Journal Pre-proofs

Review

What is the best way to present likelihood ratios? A review of past research and recommendations for future research

Geoffrey Stewart Morrison, Agnes S. Bali, Kristy A. Martire, Rebecca Hofstein Grady, William C. Thompson

PII: S1355-0306(25)00126-1
DOI: https://doi.org/10.1016/j.scijus.2025.101342
Reference: SCIJUS 101342

To appear in: *Science & Justice*

Received Date: 14 April 2025
Revised Date: 27 September 2025
Accepted Date: 1 October 2025

Please cite this article as: G.S. Morrison, A.S. Bali, K.A. Martire, R.H. Grady, W.C. Thompson, What is the best way to present likelihood ratios? A review of past research and recommendations for future research, *Science & Justice* (2025), doi: https://doi.org/10.1016/j.scijus.2025.101342

What is the best way to present likelihood ratios? A review of past research and recommendations for future research

Authors and affiliations:

Geoffrey Stewart Morrison [1,2,*]

Agnes S Bali [3]

Kristy A Martire [3]

Rebecca Hofstein Grady [4]

William C Thompson [5]

[1] Forensic Data Science Laboratory, Aston University, Birmingham, UK

[2] Forensic Evaluation Ltd, Birmingham, UK

[3] School of Psychology, University of New South Wales, Sydney, New South Wales, Australia

[4] California Digital Library, Office of the President, University of California, Oakland CA, USA

[5] Department of Criminology, Law, & Society, University of California, Irvine, Irvine CA, USA

* Corresponding author: G.S. Morrison, e-mail: geoff-morrison@forensic-evaluation.net

ORCID:

Geoffrey Stewart Morrison 0000-0001-8608-8207

Agnes S Bali 0000-0002-0166-0989

Kristy A Martire 0000-0002-5324-0732

Rebecca Hofstein Grady 0000-0002-8587-0528

William C Thompson 0000-0003-0131-7280

Author contributions:

Geoffrey Stewart Morrison: Conceptualization, Formal Analysis, Funding Acquisition, Visualization, Writing - Original Draft, Writing - Review & Editing.

Agnes S Bali: Conceptualization, Investigation, Methodology, Writing - Original Draft, Writing - Review & Editing.

Kristy A Martire: Conceptualization, Investigation, Methodology, Supervision, Writing - Original Draft, Writing - Review & Editing.

Rebecca Hofstein Grady: Writing - Review & Editing.

William C Thompson: Writing - Review & Editing.

Disclaimer:

All opinions expressed in the present paper are those of the authors, and, unless explicitly stated otherwise, should not be construed as representing the policies or positions of any organizations with which the authors are associated.

Declarations of interest:

none

Acknowledgements:

The work of Morrison, Bali, and Martire was supported in part by Research England's Expanding Excellence in England Fund as part of funding for the Aston Institute for Forensic Linguistics 2019–2024.

Highlights:

- Review of research on understandability of likelihood ratios
- CASOC indicators of comprehension
- Critical review of methodologies
- Methodological recommendations for future research

Abstract

As a first step in addressing the research question "What is the best way for forensic practitioners to present likelihood ratios so as to maximize their understandability for legal-decision makers?", this paper reviews existing empirical literature on the comprehension of likelihood ratios by laypersons. The existing literature tends to research understanding of expressions of strength of evidence in general, rather than focusing specifically on likelihood ratios. We review the literature with respect to the CASOC indicators of comprehension (particularly *sensitivity*, *orthodoxy*, and *coherence*), and compare different formats that have been used to express likelihood ratios: numerical likelihood-ratios values, numerical random-match probabilities, and verbal strength-of-support statements (none of the studies that we reviewed tested comprehension of verbal likelihood ratios). We also critically review the studies with respect to methodology, and consider additional factors that could potentially assist with communication of the meaning of likelihood ratios. We conclude that the existing literature does not answer our research question, but, based on our review, we provide recommendations for the methodology of future research aimed at addressing our research question.

Keywords

Communication; Comprehension; Likelihood ratio; Recommendation; Review; Understanding

1 Introduction

The likelihood-ratio framework is advocated as the logically correct framework for evaluation of evidence by the vast majority of experts in forensic inference and statistics, including in Aitken et al. [1], Morrison et al. [2], Morrison et al. [3], and Morrison et al. [4], with 31, 19, 20, and 57 authors and supporters respectively. Its use is also advocated by key organizations including: Association of Forensic Science

Providers of the United Kingdom and of the Republic of Ireland (AFSP [5]); Royal Statistical Society (Aitken et al. [6]); European Network of Forensic Science Institutes (Willis et al. [7]); National Institute of Forensic Science of the Australia New Zealand Policing Advisory Agency (Ballantyne et al. [8]); American Statistical Association (Kafadar et al. [9]); and Forensic Science Regulator for England & Wales [10].

There is, however, a common belief that likelihood ratios are difficult for legal-decision makers to understand (Bali et al. [11], Swofford et al. [12]), and there are many legal rulings that include misunderstandings of the meaning of likelihood ratios, the England & Wales Court of Appeal ruling in *R v T [2010] EWCA Crim 2439* being an infamous example (e.g., Aitken et al. [1], Berger et al. [13], Redmayne et al. [14], Morrison [15], Thompson [16]).

The benefits of forensic practitioners adopting the likelihood-ratio framework will not be fully realized if legal-decision makers are unable to understand the meaning of the likelihood ratios that forensic practitioners present. It is therefore important to conduct research to determine the best way for forensic practitioners to present likelihood ratios so as to maximize their understandability for legal-decision makers.

Martire [17], Thompson [18], Eldridge [19], and Martire & Edmond [20] reviewed empirical research on laypersons' understanding of forensic practitioners' expressions of strength of forensic evidence. These reviews included comparisons of understanding of likelihood ratios with understanding of other expressions of strength of evidence such as:

- categorical conclusions
  - e.g., "identification", "exclusion"
- numerical posterior probabilities
  - e.g., "95% probable that the items came from the same source"
- verbal posterior probabilities
  - e.g., "highly probable that the items came from the same source"
- vague verbal expressions
  - e.g., "consistent with", "cannot be excluded", "to a reasonable degree of scientific certainty"

Eldridge [19] concluded:

> 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 research that Martire [17], Thompson [18], Eldridge [19], and Martire & Edmond [20] reviewed was diverse in terms of conditions tested, methodologies used, and aspects of understanding tested. It was also diverse in terms of results obtained. On the basis of these reviews, it is therefore difficult to draw a clear answer to the question of whether likelihood ratios are actually harder for laypersons to understand than are other expressions of strength of evidence.

Expressions of strength of evidence other than likelihood ratios are, however, not logically tenable (e.g., Jackson [21], Kaye [22], Morrison & Thompson [23], Thompson [18]). We therefore begin with the premise that forensic practitioners should use the likelihood-ratio framework to evaluate strength of forensic evidence, and we ask the question:

- What is the best way for forensic practitioners to present likelihood ratios so as to maximize their understandability for legal-decision makers?

In the present paper, as a first step in addressing this question, we review the existing empirical literature on the comprehension of likelihood ratios by laypersons. The particular laypersons we are ultimately interested in are legal-decision makers, who could be judges or juries in the context of legal hearings, but who could also include prosecutors deciding whether to prosecute, defence attorneys deciding whether to recommend plea deals to their clients, etc.

In §2 we describe different formats that have been used to present likelihood ratios. In §3 we describe how studies were selected for inclusion in our review. In §4 we explain a key concept for our review, *effective likelihood ratio*. In §5 we critically review the studies with respect to three indicators of comprehension (*sensitivity*, *orthodoxy*, and *coherence*) and consider whether the results are informative with respect to the question of the best format for presenting likelihood ratios. In the course of reviewing studies with respect to presentation formats, we also consider the potential effect of other factors that may contribute to understandability of likelihood ratios, such as providing participants with a whole verbal scale or providing them with a table for converting from priors to posteriors. In §6, we critically review methodological issues and make recommendations for methodology in future research addressing our research question. §7 provides additional recommendations, and §8 provides a conclusion.

2 Formats for presenting likelihood ratios

2.1 Overview

There are several formats in which likelihood ratios have been presented, or it has been proposed that they be presented. These include:

- *numerical likelihood ratios*
  - e.g., the observations are 1,000 times more likely if H_1 were true than if H_2 were true[1]
- *numerical random-match probabilities*
  - e.g., the observations made on the questioned-source item and the known-source item match, the probability of observations made on an item randomly selected from the relevant population matching the observations from the questioned-source item is 1 in 1,000
- *verbal likelihood ratios*
  - e.g., the observations are much more probable if H_1 were true than if H_2 were true
- *verbal strength of support for hypotheses*[2]
  - e.g., the observations provide strong support for H_1 relative to H_2

In §2.2 and §2.3, we describe these formats and the relationships between them.

2.2 Numerical values

Numerical likelihood ratios can be calculated using relevant data, quantitative measurements, and statistical models.

If the data are discrete and have no within-source variability (as in single-source high-template DNA profiles, assuming no drop-out no drop-in, and no ambiguity as to weather a peak is allelic or stutter), then the numerator of the likelihood ratio will have a probability of either 0 or 1. If the questioned-source item and the known-source item do not have exactly the same discrete value(s), then the numerator of the likelihood

[1] H_1 and H_2 represent mutually exclusive hypotheses. In this paper, we adopt the convention that H_1 represents a same-source hypothesis and H_2 represents a different-source hypothesis.

[2] We will often abbreviate *verbal strength of support for hypotheses* to *verbal strength-of-support statement*.

ratio will be 0. If the numerator is 0, then the value of the denominator is irrelevant, the likelihood-ratio value will be 0, and, barring this result having occurred due to a mistake, one can infer that the questioned-source item and known-source items did not come from the same source. If the questioned-source item and the known-source item have exactly the same discrete value(s), then the numerator of the likelihood ratio will be 1, and the denominator will be the *random-match probability* (RMP), i.e., the probability that an item selected at random from the relevant population would have exactly the same discrete value(s) as the questioned-source item. Under the latter circumstance, instead of presenting the whole likelihood ratio, the random-match probability can be presented. Under this circumstance, the numerical likelihood ratio and the numerical random-match probability are numerically equivalent.

In contrast, if the data are discrete but have within-source variability, or if the data are continuously valued and have within-source variability (as is the case for data in most branches of forensic science), then the numerator of the likelihood ratio will not have a probability of 0 or 1. For discrete data it will be a probability value between 0 and 1, and for continuously-valued data it will be a probability-density value greater than 0. Under these circumstances, the denominator alone does not capture the same information as the whole likelihood ratio, and a numerical likelihood ratio, rather than a numerical random-match probability, should be presented as the strength-of-evidence statement.

For continuously-valued data, some publications (e.g., President's Council of Advisors on Science and Technology [24]) have proposed applying a threshold to the degree of similarity between the questioned-source item and the known-source item in order to determine whether they "match", and, if they match, then calculating and reporting a random-match probability. This procedure, however, suffers from the "cliff-edge effect"[3] and underutilizes the available data. Applying continuously-valued statistical models to calculate numerical likelihood-ratio values is a better solution (Morrison et al. [2], Morrison & Enzinger [27]).

If relevant data, quantitative measurements, and statistical models are not available, some publications advocate subjectively assigning numerical likelihood ratios. Willis et al. [7] and Marquis et al. [28] recommend that the practitioner subjectively assign a numerical value for the numerator of the likelihood ratio, subjectively assign a numerical value for the denominator of the likelihood ratio, then divide the former by the latter. Uncalibrated unvalidated subjective assignment of likelihood-ratio values has been criticized in Risinger [29], Martire et al. [30], and Morrison et al. [31]. Martire et al. [32] found that forensic handwriting practitioners' subjective assignment of probabilities of occurrence of discrete handwriting features had an average 20

[3] In the forensic-statistics literature, Evett [25] and Walsh et al. [26] attribute the term "fall-off-the-cliff effect" to personal communication from "Ken Smalldon", presumably Kenneth Wallace Smalldon.

percentage point absolute-error rate compared to the frequency of occurrence of those features in sample data.

2.3 Verbal expressions

Especially when likelihood ratios are assigned subjectively, many publications (e.g., Evett et al. [33], AFSP [5], Aitken et al. [1], Berger et al. [13], Norgaard et al. [34], Willis et al. [7], ISO 21043-4:2025 [35]) advocate expressing likelihood ratios using verbal expressions, either in addition to or instead of numerical values. These verbal expressions are arranged in ordinal scales in which each level on the scale corresponds to a range of numerical likelihood-ratio values. In such verbal scales, the correspondences between ranges of numerical values and verbal expressions are arbitrary and are simply specified by the scales. Different scales can, and do, use different verbal expressions and different ranges of numerical likelihood-ratio values.

An example of a verbal scale (based on Willis et al. [7]) is provided in Table 1. This table presents examples of expressions of *verbal likelihood ratios*, i.e., each expression has the form of a likelihood ratio:

- The observations are [qualifier] more probable if H_1 were true than if H_2 were true.

Another example of a verbal scale (also based on Willis et al. [7]) is provided in Table 2. This table presents examples of expressions of *verbal strength of support for hypotheses*, i.e., each expression has the form:

- The observations provide [qualifier] support for H_1 relative to H_2.

In the latter example, two hypotheses are mentioned, but in some verbal scales (e.g., AFSP [5]), only one hypothesis is mentioned:

- The observations provide [qualifier] support for H_1.

Verbal expressions of strength of support for hypotheses do not have the form of likelihood ratios, i.e., they do not express the probability of observing the evidence if one hypothesis were true relative to the probability of observing the evidence if the other hypothesis were true ($p(E \mid H_1)/p(E \mid H_2)$). For this reason, in our opinion, verbal strength-of-support statements are not actually expressions of likelihood ratios.

Willis et al. [7] recommends that a numerical value for a likelihood ratio be assigned first and then the corresponding verbal expression be selected, not the other way round. This (along with the recommendation to assign the numerical value of a likelihood ratio by first assigning a numerical value for the numerator and a numerical value for the denominator) is intended to ensure that users of the scale actually follow the logic of the likelihood-ratio framework.

