strength of the evidence, leading to debates over what is the most appropriate way to convey forensic results. Some scholars have advocated for a numerical scale, others for a verbal scale. Many seem to favor a likelihood ratio[5] (LR) approach, which can be presented either numerically or verbally, while still others have proposed a hybrid approach in which both verbal and numerical information is presented. In this section, we review some of the literature in which each of these approaches has been evaluated for its impact on the comprehension of the layperson.

2.1. Quantitative presentation of evidence

Forensic disciplines that have a solid statistical foundation based on population databases, such as single-source DNA comparison, often report their results using a quantitative measure, such as the Random Match Probability (RMP). The RMP is a means of expressing the chance of a coincidental match of a given set of features in a population.[6] This is often a favored mode of evidence presentation because it is measurable and provides a veneer of objectivity.

Unfortunately, multiple studies have demonstrated that laypeople struggle to understand what the RMP really means. Thompson et al. [3] found that the RMP was often so misunderstood that subjects actually interpreted it to mean the exact *opposite* of what it was intended to convey.[7] This suggests that examiners should take great care to ensure that the direction of increased strength of evidence is explicitly explained to the fact-finder to avoid misinterpretation.

In addition, both laypeople and attorneys frequently confound the RMP to mean the chance the defendant is innocent, rather than the chance of selecting a random individual from the population who possesses the same features as the defendant [4]. In point of fact, the chance of an incorrect association being reported due to examiner, system, or random error is generally orders of magnitude higher than the RMP[8] [4,6].

One challenge of the RMP and other quantitative measures is that they often require the fact-finder to perform mathematical computations. However, research has shown that laypeople struggle to perform the correct calculations to interpret quantitative testimony.

For example, McQuiston-Surrett and Saks [7] asked subjects to calculate the number of people who could share hair characteristics based upon quantitative testimony.[9] In the best performing scenario, fewer than 50% of subjects correctly answered the question. In a more difficult trial, only about 25% of subjects responded correctly. These results suggest that many laypeople are not able to extrapolate the numbers that are needed from testimony, then apply the correct mathematical function to those numbers.

A research group led by Gigerenzer has tried to simplify the quantitative presentation of evidence to improve comprehension. Gigerenzer explains that presenting the exact same RMP as a single-event probability rather than as a frequency statement can drastically impact the perception of the strength of the evidence. For example, the single-event probability statement “The probability that this match has occurred by chance is 1 in 100,000” sounds as though it is highly unlikely that someone else could be the source of the evidence. However, the frequency statement “Out of every 100,000 people, 1 will show a match” makes it sound likely that there are other matches out there, particularly if it is followed up by the information that in a city of 2 million adults, you could expect 20 of them to “match” the evidence [8]. This effect has also been demonstrated by McQuiston-Surrett and Saks [7].

Additionally, as Jackson et al. point out [9], the context in which a number is presented can change the listener's perception—a 15% probability of rain is considered a low risk, while a 15% chance of a heart attack may be perceived as a high risk. So, when numerical probabilities are presented to jurors, what will the presentation context tell them about how they should react to those numbers?

Gigerenzer's group further advocates for the use of natural frequencies to describe the strength of evidence. Natural frequencies incorporate both the chance of an error (false positive rate) and the prevalence of the features in the population (base-rate) to present the positive predictive value (the chance that a person or item is truly the source of a trace, given that an identification was made). They developed a 2-hour training program that was shown to increase laypeople's understanding of Bayesian reasoning [10], created a simple visual way of representing the math behind calculating the positive predictive value [11], and conducted

[5] The likelihood ratio (LR) is a number that expresses the examiner's assessment of how strongly the forensic findings support one of two propositions, typically represented as the prosecutor's hypothesis (e.g. the defendant left the unknown sample) and the defense's hypothesis (e.g. someone else left the unknown sample).

[6] An RMP statement might sound something like, “The probability of an unknown individual possessing these features is approximately 1 in 1 million.” Thus, if the RMP is larger (e.g., 1 in 100), you are saying that more people are likely to share these features, i.e. there is a 1 in 100 chance that if you pulled a single person at random out of the population, they would share these features, i.e. the features are more common.

[7] Some participants believed a higher RMP (e.g., 1 in 10) meant the evidence was stronger, rather than a lower RMP (e.g., 1 in 100,000).

[8] For example, consider the false positive error rate for the fingerprint comparison discipline as reported by the FBI/Noblis black box study [5] – 0.17% – compared to a typical random match probability, which is commonly smaller than 1 in 1,000,000.

[9] Participants were asked the question, “Suppose this crime occurred in a city of 500,000 people. If every person in that city were tested, approximately how many people do you think would have hair that is indistinguishably similar to the hair recovered from the crime scene?” This question was essentially a math word problem where the information necessary to do the arithmetic was provided in various forms during the testimony. The participants had to extrapolate the correct math that was needed, which involved, for example, multiplying 500,000 by 0.001 in the most difficult condition.

research specifically aimed at presenting natural frequencies in a DNA testimony context [4].

While this research is exciting and promising, it currently has limitations. The courtroom setting does not allow for a 2-hour educational program to help jurors understand statistics. And positive predictive value, while helpful to a jury, requires the prevalence of the features within a population to calculate. While this might be a good solution for single-source DNA evidence, it is not yet implementable in many other forensic disciplines where the frequency of features in a population is not known.

Gigerenzer et al. [12] also argue for the need for a reference class for quantitative information and note that quantitative scales are not more precise than qualitative ones if the reference class is not provided. For example, if a weather forecaster makes the statement, "There is a 30% probability of rain tomorrow", to what does the percentage refer? Does it mean that it will rain for 30% of the day, or in 30% of the viewing area, or that 30% of the time when this prediction is made, some amount of rainfall will occur?

A final, common concern about presenting statistical evidence is that jurors will overweight it compared to what was intended. Yet in a review of the literature by Kaye and Koehler [13], the opposite was found: jurors tended to *underweight* statistical evidence. This was also demonstrated by Martire et al., whose participants updated their beliefs in the correct direction according to the evidence presented, but at a magnitude over 350,000 times smaller than what the expert had intended [14,15] and by McQuiston-Surrett and Saks [7] whose participants' belief in guilt increased after the presentation of population rate data, but not as much as expected.

2.2. Verbal presentation of evidence

In lieu of validated statistical models capable of providing quantitative values to describe the strength of evidence, it is often suggested to use a verbal scale. On the surface, this seems like a good solution—examiners can explain how strong they feel the evidence is even though they do not have models to rely on, and jurors will not have to interpret any confusing math. However, there are pros and cons to the use of verbal scales as well. One of the biggest cons is that words can be very personal and subjective and there is a real danger that, as with quantitative methods, jurors simply will not interpret them in the way the forensic examiner intended to convey them.

One of the earliest influential articles supporting the use of a verbal scale was by Aitken and Taroni [16]. In it, the authors argue that people do not understand large numbers well and that a verbal scale can make the evidence more easily understandable. They promote the use of a logarithmic (log) scale, pointing out that people already have experience with log scales, such as the pH scale for measuring acidity, the Richter Scale for measuring the strength of an earthquake, and decibels for measuring sound; thus, they believe log scales will be easy for people to interpret. Indeed, the UK's Association of Forensic Science Providers (AFSP) uses a verbal log scale as the basis of their recommended scale [17]. However, the comprehension of a verbal log scale was not tested and jurors may struggle to understand and interpret the verbal explanation provided[10].

