EMPIRICAL ARTICLE

Diagnostic Information Produces Better-Calibrated Judgments About Forensic Comparison Evidence Than Likelihood Ratios

Gianni Ribeiro, Blake Malcolm McKimmie, and Jason Marcus Tangen
School of Psychology, The University of Queensland, Australia

Forensic expert testimony is slowly starting to reflect the uncertain nature of forensic science, but the way experts should express the uncertainty of their decisions is under debate. Here, we compare the likelihood model approach to a diagnostic approach—which provides information about performance and error rates—to determine which produces a more calibrated understanding and evaluation of the evidence. In Experiment 1 ($N = 738$), participants were more sensitive to differences in evidence strength when the evidence was expressed as diagnostic information than as a likelihood ratio, as predicted. In Experiment 2 ($N = 499$), however, when provided with both diagnostic information and a likelihood ratio, participants tended to discount the presence of the likelihood ratio in favor of the diagnostic information, which we did not predict. Together, these results suggest that providing fact-finders with diagnostic information might aid their understanding and evaluation of forensic evidence.

General Audience Summary

Forensic experts were once allowed to testify that two samples (e.g., fingerprints) "match" a defendant to the exclusion of all other people and that their decisions are infallible. However, numerous wrongful identifications and convictions have shown that these claims of infallibility simply are not true. These days, forensic testimony is slowly starting to reflect the uncertain nature of forensic decisions. One approach is known as the likelihood ratio, which conveys the probability of one hypothesis (that the fingerprints come from the same person) against another (that the fingerprints come from two different people). However, research has shown that jurors struggle to understand and interpret likelihood ratios. In this article, we present a different approach—the diagnostic information approach—which presents jurors with information about the past performance of forensic decisions (how often forensic examiners' decisions are correct and how often they are incorrect). In the first experiment, we find that presenting expert testimony in this way results in better-calibrated judgments compared to likelihood ratios. In the second experiment, we find that presenting both simultaneously does not lead to improved understanding. In conclusion, the diagnostic approach to presenting forensic testimony might be better understood, interpreted, and applied by jurors than likelihood ratios.

Keywords: forensic evidence, likelihood ratio, expert testimony, juror decision-making

Supplemental materials: https://doi.org/10.1037/mac0000062.supp

To date, the National Registry of Exonerations in the United States has documented 710 individuals who have been wrongfully convicted of crimes they did not commit on the basis of false or misleading forensic evidence, making it the fourth leading cause of wrongful convictions in their database (National Registry of Exonerations, 2022, January 31). While experts were once permitted to testify at trial that two prints, marks, or samples unequivocally "matched" the defendant, after numerous scathing reports about the state of forensic science (see, e.g., the 2009 report from the National Academy of Science and the 2016 report from the President's Council of Advisors on Science and Technology) the "match" terminology is no longer used their testimony is slowly starting to reflect the uncertain (and human) nature of their decisions. The National Academy of Science's (2009) report noted that the terminology that forensic examiners use when testifying in court can have a profound effect on how the fact-finder perceives and evaluates that evidence (p. 21) and that terminology should be standardized and presented to the court in lay terms, so that testimony is understandable and interpretable by all trial participants (p. 186). By making these improvements to communicating forensic decisions, the NAS report noted that the number of wrongful convictions should be reduced, ensuring that innocent people will not be incarcerated for crimes they did not commit while true perpetrators of the crime will not continue to commit crimes (pp. 4–5).

The Likelihood Ratio Approach to Communication Forensic Evidence

One approach to communicate the strength of forensic decisions is the likelihood ratio, which conveys the probability of an observation given two competing source hypotheses: (1) that the samples come from the same source, and (2) that the samples come from different sources. There is substantial support for the likelihood ratio approach (see Morrison's introduction to the special issue on likelihood ratios in *Science & Justice*, 2016). However, others are more skeptical, suggesting that the focus on precise, large likelihood ratios—often in the order of millions or even billions—is futile when the probability of a false positive occurring due to human error (such as mislabelling samples or accidental contamination) is far greater (Eldridge, 2019; Jackson et al., 2015; Spellman, 2018). For example, in 2011 British man Adam Scott was charged with the rape of a woman in Manchester after private forensics company LGC concluded that his DNA matched the sample taken from the victim (Peachey, 2012). However, Scott claimed, and his phone records showed, that he was at home in Plymouth more than 300 km away from the crime scene and had never been to Manchester in his life (Hall, 2017). After 5 months on remand in prison, an investigation found that Scott's DNA sample, which was taken for an unrelated incident, had contaminated the crime scene sample due to poor handling of the samples by LGC staff (Peachey, 2012).

Unfortunately, this is not an isolated incident of human error. A review of over 3,500 cases from Houston Crime Lab revealed that there were major issues in 32% of the DNA cases reviewed, including four inmates on death row (Bromwich, 2007). In 2008, the Detroit Crime Laboratory's firearms unit was shut down following a review which found that the unit complied with only 42% of the essential standards and found serious errors in 19 of a total 200 randomly selected firearms cases (Bunkley, 2008). In 2010, a review of North Carolina's state crime laboratory concluded that exculpatory evidence was either withheld or misrepresented in over 230 cases from 1987 to 2003 (Swecker & Wolf, 2010). In 2013, New York City's medical examiner's office undertook a review of over 800 rape cases over a period of 10 handled by a particular examiner who had consistently poor reviews and underperformance during her employment (Goldstein, 2013; Scott, 2013). Here, in Australia, a miscode in the computer software used to analyze DNA samples affected the probability statistics in up to 60 Queensland cases. Further, the developers of the program revealed that Queensland did not purchase the manual that accompanies the program, which may have resulted in human error (Murray, 2015).

While any approach to quantify the strength of evidence is a step in the right direction, if the ultimate fact-finder—the jury—is unable to properly understand, evaluate, and apply the information appropriately it could do more harm than good. A significant body of research in the forensic sciences is focused on how evidence is expressed in court, with a particular focus on random match probabilities (for a comprehensive review, see Eldridge, 2019). The overwhelming majority of studies find that, despite being numerically equivalent, fact-finders come to different ultimate decisions based on how the evidence is expressed to them (e.g., single-target probability vs. multitarget frequency expressions in Koehler, 2001).