Some criticisms of verbal scales appear in Mullen et al. [36], Martire & Watkins [37], Marquis et al. [28], and Morrison & Enzinger [38]. Eldridge [19] includes a review of empirical research on the understandability of verbal expressions taken from verbal scales.

Table 1. Examples of verbal likelihood ratios intended to correspond to ranges of numerical likelihood-ratio values. If the likelihood ratio is less than 1, the same verbal expressions can be used with the ratio inverted and the order of H_1 and H_2 reversed.

Ranges of numerical likelihood ratios	**Verbal likelihood ratios**
$0.5 < \Lambda < 2$	The observations are *approximately equally probable* irrespective of whether H_1 were true or whether H_2 were true.
$2 \leq \Lambda < 10$	The observations are *slightly more probable* if H_1 were true than if H_2 were true.
$10 \leq \Lambda < 100$	The observations are *more probable* if H_1 were true than if H_2 were true.
$100 \leq \Lambda < 1{,}000$	The observations are *appreciably more probable* if H_1 were true than if H_2 were true.
$1{,}000 \leq \Lambda < 10{,}000$	The observations are *much more probable* if H_1 were true than if H_2 were true.
$10{,}000 \leq \Lambda < 1{,}000{,}000$	The observations are *far more probable* if H_1 were true than if H_2 were true.
$1{,}000{,}000 \leq \Lambda$	The observations are *exceedingly more probable* if H_1 were true than if H_2 were true.

Table 2. Examples of verbal strength-of-support statements, intended to correspond to ranges of numerical likelihood-ratio values. If the likelihood-ratio value is less than 1, the same verbal expressions can be used with the ratio inverted and the order of H_1 and H_2 reversed.

Ranges of numerical likelihood ratios	**Verbal strength-of-support statements**
$0.5 < \Lambda < 2$	The observations provide *no* support for either H_1 or H_2.
$2 \leq \Lambda < 10$	The observations provide *weak* support for H_1 relative to H_2.
$10 \leq \Lambda < 100$	The observations provide *moderate* support for H_1 relative to H_2.
$100 \leq \Lambda < 1{,}000$	The observations provide *moderately strong* support for H_1 relative to H_2.
$1{,}000 \leq \Lambda < 10{,}000$	The observations provide *strong* support for H_1 relative to H_2.
$10{,}000 \leq \Lambda < 1{,}000{,}000$	The observations provide *very strong* support for H_1 relative to H_2.
$1{,}000{,}000 \leq \Lambda$	The observations provide *extremely strong* support for H_1 relative to H_2.

3 Selection of studies for inclusion in review

To be in scope for our review, a study had to report on empirical research in which numerical likelihood ratios were presented to participants. It could also include presentation of numerical random-match probabilities, verbal likelihood ratios, and/or strength-of-support statements. Commentary papers lacking primary research were excluded.

To select studies for inclusion in our review, we started with known studies on the topic

and then iteratively searched references listed in studies that we had already included.

We found 17 papers that included studies which met the criterion for inclusion. These were: Koehler (1996) [39], Taroni & Aitken (1998) [40], Nance & Morris (2002) [41], Nance & Morris (2005) [42], Langenburg et al. (2013) [43], Martire et al. (2013) [44], Martire et al. (2014) [45], Thompson & Newman (2015) [46], Bayer et al. (2016) [47], Thompson et al. (2018) [48], Ribeiro et al. (2020) [49], van Straalen et al. (2020) [50], Bali et al. (2021) [51], Ribeiro et al. (2023) [52], van Straalen et al. (2023) [53], Bali & Martire (2025) [54], Thompson et al. (2025) [55].

4 Effective likelihood ratios

A key concept in our review is that of *effective likelihood ratio*.

According to Bayes' theorem, a participant's posterior odds (their belief as to the relative probabilities of H_1 and H_2 after they have considered the likelihood ratio) should be the product of their prior odds (their belief as to the relative probabilities of H_1 and H_2 before they have considered the likelihood ratio) and the likelihood ratio, see Equation (1).

$$\text{(1)}\quad \frac{p(H_1 \mid E)}{p(H_2 \mid \mathrm{E})} = \frac{p(H_1)}{p(H_2)} \times \frac{p(E \mid H_1)}{p(E \mid H_2)}$$

$$\text{posterior odds} = \text{prior odds} \times \text{likelihood ratio}$$

If one elicits a participant's prior odds before presenting them with the likelihood ratio, and elicits their posterior odds after presenting them with the likelihood ratio, one can divide their posterior odds by their prior odds to calculate the effective likelihood ratio that they used, see Equation (2). The effective likelihood-ratio value will not necessarily equal the presented likelihood-ratio value.

$$\text{(2)}\quad \frac{\left(\frac{p(H_1 \mid E)}{p(H_2 \mid \mathrm{E})}\right)}{\left(\frac{p(H_1)}{p(H_2)}\right)} = \frac{p(E \mid H_1)}{p(E \mid H_2)}$$

$$\text{posterior odds} \div \text{prior odds} = \text{effective likelihood ratio}$$

5 Review with respect to presentation formats and CASOC indicators of

comprehension

5.1 CASOC indicators of comprehension

Empirical research into lay comprehension of expressions of strength of evidence has made use of different concepts and criteria for what constitutes comprehension. Martire [17], Martire & Edmond [20], and Bali et al. [51] developed a list of indicators of comprehension, which they called the CASOC indicators of comprehension (*Consistency*, *Sensitivity*, *Coherence*, *Ability*, *Orthodoxy*).[4] Rather than treating comprehension as a unitary construct, the CASOC framework distinguishes different types of understanding. This allows for a more nuanced synthesis of findings. Although classifying studies with respect to CASOC indicators involves a series of binary classifications based on subjective judgement and does not capture all nuances, it offers a transparent way to make sense of a conceptually and methodologically diverse field. For greater robustness in making the classifications, the following procedures were adopted: The second and third authors of the present paper independently classified each study with respect to each of the CASOC indicators. They then discussed and resolved any discrepancies in their classification. The classifications were later checked by the first author.

In this paper, we present results with respect to three of the indicators: *Sensitivity*, *Orthodoxy*, *Coherence*. These results are presented in §5.2, §5.3, and §5.4 respectively. Each of these subsections is further divided into three subsubsections:

1. Definition
2. Summary of results
3. Detailed results

In the first subsubsection, we provide the definition of the indicator from Martire & Edmond [20]. The Martire & Edmond [20] definitions cover all formats that have been used for presenting strength of evidence, not just likelihood ratios. We follow each definition from Martire & Edmond [20] with a modified definition which is specific to likelihood ratios and is adapted to the context of our review. We also provide an explanation of how we assessed studies with respect to each indicator. The second subsubsection provides a summary of the results of the review, and the third subsubsection provides detailed results. This structure is intended to help the reader understand the big picture without getting bogged down in the details, but also to have access to the details that underlie the conclusions presented in the summaries. A reader wanting just to get the big picture could read only the definition and summary

[4] The CASOC definitions of *sensitivity*, *coherence*, and *consistency* are different from the definitions that these words have when they are used in statistics.

subsubsections, and skip the details subsubsections.

Results are also presented in a series of three tables:

- Table 3 shows the formats used to present likelihood ratios, and the likelihood-ratio values presented. For expository purposes, for some papers, different experiments and/or different response conditions and/or different evidence types are listed as different "studies". None of the studies presented participants with verbal likelihood ratios. With two exceptions (Langenburg et al. [43], Thompson et al. [55]), all of the studies presented the experiments to participants in written format.

- Table 4 indicates the response conditions that were tested. Closed responses involved picking a discrete level from a multilevel scale, or making a binary choice as to which of two statements was stronger. Open responses asked participants to give a number in the form of odds or in the form of a probability. All studies that included probability responses elicited them as numbers in the range 0 to 100. Except for the binary choice in Thompson et al. [48] Study 3, all response conditions, whether open or closed, elicited some form of posterior judgement. Table 4 also indicates whether the experiment design was within-participant (each participant responded to multiple presentation formats) or between-participants (different participants responded to a different presentation format), and whether priors and posteriors were elicited from the same individual participants or whether they were elicited from different groups of participants.

- Table 5 indicates the evidence types that were tested and the demographic groups to which participants belonged.

Table 3, Table 4, and Table 5 include the classification of each study with respect to all five CASOC indicators. In the interests of brevity and relevance, this paper omits detailed discussion of the results of our review with respect to *ability* and *consistency*. We omitted detailed discussion of these results because they had limited relevance for answering our research question. Immediately below, we provide the Martire & Edmond [20] definitions of these two indicators, and an explanation of why the results with respect to these indictors have limited relevance for answering our research question.

- "*Ability* is being capable of applying statistical evidence or principles provided by a forensic scientist (or statistician) to the resolution of new problems. This is distinct from mere recognition or recollection of the statistical information. *Ability* requires an active application for the purposes of deriving information beyond what was originally provided."

    - The second and third author originally classified three studies as using *ability* as an indicator of comprehension. On close inspection, however, the first author

was not convinced that the experiments in these studies tested application of principles to a different problem (*ability*), as opposed to testing whether understanding of the presented likelihood ratios was logically correct (*coherence*). In some cases, results that the original papers counted as demonstrating lack of *ability* could have been due to participants making reasonable interpretations of what was written in the experiment questions, and these being counted as misunderstanding only because they differed from what the authors had intended when writing the questions. Irrespective of these issues, results from only three studies would not provide convincing evidence with respect to answering our research question.

- "*Consistency* is giving equal weight to evidence with quantitatively equal strength."
    - Even if there were clear evidence (which there was not) that some formats were *consistent* with one another and others were not, this alone would not help us decided which format, or which set of *consistent* formats, was the best for maximizing understandability of likelihood ratios for legal-decision makers.

Table 3. For studies reviewed: indicators of comprehension used (*Sensitivity*, *Orthodoxy*, *Coherence*, *Ability*, *Consistency*), formats used to present likelihood ratios, and the likelihood-ratio values presented. Shading is to help visually distinguish header groupings.

Indicator					Study	Likelihood-ratio format				
						Numerical		Verbal		Visual
S	O	C	A	C	Study	likelihood ratio	RMP	support statement 1 hypothesis	support statement 2 hypotheses	location on line
•				•	Koehler [39]	1000 100	1 in 1000 1 in 100			
•	•	•		•	Taroni & Aitken [40]	10k 50	1 in 10k 1 in 50			
	•	•		•	Nance & Morris [41] Principal study	25 [a]	1 in 25 [b]			
	•			•	Nance & Morris [41] Follow-up study	25 [a]				
	•			•	Nance & Morris [42]	40k [a]	1 in 40k			
					Langenburg et al. [43]	250k [c]	1 in 250k [c]			
•	•	•		•	Martire et al. [44] Experiment 1	495k 450 4.5			very strong moderately strong weak or limited [d]	
•	•	•		•	Martire et al. [44] Experiment 2	4.5 1/4.5 [e] 1/495k [e]			weak or limited in favour of H_1 [d] weak or limited in favour of H_2 very strong in favour of H_2	

Indicator					Study	Likelihood-ratio format				
						Numerical		Verbal		Visual
S	O	C	A	C	Study	likelihood ratio	RMP	support statement 1 hypothesis	support statement 2 hypotheses	location on line
●	●	●		●	Martire et al. [45]	5.5k 5.5			strong weak or limited [d,f]	"×" just past half way from midpoint to right end of line [g] "×" just to right of midpoint [d,g]
●	●	●		●	Thompson & Newman [46] odds response - DNA	1M 100	1 in 1M 1 in 100	extremely strong moderately strong		
●	●	●		●	Thompson & Newman [46] odds response - footwear	1M 100	1 in 1M 1 in 100	extremely strong moderately strong		
●	●	●		●	Thompson & Newman [46] scale response - DNA	1M 100	1 in 1M 1 in 100	extremely strong moderately strong		
●	●	●		●	Thompson & Newman [46] scale response - footwear	1M 100	1 in 1M 1 in 100	extremely strong moderately strong		
	●				Bayer et al. [47]	1M [h] 1k [h] 1/1k [e,h]				
●				●	Thompson et al. [48] Study 3	10M 100k	1 in 10M [i] 1 in 100k [i]	extremely strong very strong		
●					Ribeiro et al. [49]	5.5M 5.5k				
●		●	●	●	van Straalen et al. [50]	5M 50		extremely strong [j] moderately strong [j]		
●	●	●	●	●	Bali et al. [51] Study 2	1M 100	1 in 1M 1 in 100	extremely strong [j] moderately strong [j]		

Indicator					Study	Likelihood-ratio format				
						Numerical		Verbal		Visual
S	O	C	A	C	Study	likelihood ratio	RMP	support statement 1 hypothesis	support statement 2 hypotheses	location on line
●					Ribeiro et al. [52] Experiment 1	550k 5.5k 550 55 5				
		●	●		van Straalen et al. [53]	5M 50		extremely strong [j] moderately strong [j]		
				●	Bali & Martire [54]	1k	1 in 1k	strong		
●	●	●			Thompson et al. [55]	30 3k				

[a] In one condition, the numerical likelihood ratio was presented by itself. In another condition, a chart was also provided showing prior values (e.g., from 0% to 100% in 5 percentage-point steps) and the corresponding posterior probabilities after Bayesian updating using the presented likelihood-ratio value.

[b] In the same condition, both the RMP of "1 in 25" and "4% of the population" were presented.

[c] All of the following were presented together: The numerical likelihood ratio of 250k, the RMP of 1 in 250k, and the expected count of people in the population expected to exhibit the observed features. For the latter, values were given for Minneapolis-St Paul (10 people from a population of 1.5M), for Minnesota (21 people from a population of 5.3M), for the US (1.4k people from a population of 350M), and for the world (28k people from a population of 7B).

[d] This expression elicited the weak-evidence effect.