Despite this early recommendation, research on comprehension of the verbal scale has not been encouraging. Several research studies to test layperson comprehension of verbal scales found that while the average values of perceived strength of evidence increased with the verbal scale, there was overlap between all categories [18–20], indicating that there was not clear resolution between them. Additionally, it was found that categories at the low end of the scale (e.g., "weak" and "limited") [19] and the high end of the scale (e.g., "strong", "very strong", "extremely strong") [20] were essentially interchangeable. All but the weakest categories were undervalued in relation to the intended strength of evidence [19], which is in agreement with the findings reported above for quantitative presentations of strength of evidence.

These studies were generally unfavorable to verbal scales, yet the scales that they tested had inherent limitations. The Mullen et al. study [18] introduced an eight-interval equidistant scale, in contrast to the logarithmic six-interval scale from the AFSP [17]. This made it difficult to compare their findings to the verbal scale most typically used in testimony. The question of how many intervals a verbal scale should have is not a trivial one. Benjamin and Tullis [21] found that "decision noise" is introduced when too many options are presented in a scale, which increases the cognitive load necessary to keep track of all the options. Because of this, they suggest that any ordinal scale should be carefully balanced between cognitive load and resolution of the scale.

In addition to the question of how many intervals a verbal scale should have, in two of the above studies a line was provided along which participants were to mark the perceived strength of the evidence. Yet in one study [18], the line was blank other than a 0 at one end (no support) and a 1 at the other end (conclusive support), while the line in the other study [20] had no numbers at all. In neither of the aforementioned studies was the full verbal scale provided to give context of where the verbal expression might fall within the possible range of phrases. Given the lack of context and scale, it is perhaps unsurprising that participants did not weigh the strength of evidence as intended.

On the other hand, Sjerps and Biesheuvel [22] tested comprehension of a verbal scale and found that, even when provided the full scale, there was a large variation in interpretation between subjects. While their sample size was very small ($n = 14$), the results suggest that simply presenting the full verbal scale may not be sufficient to achieve clarity amongst jurors.

Marquis et al. caution against the presentation of the full verbal scale [23]. They feel that it presents a bias in implying that a likelihood ratio (LR) toward the top end of the range is somehow more "useful" than one lower in the range, which implies that LRs from different cases should be directly compared rather than taken only in the context of the case at hand. This argument will be discussed in greater depth later in this article.

In addition to testing specific verbal scales on potential jurors, there are broader psychological questions to consider in how verbal information is received generally. One relevant phenomenon is that while speakers prefer to *use* verbal expressions, listeners prefer to *receive* numerical expressions [24,25]. Of course, since other research has demonstrated that fact-finders struggle to interpret numerical expressions, this preference may not lead to better outcomes.

Base rate expectations can also impact the interpretation of verbal scale expressions. In a study by Wallsten, Fillenbaum, and Cox [26], subjects associated a higher number with the word "probable" to describe the chance of snow in the North Carolina mountains in December than in October. This effect could influence fact-finders in a court of law because jurors may come to a trial with preconceived notions about the likely guilt or innocence of various suspects based on demographics and may adjust their interpretation of "probable" accordingly.

Perhaps the greatest challenge for comprehension of verbal scales is not the fact that they are verbal, but the choice of the

[10] "The defendant has been selected at random from a village in which all but one of the people are guilty. The jury has then to decide if the defendant is one of the guilty people or if he is the one innocent person in the village."

words being used. Randall [27], a linguistics expert at Northeastern University, studied the effects of "legalese" and sentence structure in the context of jury instructions. She found that subject comprehension was significantly worse when instructions were complex and contained passive voice, presupposed information, and legal jargon. Furthermore, her subjects were university students, whom she notes are more likely than the average juror to comprehend complex sentence structures.

Additionally, Teigen and Brun [28,29] note that the directionality of language can have an effect on how it is taken by the listener. Terms with negative directionality, such as "unlikely" tend to focus the listener on the chance than an event does *not* occur, while terms with positive directionality, such as "likely" emphasize the chance that an event *does* occur.

Before considering any kind of testimony from expert witnesses, it may be wise to consider the lessons of linguistics and endeavor to put them into plain, understandable, and if possible, neutral English.

2.3. The likelihood ratio

While the use of a likelihood ratio can be incorporated in both quantitative and qualitative methods, its use has generated so much discussion in the literature that it warrants some separate discussion. The likelihood ratio is—as its name implies—a ratio of two likelihoods. In a forensic context, these are generally the likelihood of observing a given set of characteristics in common between two samples if they came from the same source and the likelihood of observing the same set of characteristics in common between the same two samples if they came from different sources (these are also known as the two propositions). The LR describes the strength of the evidence and can be used to update one's previous beliefs about the probability of an event (known as the prior probability) to form a new belief about the probability of the same event (known as the posterior probability). Within this framework, either quantitatively derived (through use of a statistical model) or qualitatively derived (using subjective probabilities[11] translated into numbers through the use of an equivalency chart) LRs may be used.

Many commentators have argued that the LR is the most logical way to convey the strength of forensic findings, while some [30] feel that because an LR is, by definition, a measure of *personal* belief in the likelihood of an observation, it should not be imposed on a jury. Because the expert is far better equipped than the layperson to opine on *any* strength of the evidence, a reasonable approach might be to present the personal LR of the expert, but couched in terms that explicitly state the assumptions and reasoning that went into their selection of the LR that is being presented. This transparency allows jurors to both understand *why* the expert believes in the presented LR and to adjust the strength of the evidence up or down in their own minds (effectively creating their own personal LR) if they are unconvinced by the expert's explanation. This, in essence, is exactly what has been suggested by ENFSI [31], whose guidelines provide examples of exactly how to calculate a personal LR and report on its genesis, and has been recently restated in a response to the above-mentioned critique [32].

Another area of debate regarding the LR is whether an uncertainty surrounding the number needs to be reported, or whether the uncertainty of the measurement is inherently built into the LR itself [33,34]. As argued by Martire et al. [35], an LR based on personal probabilities does not "encapsulate all uncertainty" (for example, it does not include the chance of practitioner, instrument, or random error). They further argue that the courts cannot evaluate an LR stated by the forensic scientist as a bare assertion.

Regardless of which side the reader aligns with in these debates, it is important to fully understand the recommendations that are being made for best practices in presenting LRs and to understand the literature that has evaluated how jurors receive LR information.

One of the trickiest things about using an LR is the question of the prior probability—who should assign it, and what should it be? Thompson et al. [36] address the appropriate assignment of LRs, discussing both whether a forensic scientist should assign a prior probability (generally no[12]) and if so, whether those prior odds should be 50:50 (again, typically no[13]).

