Does explaining the meaning of likelihood ratios improve lay understanding?

William C Thompson [a]

Rebecca Hofstein Grady [a,*]

Geoffrey Stewart Morrison [b,c,†]

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

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

[c] Forensic Evaluation Ltd, Birmingham, UK

[*] Now at: California Digital Library, Office of the President, University of California, Oakland CA, USA

[†] Corresponding author: G.S. Morrison, e-mail: geoff-morrison@forensic-evaluation.net. The experiments reported in this paper were prepared and conducted, and the data were collected, 2013–2015. The analysis of the data and the writing of the paper took place in 2025. Morrison's contributions in 2013 were made while he was affiliated with the Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, NSW, Australia.

Abstract

Most previous research exploring the understanding of likelihood ratios by laypersons has not provided participants with an explanation of the meaning of likelihood ratios. Although triers of fact in common-law jurisdictions usually receive oral testimony, most previous research has presented experiments in written format. The research reported in this paper presented participants with videoed testimony and tested the effect of the expert witness providing participants with an explanation of the meaning of likelihood ratios. Analysis included comparing each participant's effective likelihood ratio (the posterior odds elicited from the participant divided by the prior odd elicited from the participant) with the presented likelihood ratio. The percentage of participants whose effective likelihood ratios equalled the presented likelihood ratios was higher for participants who were provided with the explanation of the meaning of likelihood ratios than for participants who were not provided with the explanation. The difference was, however, small. The percentage of participants whose posterior odds were consistent with them having committed the prosecutor's fallacy was not lower for participants who were provided with the explanation of the meaning of likelihood ratios than for participants who were not provided with the explanation. The full set of results do not constitute convincing evidence that presenting the explanation of the meaning of likelihood ratios resulted in better understanding of likelihood ratios. We discuss whether there are factors other than participants not understanding the meaning of likelihood ratios that could have contributed to the results.

Keywords

Explanation; Forensic science; Likelihood ratio; Prosecutor's fallacy; 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] and in Morrison et al. [2] with 31 and 57 signatories 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 [3]); Royal Statistical Society (Aitken et al. [4]); European Network of Forensic Science Institutes (Willis et al. [5]); National Institute of Forensic Science of the Australia New Zealand Policing Advisory Agency (Ballantyne et al. [6]); American Statistical Association (Kafadar et al. [7]); and Forensic Science Regulator for England & Wales [8]. There is, however, a common belief that likelihood ratios are difficult for legal-decision makers to understand (Bali et al. [9], Swofford et al. [10]). The benefits of forensic practitioners adopting the likelihood-ratio framework will not be fully realized if practitioners are unable to communicate the meaning of likelihood ratios to legal-decision makers.

Martire [11], Thompson [12], Eldridge [13], and Martire & Edmond [14] reviewed empirical research on laypersons' understanding of forensic practitioners' expressions of strength of forensic evidence, including likelihood ratios.[1] Thompson [12] concluded that:

> the reporting formats that are easiest for lay people to understand are difficult to justify logically and empirically, while reporting formats that are easier to justify logically and empirically are more difficult for lay people to understand.

[1] With additional coauthors, we have conducted a review specifically with respect to layperson understanding of likelihood ratios, as opposed to also including layperson understanding of other expressions of strength of evidence such as categorical conclusions or posterior probabilities. In addition to reviewing other empirical-research studies, that review (Morrison et al. [15]) includes review of the study presented in the present paper.

Eldridge [13] 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 …

Martire & Edmond [14] concluded:

> Ultimately, the question of whether lay decision-makers appropriately comprehend statistical statements from forensic scientists is an important one, but its answer remains elusive.

In almost all the studies reviewed, however, experiments did not provide participants with an explanation of the meaning of likelihood ratios. Since they were given no explanation of the meaning of likelihood ratios, we think it unsurprising that participants had difficulty understanding likelihood ratios. We hypothesize that providing an explanation of the meaning of likelihood ratios will help laypersons (such as research participants and legal-decision makers) better understand likelihood ratios. We think it a reasonable expectation for forensic practitioners acting as expert witnesses to be asked to explain the meaning of likelihood ratios during examination in chief.

In almost all the studies reviewed, experiments were presented to participants in written format. In common-law jurisdictions, triers of fact usually receive forensic-science testimony not in written format but in oral format.

In this paper, we report on a study that provided participants with videoed testimony,[2] and that tested the effect of providing participants with a detailed explanation of the

[2] We also conducted experiments using written versions of testimony, but only in conditions that did not include the explanation of the meaning of likelihood ratios. For brevity, have not reported those experiments in the present paper.

meaning of likelihood ratios. The videoed testimony related to forensic voice comparison. In addition to testing conditions in which participants were provided with an explanation of the meaning of likelihood ratios versus not provided with the explanation, we also tested conditions designed to elicit low prior odds versus high prior odds, and conditions in which a low likelihood ratio was presented versus a high likelihood ratio was presented.

Martire [11], Martire & Edmond [14], and Bali et al. [16] developed five indicators of whether participants in research studies understood the meaning of expressions of strength of evidence, including whether they understood the meaning of likelihood ratios. In this paper, we make use of three of those indicators. The following are the definitions of the three indicators as they appear in Martire & Edmond [14]:

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

- "*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."

- "*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 [17]), directional errors (e.g., Martire et al. [18]), and aggregation errors (e.g., Koehler et al. [19])."[3]

When we use the words *sensitivity*, *orthodoxy*, and *coherence* with the definitions above, we will set the words in italics. The only particular type of *incoherence* we examine in the present paper is the prosecutor's fallacy. Note that *sensitivity*, *orthodoxy*, and *coherence* are not mutually exclusive. *Orthodox* responses would

[3] This definition of *coherence* differs from the meaning of "coherence" in statistics.

necessarily be *sensitive*, but *sensitive* responses would not necessarily be *orthodox*. *Orthodox* responses would necessarily be *coherent*; however, *unorthodox* responses would not necessarily be *incoherent*. If they were due to a particular fallacy or error in reasoning they would be *incoherent*, but if they were instead apparently random or idiosyncratic they would not be *incoherent*.

The data used in this paper, and the MATLAB and R scripts used to analyze the data, are available at https://forensic-data-science.net/communication/ and at https://osf.io/c28rt/.

2 Methodology

2.1 Conditions tested

A 2×2×2 factorial of conditions was tested:

- **Prior odds**: Participants were given written case information that in one condition was designed to elicit lower prior odds (smaller $p(H_1)/p(H_2)$) and in the other condition was designed to elicit higher prior odds (larger $p(H_1)/p(H_2)$). We adopt the convention that H_1 refers to a same-speaker hypothesis and H_2 refers to a different-speaker hypothesis.

- **Explanation of the meaning of likelihood ratios**: All participants watched 11 minutes of videoed testimony. Participants in one condition also watched an additional 9 minutes of videoed testimony during which the expert witness gave a detailed explanation of the meaning of likelihood ratios. Participants in the other condition were not provided with this explanation.

- **Likelihood-ratio value**: The final part of the videoed testimony presented participants with a numerical likelihood ratio ($\Lambda = p(E|H_1)/p(E|H_2)$) of either 30 or 3,000.

