Error Rates, Likelihood Ratios, and Jury Evaluation of Forensic Evidence

Brandon L. Garrett  
*Duke University*

William E. Crozier  
*Duke University*

Rebecca Grady  
*University of California-Irvine*

Follow this and additional works at: [https://lib.dr.iastate.edu/csafe_pubs](https://lib.dr.iastate.edu/csafe_pubs)

Part of the *Forensic Science and Technology Commons*

---

**Recommended Citation**

[https://lib.dr.iastate.edu/csafe_pubs/40](https://lib.dr.iastate.edu/csafe_pubs/40)

This Article is brought to you for free and open access by the Center for Statistics and Applications in Forensic Evidence at Iowa State University Digital Repository. It has been accepted for inclusion in CSAFE Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Error Rates, Likelihood Ratios, and Jury Evaluation of Forensic Evidence

Abstract
Forensic examiners regularly testify in criminal cases, informing the jurors whether crime scene evidence likely came from a source. In this study, we examine the impact of providing jurors with testimony further qualified by error rates and likelihood ratios, for expert testimony concerning two forensic disciplines: commonly used fingerprint comparison evidence and a novel technique involving voice comparison. Our method involved surveying mock jurors in Amazon Mechanical Turk (N = 897 laypeople) using written testimony and judicial instructions. Participants were more skeptical of voice analysis and generated fewer “guilty” decisions than for fingerprint analysis (B = 2.00, OR = 7.06, p =

Keywords
forensic science, likelihood ratios, error rates, judicial instructions, jury instructions, jury decision-making

Disciplines
Forensic Science and Technology

Comments
Brandon L. Garrett, J.D.; William E. Crozier, Ph.D.; and Rebecca Grady, Ph.D.

Error Rates, Likelihood Ratios, and Jury Evaluation of Forensic Evidence*

ABSTRACT: Forensic examiners regularly testify in criminal cases, informing the jurors whether crime scene evidence likely came from a source. In this study, we examine the impact of providing jurors with testimony further qualified by error rates and likelihood ratios, for expert testimony concerning two forensic disciplines: commonly used fingerprint comparison evidence and a novel technique involving voice comparison. Our method involved surveying mock jurors in Amazon Mechanical Turk (N = 897 laypeople) using written testimony and judicial instructions. Participants were more skeptical of voice analysis and generated fewer “guilty” decisions than for fingerprint analysis (B = 2.00, OR = 7.06, p = <0.000). We found that error rate information most strongly decreased “guilty” votes relative to no qualifying information for participants who heard fingerprint evidence (but not those that heard voice analysis evidence; B = −1.16, OR = 0.32, p = 0.007). We also found that error rates and conclusion types led to a greater decrease on “guilty” votes for fingerprint evidence than voice evidence (B = 1.44, OR = 4.23, p = 0.021). We conclude that these results suggest jurors adjust the weight placed on forensic evidence depending on their prior views about its reliability. Future research should develop testimony and judicial instructions that can better inform jurors of the strengths and limitations of forensic evidence.

KEYWORDS: forensic science, likelihood ratios, error rates, judicial instructions, jury instructions, jury decision-making

Forensic experts can provide powerful evidence in criminal cases, but as with any evidence, its evaluation is probabilistic, and errors can occur. In the United States, forensic expert testimony is regulated in federal and many state courts by the Supreme Court’s ruling in Daubert v. Merrell Dow Pharmaceuticals (1) and its progeny, as well as by revisions to Federal Rule of Evidence 702 (2) and state analogue rules regarding expert evidence. While those changes gave judges greater power to inquire into expert methods, including to consider a “known or potential rate of error,” (1, p. 594), judges have largely understood those rules as governing the initial question of the admissibility of an expert, but not the content of testimony once an expert takes the stand. However, lawyers, policymakers, and scientists have increasingly scrutinized forensic expert testimony (3). The Advisory Committee to the Rules of Evidence is considering an amendment stating federal judges should review how experts frame conclusions and testimony (4).

Although practices have changed a great deal in the past decade, fingerprint examiners had long reported error rates of zero and claimed to link evidence to a defendant, to the exclusion of all other persons in the world (5). In recent years, however, experts, judges, and prosecutors have announced new standards for testimony requiring that experts admit that error can occur, barring zero error rate claims, and requiring more cautious and noncategorical expressions of conclusions. For example, in 2018, the U.S. Department of Justice (DOJ) responded to these concerns and announced new guidelines for fingerprint testimony, that states examiners shall not assert “100% level of certainty,” but rather a “source identification” based on “extremely strong” support for the proposition that impressions came from the same source (6).