Martire et al. [37] examined the relative effect of LRs presented numerically versus verbally on laypeople's interpretation of the strength of the evidence to determine whether the intent of one presentation type was better understood than the other. This study measured the change of belief in the guilt of the accused following the presentation of an LR. They found that numerical and verbal expressions of LRs produced approximately equal changes in belief—except for an LR of low strength. In this case, the change of belief was greater for numerical than verbal presentations. Low strength verbal presentations of LR were subject to the weak evidence effect (see box 1). In all cases, the change in belief was smaller than expected by orders of magnitude, indicating that laypeople severely undervalue both numerical and verbal expressions of LR.

One possible cause of inappropriate weighting of LRs is "ratio bias" [39]. Ratio bias occurs when an event is perceived as less likely when its chance is represented by a ratio of small numbers, as opposed to a ratio of large numbers. For example, an event with a probability of 1 in 19 is seen as less likely than an event with a probability of 10 in 190, although the two proportions are mathematically equal. This may be because people tend to focus more on the numerator, and thus judge based on what looks like a larger number—10 rather than 1. Stone et al. found that this effect is eliminated when a visual representation of the ratio is presented.

With so many conflicting opinions and studies regarding the use of the LR as a means of conveying the strength of the evidence, it can seem overwhelming to know what a best practice might look like. Marquis et al. [23] have suggested a clear and logical roadmap for presenting LRs with full transparency in an effort to avoid potential misunderstandings. Here, we closely examine their suggestions.

First, they recommend that the examiner not report on anything involving a prior probability. They further caution that the examiner should report the probability of the evidence given the proposition, *not* the probability of the propositions themselves, often referred to as the transposing the conditional (see box 2). Third, they propose a verbal scale that first states the *direction* of the support (i.e. in favor of the prosecution hypothesis or in favor of the

[11] A subjective probability is an expression of an examiner's belief in a proposition based upon their experience and currently observed findings. In a forensic science context, the observed findings can be the features within the known and unknown materials that are being considered. Based upon their training and experience, the examiner might make a statement such as, "in the past, when I have seen this much agreement, and this little disagreement, between two compared entities of this type, the probability of these observations resulting from different sources has been extremely low." In this case, no calculation has been done, but the examiner has assigned a subjective probability describing their personal belief of the probability of the outcome by using their experience.

[12] Unless the forensic scientist is also the fact-finder, such as when a coroner is given full responsibility for determining cause and manner of death.

[13] This is often done in civil paternity testing, but assumes the putative father begins with a 50% probability of guilt—a figure too high to satisfy the criminal court's presumption of innocence.

Box 1
The Weak Evidence Effect.

The weak evidence effect, coined by Fernbach et al. [38], refers to the observed phenomenon that sometimes, when information is presented that only *weakly* supports one proposition, it is interpreted by the receiver of the information as supporting the alternate proposition instead. In a forensic science context, this would mean that weak evidence in favor of the prosecution would be interpreted by a jury as favoring the defense. For example, a partial DNA profile or a fingerprint with characteristics in common with the suspect, but not sufficient to identify, would be interpreted by the jury to mean that the suspect was *not* the source of the trace.

Logically, of course, this cannot be true. If you visualize reaching a verdict as a set of scales where each piece of evidence adds a stone to either the prosecution or the defense side, even a weak piece of evidence (a very small stone) placed upon one side of the scale will tip the scale, ever so slightly, in that direction. A small stone added to the prosecution side can never tip the scale back toward the defense proposition.

Box 2
Transposing the Conditional

The phenomenon of transposing the conditional has plagued the presentation of forensic science evidence for some time, and has been discussed by many authors, since the phrase "prosecutor's fallacy" was first coined by Thompson and Schumann [40]. Transposing the conditional occurs when the likelihood of observing the evidence *given* the propositions is taken to be equal to the likelihood of the propositions, thus ignoring the influence of the prior probability. In a highly simplified form of Bayes' Theorem, the likelihood of a proposition = the likelihood of observing the evidence *given* that the proposition is true x the prior probability of the proposition. Thus, in order for the first two terms to be equal, the third term (the prior probability) would have to be ignored.

In a particularly clear analogy of why this is a logical fallacy, Sjerps and Biesheuvel [22] provide the following example (p. 215):

1 "If I am a monkey, then it is highly likely that I have two eyes, two arms and two legs.

Hence,

2 If I have two eyes, two arms and two legs, then it is highly likely that I am a monkey."

From this example, it is clear that 1 does not imply 2, yet this is the same fallacious logical train of thought that is used to conclude that if features are observed in common between a trace and a suspect, it is highly likely that the suspect made the trace.

defense hypothesis) followed by the *degree* of support for that hypothesis. By using this two-pronged approach (direction plus degree), it is hoped that there will be no ambiguity about which proposition is favored; in particular, this should eliminate the weak evidence effect.

Additionally, since LRs below 1, which favor the defense proposition (Fig. 1), are difficult to understand, Marquis et al. [23] recommend inverting the proposition to yield more understandable numbers. For example, an LR of 0.1 provides weak, or limited, support in favor of the defense proposition. Rather than keeping to the usual formula of "it is <x-many times > more likely to observe these data if the prosecution proposition were true than if the defense proposition were true", which would look like "it is 0.1 times more likely to observe these data if the prosecution proposition were true than if the defense proposition were true", they propose to invert the positions of the propositions *and* the numerical value. This would look like "it is 10 times more likely to observe these data if the defense proposition were true than if the prosecution proposition were true." This makes it abundantly clear that the defense proposition is favored, and 10 times more likely is a much easier value to understand than 0.1 times more likely (which is actually less likely).

We include here the exact example text provided by Marquis et al. (p.6, Appendix B) to illustrate this point, because it shows the cleanness and elegance of the proposal. Both propositions are clearly stated as well as the direction of the support, such that there can be no confusion regarding which proposition is favored, to what degree, or what the propositions are referring to:

> "The results provide support for the proposition that an unknown person — rather than Mr. Jones — signed the contested document. This support is qualified as weak or limited, as the results are in the order of 10 times more probable given that the proposition that an unknown person signed the contested document is true, rather than given that the alternative is true (i.e., Mr. Jones signed the contested document)."

The above was translated from the French in the Marquis et al. paper, so there is some room for simplification of the wording into more plain English. Nonetheless, the structure of the statement is wonderfully clear. They first state what is supported and what is not, then they state the numerical strength of that support. Finally, they explicitly restate what proposition is supported and what proposition is not.

In addition to stating the magnitude and direction of the LR, Marquis et al. are at pains to point out that an LR stated without any explanation is of no use to anyone. A high LR in a case with an enormous prior may have little discernible effect, while a low LR in a case with a tiny prior can make a huge difference. All this to say that the magnitude of the LR (strength of the evidence) must be considered in the context of the strength of the rest of the case.