2.2 Participants

Ethics approval was obtained from the UC Irvine Institutional Review Board.

Participants were recruited using Amazon Mechanical Turk (mTurk), and the experiments were conducted using the Qualtrics online survey service.

Recruitment from mTurk has been widely used in social science (Buhrmester et al. [20]), and can offer cost-effective and high quality data (Paolacci & Chandler [21]), especially when restricted to workers from the United States (Smith et al. [22]). Thompson & Newman [23] presented a detailed demographic breakdown of participants recruited from mTurk using the same methods as employed in this study, and concluded that this population is sufficiently diverse and representative of American jurors to be suitable for a study of lay reactions to forensic science evidence.

To help ensure that our participants were representative of the jury-eligible population of the United States, we asked participants to affirm that they were US citizens over the age of 18, and the recruitment procedure included steps to exclude robotic responders and participants from outside the United States. Participants were also excluded if they had participated in a prior study run by the UC Irvine lab. A participant's responses were excluded if the participant failed to correctly answer attention-check questions, did not complete the experiment, or gave logically inconsistent answers.[4]

[4] A participant's responses were excluded if their odds responses with respect to source level were reversed relative to their percentage responses with respect to source level or relative to their responses with respect to offence level (percentage responses with respect to the source level and responses with respect to offence level are not reported in this paper). A participant's responses were also excluded if their odds responses included "%", or if their odds responses were 0. Some

After exclusions, the total number of participants was 571. The number of participants in each cell of the 2×2×2 factorial of conditions was as shown in Table 1.

Table 1. After exclusions, the number of participants in each cell of the 2×2×2 factorial of conditions.

explanation of likelihood ratios provided	presented $\Lambda = 30$		presented $\Lambda = 3,000$		total
	low prior	high prior	low prior	high prior	
yes	62	74	66	62	264
no	67	78	81	81	307

2.3 Background information about the case

2.3.1 General

Participants first read background information about the case.

The case was set in southern California. It involved credit-card fraud committed by a man with an Australian accent who made a call from a public phone in a bar. The call was recorded at the credit-card company's call centre. The questioned-speaker recording was, therefore, a recording of a telephone call made by an adult male with an Australian accent.

2.3.2 Prior-odds conditions

The written background information about the case exposed each participant to either the low-prior-odds condition or the high-prior-odds condition:

participants' odds responses were values of 95 and 98, which look suspiciously like they might have been percentages, but, since they did not include "%", we did not exclude them. More participants dropped out during the longer condition that included the explanation of the meaning of likelihood ratios than during the shorter condition that did not include the explanation.

- **Low-prior-odds condition**: In the condition intended to elicit lower prior odds, the call was made from an Australian-themed bar that was popular with visiting and resident Australians. At the time the call was made, the bar was crowded with people watching a rugby game between Australia and New Zealand. A police officer arrived at the bar two hours after the call was made, and the bartender identified a man in the bar who she thought might have made the call but she was uncertain.

- **High-prior-odds condition**: In the condition intended to elicit higher prior odds, the call was made from a Louisiana-themed bar that was popular with local residents. At the time the call was made, the bar was crowded with people watching an American-football game. A police officer arrived at the bar 15 minutes after the call was made, and the bartender identified a man in the bar who she was quite certain had made the call.

Both sets of case information included additional detail.

In both conditions, the man identified by the bartender (hereinafter "the suspect") agreed to accompany the police officer to the police station where he was interviewed and an audio recording of the interview was made. This was the known-speaker recording. The suspect spoke English with an Australian accent. A detective who listened to the recording of the questioned speaker and to the recording of the known speaker thought that the voices were sufficiently similar that it was worth submitting them for forensic comparison.

2.4 Elicitation of prior odds

After participants had read the background information about the case, prior odds were elicited from the participants in a three-stage process.

In the first stage, participants were asked the following question with the following answer options:

1. Based on the facts presented so far, do you think it is more likely that the man who called the card activation center was the suspect or someone else?

    1.1. More likely to have been the suspect than someone else.

    1.2. About equally likely to have been the suspect or someone else (about 50% chance it was the suspect).

    1.3. More likely to have been someone else than the suspect.

If a participant responded "equally likely" at the first stage, a prior-odds value of 1 was recorded and the participant did not proceed to the second and third stages.

In the second stage, participants were asked the following question with the following answer options (with the choice of "the suspect" or "someone else" depending on their answer to the first-stage question):

2. You said it was more likely that the caller was {the suspect than someone else, someone else than the suspect}. How much more likely?

    2.1. Between 1 and 10 times more likely (51%–91% chance it was {the suspect, someone else})

    2.2. Between 10 and 99 times more likely (91%–99% chance it was {the suspect, someone else})

    2.3. Between 100 and 999 times more likely (99%–99.90% chance it was {the suspect, someone else})

    2.4. Between 1,000 and 9,999 times more likely (99.90%–99.99% chance it was {the suspect, someone else})

    2.5. Between 10,000 and 99,999 times more likely (99.99%–99.999% chance it was {the suspect, someone else})

2.6. More than 100,000 times more likely (More than 99.9999% chance it was {the suspect, someone else})

In the third stage, participants were asked to respond to the following question:

3. Please use the text box below to fill in the blank with an exact number reflecting your views.

3.1. I think it is ___ times more likely that the caller was the {suspect than someone else, someone else than the suspect}.

For each participant, the prior-odds value elicited at the third stage (or at the first stage if the participant responded "equally likely" at that stage) is the value we used for subsequent analysis.

2.5 Videoed testimony

2.5.1 General

After prior odds had been elicited from the participants, they watched the videoed testimony.

In the videoed testimony, the last author of this paper played the part of an expert witness who presented testimony about forensic comparison of voice recordings, and the first author played the part of a lawyer who questioned the expert witness during examination in chief. The last author is in fact a forensic practitioner who has testified in court about forensic voice comparison, and the first author is in fact a lawyer who has examined expert witnesses in court. The expert witness and lawyer prepared for the testimony, but the questions and answers were not scripted.

The videoed testimony can be accessed at the following links:

- $\Lambda = 30$, explanation not provided: https://youtu.be/dhTUmS9_T24
- $\Lambda = 3{,}000$, explanation not provided: https://youtu.be/RM6jpRyzNiY

- $\Lambda = 30$, explanation provided: https://youtu.be/WPaS-j8SaOg
- $\Lambda = 3{,}000$, explanation provided: https://youtu.be/4VFMaJE3CPA

2.5.2 Witness qualifications and methods used

The videoed testimony included:

1. The expert witness's qualifications and experience.

    1.1. Details included that he was Director of the Forensic Voice Comparison Laboratory at the University of New South Wales in Australia. He conducted research and had published papers on forensic voice comparison. He performed casework, working about half the time under instruction from the prosecution and half the time under instruction from the defence.

2. The methods the expert witness used to evaluate the strength of evidence associated with his comparison of the questioned-speaker and known-speaker recordings.

    2.1. The expert witness made quantitative measurements of the acoustic properties of the recordings and used them to build statistical models of the two voices that were used to calculate a likelihood ratio which quantified his strength-of-evidence conclusion.