Research also suggests that fact-finders struggle to understand and interpret evidence when expressed as likelihood ratios (Martire et al., 2013, 2014). Recommendations have been made to express likelihood ratios verbally rather than numerically (Association of Forensic Science Providers, 2009; De Kinder & Olsson, 2011), perhaps because many forensic sciences lack the data necessary to compute a numerical likelihood ratio (de Keijser & Elffers, 2012). However, we need to consider whether the recipients of likelihood ratio expressions, the jury, actually interpret these numerical and verbal expressions the way that they were intended (Campbell, 2011). Research regarding probabilities and statistics in general suggests that people have difficulty understanding and converting numerical and verbal expressions (Gigerenzer & Edwards, 2003; Kahneman et al., 1982). In addition, individuals may interpret the same verbal expressions of probability differently, as the meanings attributed to words may vary between people and contexts (Budescu et al., 2009; Shaw & Dear, 1990; Wallsten & Budescu, 1995). In Martire et al.'s (2013) study, the authors manipulated whether the expert's opinion was expressed as either a verbal likelihood ratio or a numeric likelihood ratio and also varied the strength of the evidence (low, moderate, or high). They found that when the evidence strength was moderate or high, numerical and verbal expressions resulted in equal amounts of belief change. However, when evidence strength was low, there was significantly greater belief change for the numerical expression than the verbal expression. Interestingly, in the low-strength numerical condition, belief change was in the intended direction (i.e., toward guilt); however, in the low-strength verbal condition, belief change was in the opposite direction than intended (i.e., toward innocence). This finding clearly contradicts the recommendation to express likelihood ratios verbally rather than numerically. Even when both verbal and numerical likelihood ratio expressions are presented in table format among the complete range of possible expressions, Martire et al. (2014) found that belief change was in the opposite direction (i.e., toward innocence) than what the expert had intended.

Given the evidence from Martire et al.'s (2013, 2014) studies that fact-finders struggle to understand and interpret likelihood ratios; perhaps, there is an alternative approach that may help fact-finders better understand and evaluate forensic evidence.

Some have suggested that informing jurors about the accuracy and error rates of forensic decisions may help them to understand and weigh the evidence more appropriately (Garrett & Mitchell, 2013; Howes, 2015; Mnookin, 2010).

The Diagnostic Information Approach to Communicating Forensic Evidence

Like decisions made about forensic comparison evidence, many medical decisions are also binary in nature, involve human judgment, and have varying degrees of error. Similarly, the ground truth of any *particular* decision is unknowable, so the best indicator is experts' past performance on similar tasks. Tomlinson et al. (2008) conducted an experiment to determine the accuracy of a home pregnancy test (*Answer*) by administering the test to 120 pregnant women and 120 women who were not pregnant. The results are summarized in Figure 1A using natural frequencies which are demonstrably more concrete and easier to grasp than probabilistic information (see, e.g., Akl et al., 2011; Gigerenzer & Hoffrage, 1995; Hoffrage & Gigerenzer, 1998). While a woman who has just taken the *Answer*[TM] pregnancy test cannot know, on the basis of this information alone, with absolute certainty whether or not she is pregnant, she can use this information to guide her interpretation of the result of her own test.

The same diagnostic information approach can be applied to forensic evidence. A juror cannot know, with absolute certainty, whether or not a particular fingerprint found at a crime scene belongs to a defendant or not. But with information about the accuracy of fingerprint examiners' decisions from studies like Tangen et al. (2011) summarized in Figure 1B, a juror can better evaluate a forensic examiner's testimony in order to guide their interpretation of whether or not the fingerprint is likely to have come from the defendant or some other person. Taking a diagnostic information approach to forensic evidence would provide information, such as performance and error rates on previous, similar decisions, to help jurors reason about the present case. This process is often referred to as group-to-individual inference (see Faigman et al., 2014). Jurors may use this general performance information to help them weigh the evidence in a particular case and adjust their degree of belief accordingly.

Should Forensic Examiners Present Both Likelihood Ratios and Diagnostic Information Simultaneously?

Authoritative bodies, such as President's Council of Advisor's on Science and Technology (2016) and the National Research Council (NRC) (1992) have suggested that error rate estimates be provided alongside probabilistic testimony such as a random match probability or likelihood ratio. However, because error rates derived from proficiency tests provide the accuracy of an examiner in making a binary decision, while a random match probability provides the strength of the evidence without requiring a binary decision, jurors may struggle to understand and combine these two types of information. In fact, studies have demonstrated that people have difficulty combining probabilities (see, e.g., the conjunction fallacy in Tversky & Kahneman, 1983). This has been empirically demonstrated by Koehler et al. (1995), who found that the presence of a laboratory error rate (0.1% or 2%) had no impact on verdicts. Further, as previously discussed, because the chance of human error is far greater than the chance of a random match (or the inverse, a likelihood ratio), the error rate would effectively swamp the other statistic, rendering it futile.

In two preregistered experiments, we compared the likelihood ratio approach to a diagnostic information approach, which included information about accuracy and errors of fingerprint examiners' decisions. The aim of Experiment 1 was to determine whether diagnostic information would result in a more calibrated understanding and evaluation of the fingerprint evidence than a likelihood ratio, whereas the aim of Experiment 2 was to determine how presenting both a likelihood ratio and diagnostic information together would affect participants' evaluation of the evidence.

Experiment 1

The aim of Experiment 1 was to directly compare the likelihood ratio approach to the diagnostic information approach using a 2 (evidence expression: likelihood ratio vs. diagnostic information) × 5 (evidence strength: 5 levels) between-groups design.