To assist the trier of fact in comprehending this point, and appropriately applying the LR to the case at hand, Marquis et al. suggest providing the trier of fact with a table (Fig. 2) illustrating several prior probabilities, the LR being presented, and the corresponding posterior probabilities. In this way, the trier of fact can see the effect the offered LR will have on a low, medium, or high prior probability and adjust their posterior in accordance with the magnitude of their own prior in the case. Note that the trier of fact need not be adept at mathematics to perform this feat; the math has been provided for them in the chart. Even without a specific numeric prior, they should be able to recognize whether their general feeling about the guilt of the defendant prior to hearing the forensic evidence was weak, moderate or strong, and choose the corresponding lane to follow in the chart. It was further suggested

Proposition Supported	Likelihood Ratio	Verbal Expression
Prosecution	>10,000	Very strong evidence
	1,000 – 10,000	Strong evidence
	100 – 1000	Moderately strong evidence
	10 – 100	Moderate evidence
	>1 – 10	Limited evidence
1 = Inconclusive		
Defense	0.1 – <1	Limited evidence
	0.01 – 0.1	Moderate evidence
	0.001 – 0.01	Moderately strong evidence
	0.0001 – 0.001	Strong evidence
	<0.0001	Very strong evidence

Fig. 1. A likelihood ratio of exactly 1 indicates a true inconclusive – the evidence supports neither the prosecution nor the defense, or is perfectly balanced between the two. A likelihood ratio higher than 1 supports the prosecution proposition, and the magnitude of the LR indicates the degree of support for that proposition. A likelihood ratio between 0 and 1 supports the defense proposition, and the magnitude of the LR indicates the degree of support for that proposition.

Initial Degree of Belief (Prior Probability)	Strength of Evidence (LR)	Updated Degree of Belief (Posterior Probability)
0.1 – weak belief	10	0.5 – moderate belief
0.5 – moderate belief	10	0.9 – strong belief
0.9 – strong belief	10	0.99 – strong belief

Fig. 2. Table adapted from Marquis et al. (2016), showing how the application of the same LR to three different prior probabilities will result in three different posterior probabilities. Thus, an LR on its own is not informative; the context of the case and the strength of the prior probability also have an effect.

that this chart be appended to written reports to guide consumers of the information absent the testimony of an expert.

Finally, Marquis et al. suggest that a cautionary note be included in written reports to avoid an error of transposing the conditional by the reader. While, once again, the specific wording could be adjusted to convey the necessary points in more understandable plain English, the main elements are all present: a statement of what the results do *not* mean, a reference to other case information (priors) that figure into a determination of guilt, and an admonition that the duty to make a determination of guilt lies with the fact-finder, not the examiner. Once again, we reproduce the original text in its entirety here (p.7, Appendix D):

> "Note of caution: Our results do not mean that it is probably an unknown person who signed the document. Indeed, the probability that it is an unknown person (rather than Mr. Jones) who signed the contested document depends not only on the observations made on the signatures, but also on other elements (enquiry, testimony, other information). The evaluation of these other elements are of the domain of the Court and scientists should not give their opinion on the truth (or probability) of the propositions, but they should help the trier of facts by giving their probability of the observations given each proposition."

Unfortunately, the suggestions of Marquis et al. have not yet been systematically tested in a structured juror comprehension study to measure their effectiveness.

2.4. Combining multiple evidence presentation types

A growing body of research has attempted to directly compare the impact of multiple presentation types upon juror comprehension or to combine two types into a single presentation.

Martire et al. [37] compared juror belief change when presented with a verbal scale alone, a quantitative scale alone, a verbal and quantitative equivalency chart, and a visual depiction of the strength of the evidence along a line from defense to prosecution hypothesis with neutral at the midpoint. The results of this study supported that the quantitative condition produced belief changes that were most closely aligned with the intent of the examiner. The research also showed that only the quantitative condition was resistant to the weak evidence effect.

This research did contain potential weaknesses. First, the scales used in the different conditions were not fully equivalent. The verbal and quantitative equivalency chart condition presented a full range of possible conclusions that went all the way up to "> 1,000,000 times more likely" or "offers extremely strong support." However, the other conditions' ranges were capped at "1000 to 10,000 times more likely", "5500 times more likely", and "+10,000" respectively. This did not allow participants to understand the full range of possible values that was shared with them in the equivalency table condition. In addition, the visual scale was not labeled with any numbers, so participants had no way to anchor the X that was indicated against any particular value or against the full range of possible values.

Second, the wording of the equivalency chart was unclear and incomplete and may not have been clearly understood by participants. Each condition was labeled in full as, for example, "1000–10,000 times more likely if the two fingerprints originated from the same person than from different people". While we sympathize with a desire to keep the text in the box brief, this cuts out so many words that the meaning is no longer clear. For instance, *what* is 1000–10,000 times more likely? Given the propensity of laypeople, attorneys, and even forensic scientists to transpose the conditional, this is not a trivial point and should have been made explicit in the language.

Third, participants were asked to express their prior belief in the guilt of the suspect in numerical terms ("based on the available evidence I believe that it is ____ times more likely that the accused is guilty than not guilty"). Since many people lack numeracy, asking them to express their belief in this way may have resulted in inaccurate representations of the strength of participants' belief. Using something like a Likert scale may have provided a clearer indication of participants' agreement with a statement of guilt or

innocence.

Thus, while the findings of this study were that a quantitative presentation of evidence was most appropriate, additional research that addressed the above outlined concerns would be needed to further support that conclusion.

Thompson et al. [3] conducted comparative research on multiple evidence types that recognized that people are better at judging the relative strength of two stimuli than judging a single stimulus along a scale. Using online participants, they presented a series of pairs of statements and asked subjects to indicate which of the two represented a *stronger* statement in support of the prosecution hypothesis.

Six statement types[14] were tested over three experiments. The results of these studies showed, generally speaking, that the strength of statements was ordered as intended by examiners—expressions intended to convey high strength of evidence were taken as such, while expressions intended to convey low strength of evidence were likewise taken as such.

However, there were surprises illuminated by the results. For example, while an LR of "10 million times more likely" was perceived as roughly equivalent to an RMP of "1 in 10 million," which makes sense, both the statements "identification" and "individualization" were perceived as much weaker than 1 in 10 million. "Identification" was ranked as stronger than "individualization," but neither differed significantly from an RMP of "1 in 100,000." Since "identification" and "individualization" are both considered by many disciplines to be the strongest expression of same source support, this is a surprising finding and indicates that jurors are not receiving these expressions as examiners intend them.

Additionally, the term "match," which is often supposed by forensic examiners to be one of the weakest statements of association, was perceived as extremely strong evidence by the participants, who saw it as roughly equivalent to an RMP of "1 in 10 million." This can be highly problematic in practice, if examiners are reporting results as a "match," thinking that they are indicating a result of low strength, and jurors are receiving it as one of the strongest associations possible.

These results are in agreement with those of McQuiston-Surrett and Saks [7], who found that an explicit statement of the conclusion[15] only *increased* jurors' perception of the strength of the evidence in their quantitative condition, not the qualitative one. They concluded that this was because the use of the word "match" (their qualitative condition) already caused such a high perception of guilt that there was a ceiling effect by which jurors could not believe the defendant any more guilty, even when an explicit conclusion was stated.

In the broader literature on communication and decision-making, Jenkins et al. [41] found that verbal and numerical presentations of information can be made together, but that the order in which they are presented matters. For instance, one can present the verbal information followed by the numerical (e.g., unlikely (20% likelihood)) or the numerical followed by the verbal (e.g., 20% likelihood (unlikely)). They found that presenting the numerical information first led to more accurate interpretations of the presented data.