    2.2. He explained that the acoustic measurements he used were standard measurements used in speech processing and speaker recognition systems, and that they measure the frequency components of speech. He explained that these acoustic properties vary among individuals for several reasons, for example, individuals vary physiologically in the size of their vocal tracts, but they also vary in the way they have learned to speak.

    2.3. He explained that the statistical models allowed him to calculate two

probabilities: First he calculated the probability of getting the acoustical properties observed in the questioned-speaker recording if the questioned speaker were the known speaker. Second, he calculated the probability of getting the acoustical properties observed in the questioned-speaker recording if the questioned speaker were some other speaker in the relevant population. The first probability measured how *similar* the questioned speaker's voice was to the known speaker's voice. The second probability measured how *typical* the questioned speaker's voice was with respect to speakers in the relevant population. The relevant population adopted for this case was adult male speakers of Australian English.[5] The expert witness stated that the strength of evidence depends partly on how similar the voices on the questioned-speaker and known-speaker recordings are, but it also depends on how typical they are because typical voices are more likely to be similar by chance.

2.4. He explained that the statistical models do not give yes/no answers, they give probabilities, and that he used the statistical models to calculate a *likelihood ratio* which indicates how much more likely one would be to observe the acoustic properties found in the questioned-speaker recording if the questioned speaker were the known speaker than if the questioned speaker were some other adult male speaker of Australian English.

2.5.3 Explanation-of-the-meaning-of-likelihood-ratios conditions

Participants in one condition then watched additional testimony about the meaning of likelihood ratios, as summarized below. Participants in the other condition skipped this.

The expert witness explained that the likelihood ratio represented the extent to which

[5] The expert witness explained that the data he used to calculate the likelihood ratio consisted of recordings of adult male Australian English speakers, excluding speakers who sounded obviously different from the questioned speaker.

one should change one's beliefs about the relative probabilities of the same-speaker versus different-speaker hypotheses from before the evidence was presented to after the evidence was presented. To help illustrate this, he showed pictures of sets of scales with different numbers of weights on each side of the scales.

The first set of scales had 1 weight on the "different-speaker" side and 1 weight on the "same-speaker" side. He explained that this represented an example of possible relative degrees of belief in the hypotheses before the evidence was presented, and that is was purely an example and that jury members would have whatever relative degrees of belief they might have, that this was just one concrete example and was not intended to suggest that these were the numbers that jury members should use.

The expert witness used 4 as an example likelihood-ratio value. He stated that this was not the value calculated in the case, but was a number used as a concrete example. He explained that if the likelihood ratio was that "the evidence is 4 times more likely if the same-speaker hypothesis were true than if the different-speaker hypothesis were true" then one would multiply the weight on the same-speaker side of the scales by 4, resulting in an updated set of scales with 1 weight on the "different-speaker" side and 4 weights on the "same-speaker" side, and that this represented what one's relative degrees of belief in the hypotheses should logically be after considering the evidence:

> If before hearing the evidence one believed that the probabilities of the two hypotheses were equal, then, logically, after hearing the likelihood ratio expressing the strength of the evidence one should believe that the same-speaker hypothesis is 4 times more probable than the different-speaker hypothesis.

He then talked through additional examples with different prior odds (different-speaker to same-speaker ratios of 1:2, 2:1, and 8:1) showing that in each case what was done was exactly the same, the number of weights on the same-speaker size of the scales was multiplied by the likelihood ratio of 4 to arrive at an updated set of weights on the scales representing posterior odds (different-speaker to same-speaker ratios of 1:8, 2:4

equal to 1:2, and 8:4 equal to 2:1).

2.5.4 Likelihood-ratio-value conditions

Finally, the expert witness presented his conclusion, which included a likelihood ratio of either 30 or 3,000:

> On the basis of my calculations, one would be {30, 3,000} times more likely to get the acoustic properties on the questioned-speaker recording if it were a recording of the known speaker than if it were a recording of some other adult male speaker of Australian English.

2.6 Elicitation of posterior odds

After participants had watched the videoed testimony, posterior odds were elicited from them using the same three-stage process as had been used to elicit prior odds, see §2.4 above.

2.7 Elicitation of perceived quality of the testimony

In addition to responding to questions about prior odds and posterior odds, participants also responded to other questions including attention check questions, demographic questions, and questions about the quality of the testimony.

Each question about the quality of the testimony asked participants to respond using a 5-level Likert scale. Participants were asked to give their judgements about:

1. whether the expert witness was qualified
2. whether the expert witness was credible
3. whether the expert witness was trustworthy
4. whether the expert witness was biased
5. whether the methods used by the expert witness were valid

As a measure of each participant's perceived quality of the testimony, we used the mean value of their responses to the five questions, with 1 indicating lowest perceived quality and 5 indicating highest perceived quality.[6]

3 Results and Discussion

3.1 Prior odds

Figure 1 shows violin plots of the logarithms of prior-odds responses elicited in each of the prior-odds conditions. On average, participants in the high-prior-odds condition gave higher prior-odds responses than did participants in the low-prior-odds condition. This manipulation therefore appears to have been successful.[7]

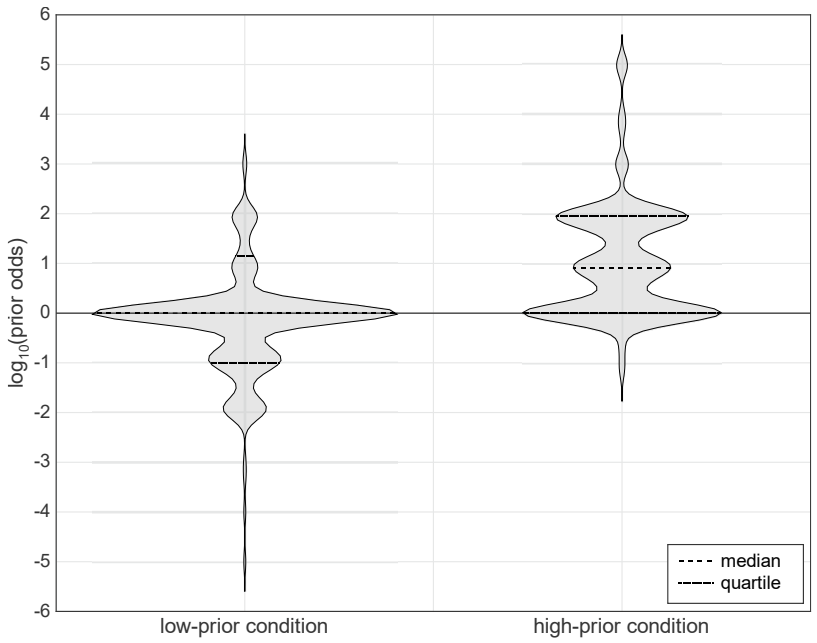

Figure 1. Prior-odds responses given each prior-odds condition.

[6] Prior to calculating the mean, we reversed the coding for the responses to the question about bias.

[7] Among the interdisciplinary group of authors, we initially had different perspectives on when to only provide descriptive statistics (as numbers or in graphical representations) and when to report the results of formal statistical tests. We came to a consensus in which we report (in §3.4 and §3.5) the results of formal statistical tests related to the core research question and report only descriptive statistics elsewhere.