Method

Participants

This experiment was preregistered prior to data collection (see http://osf.io/5xmk8). Both experiments reported in this article received ethical approval from The University of Queensland's Low & Negligible Risk Ethics Sub-Committee (Clearance Number:

Figure 1
Results of Two Experiments Shown in Natural Frequencies

(A)	Test said "pregnant"	Test said "not pregnant"		(B)	Expert said "match"	Expert said "no match"	
Pregnant	98	22	120	Same source	409	35	444
Not pregnant	2	118	120	Different source	3	441	444
	100	140			412	476	

Note. (A) Results from Tomlinson et al. (2008). (B) Results from Tangen et al. (2011).

2018000922). We recruited 740 participants for this experiment, where the sample size was determined a priori as having 95% power to detect a small-to-medium effect ($d = 0.4$) at an α level of .05. We recruited participants residing in Australia or England from the *Prolific* online testing platform who received £1.25 for their participation.

Following our preregistered exclusion plan, we excluded two participants for failing, or not responding to, the attention check question. The final sample ($N = 738$) consisted of 501 women, 231 men, three participants who identified as another gender, and three who did not report their gender with an age range of 18–88 years ($M = 37.59$, $SD = 12.32$). Most (99.5%) participants reported living in England; however, two participants did not report whether they lived in England or Australia. Other participant demographic information, such as highest level of education and prior jury service, can be viewed in the Supplemental Materials.

Procedure and Materials

Ethical clearance for this study participants completed the experiment on their own devices. They first read a short case vignette detailing the facts of a murder and were told that a latent fingerprint was retrieved from the victim's windowsill. Participants who were randomly assigned to the likelihood ratio conditions received information about the strength of evidence in a likelihood ratio in one of five strengths: 5, 55, 550, 5,500, or 550,000. These values were selected as they fall in the middle of each of the categories in the Association of Forensic Science Providers' (2009) *Standards for Numerical and Verbal Expression of Likelihood Ratios*. Below is an example of the highest strength likelihood ratio condition:

> A man with the initials B.M. is alleged to have murdered the victim in the case you just read about.
>
> During the defendant's trial, a forensic fingerprint examiner testified about the fingerprint retrieved from the victim's window sill. The forensic examiner compared the fingerprint retrieved from the victim's window sill with a fingerprint obtained from the defendant, B.M.
>
> When assessing the significance of any similarity or differences between two fingerprints, the likelihood of obtaining that similarity or difference is considered against two alternative propositions: (Hypothesis 1) the two fingerprints originated from the same person; (Hypothesis 2) the two fingerprints did not originate from the same person.
>
> The forensic examiner testified that the correspondence between the fingerprint retrieved from the victim's window sill and the fingerprint obtained from the defendant, B.M., is 5,500,000 times more likely if the two fingerprints originated from the same person (Hypothesis 1) than if the two fingerprints originated from different people (Hypothesis 2).

Whereas, participants randomly assigned to the diagnostic information conditions received information about how fingerprint examiners perform. This information was adapted from Edmond et al.'s (2014) *Guide to Interpreting Forensic Testimony*. This information was presented in one of five strengths: 50%, 63%, 75%, 88%, 99%. We chose these strengths as they represent equal intervals between chance and (near) perfect performance. Below is an example of the highest strength diagnostic information condition:

> A man with the initials B.M. is alleged to have murdered the victim in the case you just read about.
>
> During the defendant's trial, a forensic fingerprint examiner testified about the fingerprint retrieved from the victim's window sill. The examiner testified that the fingerprint retrieved from the victim's window sill matched the defendant, B.M.'s, fingerprint.
>
> A decision about whether a fingerprint found at a crime scene matches a person's fingerprint or not is based on decisions by a human examiner, not a computer.
>
> In laboratory-based experiments, a group of fingerprint examiners was tested on whether they could tell the difference between two fingerprints from the same person and two fingerprints from different people. They repeated this test 200 times.
>
> For the 100 pairs of fingerprints that matched, the examiners correctly said "match" for 99, but incorrectly said "no match" for 1.
>
> For the 100 pairs of fingerprints that did not match, the examiners correctly said "no match" for 99, but incorrectly said "match" for 1.

After reading the case vignette and the evidence information, participants were asked to respond to three key dependent variables: source likelihood (on a scale of 0–100, how likely is it that the DNA evidence retrieved from underneath the victim's fingernails belongs to B.M.?), guilt likelihood (on a scale of 0–100, how likely is it that B.M. committed this murder?), and a dichotomous verdict judgment (not guilty vs. guilty). As source and guilt likelihood measures were consistent across Experiments 1 and 2, these are reported here, whereas results for the dichotomous guilt judgment can be found in the Supplemental Materials. Finally, participants responded to one attention check question, three manipulation check questions, and five demographic questions (see the full set of experimental materials on the Open Science Framework).

Hypotheses

We predicted a main effect of how the evidence was expressed, such that participants would be more likely to conclude that the fingerprint came from the defendant, that the defendant committed the murder, and that the defendant was guilty of murder when the evidence was expressed as a likelihood ratio compared to diagnostic information (H1). We also predicted a main effect of Strength, such that participants would be more likely to conclude that the fingerprint came from the defendant, that the defendant committed the murder, and that the defendant was guilty of murder as the strength of the evidence increases (H2). Finally, we predicted an Expression × Strength interaction, such that strength will have a greater effect on the dependent variables when the evidence is expressed as diagnostic information compared to likelihood ratio (H3).

Results

Overview of Analyses

To test our predictions, in line with our preregistration we conducted 2 (evidence expression: likelihood ratio vs. diagnostic information) × 5 (evidence strength: 5 levels) between-groups factorial analyses of variance (ANOVAs) on both source and guilt likelihood. As the data violated the assumptions of normality and homogeneity of variance, we also conducted robust ANOVAs in R using 20% trimmed means (Wilcox, 2012). Unless otherwise stated,

the results of both approaches were the same, thus we will report the original ANOVA results here for ease of interpretation.