[e] Values greater than one were actually presented, but with the order of the hypotheses reversed, i.e., the value of $p(E \mid H_2)/p(E \mid H_1)$ was presented rather than the value of $p(E \mid H_1)/p(E \mid H_2)$.

[f] In one condition, the verbal strength-of-support statement was presented by itself. In another condition, the whole verbal scale and corresponding ranges of numerical likelihood-ratio values was also presented. The numerical range associated with the highest end of the scale was "> 1,000,000".

[g] This was a logarithmically scaled line covering the range log(1/10k) to log(10k), with log(1) at the midpoint (in contrast to the verbal scale for which the maximum value was > 1M). The line, however, did not include numbers indicating the scale or the range covered. Left end, midpoint, and right end were labelled "In favour of Hypothesis 2", "Neutral", and "In favour of Hypothesis 1". A line does not have the form of a likelihood ratio, and this one was not labelled with likelihood-ratio values.

[h] In one condition, the numerical likelihood-ratio value was presented by itself. In a second condition, a whole verbal scale showing single-hypothesis strength-of-support statements and corresponding ranges of numerical likelihood-ratio values was also presented. In a third condition, in addition to the numerical likelihood-ratio value and the verbal scale, a graph was also provided showing the relationship between the "number of potential offenders based on case circumstances" (this was on a logarithmic scale and was related to prior probability) and the posterior probability (as a percentage) after Bayesian updating using the presented likelihood-ratio value.

[i] Of respondents who compared RMP values, 18% interpreted "1 in 100k" as stronger than "1 in 10M".

[j] It is unclear whether the statements presented to participants included one or two hypotheses.

Table 4. For studies reviewed: indicators of comprehension used (*Sensitivity*, *Orthodoxy*, *Coherence*, *Ability*, *Consistency*), response condition, experiment design, and whether priors and posteriors were elicited from the same participants of from different participants. Shading is to help visually distinguish header groupings.

Indicator					Study	Response				Experiment design		Prior & posterior elicited from	
						Open		Closed					
S	O	C	A	C		odds	probability	multilevel scale	binary choice	within participant	between participant	same participant	different participants
•				•	Koehler [39]		•				•		
•	•	•		•	Taroni & Aitken [40]		•				•		[k]
	•	•		•	Nance & Morris [41] Principal study		•						•
	•	•			Nance & Morris [41] Follow-up study		•					•	
	•			•	Nance & Morris [42]		•						•
					Langenburg et al. [43]			• [l]					
•	•	•		•	Martire et al. [44]	•					•	•	
•	•	•		•	Martire et al. [45]	•					•	•	
•	•	•		•	Thompson & Newman [46]	•		• [m]			•	•	
	•				Bayer et al. [47]			• [n]				•	
•				•	Thompson et al. [48] Study 3				• [o]	•			
•					Ribeiro et al. [49]		•				•		

Indicator					Study	Response				Experiment design		Prior & posterior elicited from	
						Open		Closed					
S	O	C	A	C		odds	probability	multilevel scale	binary choice	within participant	between participant	same participant	different participants
●		●	●	●	van Straalen et al. [50]			● [p,q]		●	●		
●	●	●	●	●	Bali et al. [51] Study 2		●			●	●	●	
●					Ribeiro et al. [52] Experiment 1		●				●		
		●	●		van Straalen et al. [53]				● [q]	●	●		
				●	Bali & Martire [54]		[r]				●		
●	●	●			Thompson et al. [55]	●		● [s]			●	●	

[k] A prior probability of 0.6 was provided to participants.

[l] A series of multiple-choice / Likert-scale questions were asked of mock-jury members and of spectators. The questions solicited information including mock-jurors' familiarity with statistics, their opinions about the validity of fingerprint evidence in general, and their opinions of the quality of the testimony. Answer options to questions about the strength of the evidence consisted of qualitative statements that can be summarized as: definitely same source; very-likely same source, could be same source, definitely not same source. Some answer options only referred to the likelihood ratio for the fingerprint evidence, and other answer options referred to the whole of the evidence presented.

[m] 17 level scale: "certain", "9,999,999 chances in 10 million", "999,999 chances in 1 million", …, "1 chance in 1 million", "1 chance in 10 million", "impossible". Excluding the first and last level, the scale was base-ten logarithmic.

[n] Priors were elicited using an 8-level scale and posteriors were elicited using a 5-level scale. These scales tended to be worded in terms of the number of individuals who could have been the source of the questioned-source item, but the format was not consistent within or between scales.

[o] Participants were presented with two statements and asked to chose which was stronger.

[p] Participants used a 5-level Likert scale labelled "very unlikely" … "very likely" to respond to the question "How likely is it that the fingermark belongs to the suspect?"

[q] A series of yes/no/maybe questions were asked, including: "Do you think it is impossible for the finger mark to be from someone other than the suspect?" "The conclusion better fits the scenario that the finger mark belongs to the suspect than the scenario that it belongs to someone else." "There is more than a 50% chance the finger mark belongs to the suspect."

[r] Participants were asked to indicate, as a number from 0 to 100, how much weight they would give to the whole expert report in deciding the suspect's guilt.

[s] Participants first responded which hypothesis was more likely, or whether they were equally likely, then, if they did not choose equally likely, chose from a 6-level scale for how many times more likely their chosen hypothesis was than the other hypothesis: “Between 1 and 10 times more likely (51%–91% chance)”, “Between 10 and 99 times more likely (91%–99% chance)”, …, “More than 100,000 times more likely (More than 99.9999% chance)”. They then gave an open numerical response for how many times more likely their chosen hypothesis was than the other hypothesis.

Table 5. For studies reviewed: indicators of comprehension used (*Sensitivity*, *Orthodoxy*, *Coherence*, *Ability*, *Consistency*), evidence types to which strength-of-evidence statements were purported to relate, and participant demographic. Shading is to help visually distinguish header groupings.

Indicator						Evidence type				Participant demographic				
S	O	C	A	C	Study	DNA	finger-prints	footwear	voice recordings	university students	general community	jury-eligible community	former jurors / jury-pool members	criminal-justice professionals
•				•	Koehler [39]	•						•		
•	•	•		•	Taroni & Aitken [40]	•				•				•
	•	•		•	Nance & Morris [41] Principal study	•							•	
	•	•			Nance & Morris [41] Follow-up study	•							•	
	•			•	Nance & Morris [42]	•							•	
					Langenburg et al. [43]		•				•			•
•	•	•		•	Martire et al. [44]			•		•	•			
•	•	•		•	Martire et al. [45]		•			•	•			
•	•	•		•	Thompson & Newman [46]	•		•				•		
	•				Bayer et al. [47]		•	•		•				
•				•	Thompson et al. [48] Study 3	•						•		
•					Ribeiro et al. [49]	•					•			

Indicator					Study	Evidence type				Participant demographic				
S	O	C	A	C		DNA	finger-prints	footwear	voice recordings	university students	general community	jury-eligible community	former jurors / jury-pool members	criminal-justice professionals
•		•	•	•	van Straalen et al. [50]	•								•
•	•	•	•	•	Bali et al. [51] Study 2	•						•		
•					Ribeiro et al. [52] Experiment 1		•				•			
		•	•		van Straalen et al. [53]		•			•				•
				•	Bali & Martire [54]			•				•		
•	•	•			Thompson et al. [55]				•		•			

5.2 Sensitivity

5.2.1 Definition

- "*Sensitivity* is assigning greater weight to evidence of greater value, and lesser weight to evidence of lesser value."

- Participants' responses are *sensitive* if they reflect relative differences between different presented likelihood-ratio values.

In the studies we reviewed, a participant was judged to have shown *sensitivity* to likelihood-ratio values if their effective likelihood ratio or (if the latter was not calculable) if their posterior odds were further from 1 when the presented likelihood-ratio value was further from 1 than when the presented likelihood-ratio value was closer to 1.

5.2.2 Summary of results

Eleven of the papers reviewed, explicitly or tacitly, used *sensitivity* as an indicator of comprehension.

With only a few exceptions, the studies which used *sensitivity* as an indicator of comprehension found that participants were *sensitive* to differences in likelihood-ratio values across all likelihood-ratio-presentation formats that were tested, across all response conditions that were tested, across all evidence types that were tested, and across all demographic groups that were tested.

These results were not, therefore, informative with respect to the question of the best way for forensic practitioners to present likelihood ratios so as to maximize their understandability for legal-decision makers.

Analysis of results across studies suggested that it could be that participants are *sensitive* to different presented likelihood-ratio values that fall on different sides of a threshold that is somewhere between presented likelihood-ratio values of 100 and 450, but are not *sensitive* to different presented likelihood-ratio values that fall only below the threshold or that fall only above the threshold.

Earlier reviews (Martire [17], Thompson [18], Eldridge [19], Martire & Edmond [20]) concluded that participants were *sensitive* to differences in the values of presented likelihood ratios. If, however, *sensitivity* is dependent on crossing a threshold, rather than being gradient, then one could argue that this does not demonstrate appropriate understanding of the meaning of likelihood ratios.

Even if one did decide that participants were *sensitive* to differences in presented likelihood-ratio value, as a criterion for determining whether participants have understood likelihood ratios, *sensitivity* constitutes a low bar. One might consider it a

necessary but not sufficient criterion. In §5.3 below, we discuss *orthodoxy*, which constitutes a higher bar.

5.2.3 Detailed results

The papers in the review that, explicitly or tacitly, used *sensitivity* as an indicator of comprehension were: Koehler [39], Taroni & Aitken [40], Martire et al. [44], Martire et al. [45], Thompson & Newman [46], Thompson et al. [48], Ribeiro et al. [49], van Straalen et al. [50], Bali et al. [51], Ribeiro et al. [52], Thompson et al. [55].

With only a few exceptions (Koehler [39], Ribeiro et al. [49], and some conditions in Martire et al. [44], Thompson & Newman [46], and Ribeiro et al. [52] Experiment 1), the studies which used *sensitivity* as an indicator of comprehension found that participants were *sensitive* to differences in likelihood-ratio values across all likelihood-ratio-presentation formats that were tested, across all response conditions that were tested, across all evidence types that were tested, and across all demographic groups that were tested.

In Thompson & Newman [46], when participants were asked to respond in the form of odds and the evidence type was footwear marks, *sensitivity* was observed for effective likelihood-ratio values calculated when numerical random-match probabilities were presented but not when numerical likelihood ratios or when verbal strength-of-support statements were presented. In contrast, when the evidence type was DNA, *sensitivity* was observed for all three likelihood-ratio formats. Perhaps these results were due to participants having a belief about the validity of footwear-mark comparison and thus having a ceiling for how strong they believed footwear evidence could be, which may already have been reached by a presented numerical likelihood-ratio value of 100. This would explain why there was no difference in participant's responses to likelihood ratios of 100 or 1M. As noted in Thompson & Newman [46], the ceiling effect could also be due to participant's being incredulous that footwear evidence could produce a likelihood ratio as large as 1M. Participants gave greater weight to the same presented likelihood-ratio values when they were purported to relate to DNA than when they were purported to relate to footwear marks. It is not clear, however, why random-match probabilities would have been exempt from such a ceiling effect, or why the verbal strength-of-support statements "moderately strong" and "extremely strong" were interpreted differently when they were purported to relate to footwear marks then when they were purported to relate to DNA.

All except three studies that used *sensitivity* as an indicator of comprehension tested only two likelihood-ratio values. The exceptions were Martire et al. [44] Experiments 1 and 2, which each tested three levels, and Ribeiro et al. [52] Experiment 1, which tested five levels. In Martire et al. [44] Experiment 1, participants' median effective likelihood-ratio values were larger when the presented numerical likelihood-ratio value was 450 compared to when it was 4.5, but responses were not larger when the presented numerical likelihood-ratio value was 495k compared to when it was 450. In Ribeiro et

al. [52] Experiment 1, the mean posterior probabilities were approximately equal when the presented numerical likelihood-ratio values were 5 and 55, and the mean posterior probabilities were approximately equal when the presented numerical likelihood-ratio values were 550, 5.5k, and 550k, but the mean posterior probabilities were larger for the higher three presented likelihood-ratio values than for the lower two presented likelihood-ratio values. In Ribeiro et al. [49], in which *sensitivity* to the values of the presented numerical likelihood ratio was not observed, the presented values were 5.5k and 5.5M.

Considering these results, and also considering the presented numerical likelihood-ratio values in almost all the studies which presented only two values,[5] we posit a potential explanation for these observations: Participants had a floor and a ceiling for how they responded to likelihood-ratio values with a step function between the floor and ceiling occurring at a threshold somewhere between a presented likelihood-ratio value of 100 and a presented likelihood-ratio value of 450. In almost all studies that presented two likelihood-ratio values, one value was below the threshold and another above the threshold, so the plateaus at floor and ceiling were not observed.

5.3 Orthodoxy

5.3.1 Definition

- "*Orthodoxy* is used in the sense of compliance with or adherence to normative expectations, i.e., *orthodoxy* is updating beliefs in a manner that is consistent with the normative expectations derived using Bayes' theorem."

- Participants' responses are *orthodox* if they reflect use of the values of presented likelihood ratios to update priors to posteriors as per correct application of Bayes' theorem.

In the studies we reviewed, a participant's responses were judged to be *orthodox* if they indicated that the participant used the presented likelihood-ratio value to update their beliefs as would be expected if they had correctly applied Bayes' theorem, i.e., if their effective likelihood-ratio value was the same as (or close to the same as) the presented likelihood-ratio value.

5.3.2 Summary of results

Nine of the papers reviewed used *orthodoxy* as an indicator of comprehension.

For all but one of the studies that used *orthodoxy* as an indicator of comprehension, average effective likelihood ratios were always weaker than the presented likelihood

[5] The exceptions were Koehler [39], the particular instances in Thompson & Newman [46] that we discussed above, and Thompson et al. [48] Study 3, which used a different response format.

ratios, i.e., the effective likelihood ratios were closer to the neutral value of 1 than the presented likelihood ratios (the potential exception was Bayer et al. [47]). In the vast majority of cases the average effective likelihood ratios were much much weaker, e.g., a presented likelihood ratio of 1 million often resulted in a median effective likelihood ratio of less than ten. This was true for numerical likelihood ratios, numerical random-match probabilities, and for verbal strength-of-support statements.