Elicited prior-odds values tended to be at factors of ten (1/100, 1/10, 1, 10, 100, etc.), and in both conditions, the modal response was prior odds of 1 (log prior odds of 0). It is unknown whether the latter result would have differed had participants been asked to immediately give prior odds rather than going through the three-stage process. The first stage of the elicitation process may have encouraged prior odds responses of 1, and the second stage may have encouraged responses at factors of ten; however, people in general may have a natural proclivity for choosing equal priors, and, after that, for choosing factors of ten.

For both prior-odds conditions, the elicited prior odds were high – in cases assuming large populations, one might expect prior odds to be 1 in thousands or 1 in millions in favour of the different-source hypothesis. In both conditions, however, the case information was extensive, suggesting high odds in favour of the same-speaker hypothesis.

3.2 Effective likelihood ratios

If a participant correctly applied Bayes' theorem to the likelihood-ratio value that was presented to them, their posterior odds would equal their prior odds multiplied by the likelihood-ratio value that was presented, see Equation (1), in which E is the acoustic properties of the voice recordings. For each participant, their effective likelihood-ratio value, Λ, was calculated by dividing the posterior odds elicited from them by the prior odds elicited from them, see Equation (2).[8]

[8] The expert witness explicitly adopted H_1 = "the speaker on the questioned-speaker recording is the suspect", versus H_2 = "the speaker on the questioned-speaker recording is some other adult male Australian English speaker"; however, the participants were asked to give prior odds and posterior odds for H_1 = "the speaker on the questioned-speaker recording is the suspect", versus H_2 = "the

(1) $$\frac{p(H_1|E)}{p(H_2|E)} = \frac{p(H_1)}{p(H_2)} \times \frac{p(E|H_1)}{p(E|H_2)}$$

(2) $$\Lambda = \frac{p(E|H_1)}{p(E|H_2)} = \frac{\left(\frac{p(H_1|E)}{p(H_2|E)}\right)}{\left(\frac{p(H_1)}{p(H_2)}\right)}$$

Figure 2 shows violin plots of the distributions of the effective log-likelihood-ratio

speaker on the questioned-speaker recording is someone else". There therefore appears to be a mismatch in the relevant population specified as part of H_2. Before the prior odds were elicited, however, the written case scenarios had already established that both the questioned speaker and the suspect were adult male Australian English speakers. If we begin with H_2 = "some other speaker in the bar", then split the total evidence into E_a = "both the questioned speaker and the suspect are adult male Australian English speakers", which was presented in the written case scenarios, and E_b = "the acoustic properties of the voices on the questioned-speaker recording and on the known-speaker recording", which was the evidence evaluated by the expert witness, then, applying the chain rule, we get:

$$\frac{p(H_1|E_\mathrm{a}, E_\mathrm{b})}{p(H_2|E_\mathrm{a}, E_\mathrm{b})} = \frac{p(H_1|E_\mathrm{a})}{p(H_2|E_\mathrm{a})} \times \frac{p(E_\mathrm{b}|H_1, E_\mathrm{a})}{p(E_\mathrm{b}|H_2, E_\mathrm{a})}$$

$$\frac{p(H_1|E_\mathrm{a})}{p(H_2|E_\mathrm{a})} = \frac{p(H_1)}{p(H_2)} \times \frac{p(E_\mathrm{a}|H_1)}{p(E_\mathrm{a}|H_2)}$$

In which $\frac{p(E_\mathrm{b}|H_1, E_\mathrm{a})}{p(E_\mathrm{b}|H_2, E_\mathrm{a})}$ is the likelihood ratio presented by the expert witness (which includes to the right of the conditioning bar that both the questioned speaker and the suspect are adult male Australian English speakers), and $\frac{p(H_1|E_\mathrm{a})}{p(H_2|E_\mathrm{a})}$ (which also includes to the right of the conditioning bar that both the questioned speaker and the suspect are adult male Australian English speakers) is a participant's prior odds immediately before being presented with the expert witness's likelihood ratio. The effective likelihood would then be:

$$\frac{p(E_\mathrm{b}|H_1, E_\mathrm{a})}{p(E_\mathrm{b}|H_2, E_\mathrm{a})} = \frac{\left(\frac{p(H_1|E_\mathrm{a}, E_\mathrm{b})}{p(H_2|E_\mathrm{a}, E_\mathrm{b})}\right)}{\left(\frac{p(H_1|E_\mathrm{a})}{p(H_2|E_\mathrm{a})}\right)}$$

values given each combination of conditions.[9]

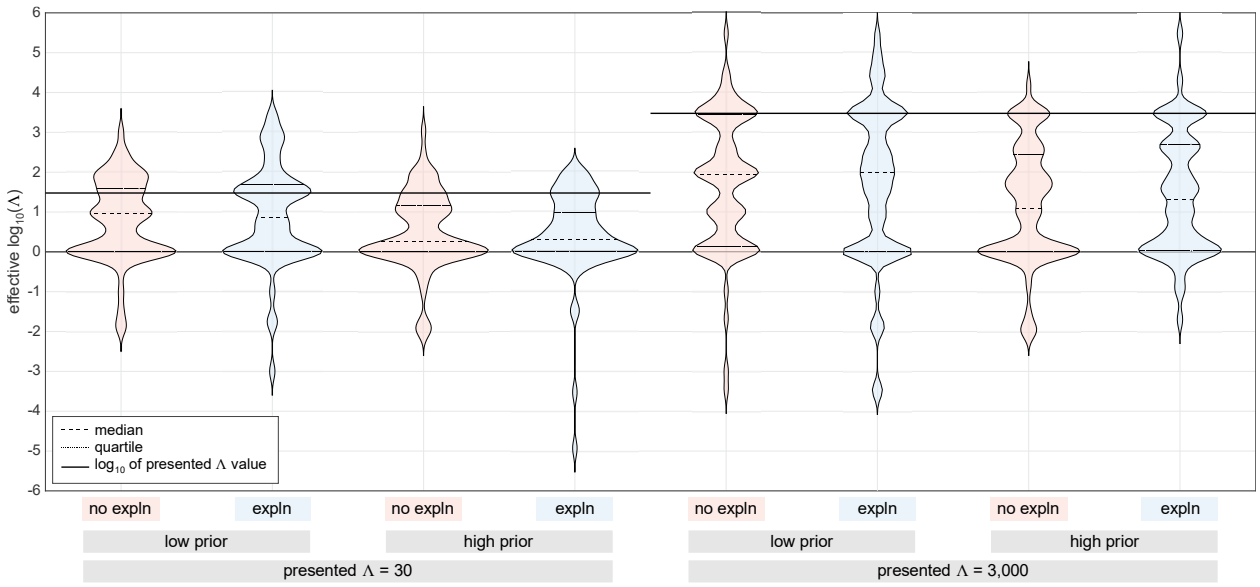

Figure 2. Effective log likelihood ratios given each combination of conditions.

3.3 Presented likelihood ratios and effective likelihood ratios

As shown in Figure 2, participants' judgements were *sensitive* to the value of the presented likelihood ratio. Effective log-likelihood-ratio values tended to be higher when the presented likelihood ratio was 3,000 than when it was 30.