Taken as a whole, the studies comparing different types of statements in support of a same source conclusion do not reach a clear conclusion of a single presentation type that is free of interpretive errors. In fact, Olson and Budescu [42] summarize their own results as "neither mode of communications is universally superior to the other." It seems, then, that no panacea has yet been identified to determine the most accurate and best-understood method of presenting the strength of the evidence.

3. Juror perceptions of credibility

While forensic science examiners are often warned of the dangers of considering factors beyond the evidence itself when rendering their decisions, it is worth noting that fact-finders are susceptible to the same effects. In a perfect world, the fact-finder would listen only to the words the expert used in describing the strength of the evidence and make their decisions based upon the content of that description. However, in reality, fact-finders may be influenced in their decision-making by a host of cues that go beyond the evidence being presented. This section illuminates some of these factors.

3.1. Perceived credibility of the expert

One factor that can influence jurors' perception of the evidence is their perception of the credibility of the expert presenting that evidence. Credibility may be judged by a host of evidence-irrelevant criteria such as appearance, tone of voice, body language, experience, degrees, or use of visual aids [43,44], which in turn can influence the listener's perception of the information received. This has implications for juror studies, since the vast majority of them have involved the interpretation of written statements, which remove this variable entirely and may not give a complete picture of the ways in which mock jurors will respond to evidence presentation.

Research has also demonstrated [45] that listeners receive information in two different modes. Central processing is used when they are engaged with the material being communicated. Central processing is marked by attention to appropriate and relevant cues such as relevant expertise and trustworthiness. Peripheral processing on the other hand, occurs when listeners are not engaged with the material (due to not understanding it, boredom, etc). Peripheral processing is exemplified by relying on non-relevant cues such as the appearance or likability of the speaker. Koehler et al. [46] performed an online experiment and a realistic in-person mock trial experiment using video-recorded testimony to investigate factors that might influence how jurors think about and use forensic science testimony. In both experiments, perceived expert experience (a peripheral factor) was found to be the most influential on jurors' perceptions of the evidence. Whether the method had been scientifically validated (a central factor) was found to modestly increase perception of the strength of the evidence in the online experiment and had no effect at all for jurors in the realistic in-person scenario.

Listeners also consider the circumstances under which the speaker came to possess the information. Listeners believe speakers may be influenced by "the information available to the speaker, the speaker's motivation, the speaker's degree of accountability, and the speaker's goals when forming his or her belief" [47]. If the listener senses that the speaker's judgment has been compromised by the system in which they are forming their judgments (e.g., by an environment with a strong pro-prosecutorial bias), they may devalue the speaker's information, reasoning that even if the speaker is presenting the information honestly, they may have deluded themselves into believing it is true.

Perceived confidence level and certainty of the expert are also

[14] The statement types included LR (e.g., 100,000 times more likely), verbal Strength of Support (e.g., very strong support), RMP (e.g., one person in 100,000), verbal statement of a Likelihood of Observed Similarity (e.g., likelihood of observing this amount of corresponding ridge detail when two fingerprints are made by different people is considered extremely low), verbal statement of Source Probability (e.g., highly probable), and Categorical Conclusions (e.g., identified).

[15] Such as stating that the suspect was, in fact, the source of the trace.

taken into account. Cramer et al. [48] found that juries find experts most credible at a medium-confidence level compared to experts exhibiting a low- or high-confidence level.[16] These findings show that not only are jurors' decisions being informed by factors that are not part of the evidence being presented, but that they prefer their experts to be a little bit accessible and relatable—they dislike or distrust experts who seem to be arrogant, are difficult to understand, or overstate their conclusions.

Fox and Irwin [44] discuss the effects of second-order uncertainty. They cite literature from the risk communication field in which if a speaker communicates their own uncertainty by citing an explicit range of probabilities, listeners judge the speaker as more honest, but also less competent. Fox and Irwin point out that these studies were in the context of information on contaminated water sites coming from the Environmental Protection Agency, so there may have been an effect based upon the public's baseline mistrust of government agencies. It is unclear whether this mistrust would extend to the expert witness testifying on behalf of the State.

Finally, listeners may be influenced by whether any uncertainty associated with a statement is expressed in internal, or external mode [44,47,50]. In internal mode, the speaker's uncertainty about a statement is based upon their *personal* knowledge or experience, such as "I am 95% certain that the subject left this trace". In external mode, the source of the uncertainty is out in the world, due to external forces or random effects, such as "It is 95% certain that the subject left this trace" or "There is a 95% probability that the subject left this trace." Listeners have a higher belief strength in statements that are presented in internal mode than those presented in external mode. This may be because statements presented in internal mode represent more certainty and a greater willingness to take responsibility for the judgment. It may also be because internal mode expressions are perceived as being based on specific scenarios, whereas external mode expressions are perceived as being based on general class or base rates.

3.2. Perceived credibility of the evidence type

It is not only the perceived credibility of the *expert* that can influence a jurors' interpretation of the strength of forensic evidence. There is research supporting that the beliefs and knowledge jurors bring with them into the courtroom about specific evidence disciplines influence their perception of the strength of that evidence.

For example, Thompson and Newman [51] found that the type of evidence (DNA vs. Shoeprint) had a greater effect on participants' judgements of the strength of evidence than the presentation type (RMP vs. LR vs. verbal equivalent) did. Belief changes after hearing DNA evidence were roughly consistent with expected values, while shoeprint evidence was significantly undervalued compared to the intent of the examiner. They attribute this effect to participants' general expectations (or prior beliefs) that DNA evidence is more discriminating than shoeprint evidence and thus less likely to risk a coincidental match.

It is clear that perceptions of credibility, whether of the expert or the evidence type, can greatly influence jurors' perceptions of the strength of evidence and that these factors should not be ignored in the design of studies on juror perception and comprehension.

3.3. Mitigating perception

Cross examination, testimony by an opposing expert, and jury instructions have all been explored as possible mechanisms to influence juror perceptions. These mechanisms have been of particular interest in situations where the forensic expert may have overstated the strength of the evidence and a course-correction might be appropriate.

Eastwood and Caldwell [52] examined the use of an opposing witness and judicial instruction to mitigate the impact of overstated forensic testimony. They found that testimony by an opposing expert to further educate the jury and point out where the initial expert had overstated the strength of the evidence resulted in a reduction in convictions. However, judicial instruction had no effect whatsoever, even when the judge went so far as to state that the initial expert "was wrong" to state the conclusions in the terms he used. Interestingly, they also found that from the control group (where no expert testimony was presented) to the condition where only the prosecution's hair evidence was provided, guilty verdicts jumped from 4% to 50%, indicating that the overstated hair evidence was very convincing to the jury.

McQuiston-Surrett and Saks [7] also used hair evidence to test juror perceptions of strength of evidence and tested both cross-examination and judicial instructions as mechanisms to introduce information about the limitations of the science. In this research, neither mode of presentation of the limitations of the science had any significant effect on the jurors' probability that the defendant was the source of the hair evidence.

The effects on juror perception of phrasing, admitting the possibility of an error, and explaining the method used have also been explored in the literature. Garrett and Mitchell [53] tested all three of these effects as related to fingerprint evidence, and found that the examiner's phrasing of the conclusion[17] had little effect on juror decisions about a suspect's guilt. If the expert admitted the possibility of an error (whether on direct or cross-examination), this reduced the weight given to the fingerprint evidence, but this effect could be mitigated by the examiner providing a detailed explanation of the method used for fingerprint examination.