Manipulation Checks

About 92.4% of participants who received evidence expressed as a likelihood ratio correctly identified the strength of the likelihood ratio that they received. Similarly, 99.5% of participants who received evidence expressed as diagnostic information correctly identified the strength of the diagnostic information that they received. As per the preregistered analysis plan, all participants were retained in the main analysis.

Source Likelihood

As illustrated in Figure 2, consistent with H1 there was a significant main effect of Expression, $F(1, 726) = 7.144$, $p = .008$, $\eta_p^2 = .010$. Participants who received the evidence expressed as a likelihood ratio ($M = 76.65$, $SD = 18.28$) were more likely to conclude that the defendant was the source of the fingerprint than participants who received the evidence expressed as diagnostic information ($M = 73.89$, $SD = 18.18$). Furthermore, consistent with H2 we found a significant main effect of evidence strength,

Figure 2
Results of Experiment 1

Note. Panels A and B depict participants' source and guilt likelihood ratings, respectively. The raincloud plots depict a half violin plot of participants' mean ratings overlaid with jittered data points from each of the participants who were randomly assigned to one of five evidence strengths (1 = weakest, 5 = strongest) along with the standard error of the mean per condition. See the online article for the color version of this figure.

$F(4, 726) = 76.99$, $p \leq .001$, $\eta_p^2 = .298$. Polynomial planned contrasts revealed a significant linear trend, $C = 20.59$, $p < .001$, showing that participants were more likely to conclude that the defendant was the source of the fingerprint as the strength of the evidence increased.

Finally, there was a significant Expression × Strength Interaction, $F(4, 726) = 27.79$, $p \leq .001$, $\eta_p^2 = .133$, where the strength of the expression influenced the likelihood that the defendant was the source of the fingerprint for both the likelihood ratio, $F(4, 726) = 14.20$, $p < .001$, and diagnostic information expressions, $F(4, 726) = 89.92$, $p < .001$. A planned linear contrast revealed that the positive linear trend was greater for the diagnostic expression, $C = 102.46$, $p < .001$, compared to the likelihood ratio expression, $C = 27.76$, $p < .001$, demonstrating that participants are more sensitive to the strength of the evidence when expressed diagnostically rather than as a likelihood ratio, which was consistent with H3.

Guilt Likelihood

There was no main effect of Expression, $F(1, 731) = 0.25$, $p = .619$, $\eta_p^2 \leq .001$, which we did not predict (H1). That is, participants were equally likely to conclude that the defendant committed the murder when the evidence was expressed as a likelihood ratio ($M = 63.67$, $SD = 20.85$) or using diagnostic information ($M = 62.87$, $SD = 21.08$). Consistent with H2, there was a main effect of evidence strength, $F(4, 731) = 6.57$, $p \leq .001$, $\eta_p^2 = .035$. Polynomial planned contrasts revealed a significant linear trend, $C = 8.21$, $p < .001$, where participants were more likely to conclude that the defendant committed the murder as the strength of the evidence increased.

Finally, there was a significant Expression × Strength Interaction, $F(4, 731) = 4.24$, $p = .002$, $\eta_p^2 = .023$. The strength of the evidence influenced the likelihood that the defendant committed the crime when expressed as diagnostic information, $F(4, 722) = 9.51$, $p < .001$, but not as a likelihood ratio, $F(4, 722) = 1.22$, $p = .299$. A planned linear contrast revealed that there was a positive linear trend for the diagnostic expression, $C = 46.07$, $p < .001$, but not the likelihood ratio expression, $C = 5.82$, $p = .439$, which was consistent with H3.
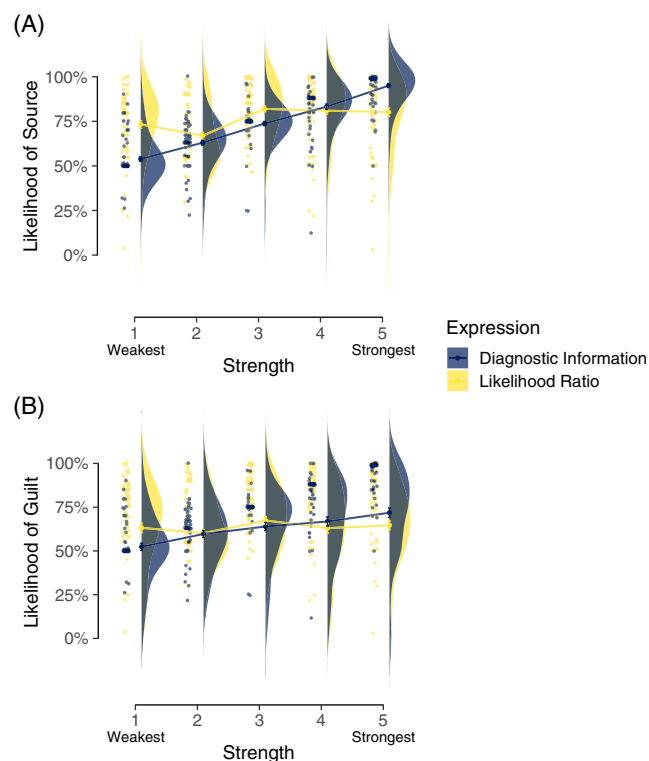
Experiment 2

In Experiment 1, we found that participants were more sensitive to the strength of the evidence when it was expressed as diagnostic information rather than as a likelihood ratio. What if participants were presented with both a likelihood ratio and diagnostic information simultaneously? As mentioned previously, providing a large likelihood ratio may not be all that useful given that the chance of a human error occurring is far greater (Eldridge, 2019; Jackson et al., 2015; Spellman, 2018). Thus, ideally, participants would disregard the likelihood ratio in favor of the diagnostic information. However, studies have demonstrated that people have difficulty combining probabilities (see, e.g., the conjunction fallacy in Tversky & Kahneman, 1983). Further, as likelihood ratios may provide the "aura of precision" (Tribe, 1971) and large likelihood ratios may be particularly compelling, it may be more likely that participants would rely more on the likelihood ratio than the diagnostic information. The aim of Experiment 2 was to determine how participants would evaluate the fingerprint evidence when presented with both a likelihood ratio as well as diagnostic information using a 2 (diagnostic information: low vs. high) × 2 (likelihood

ratio: absent vs. present) + 1 (likelihood ratio only) between-groups design.