- Some participants who were presented with a likelihood ratio of 3,000 had effective log-likelihood-ratio values that were at or close to 0, but they constituted a lower proportion than for participants who were presented with a likelihood ratio of 30.

- Some participants who were presented with a likelihood ratio of 3,000 had effective log-likelihood-ratio values that were higher than any observed for participants who were presented with a likelihood ratio of 30.

[9] Not shown, for the fourth violin plot from the right in Figure 2 and for the second panel from the right in the top row of Figure 3, is an outlier with an effective likelihood-ratio value of 40 billion.

In Figure 2, for all combinations of conditions (other than for a presented likelihood ratio of 30 and no explanation of the meaning of likelihood ratios):

- There were bulges in the violin plots around an effective $\log_{10}$-likelihood-ratio value of about 1.5 for participants presented with a likelihood ratio of 30, which corresponds to a $\log_{10}$-likelihood-ratio value of about 1.5.

- There were bulges in the violin plots around an effective $\log_{10}$-likelihood-ratio value of about 3.5 for participants presented with a likelihood ratio of 3,000, which corresponds to a $\log_{10}$-likelihood-ratio value of about 3.5.

Although they constituted a relatively small proportion of participants, these participants' effective likelihood-ratio values were the same as the presented likelihood ratio values, i.e., these results appeared to be *orthodox*, they were consistent with participants correctly applying Bayes' theorem to update their prior odds to posteriors odds given the likelihood-ratio value that was presented to them.

3.4 Explanation of the meaning of likelihood ratios and effective likelihood ratios

As shown in Figure 2, participants' judgements were *sensitive* to the value of the presented likelihood ratio both when the explanation of the meaning of likelihood ratios was provided and when it was not. If *sensitivity* is used an indicator of understanding, this suggests that participants understood the meaning of the likelihood ratios without the need for an explanation. *Sensitivity* is, however, a low bar, and we do not consider *sensitivity* alone to be a sufficient criterion for demonstrating understanding of the meaning of likelihood ratios. *Orthodoxy* constitutes a higher bar.

Turning to *orthodoxy*, Table 2 shows, for participants in each combination of conditions, the percentage of participants whose responses were as expected if they had correctly applied Bayes' theorem to the likelihood-ratio value that was presented to them. This was calculated as the percentage of effective likelihood-ratio values that were within ±10% of the presented likelihood ratio. Some participants had prior

probabilities of 1, so their posterior odds could have been due to the response being *orthodox* or due to them committing the prosecutor's fallacy, i.e., due to them transposing the conditional: If the prior odds are 1, then, according to Bayes' theorem, the posterior odds will equal the likelihood ratio. Table 2 presents results both excluding and including responses from participants whose prior odds were 1.

Table 2. For each combination of conditions, the percentage of participants whose responses were as expected if they had correctly applied Bayes' theorem to the likelihood ratio presented to them. In each cell, the first value, outside square brackets, excludes participants whose prior odds were 1, and the second value [inside square brackets] includes participants whose prior odds were 1.

explanation of likelihood ratios provided	**presented $\Lambda = 30$**		**presented $\Lambda = 3,000$**		**across all Λ and prior conditions**
	low prior	**high prior**	**low prior**	**high prior**	
yes	2.0 [19]	2.9 [9.5]	1.8 [15]	5.5 [16]	3.0 [15]
no	0 [10]	1.4 [10]	0 [15]	1.4 [12]	0.74 [12]

From Table 2 it is apparent that almost all effective likelihood-ratio values that appeared to be *orthodox* came from participants whose prior odds were 1. We cannot, therefore, be certain whether these participants correctly understood the meaning of the likelihood ratio presented to them or whether they committed the prosecutor's fallacy.[10] In order to avoid this problem in future research, we recommend using case information that is likely to elicit prior odds other than 1, e.g., information that is likely to elicit prior odds substantially less than 1.

We calculated a Bayes factor for the following hypotheses:

- H_+: The probability that a participant's effective likelihood ratio equals the

[10] An anonymous reviewer pointed out that it is also possible that some instances of responses that appeared to be *orthodox* could have been coincidental.

presented likelihood ratio is larger for participants provided with the explanation of the meaning of likelihood ratios than for participants not provided with the explanation.

- H_0: The probability that a participant's effective likelihood ratio equals the presented likelihood ratio is the same for participants provided with the explanation of the meaning of likelihood ratios as for participants not provided with the explanation.

We calculated the Bayes factor using the logit-transformation method for the A/B test, as described in Kass & Vaidyanathan [24], Gronau et al. [25], Dablander et al. [26], and Hoffmann et al. [27], and implemented in Gronau [28]. This method makes use of binomial models with logit transformations for the prior parameters. See Appendix A for details.

Excluding results from participants whose effective likelihood-ratio values were as expected if they had correctly applied Bayes' theorem but whose prior odds were 1:

- Across conditions, the number of participants whose effective likelihood-ratio values equalled the likelihood-ratio value presented to them was 7 out of 232 for those provided with the explanation ($y_{\mathrm{B}} = 7, n_{\mathrm{B}} = 232$) versus 2 out of 272 for those not provided with the explanation ($y_{\mathrm{A}} = 2, n_{\mathrm{A}} = 272$).

- The Bayes-factor value was 2.9, i.e., the count data were approximately 3 times more likely if H_+ were true than if H_0 were true.

Our prior probabilistic belief that providing an explanation would be effective was already relatively high, and this Bayes factor makes our posterior probabilistic belief somewhat higher (the odds have increased by a factor of 3). Others, however may have begun with a lower prior probabilistic belief, and (even after increasing their odds by a factor of 3) their posterior probabilistic belief may still be relatively low. Also, even if providing the explanation did have an effect, the resulting 3% of participants whose

effective likelihood-ratio values were as expected if they had correctly applied Bayes' theorem was not impressive. These results do not, therefore, constitute convincing evidence that presenting the explanation of the meaning of likelihood ratios increased understanding.

In the analysis above, we treated *orthodoxy* as categorical – we treated a participant's response as either *orthodox* or not *orthodox*. An anonymous reviewer requested that we also provide the average distance between the effective $\log_{10}$-likelihood-ratio values and the $\log_{10}$ of the presented likelihood ratios for participants provided with the explanation of the meaning of likelihood ratios versus for participants not provided with the explanation. In Table 3 we provide the median signed difference and the median unsigned distance for each combination of conditions. Since the data were not symmetrically distributed about a single mode, we provide medians rather than means. For the high-prior condition, for both presented likelihood-ratio values, both the differences and the distances were smaller in magnitude when the explanation was provided than when it was not provided (the expected result if providing the explanation was effective). This was also the case for the difference in the low-prior condition when the presented likelihood-ratio value was 3,000, but was not the case for the distance. For the low prior condition when the presented likelihood-ratio value was 30, the magnitude of both the difference and the distance were lower when the explanation was not provided than when it was provided (the opposite of the expected result if providing the explanation was effective). We think, however, that the proportion of effective likelihood ratios that have the same value as the presented likelihood ratio is more relevant for assessing *orthodoxy* than the average difference or distance between effective likelihood ratios and the presented likelihood ratio, e.g., for the low prior condition when the presented likelihood-ratio value was 30, the proportion of effective likelihood ratios that were 30 was higher when the explanation was provided than when it was not provided (the expected result if providing the explanation was effective).