Taken together, the results of the studies indicated that participants' responses were, in general, not *orthodox*. No presentation format (and no addition such as also presenting a full verbal scale) resulted in effective likelihood-ratio values that were consistently more *orthodox* than for any other presentation format. The results were not, therefore, informative with respect to the question of the best way for forensic practitioners to present likelihood ratios so as to maximize their understandability for legal-decision makers.

A priori, we would expect providing an explanation of the meaning of likelihood ratios including how to apply Bayes' theorem would lead to more *orthodox* results, but the one study in our review (Thompson et al. [55]) that tested this did not find convincing evidence that this was the case.

5.3.3 Detailed results

The papers in the review that used *orthodoxy* as an indicator of comprehension were: Taroni & Aitken [40], Nance & Morris [41], Nance & Morris [42], Martire et al. [44], Martire et al. [45], Thompson & Newman [46], Bayer et al. [47], Bali et al. [51], Thompson et al. [55].[6]

Except for Taroni & Aitken [40] (which provided participants with a prior and elicited posteriors), all studies which used *orthodoxy* as an indicator of comprehension elicited both priors and posteriors.

Table 4 indicates the response types used for eliciting priors and posteriors. In all studies which elicited probabilities, these were elicited as numbers between 0 and 100. Note that this will tend to limit the range of effective likelihood-ratio values compared to what could be calculated from elicited prior odds and posterior odds. If a participant

[6] Langenburg et al. [43] asked mock jurors to indicate their judgements as to the strength of the evidence using qualitative statements that can be summarized as: definitely same source; very-likely same source, could be same source, definitely not same source. Some answer options only referred to the likelihood ratio for the fingerprint evidence, and other answer options referred to the whole of the evidence presented. It does not appear to be possible to use this response format to calculate effective numerical likelihood ratios, so we have not counted Langenburg et al. [43] as a study that used *orthodoxy* as an indicator of comprehension.

responded using integers between 0 or 100 exclusive (1 to 99 inclusive), the lowest possible prior probability would be 1% (corresponding to prior odds of 1/99) and the highest possible posterior probability would be 99% (corresponding to posterior odds of 99), resulting in a maximum possible effective likelihood ratio of 9,801. Asking participants to provide a number within a limited range imposes a floor and ceiling, whereas asking for odds does not. The floor and ceiling of 0 and 100 may discourage responses such a 0.1 or 99.9, whereas odds of 1,000:1 in favour of H_1 or 1,000:1 in favour of H_2 would not be discouraged by a floor or ceiling.

Table 4 also indicates whether priors and posteriors were elicited from the same participants of from different participants. In studies in which priors and posteriors were elicited from different participants, responses were elicited from a group of participants who were presented with the background information about the case, but who were not presented with forensic testimony (or who were presented with testimony that was inconclusive), and these responses were used to calculate priors for groups of participants who were presented with (non-inconclusive) forensic testimony. The validity of an approach in which the priors and posteriors do not belong to the same individuals is questionable.

For studies which used *orthodoxy* as an indicator of comprehension (except for Bayer et al. [47], Bali et al. [51] Study 2, and Thompson et al. [55], which we discuss later), Figure 1 presents average effective log-likelihood-ratio values relative to the logarithms of presented likelihood ratios. We extracted median effective log-likelihood-ratio values from the text, tables, or figures provided in the papers. If it was not possible to extract median values, we extracted means.[7] From other information provided in these papers, it was usually clear that medians or means obscured substantial between-participant variability and that between-participant distributions were unlikely to be Gaussian. The medians and means were, therefore, poor summary statistics, but we use them for want of any better way of comparing results within and between these studies.

[7] In Martire et al. [44], Martire et al. [45], and Thompson & Newman [46], effective likelihood-ratio values were provided. In Taroni & Aitken [40], Nance & Morris [41], and Nance & Morris [42], average prior probabilities and posterior probabilities were provided, and from these we calculated the average effective log-likelihood-ratio values. For Taroni & Aitken [40] and for Thompson & Newman [46] median values were not provided but mean values were, so we used mean values.

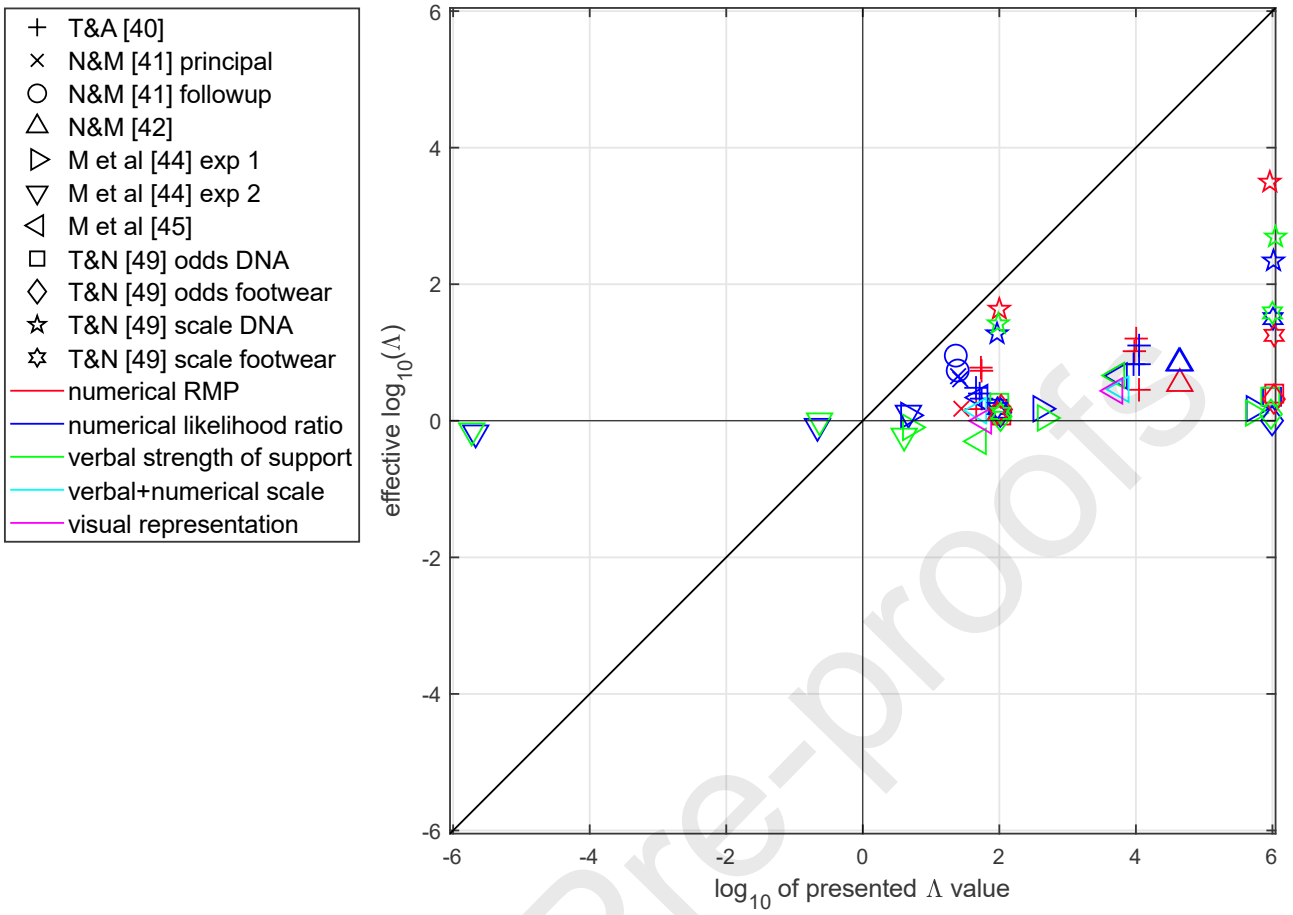

Figure 1. Average effective log-likelihood-ratio values relative to logarithms of presented likelihood ratios in selected studies that used *orthodoxy* as an indicator of comprehension. To increase visual separation between symbols, plotted values for logarithms of presented likelihood-ratio values have been jittered. Symbols that have approximately the same value on the *x* axis represent the logarithm of exactly the same presented likelihood-ratio value.

On examining Figure 1, it is immediately obvious that the average effective likelihood ratios were always weaker than the presented likelihood ratios (the effective log likelihood ratios were closer to the neutral value of 0 than the logarithms of the presented likelihood ratios), and in the vast majority of cases they were much much weaker, e.g., a presented likelihood ratio of 1 million often resulted in a median effective likelihood ratio of less than ten. This was true for numerical likelihood ratios, numerical random-match probabilities, and for verbal strength-of-support statements (and for the visual representation in Martire et al. [45]). None of these studies obtained average effective likelihood ratios that could be said to be *orthodox*.

In Thompson & Newman [46], the mean of the responses was closer to being *orthodox* when DNA evidence was purportedly presented and responses were collected using a multilevel scale compared to when DNA evidence was purportedly presented and responses were collected as odds, and compared to when footwear evidence was purportedly presented and either response format was used. The scale had 17 levels,

which (excluding the first and last level) had order-of-magnitude steps (or log-base-ten steps):

- Certain to be guilty
- About 9,999,999 chances in 10 million that he is guilty
- About 999,999 chances in 1 million that he is guilty

  …

- One chance in 2 (fifty-fifty chance) that he is guilty

  …

- About 1 chance in 1 million that he is guilty
- About 1 chance in 10 million that he is guilty
- Impossible that he is guilty

Thompson & Newman [46] called this a "log scale". Using this scale, mean responses to numerical random-match probabilities were closer to being *orthodox* than mean responses to numerical likelihood ratios and than mean responses to verbal strength-of-support statements.

Since the combination of DNA evidence and responses collected using odds did not lead to responses that were anywhere near *orthodox*, we conclude that the more *orthodox* results were due to using the multilevel scale to collect responses. Thompson & Newman [46] p. 344 noted: "People may simply find it easier to give high estimates on the log scale, where they must check a box to indicate their answer, than on the odds scale, where they must generate a number on their own."

The question arises, however, of whether the results elicited by selecting a level on a multilevel scale are indicative of better understanding of the meaning of a likelihood ratio, or whether they are an artifact of using the scale – perhaps the participants were just picking a relative level on the scale irrespective of the numbers written on the levels. On a scale with the same number of levels but with different numbers written on the levels, participants might have selected the same relative levels. As in Basu et al. [56], odds can be elicited by asking a participant to enter a number 1 or greater in either of two boxes, one for $H_1 \geq H_2$ and the other for $H_2 \geq H_1$, thus there is no limit on the maximum odds value that can be elicited in either direction. In contrast, a scale is finite and suggests that the top and bottom levels are as strong as the evidence can get in support of each of the hypotheses. Changing the most extreme values written on the scale from 10 million to 1 million, or from 10 million to 1 billion might not affect the relative levels that participants select.

When participants were presented with the same likelihood-ratio values but the evidence type was purported to be footwear instead of DNA, the levels on the scale that participants selected were closer to the neutral level in the middle of the scale. This may suggest that the participants did not believe that footwear evidence could be very strong (which Thompson & Newman [46] discussed as "credibility of the evidence"), and that they therefore selected levels closer to the neutral level, not because of a relation between the likelihood-ratio value presented and the numbers written on the levels, but because of the relative locations of the levels on the scale.[8]

Bayer et al. [47] included three conditions: In one condition, the numerical likelihood-ratio value was presented by itself. In a second condition, a whole verbal scale showing single-hypothesis strength-of-support statements and corresponding ranges of numerical likelihood-ratio values was also presented. In a third condition, in addition to the numerical likelihood-ratio value and the verbal scale, a graph was also provided showing the relationship between the number of potential offenders who could have been the source of the questioned-source item (this was on a logarithmic scale) and the posterior probability (as a percentage) after Bayesian updating using the presented likelihood-ratio value – the *x* axis was not explicitly prior probability and the two axes were not scaled the same way. Bayer et al. [47] elicited priors using an 8-level scale and posteriors using a 5-level scale – the scales were inconsistent with one another. The levels on the response scales tended to be worded in terms of the number of individuals who could have been the source of the questioned-source item, but the scales were internally inconsistent. We do not understand how effective likelihood ratios were calculated from these inconsistent scales, but from the presentation of the results it was apparent that the resolution was single orders of magnitude. The median values in the results suggested that providing the verbal scale resulted in a larger proportion of *orthodox* responses,[9] but the boxplots used to present the results did not effectively convey the relative proportion of responses at each order of magnitude.

In Bali et al. [51] Study 2, prior and posterior probabilities (as numbers between 0 and 100) were elicited from the same participants before and after presentation of forensic testimony about each of a series of a pieces of evidence, each piece of evidence relating to a different suspect. Within-participant comparisons were performed between different formats for presentation of (purportedly) the same likelihood-ratio value: a numerical random-match probability of 1 in 100; a numerical likelihood ratio of 100;

[8] Even if one concluded that such a response scale did aid understanding, it seems an unrealistic proposition that juries, or even judges, would use them in a courtroom.

[9] Martire et al. [45] also tested a condition in which participants were presented with the whole verbal

and "moderate" strength of support. Figure 2 shows violin plots of the effective log likelihood ratios which we calculated using the prior-probability and posterior-probability responses from Bali et al. [51] Study 2.[10] Substantial between-participant variability is apparent, but participants' effective log-likelihood-ratio values tended to be closer to the neutral value of 0 than the logarithm of the presented likelihood-ratio value. Although the presented likelihood ratio value was (purportedly) 100 irrespective of the presentation format, the median effective likelihood-ratio value for each presentation format was less than 10. No presentation format resulted in distributions of elicited likelihood-ratio values that were obviously more *orthodox* than those for the other presentation formats.