Interestingly, Schweitzer [54] found that whether the examiner's concession that errors were possible affected juror confidence in the forensic evidence varied depending on the evidence type. When this concession was made around bitemark evidence, juror confidence in the evidence decreased. However, when the concession was made regarding fingerprint evidence, juror confidence was actually slightly strengthened. Schweitzer concludes (like Thompson and Newman [51] before him) that this effect is attributable to jurors' prior beliefs about the strength and trustworthiness of the discipline.

These studies seem to indicate that the most effective mechanisms for correcting an overstatement of the strength of the evidence are introducing an opposing expert or getting the expert to admit to the possibility of an error. These mechanisms, however, appear to be dependent on evidence type and thus the efficacy of any given mechanism may vary considerably from case to case.

4. Approaches for teaching through testimony

Experts try to present complex forensic or statistical

[16] One criterion used to distinguish the medium-confidence expert from the high-confidence expert was the use of "formal speech" (includes use of lay terminology and understandable vocabulary) versus "hypercorrect speech" (marked by technical terminology, pedantic word choice, and use of impersonal labels for people, such as "the client") [49]. These speech patterns were incorporated into the video performances used in the study. Other distinguishing criteria included posture, certainty of conclusions, and rate and flow of speech.

[17] 15 variations in fingerprint testimony were used in this study. They were modeled after fingerprint conclusion statements found in court transcripts or standards/guidelines available to the fingerprint community and were categorized by type as: Simple Positive Match, Bolstered Positive Match, Qualified or Inconclusive Match, or Exclusion statements.

information to a fact-finder who has no background in either topic. They are then expected to understand it, then immediately and accurately apply their newly acquired knowledge to reach a rational decision. Forensic scientists spend years studying and practicing their fields and then years more trying to understand a statistical framework in which to present their findings. Then they go to court and spend between a few minutes and a few hours trying to convey that information to a jury in a highly constrained format in which the examiner does not drive the conversation. Is it any wonder that jurors often take home a different message than the one the examiner intended?

To address this problem, research should focus on how to exploit what is known in cognitive science about knowledge transfer and memory to allow an examiner to capture the jury's attention, distill the necessary information, and impart it in a way that will be accurately received and retained. The examiner should leave the stand confident that the fact-finder will be successful in applying the necessary information during deliberations.

Jackson et al. [9] summarize the challenge well, describing how gist and verbatim memory are activated during learning and how the prior beliefs and experiences of each juror will color how they interpret what they learn. Essentially, when a listener hears some information (such as the probability of observing a set of features under the same source proposition versus the different source proposition), they first store that exact information in verbatim memory. Then they distill the essence of the information in gist memory. Mapping between the verbatim information and gist information will vary according to the juror's background, experience, and education.

We have noted several times that jurors' prior experiences and beliefs influence the way they interpret evidence, and Jackson et al. provide a clear example of this: if someone is told a car costs $100,000, they will remember that exact price (verbatim memory), but will distill the essence of it (gist memory) according to their life experiences. This price may be recorded as "exorbitant" to a middle-class individual but may be considered "pocket change" to a billionaire. Conversely, if someone is told a car costs "a lot," the middle-class individual may interpret this to mean around $50,000, while to the billionaire, it may take $500,000 to reach their personal threshold of "a lot."

Jackson et al. argue that in order to effectively teach jurors what they need to know quickly, both gist and verbatim memory need to be triggered. They suggest achieving this by presenting both precise (verbatim) and bottom-line (gist) information and including a scale of reference to properly frame the problem as the examiner intends, rather than relying on the juror to provide their own framing, which will be based on their own prior experiences. They recommend that this be done in a way that makes relationships between different pieces of information clear.

While these recommendations appear to be grounded in sound cognitive principles, Jackson et al. acknowledge that research is needed to determine how best to implement it in a courtroom.

Another suggestion based upon principles of cognitive psychology is provided by Spellman [55]. She argues that, rather than torturing jurors with incomprehensible statistics, examiners should leverage knowledge the jurors already have to develop alternative tactics. One technique she suggests is attribute substitution.

In attribute substitution, if someone doesn't know the answer to a difficult question, they will substitute an easier question (even if subconsciously) and answer that instead. She posits that in the case of forensic evidence, surprise may be an acceptable substitution for probability and that instead of presenting a statistic, examiners could express verbally the level of surprise they would feel if they were proven wrong (e.g., "I'd be extremely EXTREMELY surprised to learn that it was not the suspect's fingerprint") [55].

The application of cognitive science principles to teaching complex concepts in a courtroom is in its infancy. This area of research is wide open and specific solutions need to be proposed, and tested, to determine the best steps for moving forward.

5. Themes in the literature and suggestions for future research

A review of the literature into juror comprehension and interpretation of forensic strength of evidence testimony has revealed some apparent universal truths. Jurors do not, as a rule, interpret forensic findings in the way examiners intend them. They often undervalue evidence, particularly if it is in a discipline that they may have previously considered to be less discriminating. They do not understand numerical testimony well, although they may prefer to hear it, and they vary widely in their interpretation of verbal expressions, although they do tend to rank them in approximately the correct order. The one verbal cue they seem to agree on is that the term "match" is extremely strong, in contrast to how it is typically perceived by the forensic science community. The terms "identification" and "individualization," on the other hand, tend to be undervalued along with the quantitative expressions.

Throughout the reviewed literature there is a theme of jurors struggling with the directionality of strength of evidence. They tend to transpose the conditional. They often invert the RMP (believing that a larger RMP is more indicative of guilt rather than a smaller one). They are often subject to the weak evidence effect.

Despite these clear themes in the literature, other effects appear to be contradictory or are simply unknown. These must be thoroughly explored through carefully structured research. It would be premature to make recommendations for changing the way testimony is presented that go beyond the above list of commonalities—at this juncture, it would be little more than an intuitive guess at what might work based on narrow observations. It would certainly be a mistake to assume that a presentation that worked well under a specific set of research conditions is generalizable to *all* forensic testimony.

The three conclusions that can be safely drawn from the above observations are as follows:

1) Examiners should be wary of using the term "match". It is perceived much more strongly than intended and carries a high risk of being misinterpreted and overvalued.
2) Understanding of key terms and concepts, especially the directionality of support, should never be assumed. More explanation is always better.[18]
3) Visual aids, especially to represent probabilistic concepts, are effective.

To draw conclusions beyond these recommendations, further research is needed. Some suggested areas and themes are discussed here.

[18] For example, whatever presentation form is used, if an RMP is presented, it should be accompanied by a clear explanation of what it means and what it does not mean. If an LR is presented, both propositions should be declared along with the direction and magnitude of support the evidence provides. Care should be taken to explain the danger of transposing the conditional and examples should be provided. This general principle of careful explanation of ambiguous terms and concepts should be applied universally and not limited to the examples provided here.