In this study, we examine the impact of providing jurors with testimony that attempts to further express limitations of a method in two ways: error rates and likelihood ratios, and for two types of forensic techniques: fingerprint comparison evidence and voice comparison evidence. First, scientists have endorsed the presentation of information in reports and testimony disclosing the overall rate of false-positive conclusions that incorrectly link evidence to a defendant. The Presidential Council of Advisors on Science and Technology (PCAST) report stated: “The expert should report the overall false-positive rate and sensitivity for the method established in the studies of foundational validity and should demonstrate that the samples used in the foundational studies are relevant to the facts of the case.” (3, p. 6). They recommended forensic evidence should only be presented in court with upper- and lower-bound estimates of error rates (referring to false-positive rates) in carefully done validation studies.

A second approach is to present a forensic conclusion not in categorical form, but as a likelihood ratio that incorporates information about the probability of seeing a given correspondence in order to express the probative value of a given conclusion. For example, if one pair of samples shared a very common feature, and one pair shared a very rare feature, they might both be classified as “matching,” but the latter would have a greater

1School of Law, Duke University, 210 Science Drive, Durham, NC.
2Department of Psychological Science and Department of Criminology, Law & Society, University of California-Irvine, 2340 Social Ecology II, Irvine, 92617, CA.
3Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the CSAFE.

*Supported by a grant from CSAFE under Grant Number SES-0922314.
†Corresponding author: Brandon L. Garrett, J.D. E-mail: bgarrett@law.duke.edu

likelihood ratio because the likelihood of seeing the former match just by chance is much greater. In recent years, some practitioners and researchers have advocated the use of likelihood ratios in both fingerprint comparison (7,8) and voice comparison (9). Importantly, likelihood ratios use numbers to describe the strength of a conclusion, although they are different types of numbers than error rates. Error rates communicate information about the accuracy of a categorical conclusion, while likelihood ratios typically describe the strength of evidence without requiring the examiner to make a categorical, match/no match, determination. As such, a juror hearing both types of information may find it difficult to reconcile how the two reliability factors fit together.

We do not know how jurors might react to hearing quantitative information about error rates along the lines of what the PCAST Report recommends, or second, likelihood ratio information. To date, courts in the United States have not insisted that jurors hear about such statistical information, whether in the form of an error rate or a likelihood ratio. A proposed amendment to Rule 702 would permit judges to require that a forensic expert state “the probable value” or “any similarity of match” with an empirical basis (4). Forensic examiners increasingly acknowledge verbally that no technique is fool-proof, but without explaining what the error rates or limitations of the techniques are in numerical terms. Thus, in latent fingerprint comparison, fingerprint examiners are cautioned by the International Association for Identification (IAI) that they may not assert an error rate of zero (10). Practitioners in many forensic science disciplines increasingly acknowledge that errors are possible in their standardized reporting language and that error rates are calculable and observable (11). One laboratory, the Defense Forensic Science Center, reports fingerprint results in reports and in trial testimony using a probabilistic method (12). Most recently, the Department of Justice, as noted, made a series of recommendations concerning language to be used in forensic techniques, such as latent fingerprint comparisons and firearms comparisons. The DOJ said that examiners should not assert “100% level of certainty” or claim that they have a zero error rate (13).

We examine how lay participants respond to forensic evidence when presented using a quantified error rate, a likelihood ratio instead of categorical conclusion, or both, using language for the error rates specifically as recommended by the PCAST report. The PCAST Report recommended jurors be told upper-bound false-positive error rates for latent fingerprint evidence are, based on current research, as high as one in 18 and one in 306 (3). It is not known how jurors will incorporate such information into their decision-making. In this study, we provide error rate information to laypeople, in conjunction with either a traditional identification conclusion or as a likelihood ratio of 3000. We note the goal is to examine whether judges can educate jurors by providing instructions following expert testimony, not to compare error rate versus likelihood ratio information because the two cannot be directly compared.

We seek to assess how such information affects jurors in two very different disciplines. First, latent fingerprint evidence has been presented in courts in the United States for over a hundred years, it is widely depicted in the media, and laypeople place great weight on it (5). Second, voice comparison evidence, in contrast, had been discontinued by the FBI after a critical National Academy of Sciences Report in the late 1970s, and only recently have new methods, not yet fully validated, been developed to compare human voice recordings (14).