Figure 2. For each of the three presentation formats, violin plots of effective log-likelihood-ratio

scale, but this did not lead to *orthodox* results.

[10] Figure 2 excludes 15 effective log likelihood ratios from when participants gave a prior-probability response or a posterior-probability response that was either 0% or 100%. These would have resulted in effective log-likelihood-ratio values of plus infinity or minus infinity.

values calculated using the response data from Bali et al. [51] Study 2.

In Thompson et al. [55], videoed testimony was presented to participants. Figure 3 shows violin plots of effective log-likelihood-ratio values given each combination of conditions tested: a condition designed to elicit higher prior odds versus a condition designed to elicit lower prior odds; a presented numerical likelihood ratio of 3,000 versus 30; and provision of an explanation of the meaning of likelihood ratios versus no provision of an explanation.

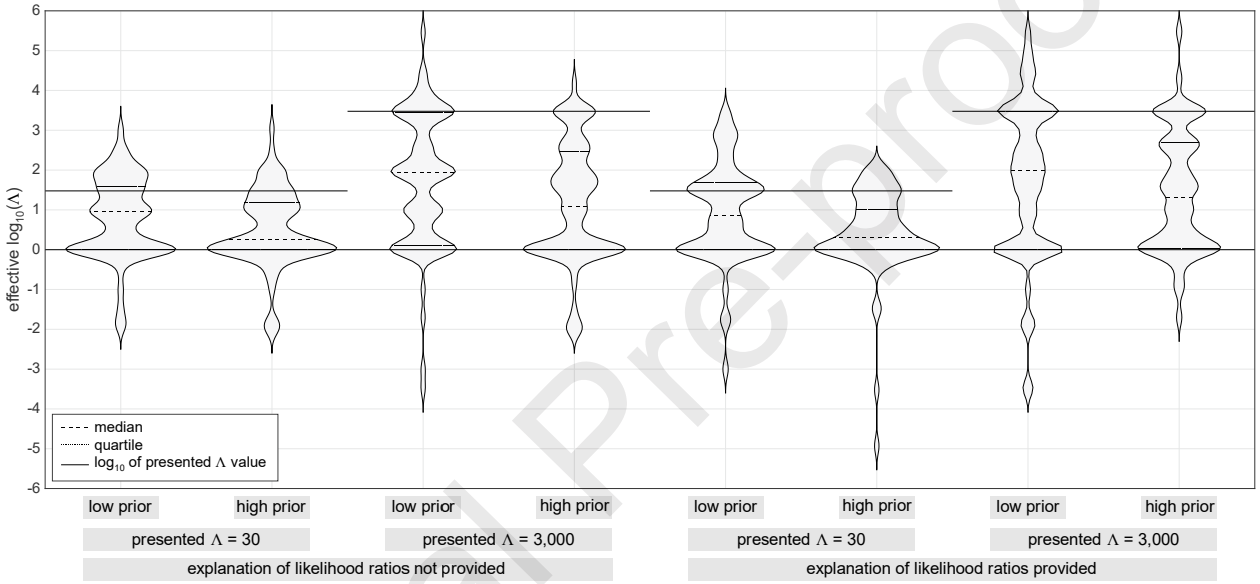

Figure 3. Effective log likelihood ratios given each combination of conditions in Thompson et al. [55].

The results shown in Figure 3 indicate *sensitivity*: Participants tended to have higher effective log likelihood ratios when the presented numerical likelihood ratio was 3,000 compared to when it was 30. With respect to *orthodoxy*, however, most participants had effective log-likelihood-ratio values that were lower than the logarithm of the presented likelihood-ratio value.

Interpretation of results was complicated by the fact that, across conditions, a

substantial proportion of participants' prior odds were 1,[11] and one could not determine whether effective likelihood-ratio values that equalled the presented likelihood-ratio value were due to correct application of Bayes' theorem to the presented likelihood-ratio value, or due to the prosecutor's fallacy (Thompson and Schumann [57], the prosecutor's fallacy is discussed in §5.4 below). Excluding participants whose effective likelihood-ratio values appeared to be *orthodox* but whose prior odds were 1, across prior-odds and presented-likelihood-ratio conditions, the number of participants whose effective likelihood-ratio values were *orthodox* was 7 out of 232 (3.0%) for those provided with the explanation of the meaning of likelihood ratios, and 2 out of 272 (0.74%) for those not provided with the explanation. The proportion of participants whose effective likelihood-ratio values equalled the presented likelihood-ratio values was higher for participants provided with the explanation than for participants not provided with the explanation, but (as discussed in Thompson et al. [55]) even if this were taken as evidence that providing the explanation of the meaning of likelihood ratios resulted in a higher proportion of *orthodox* responses, the proportion of participants whose responses were *orthodox* was still small. Thompson et al. [55] concluded that this did not constitute convincing evidence that providing the explanation of the meaning of the likelihood ratio led to more *orthodox* responses.

5.4 Coherence

5.4.1 Definition

- "*Coherence* is responding to evidence in a logical manner." "This definition excludes a range of potentially 'incoherent' lay responses to statistical statements that are incompatible with genuine comprehension such as the Prosecutor's and Defense Attorney's Fallacies (e.g., Thompson and Schumann [57]), directional errors (e.g., Martire et al. [44]), and aggregation errors (e.g., Koehler et al. [58])."

- Participants' responses are *coherent* if they reflect logically correct

[11] In Thompson et al. [55], participants gave open prior odds responses (and later open posterior odds responses), but this was the third of three elicitation stages. In the first stage, participants responded to a forced-choice question about whether the speaker of questioned identity was more likely to be the suspect, equally likely to be the suspect or someone else, or more likelihood to be someone else. If a participant responded "equally likely", the odds were recorded as 1, and the participant did not proceed to the second and third stages. This procedure might have induced a larger proportion of prior odds responses of 1 than if only an open elicitation of prior odds had been used. (The second stage asked participants to respond using a six-level scale in which the levels were at orders of magnitude, which might have influenced the values of the open odds responses that participants provided at the third stage.)

interpretation of likelihood ratios, i.e., if they indicate that participants have avoided reasoning errors and logical fallacies.

In the studies we reviewed, participants were judged to have understood the meaning of the likelihood ratios if their responses indicated that they avoided reasoning errors and logical fallacies.

Diversity within and between studies with respect to which reasoning errors and logical fallacies were investigated and how they were investigated makes summarizing the full range of results with respect to *coherence* difficult. Instead, we focused only on the *weak-evidence effect* (Martire et al. [44]), and on the most widely discussed fallacy with respect to interpretation of likelihood ratios: the transposition of the conditionals, also known as the *prosecutor's fallacy* (Thompson & Schumann [57]).

The weak-evidence effect occurs when, for example, "weak support" for H_1 is interpreted as support for H_2, or a numerical likelihood ratio that is a little larger than 1 is interpreted as if it were a numerical likelihood ratio that is less than 1.

The prosecutor's fallacy occurs when a likelihood ratio (the relative probabilities of the evidence given the hypotheses, $p(E \mid H_1)/p(E \mid H_2)$) is interpreted as if it were the posterior odds (the relative probabilities of the hypotheses given the evidence, $p(H_1 \mid E)/p(H_2 \mid E)$), i.e., when the effect of the prior odds are ignored.

5.4.2 Summary of results

Nine of the papers reviewed used *coherence* as an indicator of comprehension.

The weak-evidence effect was much more prevalent for verbal strength-of-support statements than for numerical likelihood ratios. Providing participants with the whole verbal scale reduced the prevalence of the weak-evidence effect for strength-of-support statements.

The prosecutor's fallacy appeared to occur more frequently when participants were presented with numerical likelihood ratios than when they were presented with numerical random-match probabilities, and much more frequently than when they were presented with verbal strength-of-support statements, but these results may have been an artifact of experiment design.

In studies that asked participants to identify fallacies in written statements, the rate at which participants failed to identify statements that contained the prosecutor's as fallacious, was much higher than the rate of spontaneous occurrence of the prosecutor's fallacy in studies in which priors were elicited, likelihood-ratio values were presented, and posteriors were elicited. The high rates observed for the prosecutor's fallacy when written statement were presented might have been induced by the experiment design itself. The low rates observed for spontaneous occurrence of the prosecutor's fallacy

might suggest that the prosecutor's fallacy is not as prevalent as it is often feared to be.

The results suggest that numerical likelihood ratios are better for avoiding the weak-evidence effect, but that verbal strength-of-support statements are better for avoiding the prosecutor's fallacy (although the latter comes with the caveat that the results may have been artifacts of experiment design). These results suggest a trade-off with respect to the question of the best way for forensic practitioners to present likelihood ratios so as to maximize their understandability for legal-decision makers.

5.4.3 Detailed results

5.4.3.1 Papers

The papers in the review that used *coherence* as an indicator of comprehension were: Taroni & Aitken [40], Nance & Morris [41], Martire et al. [44], Martire et al. [45], Thompson & Newman [46], van Straalen et al. [50], Bali et al. [51] Study 2, van Straalen et al. [53], Thompson et al. [55].

5.4.3.2 Weak-evidence effect

As is apparent from examination of Figure 1, in Martire et al. [44] Experiment 1, Martire et al. [44] Experiment 2, and Martire et al. [45], verbal strength-of-support statements of "weak or limited support" for H_1 resulted in median effective log-likelihood-ratio values that were negative. In Martire et al. [44] Experiment 2, however, the weak-evidence effect was not observed for "weak of limited" support for H_2, i.e., the median effective log-likelihood-ratio value was not positive.

Martire et al. [45] included a condition in which a whole verbal scale including verbal strength-of-support statements and corresponding ranges of numerical likelihood-ratio values was presented.[12] Presenting the whole verbal scale has been recommended in AFSP [5], Jackson et al. [59], and ISO 21043-5:2025 [60], although Marquis et al. [28] argued against this practice. As is apparent from examination of Figure 1, the results in Martire et al. [45] indicated that presenting the whole verbal scale (at least partially) alleviated the weak-evidence effect observed when only a verbal strength-of-support

[12] Martire et al. [45] presented 55 and 5,500 as numerical likelihood ratios. The corresponding numerical likelihood-ratio ranges on the verbal scale were 10–100 and 1,000–10,000. For the strength-of-support statements (with or without the whole verbal scale), the values used for comparison with the effective likelihood-ratio values were 55 and 5,500. The stated logic was that 55 and 5,500 were in the middle of the numerical likelihood-ratio ranges on the verbal scale. On the logarithmic scaling of the verbal scales the middles of the ranges were actually 32 and 3,162, but using these values instead would not have affected the conclusion that the responses were not *orthodox*.

statement was presented; however, they also indicated that presenting the whole verbal scale did not produce more *orthodox* responses compared to when only a numerical likelihood ratio is presented.

More fine-grained analysis revealed that, although the weak-evidence effect was strongest for verbal strength-of-support statements, it also occurred to lesser extents for other formats for presenting likelihood ratios. Martire et al. [45] reported that the weak-evidence effect was exhibited in the responses of 64% of participants in its verbal-strength-of-support condition, 32% of participants in its verbal-strength-of-support + verbal-scale condition, 38% of participants in its visual-representation condition, and 13% of participants in its numerical-likelihood-ratio condition.

Negative log-likelihood-ratio values in Figure 2 and Figure 3 also reveal small proportions of responses in Bali et al. [51] Study 2 and in Thompson et al. [55] exhibiting the weak-evidence effect for numerical random-match probabilities, numerical likelihood ratios, and verbal strength-of-support statements. As noted in Thompson et al. [55], however, these results could have been due to participants in online experiments not paying attention.

5.4.3.3 Prosecutor's fallacy

In Nance & Morris [41] the random-match probability was 1 in 25, or 4%. Nance & Morris [41] therefore suggested that a posterior-probability response of 96% could be considered an indication that the participant had committed the prosecutor's fallacy.[13] Assuming this to be true, participants presented with the numerical random-match probability committed the prosecutor's fallacy at a rate of 2%, participants presented with the equivalent numerical likelihood ratio at a rate of 8%, and participants presented with both the numerical likelihood ratio and a chart for converting from prior probabilities to posterior probabilities at a rate of 9%. Nance & Morris [41] noted that the testimony using the numerical random-match probability was designed to reduce the incidence of the prosecutor's fallacy. If these results are indeed due to erroneous understanding, the rates of occurrence of the prosecutor's fallacy was relatively low.

In Thompson et al. [55], participants were presented with numerical likelihood ratios. Excluding participants whose posterior odds were consistent with them having committed the prosecutor's fallacy but whose prior odds were 1, the number of participants who committed the prosecutor's fallacy was 31 out of 232 (13%) for those provided with an explanation of the meaning of likelihood ratios versus 47 out of 272 (17%) for those not provided with the explanation. Thompson et al. [55] concluded that, based on these results, there was no evidence that providing the explanation of the

[13] The logic is, perhaps, more easily understood in terms of likelihood ratios and odds: If the likelihood ratio of 25 is misinterpreted as posterior odds of 25, the posterior probability is, to two significant figures, $25/(1+25) = 25/26 = 0.96$.

meaning of likelihood ratios reduced the prevalence of the prosecutor's fallacy.

The rates of occurrence of the prosecutor's fallacy in Thompson et al. [55] were somewhat higher than those in Nance & Morris [41]. Since Thompson et al. [55] only presented participants with numerical likelihood ratios, the results are not informative with respect to relative *coherence* for different likelihood-ratio-presentation formats.

In Taroni & Aitken [40], participants were presented with twelve excerpts of transcripts from real cases and asked whether the statements made in the excerpts were correct or incorrect. All twelve statements involved correct or incorrect interpretations of random-match probabilities. Three statements included the prosecutor's fallacy. On average forensic-medicine students responded that the latter statements were correct at a rate of 69%, forensic-science students at a rate of 32%, and criminal-justice professionals at a rate of 15%. Although the criminal-justice professionals performed the best, their failure to recognize the prosecutor's fallacy at a rate of 15% may still be of concern. Since the statements in Taroni & Aitken [40] only involved random-match probabilities, the results are not informative with respect to relative *coherence* for different likelihood-ratio-presentation formats.