Table 3. For each combination of conditions, the median signed difference between the effective $\log_{10}$-likelihood-ratio values and the $\log_{10}$ of the presented likelihood ratio, and [in square brackets] the median unsigned distance between the effective $\log_{10}$-likelihood-ratio values and the $\log_{10}$ of the presented likelihood ratio.

explanation of likelihood ratios provided	**presented $\Lambda = 30$**		**presented $\Lambda = 3,000$**	
	low prior	**high prior**	**low prior**	**high prior**
yes	−0.61 [1.14]	−1.18 [1.18]	−1.48 [1.65]	−2.16 [2.18]
no	−0.52 [1.00]	−1.23 [1.29]	−1.55 [1.55]	−2.38 [2.38]

3.5 Explanation of the meaning of likelihood ratios and the prosecutor's fallacy

Turning to *coherence*, if participants responded with posterior odds equal to the presented likelihood ratio, this could indicate that they had committed the prosecutor's fallacy. If a participant's prior odds were 1, then their posterior odds equalling the presented likelihood-ratio value could either be due to them having committed the prosecutor's fallacy or due to them having correctly applied Bayes' theorem to the presented likelihood ratio. For each combination of conditions, Table 4 shows the percentage of participants whose posterior odds were the same as the presented likelihood ratio, separated according to whether their prior odds were 1 or not.

Table 4. For each combination of conditions, the percentage of participants whose posterior odds exactly equaled the presented likelihood ratio. In each cell, the first value, outside square brackets, excludes participants whose prior odds were 1, and the second value [inside square brackets] includes participants whose prior odds were 1.

explanation of likelihood ratios provided	presented $\Lambda = 30$		presented $\Lambda = 3,000$		across all Λ and prior conditions
	low prior	high prior	low prior	high prior	
yes	9.8 [26]	12 [18]	12 [24]	20 [29]	13 [24]
no	12 [21]	18 [26]	7.2 [21]	30 [38]	17 [27]

Using the same procedures as described in §3.4 above, we calculated Bayes factors for the following hypotheses:

- H_{-}: The probability that a participant's posterior odds would equal the presented likelihood ratio is smaller for participants who were provided with the explanation of the meaning of likelihood ratios than for participants who were not provided with the explanation.

- H_0: The probability that a participant's posterior odds would equal the presented likelihood ratio is the same for participants who were provided with the explanation of the meaning of likelihood ratios as for participants who were not provided with the explanation.

Excluding results from participants whose posterior odds equalled the presented likelihood-ratio value but whose prior odds were 1:

- Across conditions, the number of participants whose posterior odds equalled the likelihood-ratio value presented to them was 31 out of 232 for those provided with the explanation of the meaning of likelihood ratios ($y_{\mathrm{B}} = 31, n_{\mathrm{B}} = 232$) versus 47 out of 272 for those not presented with the explanation ($y_{\mathrm{A}} = 47, n_{\mathrm{A}} = 272$).

- The Bayes-factor value was 0.83, i.e., the count data were approximately 1.2 times more likely if H_0 were true than if $H_{\_}$ were true.

The Bayes factor is so close to 1 that it has not meaningfully altered our probabilistic belief with respect to whether providing the explanation of the meaning of likelihood ratios would reduce the prevalence of the prosecutor's fallacy or not. The results do not support the hypothesis that providing the explanation would reduce the prevalence of the prosecutor's fallacy.

A potential contributing factor to participants in the explanation condition committing the prosecutor's fallacy could be that, in the experts witness's explanation, the first example used prior odds of 1. In this example, the posterior-odds value did equal the presented likelihood-ratio value. In future research we recommend testing conditions in which only explanations that do not use prior odds of 1 are presented to participants.

3.6 Perceived quality of testimony and effective likelihood ratios

Returning to *orthodoxy*, it is not necessarily a reasonable expectation that legal-decision makers directly use the likelihood-ratio value presented by an expert witness. Legal-decision makers may assign their own weight (their own strength of evidence) to the strength-of-evidence statement presented by the expert witness. In likelihood ratio terms, legal-decision makers may assign their own likelihood-ratio value in which, for them, the E of Equations (1) and (2) is not the acoustic properties of the voice recordings but the likelihood-ratio value presented by the expert witness, i.e., Equation (3), in which $\Lambda_{\mathbb{d}}$ is the legal-decision maker's (or participant's) likelihood ratio, $\Lambda_{\mathbb{w}}$ is the likelihood ratio presented by the expert witness, and $I_{\mathbb{d}}$ is other information (or other factors) that the legal-decision maker (or participant) has considered in assigning their likelihood-ratio value.

$$(3) \quad \Lambda_{\mathbb{d}} = \frac{p(\Lambda_{\mathbb{w}}|H_1, I_{\mathbb{d}})}{p(\Lambda_{\mathbb{w}}|H_2, I_{\mathbb{d}})}$$

As discussed in Gittelson et al. [29] and Martire & Edmond [14], rather than using the presented likelihood ratios ($\Lambda_{\mathbb{w}}$) directly, participants could have weighted them based on other factors ($I_{\mathbb{d}}$). Participants could have understood the meaning of likelihood ratios and correctly apply Bayes' theorem, but applied it to their weighted values, $\Lambda_{\mathbb{d}}$, rather than to the presented likelihood-ratio values, $\Lambda_{\mathbb{w}}$. Their effective likelihood ratios would then differ from the presented likelihood ratios, giving the false impression that they did not understand the meaning of likelihood ratios.

As reported in §3.33.3, very few participants had effective likelihood-ratio values that equalled the presented likelihood-ratio value. This result could, in part, have been due to participants who weighted the presented likelihood-ratio value based on other factors, such as how valid they perceived the testimony to be. This could be considered a rational thing to do and to not be indicative of misunderstanding the meaning of likelihood ratios.

In future research, we recommended that participants be asked what their posterior odds would be if they applied Bayes' theorem to the likelihood-ratio value presented, and also be asked what their posterior odds actually are, and, if the latter differs from the former, be asked why they made the latter different.

One factor which could have contributed to participants weighting the presented likelihood-ratio value, rather then using it directly, could be their perception of the quality of the testimony. We investigated the relationship between the participants' effective likelihood ratios and their perceived quality of the testimony. As described in §2.7, for each participant, their perceived quality of the testimony was calculated as the mean of their response values from five 5-level Likert scales. The higher the value the higher the perceived quality of the testimony.

Figure 3 shows, for each combination of conditions, individual participant's effective log-likelihood-ratio values plotted against their perceived quality of the testimony. If the effective log-likelihood-ratio values increased as the perceived quality of the

testimony increased, this would suggest that participants weighted the presented likelihood-ratio value based on how valid they perceived the testimony to be. Such a pattern of results was not, however, obvious in Figure 3. The obvious pattern was one of high inter-participant variability.