The existing literature suggests jurors place great weight on forensic evidence in criminal cases. Several studies have asked laypeople for their general view on the accuracy of forensic evidence and found such views are not always accurate. Ribeiro, Tangen, and McKimmie (15) found 101 Australian lay participants' belief in the accuracy of forensic techniques varied, with the highest accuracy assigned to DNA testing and lowest accuracy assigned to document comparison. Koehler similarly found that for shoeprint comparison testimony, an admission that an error could occur (although not disclosure of an error rate) led to less weight placed on the evidence by mock jurors (16). Mitchell and Garrett found that laypeople, recruited using Amazon MTurk, reported a greater belief in the uniqueness of fingerprint comparison than of uniqueness of DNA evidence (5).

However, informing jurors about the existence of error rates may impact jurors and reduce the weight jurors place on forensic evidence. Mitchell and Garrett found that an admission by the fingerprint expert that errors can occur significantly reduced the weight that laypeople give to the evidence in a criminal trial (5).

A few studies have focused on error rates presented in a quantified manner, rather than a qualitative admission that errors are possible. Mitchell and Garrett (17) found that information about the proficiency of a fingerprint expert, or errors made in a routine test designed to assess the accuracy of the individual expert, had a consistent impact on jurors, reducing the weight that they place on the expert’s testimony. However, jurors did not distinguish between false-negative and false-positive errors; they aggregated both types of error rates.

Research on juror comprehension of likelihood ratios suggests that how jurors incorporate such statistical information may depend on which forensic technique is being used. Thompson and Newman examined how laypeople respond to two types of evidence, DNA evidence and shoeprint evidence, when experts provided likelihood ratios, random match probabilities, and verbal equivalents of likelihood ratios. They found that jurors adjusted the weight given to DNA evidence when they heard about sources of potential error. However, for shoeprint evidence, the additional information concerning error was not well incorporated. The authors suggested this might occur if laypeople placed far less weight on shoeprint evidence at the outset (18). Thus, laypeople attended to what experts said, but also incorporated their prior beliefs concerning the accuracy of the evidence.

A large body of research has focused on how jurors evaluate DNA evidence, which is presented in a quantitative fashion. Koehler (19) examined how jurors use either frequencies, likelihood ratios, or error rate information in the context of DNA evidence. Koehler found that likelihood ratios had the potential to confuse jurors, who may incorrectly treat them like posterior odds, or the probability that the defendant is the source. Scharl and Diamond (20) found that jurors made errors in evaluating probabilistic DNA evidence, in which both a likelihood ratio and random match probability were presented, but that also found jurors have their own prior views concerning probability of error, and will therefore make assumptions about the probative value of the forensic analysis whether they are presented with one or not.

Relatedly, the juror’s prior beliefs, based on other evidence they may have heard, about the likelihood the evidence originated from the defendant = in a case may also affect the impact of error information on their verdicts. Thompson, Kassa, and Peterson (21) examined how potential jurors assessing DNA evidence responded to information about the random match
probability as well as about a false report probability. They found that in one scenario, the error rate information had an impact, but that in a second scenario in which the defendant had a strong alibi, error rate information had no impact, perhaps since jurors were already skeptical of the case.

More research has focused on qualitative language that experts use that may affect jurors. These studies do not find that jurors draw sharp differences between terms used, perhaps similarly reflecting their prior views concerning the accuracy of the evidence. For example, Thompson, Grady, Lai, and Stern (22) found jurors weighted categorical “match” language as similar to a random match probability of one in ten million. In the area of latent fingerprint testimony, prior research suggests that there are not meaningful differences in the weight that laypeople place on different formulations of conclusions. In two studies conducted using lay participants recruited online using Amazon MTurk, Mitchell and Garrett (5) found that jurors consistently place great weight on fingerprint evidence. However, the authors found no statistically significant differences between a wide range of different commonly used ways that latent fingerprint experts express conclusions, using terms like “came from the same source,” “made by the same individual,” and “identification” and “individualization” as well as less forceful terms like “very likely,” “I am confident,” “practical certainty” and “reasonable scientific certainty” and “match.” One reason for that result may have been the strong association that laypersons have with fingerprint evidence and the presentation of infallible “unique” results. In contrast, the expert’s acknowledgment that error can occur had a significant impact on juror evaluations, because it challenged their view that such evidence is infallibly reliable.

Materials and Methods

This study used a mock trial format to study lay assessments of forensic evidence. In order to avoid corroborating nonforensic evidence affecting the weight that laypeople place on the forensic evidence, the scenario was intentionally kept simple and consisted primarily of a single forensic link between the criminal defendant and the crime. Excerpts from the written materials are set out in an Appendix A and complete materials are available on an Open Science Foundation website (and this link anonymizes all author information) at https://osf.io/6avzn/?view_only=c6874667b0d54b46b37028047209eb2. This study was approved by the Duke University Institutional Review Board, and we protected the rights of all human participants throughout the experiment.