Thompson & Newman [46] gave participants three written texts interpreting the meaning of presented strength-of-evidence statements and asked them to respond whether the interpretations were correct or incorrect. Two of these interpretations committed the prosecutor's fallacy, one stated in odds format, e.g., "it is 100 times more likely that the DNA came from the defendant than from a random person", and the other stated in frequency format, e.g., "there is one chance in 100 that the DNA came from any person other than the defendant". The percentage of participants who responded that at least one of these interpretations was correct was 86% for participants who were presented with numerical likelihood ratios, 78% for participants presented with numerical random-match probabilities, and 26% for participants presented with verbal strength-of-support statements.

Participants who were presented with numerical likelihood ratios were more likely to respond that the odds-format version of the prosecutor's fallacy was correct, and participants presented with numerical random-match probabilities were more likely to respond that the frequency-format version of the prosecutor's fallacy was correct. As discussed in Thompson & Newman [46], participants may simply have agreed with wording that was superficially similar to the wording of the strength-of-evidence format that they were presented with, i.e., they may not have carefully read the interpretations and noticed that they differed from the strength-of-evidence statements but still decided they had the same meaning. The results may have been induced by the format of the task rather than actually representing underlying understanding. Neither of the wordings of the prosecutor's fallacy were similar to the verbal strength-of-support format, and participants who were presented with that format responded that the prosecutor's fallacy interpretations were correct at a much lower rate.

The results may appear to suggest that the numerical likelihood-ratio format is more prone to eliciting the prosecutor's fallacy than the numerical random-match probability format which in turn is more prone to eliciting the prosecutor's fallacy than the verbal strength-of-support format, but this may be an artifact of the design of the experiment.

Bali et al. [51] Study 2 presented participants with a written statement containing the prosecutor's fallacy and asked whether it was correct or incorrect. Participants who had earlier in the experiment been presented with numerical likelihood ratios failed to recognize that the statement was incorrect at a rate of 93%, participants presented with numerical random-match probabilities at a rate of 74%, and participants presented with verbal strength-of-support statements at a rate of 84%. The rate for verbal strength-of-support statements was much higher than in Thompson & Newman [46].

The error rates for forensic-medicine students and forensic-science students in Taroni & Aitken [40] and for members of a jury-eligible communities in Thompson & Newman [46] and in Bali et al. [51] Study 2 were much higher than those observed for former jurors / jury-pool members in Nance & Morris [41] and for members of the general population in Thompson et al. [55]. The tasks that participants were asked to perform were, however, very different. Whereas Nance & Morris [41] and Thompson et al. [55] may have uncovered spontaneous instances of the prosecutor's fallacy, the much higher rates in Taroni & Aitken [40], Thompson & Newman [46], and Bali et al. [51] Study 2 may have been an artifact of the experiment task: Participants may have agreed to statements that appeared to be superficially similar to testimony that was presented to them, rather than carefully reading the testimony and the statements and noticing that they differed but still deciding that they meant they same thing, i.e., the results may have been due to lack of attention to detail.

In van Straalen et al. [50], and in van Straalen et al. [53], participants were asked to respond whether a number of statements about the presented testimony were correct or incorrect. One of these statements was "There is more than a 50% chance the fingermark belongs to the suspect." Table 6 shows rates at which participants responded that this was correct.[14] Van Straalen et al. [50] claimed that participants who responded that this was correct committed the prosecutor's fallacy because the testimony presented the probability of the evidence given the hypotheses, not the probability of the hypotheses. At the higher strength of evidence, participants who were presented with a numerical likelihood ratio of 5M therefore appeared to commit the prosecutor's fallacy at a higher rate than participants who were presented with a verbal strength-of-evidence statement of "extremely strong". It seems reasonable, however, that in answering this question participants could have combined the presented strength of evidence with their own priors to arrive at their own posteriors, and then answered on

[14] For simplicity, in the present paper, when summarizing multiple studies or multiple conditions, if multiple studies or conditions results in different values but those differences are not important, we write the range of values from across the studies or conditions (e.g., 40–42).

the basis of whether their own posterior probability was greater than 50%, thus accounting for why participants who were presented with stronger strengths of evidence responded "correct" at higher rates.

Table 6. Percentages of participants in van Straalen et al. [50] and in van Straalen et al. [53] who responded "correct" to "There is more than a 50% chance the fingermark belongs to the suspect."

Strength of evidence	Numerical likelihood ratio	Verbal strength of support
5M or "extremely strong"	40–42	31
50 or "moderately strong"	10–11	8–9

6 Review and recommendations with respect to methodological issues

6.1 Overview

Many of the studies that we reviewed suffered from weaknesses in experimental design, and all the studies that we reviewed (except Thompson et al. [55]) were intended to answer research questions that differed from our own. This often led to research designs that were suboptimal for addressing our research question. In the present section, we focus on methodological issues, and make recommendations for methodology in future research that addresses our research question.

6.2 Presented likelihood-ratio values

In §5.2.3, based on consideration of sensitivity results across studies, we hypothesized that participants were not *sensitive* to differences between presented likelihood-ratio values that all fell below a threshold, or *sensitive* to differences between presented likelihood-ratio values that all fell above the threshold, but only *sensitive* to differences between presented likelihood-ratio values if one fell below the threshold and another fell above the threshold. The posited threshold was somewhere between a presented likelihood-ratio value of 100 and a presented likelihood-ratio value of 450. To test this hypothesis, we recommend testing presentation of at least two likelihood-ratio values on each side of the posited threshold.

With the exception of Martire et al. [44] Experiment 2, all studies only presented likelihood ratios for which $p(E \mid H_1)/p(E \mid H_2) > 1$, and did not present likelihood ratios for which $p(E \mid H_1)/p(E \mid H_2) < 1$. Martire et al. [44] found a differential effect with respect to the weak-evidence effect. To further investigate the understanding of

both $p(E \mid H_1)/p(E \mid H_2) > 1$ and $p(E \mid H_1)/p(E \mid H_2) < 1$, we recommend presenting both. $p(E \mid H_1)/p(E \mid H_2) < 1$ would be presented as numbers greater than 1, but with the hypotheses inverted, i.e., $p(E \mid H_2)/p(E \mid H_1) > 1$.

6.3 Elicitation of priors and posteriors

For reasons explained in §5.3.3 we recommend:

- that prior and posterior responses be elicited in open odds format, not as probabilities and not using a multilevel scale;[15] and
- that prior odds and posterior odds be elicited from the same individuals, not from different groups.

As discussed in §5.3.3 and §5.4.3.3, if the elicited prior odds are 1 and the elicited posterior odds equal the presented likelihood-ratio value, one cannot distinguish whether the latter is due to the participant correctly applying Bayes' theorem to the presented likelihood-ratio value or whether it is due to them committing the prosecutor's fallacy. We therefore recommend that case scenarios be designed with the intent of eliciting prior odds that are substantially different from 1, e.g., substantially less than 1.

6.4 Source level versus offence level

Although all of the studies in our review presented likelihood-ratio values addressing source-level hypotheses, seven papers (Taroni & Aitken [40], Nance & Morris [41], Nance & Morris [42], Martire et al. [44], Martire et al. [45], Thompson & Newman [46], Bali & Martire [54]) did not elicit probabilistic responses phrased in terms of source-level hypotheses, but, instead, elicited probabilistic responses phrased in terms of guilt, i.e., offence-level hypotheses.[16]

A participant could potentially have an *orthodox* interpretation of a source-level

[15] A reviewer pointed out that many people are more familiar with percentages than with odds; however, we think the disadvantages of percentages outweigh any benefit that may be due to their familiarity. The form of Bayes' theorem that uses prior odds and posterior odds is simpler, and thus is expected to be easier to understand than the form that uses prior probability and posterior probability.

[16] Some studies elicited binary "guilty"/"not guilty" responses. We did not consider these in our

likelihood ratio, but (quite reasonably) consider other information or other factors when asked about their probabilistic beliefs with respect to offence-level hypotheses. The participant's effective offense-level likelihood ratio would then differ from their effective source-level likelihood ratio. It would also differ from the presented likelihood-ratio values, giving the false impression that they did not understand the meaning of likelihood ratios.

We recommend that priors and posteriors be elicited using questions that are clearly phrased in terms of source level, not offence level.[17]

6.5 Extraneous information

Eleven studies (Koehler [39], Taroni & Aitken [40], Nance & Morris [41], Nance & Morris [42], Langenburg et al. [43], Martire et al. [44], Martire et al. [45], Thompson & Newman [46], Bali et al. [51], Bali & Martire [54], Thompson et al. [55]) presented elaborate case scenarios.[18]

In these case scenarios, extraneous information that was unrelated to controlled experimental factors may have affected the results in unanticipated ways. It may have affected participants' assessments of the presented likelihood-ratio values and thus affected the strength of their effective likelihood ratios. Rather than updating their beliefs based solely on the likelihood-ratio value presented and on controlled experimental factors, participants' updating of beliefs may have also taken account of extraneous information.

Although related to an experimental factor rather than to extraneous case information, a clear example of context affecting participants' responses occurred in Thompson & Newman [46]. In Thompson & Newman [46], participants' responses gave more weight to the same likelihood-ratio values when they were purported to relate to DNA than when they were purported to relate to footwear marks.[19] A possible explanation is that participants did not believe that footwear-mark evidence could be as strong as DNA evidence and so downweighed the footwear-mark evidence.

review. For this reason, we did not include Garrett et al. [61] in our review.

[17] If activity-level testimony is presented, then priors and posteriors should be elicited using questions that are clearly phrased in terms of activity level, not offence level.

[18] Some other studies lacked details of what was presented as case scenarios. Van Straalen et al. [50] explicitly presented casework reports with minimal information.

[19] Similar results occurred in Garrett et al. [61], in which participants gave more "guilty" responses to the same likelihood-ratio values when they were purported to relate to fingermark-fingerprint evidence than when they were purported to relate to voice-recording evidence.

Similarly, extraneous case information could have affected participants' beliefs about the validity or trustworthiness of the presented likelihood-ratio value or could otherwise have biased how they interpreted the presented likelihood-ratio value.

To focus on the interpretation of the likelihood ratio, and to minimize the potential impact of extraneous information, we recommend that case information provided, apart from that related to experimental factors, be restricted to the minimum necessary to inform prior odds.

Psychologists have struggled for decades with what is sometimes called the "real-world or the lab dilemma" of whether it is better to study social phenomena in contexts as close as possible to real-world settings or in laboratory settings that are contrived to allow a greater level of uniformity and experimental control (Hollerman et al. [62]). We believe that, at this juncture, research on understanding of likelihood ratios would benefit greatly from methods that allow higher levels of uniformity and control of extraneous variables. Although it may eventually be helpful to study the potential effect that more "ecological" contexts and factors have on participants' responses to presented likelihood ratios, we believe the first goal should be to better understand factors that, in controlled settings, affect participants' responses to presented likelihood ratios.

6.6 Perceived quality of testimony

Thompson et al. [55] elicited participants' judgements about the quality of the presented testimony, collecting Likert-scale responses to questions about whether the expert witness was qualified, whether the expert witness was credible, whether the expert witness was trustworthy, whether the expert witness was biased, and whether the methods used by the expert witness were valid. Thompson et al. [55] found a positive correlation between participants' judgements about the quality of the testimony and their effective likelihood-ratio values. This suggested that participants were weighting the presented likelihood-ratio values based on their judgements of the quality of the testimony. As discussed in Thompson et al. [55], this may be a perfectly reasonable thing to do, but it would result in effective likelihood-ratio values that differed from the presented likelihood-ratio values, giving the false impression that participants did not understand the meaning of likelihood ratios.[20]

To reduce the probability that participants downweight the likelihood-ratio value because they perceive the validity of a particular branch of forensic science to be low, we recommend that conditions be tested in which the technology used and the decisions

[20] Similarly Thompson et al. [55] found negative correlations between participants' prior odds and their effective likelihood ratios. Thompson et al. [55] speculated that this could be due to participants having a bias against arriving at extremely high posterior odds. Having a such a conservative bias might be considered reasonable, but it would lead to results that were not *orthodox*.

made in calculating the likelihood-ratio value are (at a high level) explained (this was done in Thompson et al. [55]), and in which validation of the forensic-evaluation system under the casework conditions is explained and the results presented (see Morrison et al. [3]), and in which the presented likelihood-ratio values is clearly supported by the validation results. These are all thing which we believed should be standard as part of a forensic practitioner's expert testimony.

Also (as was done in Thompson et al. [55]), we recommend eliciting participants' judgements about the quality of the testimony so that these judgement can be compared with participants' effective likelihood-ratio values.

We also recommend that, when eliciting posterior odds, participants be asked not only what value they actually assigned for their posterior odds, but also what the posterior odds would be if they had applied Bayes' theorem to the presented likelihood-ratio value. This would allow researchers to determine whether participants understand the meaning of the presented likelihood ratio in terms of being able to correctly apply Bayes' theorem to the presented likelihood ratio.

6.7 Manipulating quality of testimony

Four studies (Nance & Morris [41], Nance & Morris [42], Ribeiro et al. [52], Bali & Martire [54]) included conditions in which, in addition to presenting likelihood-ratio values, classification-error rates were also presented. Ribero et al. (2020) also presented alibis of different strengths, and Bali & Martire [54] also included weaknesses in other parts of the report. These were experimental factors which were expected to affect participant's responses, but they made it difficult to distinguish effects related to interpretation of formats for presentation of likelihood ratios versus effects related to interpretation of classification-error rates, alibis, or weaknesses.

Also, although probability of a match due to random selection from a population and probability of a match due to an error have the same format and thus could be considered commensurate, for continuously-valued data, numerical likelihood-ratio values and classification-error rates are not commensurate. A likelihood ratio is not a categorical decision based on thresholding posterior odds, so the performance of systems that output likelihood ratios cannot be assessed in terms of classification-error rates (Morrison [63]).