Method

Participants

This experiment was preregistered prior to data collection (see http://osf.io/n27p8). We recruited 500 participants for this experiment, where the sample size was determined a priori as having 95% power to detect a small-to-medium effect ($d = 0.4$) at an α level of .05. We recruited participants residing in Australia or England from the *Prolific* online testing platform who agreed to receive £1.25 for their participation.

Following our preregistered exclusion plan, we excluded one participant for failing the attention check question. The final sample ($N = 499$) consisted of 312 women, 185 men, and two participants who did not report their gender, with an age range of 18–72 years ($M = 35.43$, $SD = 13.10$). 99.4% of participants reported living in England, 0.4% of participants reported living in Australia, and one participant did not report whether they lived in England or Australia. Other participant demographic information, such as highest level of education and prior jury service, can be viewed in the Supplemental Materials.

Procedure and Materials

Participants completed the experiment on their own devices. Participants first read a short case vignette (same as Experiment 1). Participants were then randomly assigned to one of five experimental conditions. In the likelihood ratio present condition, participants received a likelihood ratio value of 5,500,000—which falls in the "extremely strong" category of the Association of Forensic Science Providers' (2009) *Standards for Numerical and Verbal Expression of Likelihood Ratios*—whereas, participants in the likelihood ratio absent condition received no likelihood ratio information at all. Participants in the diagnostic information condition received information about fingerprint examiners' performance, presented similarly to that in Experiment 1. In the low condition, this information conveyed an overall accuracy of 63%, whereas, in the high condition, this information conveyed an overall accuracy of 99%. Finally, participants in the likelihood ratio only condition received only the likelihood ratio information, but no diagnostic information.

After reading the scenario and the evidence information, participants were asked to respond to three key dependent variables: source likelihood (on a scale of 0–100, how likely is it that the fingerprint retrieved from the victim's window sill belongs to B.M.?), guilt likelihood (on a scale of 0–100, how likely is it that B.M. committed this murder?), and error likelihood (on a scale of 0–100, how likely is it that there could have been an error during the analysis and comparison of B.M.'s fingerprint to the fingerprint retrieved from the victim's window sill?). As source and guilt likelihood measures were consistent across Experiments 1 and 2, these are reported here, whereas results for the error likelihood measure can be found in the Supplemental Materials. Finally, as in Experiment 1, participants responded to one attention check question, three manipulation check questions, and five demographic questions (see full experimental materials on the Open Science Framework).

Hypotheses

We predicted a main effect of likelihood ratio, such that participants would be more likely to conclude that the evidence came from the defendant and that the defendant committed the murder, but less likely to conclude that an error could have occurred, when a likelihood ratio was present compared to absent (H1). We also predicted a main effect of diagnostic information, such that participants would be more likely to conclude that the evidence came from the defendant and that the defendant committed the murder, but less likely to conclude that an error could have occurred, when diagnostic information was high compared to low (H2). Further, we predicted a Likelihood Ratio × Diagnostic Information interaction, such that we expect diagnostic information to have a greater effect on the dependent variables when likelihood ratio is absent compared to present (H3).

Finally, if participants are not sensitive to the diagnostic information and are instead focused solely on the likelihood ratio information, then the floating control condition (likelihood ratio only) should only differ significantly from the conditions where a likelihood ratio is absent and only diagnostic information is presented, but not when a likelihood ratio is present alongside diagnostic information (H4).

Results

Overview of Analyses

To test our predictions, we conducted 2 (diagnostic information: low vs. high) × 2 (likelihood ratio: absent vs. present) between-groups factorial ANOVAs on each of our dependent measures: source, guilt, and error likelihood. As the data violated the assumptions of normality and homogeneity of variance, we also conducted robust ANOVAs in R using 20% trimmed means (Wilcox, 2012). Unless otherwise stated, the pattern of results of both approaches was the same, thus we will report the original ANOVA results here for ease of interpretation. We also conducted a one-way between-groups ANOVA on the three likelihood ratio conditions: likelihood ratio with high and low diagnostic information, and likelihood ratio only.

Manipulation Checks

In the likelihood ratio absent condition, 87.88% of participants correctly identified that they did not receive a likelihood ratio, whereas 88.33% of participants in the present condition correctly identified that they received a likelihood ratio of 5,500,000. For the diagnostic information manipulation, 94% of participants in the low diagnostic condition and 97.9% of participants in the high diagnostic condition correctly identified the level of diagnostic information they were presented in the vignette. As per the preregistered analysis plan, all participants were retained in the main analysis.

Source Likelihood

As illustrated in Figure 3, there was no main effect of likelihood ratio, $F(1, 395) = 2.71$, $p = .10$, $\eta_p^2 = .007$; therefore, participants were equally likely to conclude that the defendant was the source of the fingerprint when a likelihood ratio was present ($M = 78.26$, $SD = 20.27$) or absent ($M = 75.19$, $SD = 21.14$), which we did not predict (H1). However, consistent with H2, there was a main effect of diagnostic information, $F(1, 395) = 212.54$, $p \leq .001$, $\eta_p^2 = .350$.