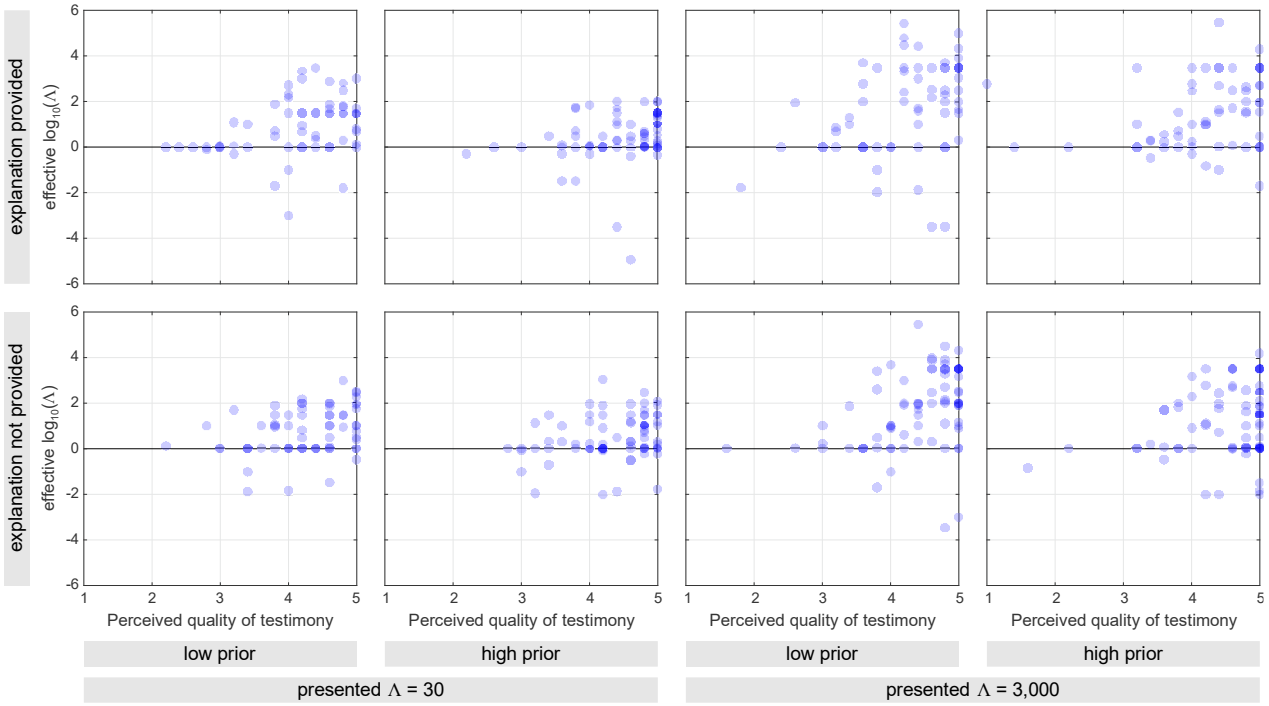

Figure 3. For each combination of conditions, individual participants' effective log-likelihood-ratio values plotted against their perceived quality of the testimony.

3.7 Caveat

The data were collected using an online platform with participants recruited from an online labour pool. Although this procedure has been widely used in the social sciences for studies of human decision making [20], it raises some concerns about data quality, such as whether participants were truly qualified and whether they were taking the task seriously.

As described in §2.2, to help ensure that participants were qualified, we used a previously-developed recruitment procedure that has been shown to be successful at identifying participants who are broadly representative of the jury-eligible population

of the United States[23].

To help ensure participants were taking the study seriously, we included a number of attention and comprehension checks, and excluded participants who failed them. We also excluded participants who gave inconsistent or nonsensical answers. We cannot be certain, however, that participants took the task as seriously as legal-decision makers would if they were evaluating actual evidence in an actual trial, or even as seriously as if they were in an experimental environment where they were monitored by researchers.

Consequently, although it is common and reasonable to conduct exploratory studies inexpensively using online platforms and convenient populations, we recommend that key conclusions derived from studies like this one be confirmed or disconfirmed in studies that use actual legal-decision makers (participants recruited directly from jury pools, lawyers, judges, etc.) who are monitored while participating in experiments.

4 Conclusion

The research presented in this paper found that participants were *sensitive* to the value of the likelihood ratio presented by the expert witness:

- Participants' effective likelihood ratios were higher when the presented likelihood ratio was 3,000 than when it was 30.

The research did not, however, find convincing evidence that providing an explanation of the meaning of likelihood ratios helped participants better understand likelihood ratios:

- With respect to *orthodoxy*: The results supported the hypothesis that providing the explanation of the meaning of likelihood ratios would increase the probability that participants' effective likelihood ratios would be as expected if they had correctly applied Bayes' theorem, but the size of the difference between the explanation versus no-explanation conditions was small, and the percentage of *orthodox*

responses for the explanation condition was small in absolute terms.

- With respect to *coherence*: The results did not support the hypothesis that providing the explanation of the meaning of likelihood ratios would decrease the probability that participants' responses would be as expected if they committed the prosecutor's fallacy (posterior odds equal to presented likelihood ratio).

A substantial proportion of participants had prior odds of 1 and had posterior odds that equalled the presented likelihood ratio. For these participants, it was not possible to distinguish whether they correctly applied Bayes' theorem to the presented likelihood ratio or whether they committed the prosecutor's fallacy.

Participants could have weighted the values of the likelihood ratios presented to them based on other factors. Even if participants had correctly applied Bayes' theorem, they would have applied it to their weighted values rather than to the presented likelihood-ratio values, which would have resulted in effective likelihood ratios that differed from the presented likelihood ratios. This would have given the false impression that the participants did not understand the meaning of likelihood ratios.

We tested the effect of only one explanation of the meaning of likelihood ratios. It could be that some other explanation or some other way of explaining the meaning of likelihood ratios would lead to better understanding, for example, instead of using an arbitrary example likelihood-ratio value in the explanation, the presented likelihood-ratio value could be used. Also, it may help if prior odds of 1 are not used in any part of the explanation.

Despite the results obtained in this study, our priors are such that we still believe that explaining the meaning of likelihood ratios to legal-decision makers will improve their understanding of likelihood ratios. The explanation used in the present study does not appear to have been effective, but we recommend conducting additional research aimed at finding more effective ways of explaining the meaning of likelihood ratios to legal-decision makers.

Appendix A: Logit-transformation method for the A/B test

The logit transformation method for the A/B test is used to calculate a Bayes factor for the comparison of two sets of count data. The method is described in Kass & Vaidyanathan [24], Gronau et al. [25], Dablander et al. [26], and Hoffmann et al. [27], and has been implemented in Gronau [28]. What follows is abridged from information provided in the latter references. Please see the latter references for further details.