We recommend that manipulated conditions related to the quality of the testimony focus exclusively on the performance of the system (purportedly) used to generate the presented likelihood ratio, and that procedures used for assessing performance and metrics and graphics used to represent system performance be commensurate with the likelihood-ratio framework, e.g., log-likelihood-ratio costs (C_{llr}) and Tippett plots. For guidance on how to assess the performance of systems that output likelihood ratios (and for explanations of C_{llr} and Tippett plots), see Morrison et al. [3]. We further recommend that any likelihood-ratio values presented be supported by the validation

results. For two well-calibrated systems, if one system has poorer performance than the other system, then the likelihood ratios output by the poorer-performing system will tend to be closer to the neutral likelihood-ratio value of 1 than is the likelihood ratios output by the better-performing system (Morrison [64]).

6.8 Explanation of the meaning of likelihood ratios

With the exception of Langenburg et al. [43] and Thompson et al. [55], all the studies we reviewed presented likelihood ratios without any explanation of their meaning. A priori, we expect that providing an explanation of the meaning of likelihood ratios would help legal-decision makers understand the meaning of likelihood ratios, and believe that it would be appropriate for forensic practitioners acting as expert witnesses to be asked to explain the meaning of likelihood ratios during examination in chief.

Thompson et al. [55] did not find convincing evidence that providing an explanation of the meaning of likelihood ratios increased the proportion of participants whose responses were *orthodox*, and (with respect to *coherence*) did not find evidence that it reduced the prevalence of the prosecutor's fallacy; however, these were the results from only one study, and our priors are such that we recommend that future studies further investigate the effect of providing explanations of the meaning of likelihood ratios.

The explanation of the application of Bayes' theorem in Thompson et al. [55] used an example likelihood-ratio value of 4, which differed from the presented value (which was 30 or 3,000). As suggested in Thompson et al. [55], we recommend instead using the presented likelihood-ratio value in the explanation. This would focus on the question of whether participants understand the likelihood-ratio value actually presented to them, rather than on the question of whether participants were able to generalize principles of understanding from one value to another. Although *ability* might be considered the gold standard for demonstrating understanding, in the context of a case, a legal-decision maker does not have to generalize principles of understanding to likelihood ratios in general, but does have to understand the meaning of the particular likelihood-ratio value that is actually presented to them.

As discussed in Thompson et al. [55], in that study, the first example in the explanation of the application of Bayes' theorem used prior odds of 1, which may have mislead participants into committing the prosecutor's fallacy. As suggested in Thompson et al. [55], we recommend that explanations of Bayes' theorem avoid using examples in which the prior odds are 1.

6.9 Charts or graphics for converting priors to posteriors

In three studies (Nance & Morris [41], Nance & Morris [42], Bayer et al. [47]), in addition to being presented with a numerical likelihood ratio, participants were also presented with a chart or a graphic for converting from priors to posteriors given the presented likelihood-ratio value. These could be viewed as attempts to explain the

meaning of the presented likelihood-ratio value.

The charts in Nance & Morris [41] and Nance & Morris [42] listed a selection of prior probabilities and posterior probabilities, and the graphic in Bayer et al. [47] related number of potential offenders to posterior probabilities. Compared to a chart or graphic relating prior odds to posterior odds, these formats would have made it difficult to understand and generalize the meaning of a likelihood ratio as the amount by which one should update ones prior beliefs about the relative probabilities of the hypotheses so as to arrive at posterior beliefs about the relative probabilities of the hypotheses, i.e., by multiplying ones prior odds by the numerical likelihood-ratio value to arrive at ones posterior odds.

In Nance & Morris [41] and in Nance & Morris [42], providing the chart or the graphic did not results in a larger proportion of *orthodox* responses: In Nance & Morris [41], for a presented numerical likelihood-ratio value of 25, median effective likelihood-ratio values were 4.5–9.0 when the numerical likelihood ratio alone was presented, and 4.1–5.4 when the chart was also presented. In Nance & Morris [42], for a presented numerical likelihood-ratio value of 4,000, the median effective likelihood-ratio value was 7.0 when the numerical likelihood ratio alone was presented, and was also 7.0 when the chart was also presented.

In Bayer et al. [47], when, in addition to the numerical likelihood ratio, a verbal scale was presented or a verbal scale and a graphic for converting from priors to posteriors were presented, the median order of magnitude of the effective likelihood-ratio values was the same as for the presented likelihood-ratio values. The results may already have asymptoted after adding the verbal scale, so the effect of adding the chart could not be determined (the presentation of the results in Bayer et al. [47] did not allow more fine-grained analysis than at the level of the median order of magnitude).

We recommend that any tables, chart, or explanations for how to update priors to posteriors using a likelihood ratio be formatted in terms of odds. We recommend that explanation of the meaning of likelihood ratios be based on the odds form of Bayes' theorem.

7 Additional recommendations

7.1 Overview

The recommendations in this section come not directly from our review of published studies, but from our insights based on broader reflection on the research question and related methodological issues.

7.2 Best format from a theoretical perspective

Most research included in our review was agnostic with respect to theoretical issues related to the best way to present likelihood ratios, and tested multiple presentation formats that are in use. From a theoretical perspective, however, random-match probabilities cannot be used if, as is the case in most branches of forensic science, the data to be interpreted are continuously valued and have within-source variability. Also, verbal likelihood ratios and verbal strength-of-support statements have no intrinsic meaning, and the only way to give them specific definitions would be by reference to the ranges of numerical likelihood ratios to which they are arbitrarily associated in verbal scales.

Given these problems with the other formats, we recommend that future research focus on presentation of numerical likelihood ratios. Presenting a verbal expression in addition to the presented a numerical likelihood ratio could be an experimental condition (it could be hypothesized that providing a verbal expressions in addition to a numerical likelihood-ratio value would aid understanding), but we do not recommend presenting verbal expressions by themselves.

7.3 Written versus oral presentation of testimony

To better reflect how triers of fact in common-law jurisdictions usually receive forensic-science testimony, rather than presenting the experiments entirely in writing, we recommend that, as was done in Thompson et al. [55], testimony (including explanations of the meaning of likelihood ratios) be presented via video recordings. Video recording, rather than live acting, will maintain consistency when the same testimony is presented at different times to different participants. We recommend that participants also be provided with a transcript of the testimony. Access to a transcript of testimony would be a reasonable expectation for legal-decision makers. It would allow participants to review the testimony in detail.

In some jurisdictions or contexts, legal-decision makers make decisions on the basis of written reports, so the written format is not invalid, but we recommend that the scope of applicability of written-format experiments and video-format experiments be made clear.

7.4 Individual participants versus collaborating groups of participants

In all the studies we reviewed, responses were collected from individual participants. This may be informative with respect to understanding by individual judges or individual lawyers, but not with respect to understanding by juries who are groups of collaborating individuals. We therefore recommend that future research include tests of the understanding of likelihood ratios by groups of collaborating individuals. Procedures could be similar to those used in Bali et al. [65] to test speaker-identification performance by groups of collaborating listeners.

7.5 Online versus in-person experiments, and representativeness of participants

As indicated in Table 5, although some studies included in our review recruited participants who were jury eligible or former jury members, or who were criminal-justice professionals, other studies recruited participants from more convenient pools, such as university students. In recent years, it has become increasingly common to conduct participant-response experiments using online platforms, and this is reflected in the studies in our review. Recruiting participants and running experiments online has the great advantages of being able to obtain responses from large numbers of participants quickly and cheaply, but has the disadvantage that some participants might not perform the task as conscientiously as participants who are invigilated during in-person experiments (or even who would be willing to take part in in-person experiments). Recruiting participants who are judges or lawyers, or to a lesser extent who are jury eligible, is more difficult, and running in-person invigilated experiments is more difficult. These have the disadvantages of being more costly, more time consuming, and of researchers not being able to recruit as many participants (the geographically local pool of willing volunteers may be quite small).

Practically, for future research, it would make sense to conduct early experiments with participants from convenient populations who are recruited via online platforms and who participate as individuals in online experiments, but ultimately it would be desirable to recruit actual legal-decision makers to participate as individuals in in-person experiments and to recruit jury-eligible participants as members of groups whose members collaborate in in-person experiments.

8 Conclusion

We began with the premise that forensic practitioners should use the likelihood-ratio framework to evaluate strength of forensic evidence, and our research question was:

- What is the best way for forensic practitioners to present likelihood ratios so as to maximize their understandability for legal-decision makers?

We reviewed studies in which participants responded to different formats for presentation of likelihood-ratio values:

- numerical likelihood ratios
- numerical random-match probabilities
- verbal strength-of-support statements

None of the studies we reviewed presented participants with verbal likelihood ratios.

In general, participants were *sensitive* to all three formats for presentation of likelihood-ratio values, i.e., they gave more weight to stronger likelihood-ratio values than to weaker likelihood-ratio values. Considered across studies, however, the results suggested that, rather than responding in a gradient manner, participants might have responded differently to likelihood ratios that were below a threshold value compared to likelihood ratios that were above the threshold value. If so, this would demonstrate a lack of understanding of the meaning of likelihood ratios. Since the *sensitivity* results were similar for all three formats, they do not help answer our research question.

In general, participants' responses to all three formats were not *orthodox* compared to updating of beliefs as per Bayes' theorem. For all three formats, average effective likelihood-ratio values were much weaker than the presented likelihood-ratio values. Instead of directly using the values of the presented likelihood ratio, however, participants might have weighted them based on other information provided, or on their perception of the validity of the branch of forensic science to which the testimony was related, or based on their perception of the quality of the particular testimony. This would have led to their effective likelihood-ratio values differing from the presented likelihood-ratio values. Since the *orthodoxy* results were similar for all three formats, they do not help answer our research question.

With respect to *coherence*, the weak evidence effect occurred more frequently given verbal strength-of-support statements, and the prosecutor's fallacy occurred more frequently given numerical likelihood ratios, but many of the results with respect to the prosecutor's fallacy may have been an artifact of the experiment design rather than being actually indicative of errors of understanding. The *coherence* results are not, therefore, particularly helpful in answering our research question.

Providing a table of chart for converting from prior to posteriors, or providing an explanation of the meaning of likelihood ratios, did not results in clearly better understanding of likelihood ratios, and improvement from providing the whole verbal scale was restricted to reducing the weak-evidence effect for strength-of-support statements.

With the exception of one study (Thompson et al. [55]), none of the studies we reviewed set out to address our specific research question, and, based on our review, we conclude that the existing literature does not provide an answer to our research question. We did, however, identify multiple methodological weaknesses in the studies, weaknesses that could have affected the results. In response, we generated a number of recommendations for methodology in future research. These recommendations could be followed in a series of experiments that systematically examines understanding of likelihood ratios by laypersons. We plan to conduct future research that follows the methodology recommendations made in the present paper.

9 References

[1] Aitken C.G.G., Berger C.E.H., Buckleton J.S., Champod C., Curran J.M., Dawid A.P., Evett I.W., Gill P., González-Rodríguez J., Jackson G., Kloosterman A., Lovelock T., Lucy D., Margot P., McKenna L., Meuwly D., Neumann C., Nic Daéid N., Nordgaard A., Puch-Solis R., Rasmusson B., Redmayne M., Roberts P., Robertson B., Roux C., Sjerps M.J., Taroni F., Tjin-A-Tsoi T., Vignaux G.A., Willis S.M., Zadora G. (2011). Expressing evaluative opinions: A position statement. *Science & Justice*, 51, 1–2. https://doi.org/10.1016/j.scijus.2011.01.002

[2] Morrison G.S., Kaye D.H., Balding D.J., Taylor D., Dawid P., Aitken C.G.G., Gittelson S., Zadora G., Robertson B., Willis S.M., Pope S., Neil M., Martire K.A., Hepler A., Gill R.D., Jamieson A., de Zoete J., Ostrum R.B., Caliebe A. (2017). A comment on the PCAST report: Skip the "match"/"non-match" stage. *Forensic Science International*, 272, e7–e9. http://dx.doi.org/10.1016/j.forsciint.2016.10.018

[3] Morrison G.S., Enzinger E., Hughes V., Jessen M., Meuwly D., Neumann C., Planting S., Thompson W.C., van der Vloed D., Ypma R.J.F., Zhang C., Anonymous A., Anonymous B. (2021). Consensus on validation of forensic voice comparison. *Science & Justice*, 61, 229–309. https://doi.org/10.1016/j.scijus.2021.02.002

[4] Morrison G.S., Biedermann A., Tart M., Meuwly D., Berger C.E.H., Guiness J., Houck M.M., Gibb C., Dawid A.P., Kotsoglou K.N., Kaye D.H., Rose P., Taroni F., Kokshoorn B., Saks M.J., Buckleton J.S., Curran J.M., Taylor D., Zhang C., Vuille J., Champod C., Simonsen B.T., Mattei A., Lucena-Molina J.J., Zabell S., Chin J.M., Gallidabino M., Wevers G., Moreton R., Eldridge H., Martire K.A., Aitken C.G.G., Cole S.A., González-Rodríguez J., Smithuis M., Edvardsen T., Wilson-Wilde L., Zadora G., Gittelson S., Jackson G., Sjerps M.J., Brard F., Hicks T., Kennedy J., Latten B.G.H., Weber P., Willis S., Ramos D., Koehler J.J., Ribeiro R.O., Crispino F., Basu N., Meakin G.E., Kirkbride K.P., Tully G., Jessen M., Syndercombe Court D. (2025). A response to EA-4/23 INF:2025 "The Assessment and Accreditation of Opinions and Interpretations using ISO/IEC 17025:2017". *Forensic Science International*. https://doi.org/10.1016/j.forsciint.2025.112589

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

[6] Aitken C.G.G., Roberts P., Jackson G. (2010). *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses*. London, UK: Royal Statistical