1) Noticeable by its absence is any discussion of linguistics. While there are many studies discussing the choice of terms from among favorites commonly in use or currently proposed by forensic practitioners or statisticians, there has been no analysis of these terms from a cognitive psychology perspective. Nobody has measured the impact of using technically correct, jargon-filled, and complex presentations of evidence against stating conclusions in plain English.
2) Another theme in the literature has been a lack of clearly defined numerical scales, or just detailed explanations in general. Many of the existing studies relied on number lines with no numbers on them, or log scales or other quantitative presentations with little or no explanation. There have not been studies in which the scale, or quantitative concept, has been clearly and carefully *explained*. This would allow measurement of whether jurors are able to learn enough about the scales to understand them and appropriately apply them.

On a related topic, research should be done in which jurors are not expected to do any math at all (e.g., to calculate the number of potential donors from a RMP), but are This will allow an evaluation of whether they are able to weight it properly when they do not have to guess what equation is needed and do math themselves.

3) Marquis et al. [23] suggested to clearly and unambiguously present *all* the information that might be needed for comprehension (e.g., a chart of potential prior probabilities, the LR presented in the case, and the updated posterior probabilities). This should be tested on potential jurors to find out whether it is successful in helping jurors to correctly weight the evidence.

Similarly, other suggestions in the literature, such as presenting warnings against transposing the conditional or avoiding common fallacies, should be tested for efficacy. Analogies that may be universal should be developed and likewise, tested for efficacy.

4) Spellman [55] suggests not presenting statistical information at all and relying on alternate strategies to convey the strength of the evidence, several of which she lists. Again, these strategies should be tested on potential jurors to measure their likely impact.
5) Each study tended to focus on one, or a handful, of evidence types or crime types. While there are obvious reasons this was done, the literature has indicated that jurors interpret evidence differently according to the type of evidence and possibly according to the crime type, dependent on their own prior beliefs and values. Comparative studies are needed to explicitly determine which observed effects are universal for all crime/evidence combinations, which ones may vary, and how they vary.
6) The “juries” tested in many (though not all) of the discussed studies have been comprised of university students. This is not representative of the population of potential jurors. University students, on average, are going to be richer, younger, more technologically literate, and better educated than many potential jurors. These attributes will color their prior beliefs and also affect how well they are able to comprehend complex sentence structures and mathematical concepts. To get a balanced view of how jurors interpret evidence and to test whether proposed solutions will improve the situation, a much better effort needs to be made to test a widely diverse population in age, race, educational level, and socioeconomic status.
7) The vast majority of the studies focused on questioning single individuals. This is not how jurors reach decisions. The deliberative process was largely left out of the literature, yet the limited studies that included jury deliberations [54,56] found that this process improved juror interpretation in some situations when other jurors helped to clarify or illuminate important points. Future research studies should incorporate the deliberation process in order to get a more complete picture of the final conclusion jurors might reach in a real-world situation.
8) Nearly (again, not all) of the studies involved querying potential jurors through written or online questionnaires. This cuts out all of the peripheral and non-verbal cues that real-life jurors may subconsciously rely on in their decision-making. An effort should be made to construct studies in a closer to real-world environment where these factors can be taken into account.
9) Finally, future research studies should leverage knowledge on memory and teaching from the cognitive science domain around how best to convey information so that it is properly encoded and retrievable.

The overall theme of these suggestions for future research has been testing for implementation. There is a body of research making suggestions for how to better present forensic evidence so that it will be understood. These suggestions need to be tested under a variety of conditions, including varied crime types, evidence types, and juror types. Only then will it be reasonable to conclude that they should be implemented and expect that they will improve juror comprehension and interpretation.

Declarations of interest

None.

Funding

This project was supported by Award No. 2016-MU-BX-K110, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication/program/exhibition are those of the author(s) and do not necessarily reflect those of the Department of Justice.

Acknowledgements

Thanks to Ashley Cochran who provided valuable input on a draft of the manuscript and assistance in producing the figures used in the article. Thanks also to two anonymous reviewers whose insightful suggestions materially improved the manuscript.

References

[1] National Research Council, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington, D.C, 2009.
[2] AAAS, Forensic science assessments: A quality and gap analysis - latent fingerprint examination, in: Report Prepared by William Thompson, John Black, Anil Jain, and Joseph Kadane, 2017.
[3] W.C. Thompson, et al., Perceived strength of forensic scientists' reporting statements about source conclusions, Law Probab. Risk 17 (2) (2018) 133–155.
[4] S. Lindsey, R. Hertwig, G. Gigerenzer, Communicating statistical DNA evidence, Jurimetrics 43 (2003) 147–163.
[5] B.T. Ulery, et al., Accuracy and reliability of forensic latent fingerprint decisions, Proc. Natl. Acad. Sci., U.S.A. 108 (19) (2011) 7733–7738.
[6] J.J. Koehler, On Conveying the Probative Value of DNA Evidence: Frequencies, Likelihood Ratios, and Error Rates, vol. 67, University of Colorado Law Review, 1996, pp. 859–886.
[7] D. McQuiston-Surrett, M.J. Saks, The testimony of forensic identification science: what expert witnesses say and what factfinders hear, Law Hum. Behav. 33 (5) (2009) 436–453.
[8] G. Gigerenzer, How I got started: teaching physicians and judges risk literacy, Appl. Cognit. Psychol. 28 (4) (2014) 612–614.
[9] G. Jackson, et al., Communicating the results of forensic science examinations, in: Cedric Neumann, Anjali Ranadive, David H. Kaye (Eds.), Final Technical Report for NIST Award 70NANB12H014, 2015, 2015; Penn State Law Research