The Bayes factor is calculated as in Equation (4), in which $\mathcal{D}$ are the sample data, $p(\mathcal{D}|\theta, H)$ are likelihoods, $p(\theta|H)$ are prior distributions, θ, θ_{A}, and θ_{B} are probabilities ($\theta \equiv \theta_{\mathrm{A}} = \theta_{\mathrm{B}}$), H_0 is the hypothesis of $\theta_{\mathrm{A}} = \theta_{\mathrm{B}}$, and H_+ is the hypothesis of $\theta_{\mathrm{B}} > \theta_{\mathrm{A}}$. Subscript A indicates the condition "explanation of the meaning of likelihood ratios not presented" and subscript B indicates the condition "explanation of the meaning of likelihood ratios presented". $\mathcal{D}$ consist of the count of positive responses, y (the participants' effective likelihood ratios equalled the presented likelihood ratio, or the participants' responses were as expected if they had committed the prosecutor's fallacy), in each condition, and the count of all responses, n, in each condition, i.e., $\mathcal{D} = (y_{\mathrm{A}}, n_{\mathrm{A}}, y_{\mathrm{B}}, n_{\mathrm{B}})$.

$$(4) \quad \frac{p(\mathcal{D}|H_+)}{p(\mathcal{D}|H_0)} = \frac{\int_{\theta_{\mathrm{A}}} \int_{\theta_{\mathrm{B}}} p(\mathcal{D}|\theta_{\mathrm{A}}, \theta_{\mathrm{B}}, H_+) p(\theta_{\mathrm{A}}, \theta_{\mathrm{B}}|H_+)\, d\theta_{\mathrm{A}}\, d\theta_{\mathrm{B}}}{\int_{\theta} p(\mathcal{D}|\theta, H_0) p(\theta|H_0)\, d\theta}$$

The count data are assumed to have binomial distributions, see Equation (5).

$$(5) \quad Y_{\mathrm{A}} \sim \mathrm{Bin}(n_{\mathrm{A}}, \theta_{\mathrm{A}})$$

$$Y_{\mathrm{B}} \sim \mathrm{Bin}(n_{\mathrm{B}}, \theta_{\mathrm{B}})$$

Rather than assigning prior distributions directly to θ_{A} and θ_{B}, a logit transformation is applied to θ_{A} and θ_{B}, and prior distributions are assigned to the transformed parameters. The transformed parameters, β and ψ, are calculated as in Equation (6).

(6) $$\beta = \frac{1}{2}\left(\log\left(\frac{\theta_A}{1-\theta_A}\right) + \log\left(\frac{\theta_B}{1-\theta_B}\right)\right)$$

$$\psi = \log\left(\frac{\theta_B}{1-\theta_B}\right) - \log\left(\frac{\theta_A}{1-\theta_A}\right)$$

Note that β is the mean of the logits of θ_A and θ_B, and ψ is the signed difference between the logits of θ_A and θ_B.

We assigned relatively uninformative standard Gaussian prior distributions that are the defaults described in the papers cited above. Specifically:

- For all hypotheses, we assigned $\beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2)$ with $\mu_\beta = 0$ and $\sigma_\beta^2 = 1$.
- For $H_0: \theta_A = \theta_B$, we assigned $\psi \sim \mathcal{N}(\mu_\psi, \sigma_\psi^2)$ with $\mu_\psi = 0$ and $\sigma_\psi^2 = 1$.
- In §3.4, testing *orthodoxy*, for $H_+: \theta_B > \theta_A$, we assigned $\psi \sim \mathcal{N}(\mu_\psi, \sigma_\psi^2)\big|_{\psi \geq 1}$ (i.e., a truncated Gaussian distribution with a lower bound at 0) with $\mu_\psi = 0$ and $\sigma_\psi^2 = 1$.
- In §3.5, testing potential incidences of the prosecutor's fallacy, instead of testing H_+, we tested $H_-: \theta_B < \theta_A$, for which we assigned $\psi \sim \mathcal{N}(\mu_\psi, \sigma_\psi^2)\big|_{\psi \leq 1}$ (i.e., a truncated Gaussian distribution with an upper bound at 0) with $\mu_\psi = 0$ and $\sigma_\psi^2 = 1$.

The solutions to the integrals in Equation (4) are complex, and we do not provide details here; see Gronau et al. [25] for details. Under H_0, the marginal likelihood is calculated using Laplace approximations. Under H_+ or H_-, the marginal likelihood is calculated using a Monte Carlo method.

Author contributions

William C Thompson: Conceptualization, Investigation, Methodology, Supervision, Writing - Original Draft, Writing - Review & Editing. **Rebecca Hofstein Grady:** Conceptualization,

Investigation, Methodology, Writing - Original Draft, Writing - Review & Editing. **Geoffrey Stewart Morrison:** Conceptualization, Formal Analysis, Writing - Original Draft, 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 Thompson and Grady was supported by the Center for Statistical Applications in Forensic Evidence (CSAFE), which in turn was supported by the National Institute of Standards and Technology (NIST) through Cooperative Agreement #70NANB15H176 between NIST and Iowa State University, which included activities carried out at Carnegie Mellon University, University of California, Irvine, and University of Virginia. The work of Morrison was supported in part by the Australian Research Council, Australian Federal Police, New South Wales Police, Queensland Police, National Institute of Forensic Science, Australasian Speech Science and Technology Association, and the Guardia Civil through Linkage Project LP100200142.

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., 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., 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*, 376, 112589. https://doi.org/10.1016/j.forsciint.2025.112589

[3] 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

[4] 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/

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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

[14] 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

[15] Morrison G.S., Bali A.S., Martire K.A., Grady R.H., Thompson W.C. (2025). What is the best way to present likelihood ratios? A review of past research and recommendations for future research. *Science & Justice*, 65, 101342. https://doi.org/10.1016/j.scijus.2025.101304

[16] 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

[17] 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

[18] 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

[19] 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

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

[21] Paolacci G., Chandler J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, 23, 184–188. https://doi.org/10.1177/0963721414531598

[22] Smith S.M., Roster C.A., Golden L.L., Albaum G.S. (2016). A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research*, 69, 3139–3148. https://doi.org/10.1016/j.jbusres.2015.12.002

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

[24] Kass R.E., Vaidyanathan S.K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society, Series B*, 54, 129–144. https://doi.org/10.1111/j.2517-6161.1992.tb01868.x

[25] Gronau Q.F., Raj K.N. A., Wagenmakers E.-J. (2021). Informed Bayesian inference for the A/B test. *Journal of Statistical Software*, 100, 17. https://doi.org/10.18637/jss.v100.i17

[26] Dablander F., Huth K., Gronau Q.F., Etz A., Wagenmakers E.-J. (2022). A puzzle of proportions: Two popular Bayesian tests can yield dramatically different conclusions. *Statistics in Medicine*, 41, 1319–1333. https://doi.org/10.1002/sim.9278

[27] Hoffmann T., Hofman A., Wagenmakers E.J. (2022). Bayesian tests of two proportions: A tutorial with R and JASP. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 18, 239–277. https://doi.org/10.5964/meth.9263

[28] Gronau Q.F. (2019). *Abtest: Bayesian A/B testing*. R Foundation for Statistical Computing. https://CRAN.R-project.org/package=abtest

[29] Gittelson S., Berger C.E.H., Jackson G., Evett I.W., Champod C., Robertson B., Curran J.M., Taylor D., Weir B.S., Coble M.D., Buckleton J.S. (2018). A response to “Likelihood ratio as weight of evidence: A closer look” by Lund and Iyer. *Forensic Science International*, 228, e15–e19. https://doi.org/10.1016/j.forsciint.2018.05.025