Society. https://rss.org.uk/news-publication/publications/law-guides/

[7] Willis S.M., McKenna L., McDermott S., O'Donnell G., Barrett A., Rasmusson A., Nordgaard A., Berger C.E.H., Sjerps M.J., Lucena-Molina J.J., Zadora G., Aitken C.G.G., Lunt L., Champod C., Biedermann A., Hicks T.N., Taroni F. (2015). *ENFSI Guideline for Evaluative Reporting in Forensic Science*. Wiesbaden, Germany: European Network of Forensic Science Institutes. http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf

[8] Ballantyne K.N., Bunford J., Found B., Neville D., Taylor D., Wevers G., Catoggio D. (2017). *An Introductory Guide to Evaluative Reporting*. Melbourne, VIC, Australia: National Institute of Forensic Science of the Australia New Zealand Policing Advisory Agency. https://www.anzpaa.org.au/nifs/publications/general

[9] Kafadar K., Stern H., Cuellar M., Curran J., Lancaster M., Neumann C., Saunders C., Weir B., Zabell S. (2019). *American Statistical Association Position on Statistical Statements for Forensic Evidence*. Alexandria, VA: American Statistical Association. https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf

[10] Forensic Science Regulator (2021). *Codes of Practice and Conduct: Development of Evaluative Opinions* (FSR-C-118 Issue 1). Birmingham, UK: Forensic Science Regulator. https://www.gov.uk/government/publications/development-of-evaluative-opinions

[11] Bali A.S., Edmond G., Ballantyne K.N., Kemp R.I., Martire K.A. (2020). Communicating forensic science opinion: An examination of expert reporting practices. *Science & Justice*, 60, 216–224. https://doi.org/10.1016/j.scijus.2019.12.005

[12] Swofford H., Cole S., King V. (2021). Mt. Everest – we are going to lose many: A survey of fingerprint examiners' attitudes towards probabilistic reporting. *Law, Probability and Risk*, 19, 255–291. https://doi.org/10.1093/lpr/mgab003

[13] Berger C.E.H., Buckleton J., Champod C., Evett I.W., Jackson G. (2011). Evidence evaluation: A response to the Court of Appeal judgment in R v T. *Science & Justice*, 51, 43–49. https://doi.org/10.1016/j.scijus.2011.03.005

[14] Redmayne M., Roberts P., Aitken C.G.G., Jackson G. (2011). Forensic science evidence in question. *Criminal Law Review*, 2011(5), 347–356.

[15] Morrison, G.S. (2012). The likelihood-ratio framework and forensic evidence in court: A response to R v T. *International Journal of Evidence and Proof*, 16, 1–29. https://doi.org/10.1350/ijep.2012.16.1.390

[16] Thompson W.C. (2012). Discussion paper: Hard cases make bad law – reactions to R v T. *Law, Probability and Risk*, 11, 347–359. https://doi.org/10.1093/lpr/mgs020

[17] Martire K.A. (2018). Clear communication through clear purpose: Understanding statistical statements made by forensic scientists. *Australian Journal of Forensic Sciences*, 50, 619–627. https://doi.org/10.1080/00450618.2018.1439101

[18] Thompson W.C. (2018). How should forensic scientists present source conclusions? *Seton Hall Law Review*, 48, 773–813. https://scholarship.shu.edu/shlr/vol48/iss3/9

[19] Eldridge H. (2019). Juror comprehension of forensic expert testimony: A literature review and gap analysis. *Forensic Science International: Synergy*, 1, 24–34. https://doi.org/10.1016/j.fsisyn.2019.03.001

[20] Martire K.A., Edmond G. (2020). How well do lay people comprehend statistical statements from forensic scientists? In Banks D., Kafadar K., Kaye D.H., Tackett M. (Eds.), *Handbook of Forensic Statistics*, pp. 201–224. Boca Raton, FL: CRC. https://doi.org/10.1201/9780367527709

[21] Jackson G. (2009). Understanding forensic science opinions. In Fraser J., Williams R. (Eds.), *Handbook of Forensic Science*, pp. 419–445. Cullompton, UK: Willan. https://doi.org/10.4324/9781843927327

[22] Kaye D.H. (2015). Presenting forensic identification findings: The current situation. In Neumann C., Ranadive A., Kaye D.H. (Eds.), *Communicating the Results of Forensic Science Examinations: Final Technical Report for NIST Award Number 70NANB12H014*, pp. 12–30. https://ssrn.com/abstract=2690899

[23] Morrison G.S., Thompson W.C. (2017). Assessing the admissibility of a new generation of forensic voice comparison testimony. *Columbia Science and Technology Law Review*, 18, 326–434. https://doi.org/10.7916/stlr.v18i2.4022

[24] President's Council of Advisors on Science and Technology (2016). *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports/

[25] Evett I.W. (1991). Interpretation: A personal odyssey. In Aitken C.G.G., Stoney D.A., (Eds.) *The Use of Statistics in Forensic Science*, pp 9–22. Chichester, UK: Ellis Horwood.

[26] Walsh K.A.J., Buckleton J.S., Triggs C.M. (1996). A practical example of the

interpretation of glass evidence. *Science & Justice*, 36(4), 213–218. https://doi.org/10.1016/S1355-0306(96)72607-2

[27] Morrison G.S., Enzinger E. (2018). Score based procedures for the calculation of forensic likelihood ratios – Scores should take account of both similarity and typicality. *Science & Justice*, 58, 47–58. http://doi.org/10.1016/j.scijus.2017.06.005

[28] Marquis R., Biedermann A., Cadola L., Champod C., Gueissaz L., Massonnet G., Mazzella W.D., Taroni F., Hicks, T. (2016). Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings. *Science & Justice*, 56, 364–370. https://doi.org/10.1016/j.scijus.2016.05.007

[29] Risinger D.M. (2013). Reservations about likelihood ratios (and some other aspects of forensic 'Bayesianism'). *Law, Probability and Risk*, 12, 63–73, https://doi.org/10.1093/lpr/mgs011

[30] Martire K.A., Edmond G., Navarro D.J., Newell B.R. (2017). On the likelihood of "encapsulating all uncertainty". *Science & Justice*, 57(1), 76–79. https://doi.org/10.1016/j.scijus.2016.10.004

[31] Morrison G.S., Ballantyne K., Geoghegan P.H. (2018). A response to Marquis et al (2017) What is the error margin of your signature analysis? *Forensic Science International*, 287, e11–e12. https://doi.org/10.1016/j.forsciint.2018.03.009

[32] Martire K.A., Growns B., Navarro D.J. (2018). What do the experts know? Calibration, precision, and the wisdom of crowds among forensic handwriting experts. *Psychonomic Bulletin & Review*, 25, 2346–2355, https://doi.org/10.3758/s13423-018-1448-3

[33] Evett I.W., Jackson G., Lambert J.A., McCrossan S. (2000). The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice*, 40, 233–239. https://doi.org/10.1016/S1355-0306(00)71993-9

[34] Nordgaard A., Ansell R., Drotz W., Jaeger L. (2012). Scale of conclusions for the value of evidence. *Law, Probability, and Risk*, 11, 1–24. https://doi.org/10.1093/lpr/mgr020

[35] ISO 20143-4:2025 *Forensic Sciences – Part 4: Interpretation*. Geneva, Switzerland: International Organization for Standardization. https://www.iso.org/obp/ui/en/#iso:std:iso:21043:-4:ed-1:v1:en

[36] Mullen C., Spence D., Moxey L., Jamieson A. (2014). Perception problems of the verbal scale. *Science & Justice*, 54, 154–158.

https://doi.org/10.1016/j.scijus.2013.10.004

[37] Martire K.A., Watkins I. (2015). Perception problems of the verbal scale: A reanalysis and application of a membership function approach. *Science & Justice*, 55, 264–273. https://doi.org/10.1016/j.scijus.2015.01.002

[38] Morrison G.S., Enzinger E. (2016). What should a forensic practitioner's likelihood ratio be? *Science & Justice*, 56, 374–379. https://doi.org/10.1016/j.scijus.2016.05.007

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

[40] Taroni F., Aitken C.G.G. (1998). Probabilistic reasoning in the law: Part 1: assessment of probabilities and explanation of the value of DNA evidence. *Science & Justice*, 38, 165–177. https://doi.org/10.1016/S1355-0306(98)72101-X

[41] Nance D.A., Morris S.B. (2002). An empirical assessment of presentation formats for trace evidence with a relatively large and quantifiable random match probability. *Jurimetrics*, 42, 403–448. https://www.jstor.org/stable/29762779

[42] Nance D.A., Morris S.B. (2005). Juror understanding of DNA evidence: An empirical assessment of presentation formats for trace evidence with a relatively small random-match probability. *Journal of Legal Studies*, 34(2), 395–444. https://doi.org/10.1086/428020

[43] Langenburg G., Neumann C., Meagher S.B., Funk C., Avila J.P. (2013). Presenting probabilities in the courtroom: A moot court exercise. *Journal of Forensic Identification*, 63(4), 424–488.

[44] Martire K.A., Kemp R.I., Watkins I., Sayle M.A., Newell B.R. (2013). The expression and interpretation of uncertain forensic science evidence: Verbal equivalence, evidence strength, and the weak evidence effect. *Law & Human Behavior*, 37(3), 197–207. https://doi.org/10.1037/lhb0000027

[45] Martire K.A., Kemp R.I., Sayle M., Newell B.R. (2014). On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect. *Forensic Science International*, 240, 61–68. https://doi.org/10.1016/j.forsciint.2014.04.005

[46] Thompson W.C., Newman E.J. (2015). Lay understanding of forensic statistics: Evaluation of random match probabilities, likelihood ratios, and verbal equivalents. *Law and Human Behavior*, 39(4), 332–349. https://doi.org/10.1037/lhb0000134

[47] Bayer D., Neumann C., Ranadive A. (2016). Communication of statistically based conclusions to jurors–A pilot study. *Journal of Forensic Identification*, 66(5), 405–427.

[48] Thompson W.C., Grady R.H., Lai E., Stern H.S. (2018). Perceived strength of forensic scientists' reporting statements about source conclusions. *Law, Probability and Risk*, 17(2), 133–155. https://doi.org/10.1093/lpr/mgy012

[49] Ribeiro G., Tangen J., McKimmie B. (2020). Does DNA evidence in the form of a likelihood ratio affect perceivers' sensitivity to the strength of a suspect's alibi? *Psychonomic Bulletin and Review*, 27, 1325–1332. https://doi.org/10.3758/s13423-020-01784-x

[50] van Straalen E.K., de Poot C.J., Malsch M., Elffers H. (2020). The interpretation of forensic conclusions by criminal justice professionals: The same evidence interpreted differently. *Forensic Science International*, 313, 110331. https://doi.org/10.1016/j.forsciint.2020.110331

[51] Bali A.S., Martire K.A., Edmond G. (2021). Lay comprehension of statistical evidence: A novel measurement approach. *Law & Human Behavior*, 45, 370–390. https://doi.org/10.1037/lhb0000457

[52] Ribeiro G., McKimmie B.M., Tangen J.M. (2023). Diagnostic information produces better-calibrated judgments about forensic comparison evidence than likelihood ratios. *Journal of Applied Research in Memory and Cognition*, 12, 412–420. https://doi.org/10.1037/mac0000062

[53] van Straalen E.K., de Poot C.J., Malsch M., Elffers H. (2023). The interpretation of forensic conclusions by professionals and students: Does experience matter? *Forensic Science International: Synergy*, 100437. https://doi.org/10.1016/j.fsisyn.2023.100437

[54] Bali A.S., Martire K.A. (2025). Exploring mock juror evaluations of forensic evidence conclusion formats within a complete expert report. *Forensic Science International: Synergy*, 10, 100564. https://doi.org/10.1016/j.fsisyn.2024.100564

[55] Thompson W.C., Grady R.H., Morrison G.S. (2025). Does explaining the meaning of likelihood ratios improve lay understanding? Manuscript submitted for publication. Preprint at https://osf.io/c28rt/

[56] Basu N., Bali A.S., Weber P., Rosas-Aguilar C., Edmond G., Martire K.A., Morrison G.S. (2022). Speaker identification in courtroom contexts – Part I: Individual listeners compared to forensic voice comparison based on automatic-speaker-recognition technology. *Forensic Science International*, 341, 111499. https://doi.org/10.1016/j.forsciint.2022.111499

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

[58] Koehler J.J., Chia A., Lindsey S. (1995). The random match probability in DNA evidence: Irrelevant and prejudicial. *Jurimetrics*, 35(2), 201–219. https://www.jstor.org/stable/29762371

[59] Jackson G., Evett I.W., Champod C., Buckleton J. (2014). Letter to the editor. *Science and Justice*, 54, 180. https://doi.org/10.1016/j.scijus.2014.02.001

[60] ISO 21043-5:2025 *Forensic Sciences – Part 5: Reporting*. Geneva, Switzerland: International Organization for Standardization. https://www.iso.org/obp/ui/en/#iso:std:iso:21043:-5:ed-1:v1:en

[61] Garrett B.L., Crozier W.E., Grady R. (2020). Error rates, likelihood ratios, and jury evaluation of forensic evidence. *Journal of Forensic Sciences*, 65, 1199–1209. https://doi.org/10.1111/1556-4029.14323

[62] Hollerman G., Hooge I., Kemner C., Hessels R.S. (2020). The 'real-world approach' and its problems: A critique of the term ecological validity. *Frontiers of Psychology*, 11. https://doi.org/10.3389/fpsyg.2020.00721

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

[64] Morrison G.S. (2024). Bi-Gaussianized calibration of likelihood ratios. *Law, Probability & Risk*, 23, mgae004. https://doi.org/10.1093/lpr/mgae004

[65] Bali A.S., Basu N., Weber P., Rosas-Aguilar C., Edmond G., Martire K.A., Morrison G.S. (2024). Speaker identification in courtroom contexts – Part III: Groups of collaborating listeners compared to forensic voice comparison based on automatic-speaker-recognition technology. *Forensic Science International*, 360, 112048. https://doi.org/10.1016/j.forsciint.2024.112048