We recruited 1428 participants, aged 18–78 (median = 34) using Amazon Mechanical Turk, who completed the survey online. Participants were excluded from analyses if they did not pass an attention or reading checks, if they were not a U.S. citizen, or if they came from suspicious, duplicate geolocations. At the time of data collection, there was a problem of respondents taking large numbers of social science surveys from the same geolocation and providing poor-quality data (23). We excluded three geolocations identified by TurkPrime as contributing most of the bad responses. This led to the greater rate of exclusions than in a typical MTurk study, and is the reason we kept our inclusion criteria strict by requiring that participants pass every attention and reading check. After applying these exclusion criteria, our sample consisted of 897 participants. Overall, the study was balanced with respect to gender (52.7%, self-identified as male; 47.3% as female). Participants self-reported as 9.6% Black, 7.6% Asian, 80.7% White, 6.2% Latino, 1.4% Native American or Pacific Islander, and 0.3% Arab or Middle Eastern. Participants could select more than one race/ethnicity, so percentages sum to more than 100%. By education, 51.3% had at least 4-year college degree, 48.7% did not, and 11.7% had a postgraduate degree.

Participants were asked to self-identify their political preferences: 43.5% stated they affiliate with the Democratic party, 22.5% stated they affiliate with the Republican party, 17.4% identified a different party, and 16.8% stated no party preference. Twelve percent had a self-reported annual household income of less than $20,000 and 6.6% had an income above $120,000. The majority of the sample (50.9%) fell between $20,000 and $60,000 per year. About 18.4% of the participants reported having served previously on a jury. Each participant received $1 for completing the survey, which participants completed in a median of 8.7 min.

The participants first provided informed consent and responded to basic eligibility questions such as age and U.S. citizenship. Each participant then read, as described, a written scenario describing a simple criminal offense and police investigation. All participants read a brief description of a convenience store robbery, in which the store clerk testified she could not see the robber’s face but that the person had pulled out a gun and demanded money. The robber rushed out of the store and dropped his gun when his hand caught on the door. A detective arrested a person, the defendant in this case, who was found in the vicinity shortly after the robbery, without any of the stolen proceeds of the crime.

Following that scenario, the participants were randomly assigned to one of eight different conditions, with over one hundred participants in each. The study used a 2 (Evidence: Fingerprint vs. Voice Comparison) × 2 (Identification: Categorical or Likelihood Ratio) × 2 (Instructions: Generic vs. Error rate) between-subjects design. Each participant read a report by either a fingerprint or voice comparison expert. The participants who heard about fingerprint evidence were told police located a fingerprint on the handle of the gun, which was then compared with the known fingerprints taken from the defendant on an inked card. The fingerprint examiner’s report was then provided, which, based on authoritative descriptions of the methods used in the latent fingerprinting, briefly explained the method used in latent fingerprint comparisons, known as the ACE-V method.

In voice comparison conditions, the participants were told that the police obtained a recording of the event captured on a security camera within the store that took audio and video of the incident. A voice comparison expert compared a recording taken of the defendant’s voice to the audio recording of the robber at the convenience store. The jurors read the voice examiner’s report describing the methods used.

In the categorical conclusion conditions, a simple conclusion of identification was presented. The examiner testified, following recommended uniform language adopted by the U.S. Department of Justice, that fingerprints or voice samples “originated from the same source” (13). In the other conditions, a likelihood ratio was presented. The examiner explained what a likelihood ratio is and concluded: “On the basis of my comparisons, it would be 3000 times more likely to get the fingerprint details observed on the latent fingerprint (or acoustic properties on the offender recording) had they been produced by the Defendant than had they been produced by some other person from the relevant population.” This number was chosen to be reasonably strong to match the strength of the identification conclusion.
The jurors were then provided with judicial instructions, which in the basic conditions asked jurors to assess the expert testimony, including by taking note of the fact that any analysis involves some error rate greater than zero. However, the judge noted there are no studies in the field measuring actual error rates. In the error rate condition, an additional portion of the judicial instructions was provided, following the recommendation of the PCAST Report, and stating two studies of an “upper bound” on the “false-positive rate” in the evidence type have been done, and that one study’s estimate was “1 in 306,” and the other study’s estimate was “1 in 18.”