Figure 3
Results of Experiment 2
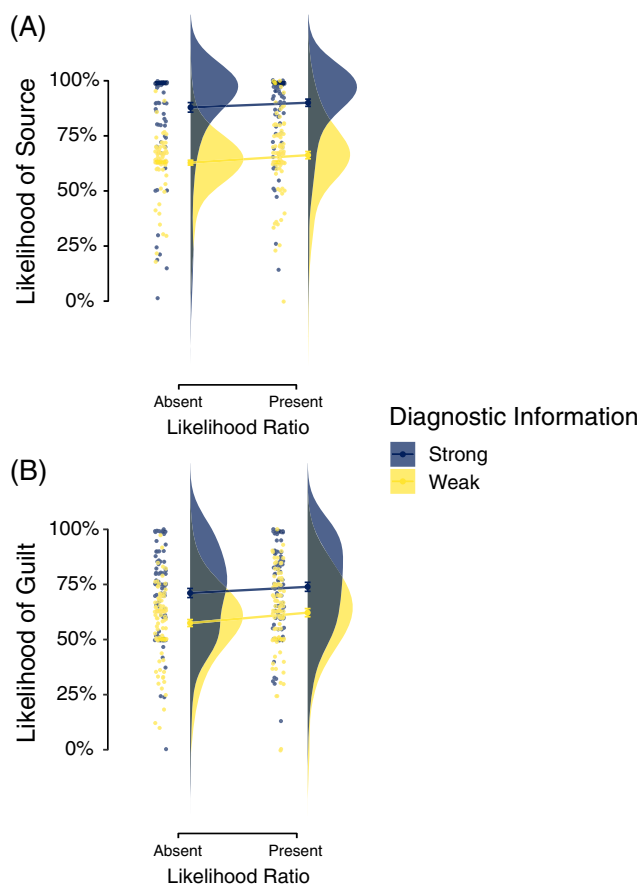
Note. Panels A and B depict participants' source and guilt likelihood ratings, respectively. The raincloud plots depict a half violin plot of participants' mean ratings overlaid with jittered data points from each of the participants along with the standard error of the mean per condition. See the online article for the color version of this figure.

Participants were more likely to conclude that the defendant was the source of the fingerprint when diagnostic information was high ($M = 89.00$, $SD = 19.10$) compared to low ($M = 64.53$, $SD = 14.05$). Contrary to our prediction (H3), there was no significant Likelihood Ratio × Diagnostic Information interaction, $F(1, 395) = .155$, $p = .694$, $\eta_p^2 \leq .001$, indicating that the effect of diagnostic information on participants' judgments of whether the defendant was the source of the fingerprint did not differ depending on whether participants received a likelihood ratio or not.

Guilt Likelihood

Consistent with H1, there was a significant main effect of likelihood ratio, $F(1, 386) = 3.94$, $p = .048$, $\eta_p^2 = .010$; therefore, participants were slightly more likely to conclude that the defendant committed the murder when a likelihood ratio was present ($M = 68.03$, $SD = 20.03$) or absent ($M = 64.20$, $SD = 19.33$). However, this effect was not significant with a robust ANOVA, $Q = 346$, $p = .065$. Furthermore, consistent with H2, there was a main effect of diagnostic information, $F(1, 386) = 44.86$, $p \leq .001$, $\eta_p^2 = .104$,

where participants were more likely to conclude that the defendant committed the murder when diagnostic information was high ($M = 72.52$, $SD = 20.19$) compared to low ($M = 59.84$, $SD = 17.18$). However, contrary to our prediction (H3), there was no significant Likelihood Ratio × Diagnostic Information interaction, $F(1, 386) = .254$, $p = .615$, $\eta_p^2 \leq .001$, indicating that the effect of diagnostic information on participants' judgments of whether the defendant committed the murder did not differ depending on whether participants received a likelihood ratio or not.

Comparing Likelihood Ratio Conditions

Finally, we compared the floating control condition, likelihood ratio only, to the other conditions in the experimental design. A one-way ANOVA comparing all five experimental conditions revealed a significant difference between conditions on the measure of source likelihood, $F(4, 498) = 55.79$, $p < .001$. As per H4, we hypothesized that the likelihood ratio only condition would only differ from conditions where a likelihood ratio was absent (i.e., diagnostic information only) compared to present. However, contrary to our hypothesis, post hoc multiple comparisons using Tukey's range test revealed that participants in the likelihood ratio only condition ($M = 83.67$, $SD = 17.93$) were significantly more likely to conclude that the defendant was the source of the evidence than participants who received a likelihood ratio with weak diagnostic information ($M = 66.25$, $SD = 16.38$), $p < .001$, and participants who only received weak diagnostic information ($M = 62.83$, $SD = 11.12$), $p < .001$, but not participants who received a likelihood ratio with strong diagnostic information ($M = 90.03$, $SD = 16.49$), $p = .063$, or participants who received only strong diagnostic information ($M = 83.67$, $SD = 17.93$), $p = .396$.

A similar pattern of results was found for the measure of guilt likelihood, $F(4, 487) = 12.83$, $p < .001$. Post hoc multiple comparisons using Tukey's range test revealed that participants in the likelihood ratio only condition ($M = 70.36$, $SD = 20.35$) were significantly more likely to conclude that the defendant was the source of the evidence than participants who received a likelihood ratio with weak diagnostic information ($M = 62.20$, $SD = 18.28$), $p = .024$, and participants who only received weak diagnostic information ($M = 57.50$, $SD = 15.76$), $p < .001$, but not participants who received a likelihood ratio with strong diagnostic information ($M = 73.92$, $SD = 20.08$), $p = .687$, or participants who received only strong diagnostic information ($M = 71.11$, $SD = 20.30$), $p = .999$. Together, these findings demonstrate that participants' conclusions about the likelihood that the fingerprint belongs to the defendant and the likelihood that the defendant committed the crime is driven largely by the diagnostic information and not the likelihood ratio.

Discussion

There is wide support for forensic examiners adopting likelihood ratios to communicate the uncertainty of forensic decisions (Morrison, 2016); however, research has shown that mock jurors struggle to understand and interpret likelihood ratios (Martire et al., 2013, 2014). The aim of the current article was to compare the use of likelihood ratios to a diagnostic information approach, where jurors would be presented with information regarding forensic examiner's past performance—the number of correct and incorrect decisions—in natural frequencies, which are demonstrably better

understood than probabilistic information (see, e.g., Akl et al., 2011; Gigerenzer & Hoffrage, 1995; Hoffrage & Gigerenzer, 1998).