Paper No. 22-2015. Available at: SSRN: http://ssrn.com/abstract=2690899.
[10] P. Sedlmeier, G. Gigerenzer, Teaching Bayesian reasoning in less than two hours, J. Exp. Psychol. Gen. 130 (3) (2001) 380–400.
[11] G. Gigerenzer, et al., Helping doctors and patients make sense of health statistics, Psychol. Sci. Publ. Interest: J. Am. Psychol. Soc. 8 (2) (2007) 53–96.
[12] G. Gigerenzer, et al., A 30% chance of rain tomorrow": how does the public understand probabilistic weather forecasts? Risk Anal. 25 (3) (2005) 623–629.
[13] D.H. Kaye, J.J. Koehler, Can jurors understand probabilistic evidence? J. R. Stat. Soc. Ser. A (Stat. Soc) 154 (1) (1991) 75–81.
[14] K.A. Martire, R.I. Kemp, B.R. Newell, The psychology of interpreting expert evaluative opinions, Aust. J. Forensic Sci. 45 (3) (2013) 305–314.
[15] K.A. Martire, et al., The expression and interpretation of uncertain forensic science evidence: verbal equivalence, evidence strength, and the weak evidence effect, Law Hum. Behav. 37 (3) (2013) 197–207.
[16] C.G.G. Aitken, F. Taroni, A verbal scale for the interpretation of evidence, Sci. Justice 38 (4) (1998) 279–283.
[17] Association of Forensic Science Providers, Standards for the formulation of evaluative forensic science expert opinion, Sci. Justice 49 (3) (2009) 161–164.
[18] C. Mullen, et al., Perception problems of the verbal scale, Sci. Justice 54 (2) (2014) 154–158.
[19] K.A. Martire, I. Watkins, Perception problems of the verbal scale: a reanalysis and application of a membership function approach, Sci. Justice 55 (4) (2015) 264–273.
[20] E. Arscott, et al., Understanding forensic expert evaluative evidence: a study of the perception of verbal expressions of the strength of evidence, Sci. Justice 57 (3) (2017) 221–227.
[21] A.S. Benjamin, J.G. Tullis, J.H. Lee, Criterion noise in ratings-based recognition: evidence from the effects of response scale length on recognition accuracy, J. Exp. Psychol. Learn. Mem. Cogn. 39 (5) (2013) 1601–1608.
[22] M. Sjerps, D.B. Biesheuvel, Interpretation of conventional and 'bayesian' verbal scales for expressing expert opinion: a small experiment among jurists, Forensic Linguist. 6 (2) (1999) 214–227.
[23] R. Marquis, et al., Discussion on how to implement a verbal scale in a forensic laboratory: benefits, pitfalls and suggestions to avoid misunderstandings, Sci. Justice 56 (5) (2016) 364–370.
[24] T.S. Wallsten, et al., Preferences and reasons for communicating probabilistic information in verbal or numerical terms, Bull. Psychon. Soc. 31 (2) (1993) 135–138.
[25] W. Brun, K.H. Teigen, Verbal probabilities: ambiguous, context-dependent, or both? Organ. Behav. Hum. Decis. Process. 41 (3) (1988) 390–404.
[26] T.S. Wallsten, S. Fillenbaum, J.A. Cox, Base rate effects on the interpretations of probability and frequency expressions, J. Mem. Lang. 25 (5) (1986) 571–587.
[27] J. Randall, Tackling 'legalese': how linguistics can simplify legal language and increase access to justice, in: M.J.J. Emonds (Ed.), Language Use and Linguistic Structure, Univerzita Palackeho, Olomouc, 2014.
[28] K.H. Teigen, W. Brun, Yes, but it is uncertain: direction and communicative intention of verbal probabilistic terms, Acta Psychol. 88 (3) (1995) 233–258.
[29] K.H. Teigen, W. Brun, The directionality of verbal probability expressions: effects on decisions, predictions, and probabilistic reasoning, Organ. Behav. Hum. Decis. Process. 80 (2) (1999) 155–190.
[30] S.P. Lund, H. Iyer, Likelihood ratio as weight of forensic evidence: a closer look, J. Res. Natl. Inst. Stand. Technol. 122 (27) (2017).
[31] S. Willis, et al., ENFSI Guideline for Evaluative Reporting in Forensic Science, European Network of Forensic Science Institutes, Dublin, 2015.
[32] S. Gittelson, et al., A response to "Likelihood ratio as weight of evidence: a closer look" by Lund and Iyer, Forensic Sci. Int. 288 (2018) e15–e19.
[33] C.E.H. Berger, K. Slooten, The LR does not exist, Sci. Justice 56 (5) (2016) 388–391.
[34] A. Biedermann, et al., Reframing the debate: a question of probability, not of likelihood ratio, Sci. Justice 56 (5) (2016) 392–396.
[35] K.A. Martire, et al., On the likelihood of "encapsulating all uncertainty, Sci. Justice 57 (1) (2017) 76–79.
[36] W.C. Thompson, et al., The role of prior probability in forensic assessments, Front. Genet. 4 (220) (2013).
[37] K.A. Martire, et al., On the interpretation of likelihood ratios in forensic science evidence: presentation formats and the weak evidence effect, Forensic Sci. Int. 240 (2014) 61–68.
[38] P.M. Fernbach, A. Darlow, S.A. Sloman, When good evidence goes bad: the weak evidence effect in judgment and decision-making, Cognition 119 (3) (2011) 459–467.
[39] E.R. Stone, A.M. Parker, L.D. Townsend, Distinguishing the ratio bias from unsystematic error: situation and individual-difference effects, J. Behav. Decis. Mak. 31 (2018) 587–601.
[40] W. Thompson, E.L. Schumann, Interpretation of statistical evidence in criminal trials - the Prosecutor's Fallacy and the *Defense Attorney*'s Fallacy, Law Hum. Behav. 11 (3) (1987) 167–187.
[41] S.C. Jenkins, A.J.L. Harris, R.M. Lark, Understanding 'unlikely (20% likelihood)' or '20% likelihood (unlikely)' outcomes: the robustness of the extremity effect, J. Behav. Decis. Mak. 31 (4) (2018) 572–586.
[42] M.J. Olson, D.V. Budescu, Patterns of preference for numerical and verbal probabilities, J. Behav. Decis. Mak. 10 (1997) 117–131.
[43] A. McCarthy Wilcox, N. NicDaeid, Jurors' perceptions of forensic science expert witnesses: experience, qualifications, testimony style and credibility, Forensic Sci. Int. 291 (2018) 100–108.
[44] C.R. Fox, J.R. Irwin, The role of context in the communication of uncertain beliefs, Basic Appl. Soc. Psychol. 20 (1) (1998) 57–70.
[45] R. Petty, J. Cacioppo, Communication and persuasion: central and peripheral routes to attitude change, Springer Ser. Soc. Psychol. Chapter 6 (1986) 141–172. New York: Springer-Verlag.
[46] J.J. Koehler, et al., Science, technology, or the expert witness: what influences jurors' judgments about forensic science testimony? Psychol. Publ. Pol. Law 22 (4) (2016) 401–413.
[47] C.R. Fox, B.F. Malle, On the Communication of Uncertainty: Two Modes of Linguistic Expression, 1997 (Unpublished manuscript).
[48] R.J. Cramer, S.L. Brodsky, J. DeCoster, Expert witness confidence and juror personality: their impact on credibility and persuasion in the courtroom, J. Am. Acad. Psychiatr. Law 37 (2009) 63–74.
[49] W.M. O'Barr, Linguistic Evidence: Language, Power, and Strategy in the Courtroom, Academic Press, New York, 1982.
[50] E. Løhre, K.H. Teigen, There is a 60% probability, but I am 70% certain: communicative consequences of external and internal expressions of uncertainty, Think. Reas. 22 (4) (2015) 369–396.
[51] W.C. Thompson, E.J. Newman, Lay understanding of forensic statistics: evaluation of random match probabilities, likelihood ratios, and verbal equivalents, Law Hum. Behav. 39 (4) (2015) 332–349.
[52] J. Eastwood, J. Caldwell, Educating jurors about forensic evidence: using an expert witness and judicial instructions to mitigate the impact of invalid forensic science testimony, J. Forensic Sci. 60 (6) (2015) 1523–1528.
[53] B. Garrett, G. Mitchell, How jurors evaluate fingerprint evidence: the relative importance of match language, method information, and error acknowledgment, J. Empir. Leg. Stud. 10 (3) (2013) 484–511.
[54] N.J. Schweitzer, Communicating Forensic Science, National Institute of Justice, 2016.
[55] B.A. Spellman, Communicating forensic evidence: lessons from psychological science, Seton Hall Law Rev. 48 (2018) 827–840.
[56] A. Wilcox, Just a juror's perception, in: J. Morgan (Ed.), Just Science, RTI International, 2018.