Experiment 1 compared information about the strength of evidence expressed either as a likelihood ratio (which communicates the strength of the observations between the two fingerprints) or as diagnostic information (which communicates the strength of fingerprint examiners' conclusions through information about examiners' performance). We predicted that participants would be more sensitive to the strength of the evidence when it was presented diagnostically, including information about accuracy and error, rather than as a likelihood ratio. This hypothesis was supported, as participants' judgments about whether the defendant was the source of the evidence and committed the murder increased linearly as the strength of the evidence increased in the diagnostic expression condition, but not the likelihood ratio expression condition. Participants' insensitivity to the strength of the evidence, when presented as a likelihood ratio, is consistent with Martire et al.'s (2013, 2014) findings that jurors struggle to interpret the strength of a likelihood ratio, even when it is presented in the context of a table where the full range of possible likelihood ratio values was presented. Our findings suggest that expressing the strength of forensic evidence diagnostically, rather than as a likelihood ratio, may result in greater sensitivity to the strength of the evidence by fact-finders and aid their interpretation. It is important to note that we do not intend to claim that the likelihood ratios and diagnostic information presented for each strength are equivalent (e.g., a likelihood ratio of 5 is not equivalent to an overall accuracy of 50% in the lowest strength condition), hence why we conducted analyses *within* each expression type rather than *between* expression types.

In Experiment 2, we wanted to determine whether providing participants with a likelihood ratio in addition to the diagnostic information would help or hinder their judgments. As the chance of an error having occurred outweighs the strong likelihood ratio, participants should ideally disregard the likelihood ratio in favor of the diagnostic information. However, previous research suggests that people combine probabilities poorly (e.g., Tversky & Kahneman, 1983) and, as the strong likelihood ratio may seem overly compelling and precise, participants may favor the likelihood ratio over the diagnostic information. Thus, we predicted that participants would only be sensitive to the diagnostic information when it was presented alone rather than presented alongside a likelihood ratio, which the data did not support. The only effect was a medium-to-large sized main effect of diagnostic information, where participants were more likely to conclude that the defendant was the source of the evidence and committed the murder when the diagnostic information was high compared to low, regardless of whether the diagnostic information was accompanied by a likelihood ratio or not. Thus, there appears to be no added benefit to providing participants with a likelihood ratio in addition to the diagnostic information, but it does not appear to hinder their ability to interpret the diagnostic information.

These results appear to contradict the existing literature which suggests that people have difficulty combining probabilities. For example, Koehler et al. (1995) found that the presence of a random match probability, whether accompanied by an error rate or not, increased participants' guilty verdicts and that the error rate information had no impact on participants' verdicts. Whereas, the results of our experiment are the opposite: The presence of a likelihood ratio had no effect, but the error rate information did. One possible explanation for this difference is that Koehler et al.'s (1995) study

presented error rate information very briefly; participants were simply told that "the probability that a human error occurred which may have led to an incorrect report was about 1 in 1,000" (p. 219) or "1 in 50." Whereas, in our experiment, participants were told about a test that fingerprint examiners have taken involving pairs of known matching and nonmatching prints and are provided with the results of that test, indicating their overall accuracy and error rate. It may be that providing participants with this added information about how the error rate was derived makes it easier to understand and, in turn, leads to increased sensitivity to the strength of the evidence.

Further, Koehler et al.'s (1995) study used a random match probability as the evidence statistic, compared to a likelihood ratio in our experiment. While a likelihood ratio is the inverse of a random match probability, it is possible that participants may interpret these statistics differently. Previous research seems to support this argument, as studies have found that people give more weight to the evidence when the random match probability is low compared to high (e.g., Goodman, 1992; Smith et al., 1996), whereas people do not appear to be sensitive to the value of the likelihood ratio (Martire et al., 2013, 2014). It is also worth that we did not assess participants' jury eligibility in their country of residence, thus we could not exclude any jury ineligible participants which may affect the ecological validity of our findings.

Taken together, the results of both experiments suggest that people are more sensitive to the strength of evidence when presented with diagnostic information rather than a likelihood ratio and, when presented with both, rely primarily on the diagnostic information to inform their judgments. If this is the case, then forensic examiners should strongly consider adopting a diagnostic approach to communicate the strength of their evidence.

References

Akl, E. A., Oxman, A. D., Herrin, J., Vist, G. E., Terrenato, I., Sperati, F., Costiniuk, C., Blank, D., & Schünemann, H. (2011, March 16). *Using different statistical formats for presenting health information*. https://www.cochrane.org/CD006776/COMMUN_using-different-statistical-formats-for-presenting-health-information

Association of Forensic Science Providers. (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, *49*(3), 161–164. https://doi.org/10.1016/j.scijus.2009.07.004

Bromwich, M. R. (2007). *Final report of the independent investigator for the Houston police department crime laboratory and property room*. http://www.hpdlabinvestigation.org/reports/070613report.pdf

Budescu, D. V., Broomell, S., & Por, H.-H. (2009). Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. *Psychological Science*, *20*(3), 299–308. https://doi.org/10.1111/j.1467-9280.2009.02284.x

Bunkley, N. (2008). *Detroit police lab is closed after audit finds serious errors in many cases*. https://www.nytimes.com/2008/09/26/us/26detroit.html?_r=0

Campbell, A. (2011). *The fingerprint inquiry report*. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.361.380&rep=rep1&type=pdf

de Keijser, J., & Elffers, H. (2012). Understanding of forensic expert reports by judges, defense lawyers and forensic professionals. *Psychology, Crime & Law*, *18*(2), 191–207. https://doi.org/10.1080/10683161003736744

De Kinder, J., & Olsson, T. (2011). Expressing evaluative opinions: A position statement. *Science & Justice*, *51*(1), 1–2. https://doi.org/10.1016/j.scijus.2011.01.002

Edmond, G., Thompson, M. B., & Tangen, J. M. (2014). A guide to interpreting forensic testimony: Scientific approaches to fingerprint evidence. *Law Probability and Risk*, *13*(1), 1–25. https://doi.org/10.1093/lpr/mgt011

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

Faigman, D. L., Monahan, J., & Slobogin, C. (2014). Group to individual (G2i) inference in scientific expert testimony. *The University of Chicago Law Review. University of Chicago. Law School*, *81*(2), 417–480. https://doi.org/10.2139/ssrn.2298909

Garrett, B. L., & Mitchell, G. (2013). How jurors evaluate fingerprint evidence: The relative importance of match language, method information, and error acknowledgment. *Journal of Empirical Legal Studies*, *10*(3), 484–511. https://doi.org/10.1111/jels.12017

Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: From innumeracy to insight. *British Medical Journal*, *327*(7417), 741. https://doi.org/10.1136/bmj.327.7417.741

Gigerenzer, G., & Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Goldstein, J. (2013). *New York examines over 800 rape cases for possible mishandling of evidence*. https://www.nytimes.com/2013/01/11/nyregion/new-york-reviewing-over-800-rape-cases-for-possible-mishandling-of-dna-evidence.html

Goodman, J. (1992). Jurors' comprehension and assessment of probabilistic evidence. *American Journal of Trial Advocacy*, *16*(2), 361. https://heinonline.org/HOL/LandingPage?handle=hein.journals/amjtrad16&div=18&id=&page=

Hall, L. (2017, February 5). *DNA profiling is not infallible and can lead to innocent people being wrongly convicted: Report*. https://www.smh.com.au/national/nsw/dna-profiling-is-not-infallible-and-can-lead-to-innocent-people-being-wrongly-convicted-report-20170202-gu3rwe.html

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73*(5), 538–540. https://doi.org/10.1097/00001888-199805000-00024

Howes, L. M. (2015). A step towards increased understanding by non-scientists of expert reports: Recommendations for readability. *Australian Journal of Forensic Sciences*, *47*(4), 456–468. https://doi.org/10.1080/00450618.2015.1004194

Jackson, G., Kaye, D. H., Neumann, C., Ranadive, A., & Reyna, V. F. (2015). *Communicating the results of forensic science examinations*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2690899

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: heuristics and biases*. Cambridge University Press.

Koehler, J. (2001). The psychology of numbers in the courtroom: How to make DNA-match statistics seem impressive or insufficient. *Southern California Law Review*, *74*(5), 1275–1305. https://www.scholars.northwestern.edu/en/publications/the-psychology-of-numbers-in-the-courtroom-how-to-make-dna-match-

Koehler, J. J., Chia, A., & Lindsey, S. (1995). The random match probability in DNA evidence: Irrelevant and prejudicial? *Jurimetrics Journal of Law, Science and Technology*, *35*(2), 201–219. https://www.scholars.northwestern.edu/en/publications/the-random-match-probability-rmp-in-dna-evidence-irrelevant-and-p

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

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 and Human Behavior*, *37*(3), 197–207. https://doi.org/10.1037/lhb0000027

Mnookin, J. L. (2010). The courts, the NAS, and the future of forensic science. *Brooklyn Law Review*, *75*, 1209–1489. https://brooklynworks.brooklaw.edu/blr/vol75/iss4/10/

Morrison, G. S. (2016). Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate. *Science & Justice*, *56*(5), 371–373. https://doi.org/10.1016/j.scijus.2016.05.002

Murray, D. (2015). *Queensland authorities confirm 'miscode' affects DNA evidence in criminal cases*. https://www.couriermail.com.au/news/queensland/queensland-authorities-confirm-miscode-affects-dna-evidence-in-criminal-cases/news-story/833c580d3f1c59039efd1a2ef55af92b

National Academy of Science. (2009). *Strengthening forensic science in the United states: A path forward*. The National Academies Press.

National Registry of Exonerations. (2022, January 31). *% exonerations by contributing factor*. https://www.law.umich.edu/special/exoneration/Pages/ExonerationsContribFactorsByCrime.aspx

National Research Council. (1992). *DNA technology in forensic science*. https://www.nap.edu/catalog/1866/dna-technology-in-forensic-science

Peachey, P. (2012, October 2). *Rape accused Adam Scott was victim of forensics error, regulator finds*. https://www.independent.co.uk/news/uk/crime/rape-accused-adam-scott-was-victim-of-forensics-error-regulator-finds-8193163.html

President's Council of Advisor's on Science and Technology. (2016). *Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods*. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf

Scott, C. L. (2013). *Investigation into the New York City Office of Chief Medical Examiner: Department of Forensic Biology*. https://ig.ny.gov/sites/g/files/oee571/files/2016-12/OCMEFinalReport.pdf

Shaw, N. J., & Dear, P. R. (1990). How do parents of babies interpret qualitative expressions of probability? *Archives of Disease in Childhood*, *65*(5), 520–523. https://doi.org/10.1136/adc.65.5.520

Smith, B. C., Penrod, S. D., Otto, A. L., & Park, R. C. (1996). Jurors' use of probabilistic evidence. *Law and Human Behavior*, *20*(1), 49–82. https://doi.org/10.1007/BF01499132

Spellman, B. A. (2018). Communicating forensic evidence: Lessons from psychological science. *Seton Hall Law Review*, *48*(3), 827–840.

Swecker, C., & Wolf, M. (2010). *An independent review of the SBI forensic laboratory*. https://forensicresources.org/wp-content/uploads/2019/09/Swecker_Report.pdf

Tangen, J. M., Thompson, M. B., & McCarthy, D. J. (2011). Identifying fingerprint expertise. *Psychological Science*, *22*(8), 995–997.

Tomlinson, C., Marshall, J., & Ellis, J. E. (2008). Comparison of accuracy and certainty of results of six home pregnancy tests available over-the-counter. *Current Medical Research and Opinion*, *24*(6), 1645–1649. https://doi.org/10.1185/03007990802120572

Tribe, L. H. (1971). Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review*, *84*(6), 1329–1393. https://doi.org/10.2307/1339610

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *The Knowledge Engineering Review*, *10*(1), 43–62. https://doi.org/10.1017/S0269888900007256

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Academic Press.