We note an excellent point raised by a reviewer that because error rate information and likelihood ratio conclusions communicate incompatible information, the two cells that contain a likelihood ratio conclusion and error rate instructions may be unrealistic. We stand by this design and subsequent analyses for several reasons. First and foremost, we wanted to be able to compare and tease apart the effects of both types of information independently, and in conjunction with each other—which a fully crossed design allows us to do. This design allows us to observe whether participants are critically considering error rate information or likelihood conclusions. If, as logic would suggest, the two types of information are incompatible, then cells that include both should yield no additional reduction in guilty decisions, compared to cells in which only one is presented. If, on the other hand, jurors are not logically considering the meaning of such information, we may see an unwarranted decrease in guilty decisions.

Second, although the two types of information are not compatible, it is not impossible that jurors could hear both. Although not an element of our design here, it is possible that both types of information could be raised at different parts of trial (e.g., a likelihood ratio conclusion on direct examination and error rate information offered in response to questions regarding error rate studies on cross). Third, reporting and analyzing these data with two missing cells by excluding the unrealistic ones could lead to false positives in our statistical reporting (24).

In all of eight conditions, the written materials concluded as follows: “The Defendant maintains that he has nothing to do with the robbery and was just walking from his home to the grocery store at the time of the robbery.”

After reading the materials and jury instructions, participants were provided with a final instruction from the judge concerning a “beyond-a-reasonable-doubt” standard and were asked whether they would vote to find the defendant guilty: our primary dependent measure. Participants were also asked about the type of crime in the case and whether the eyewitness identified the defendant as the culprit in the case (which the eyewitness did not). Given that these were basic case facts, participants who answered either incorrectly were excluded from analysis.

Additional exploratory measures for which we have no specific hypotheses included asking them to rate the strength of the case (1 “Extremely weak” to 5 “Extremely strong”), the accuracy of DNA, fingerprint, and voice comparison evidence in general (1 “Not reliable at all” to 5 “Extremely reliable”) and the likelihood that the defendant is actually the person that committed the robbery (0–100% chance). Participants were asked whether they thought falsely convicting an innocent person was the worse error, failing to convict a guilty person was a worse error, or whether they were both equally bad.

Finally, participants were asked a series of demographic questions, such as gender, race, political views, education, income, previous jury service, or previous experience with criminal justice supervision or a sexual assault. All study materials, as well as underlying data, are available on an Open Science Foundation website; the mock trial materials are also set out in the accompanying Appendix A.

Overall, the study took approximately 6–12 min to complete.

We tested three primary hypotheses in this experiment. First, we hypothesized jurors would place less weight on voice comparison testimony (as shown by lower guilty vote rates) as they would for fingerprint testimony, given the longstanding use of fingerprint testimony and wide public knowledge of the use of fingerprints and the claims that fingerprints are unique. In contrast, participants should be less familiar with voice analysis and naturally put less weight on forensic testimony relying on the method. Second, we hypothesized participants who heard jury instructions that provide error rate information would put less weight on forensic evidence (and thus vote guilty less often) than participants who hear traditional and generic instructions that lack such error rates, and instead merely inform that experts may testify to opinions and should be evaluated based on their experience and training. Third, we hypothesized that providing a likelihood ratio will reduce the weight that jurors place on forensic expert testimony, as compared with testimony offering an unequivocal and categorical conclusion of an identification or match.

Finally, we make no firm predictions about the interactions between these variables and treat them as exploratory. We were not sure whether the combination of error rate and likelihood information would have a greater impact on jurors, given the differences between those concepts (and findings in prior studies regarding DNA evidence). Nor were we sure whether error rate information would be more salient for a familiar type of forensic evidence perceived to be extremely strong, like fingerprint evidence, versus a relatively unfamiliar voice comparison.

Results

On the primary measure, whether lay participants voted the defendant guilty of the crime, we found an overall guilty vote rate of 33.78%. Recall that the evidence in this mock case almost entirely consisted in a forensic comparison, with no other evidence linking the defendant to the crime. However, the guilty vote rate varied greatly by condition, as tested by logistic regression of guilty vote rates by the three experimental manipulations, shown in Table 1 (with an overall significant model, $\chi^2 = 98.686, p < 0.001$). All guilty vote rates by condition are shown in Fig. 1, displayed below. Next, Table 1 displays the regression results.

There was a significant three-way interaction between conclusion type, error rate, and evidence type ($p = 0.021$). This interaction is based on the differential effect of error rates on verdicts between evidence types: essentially, a significant interaction between conclusion type and error rate when hearing fingerprint evidence, but not when hearing voice analysis. Specifically, for fingerprint evidence, error rate information decreases guilty verdicts for an ID conclusion, but has little to no effect when the expert is using a likelihood ratio. However, neither the error rate nor likelihood ratio information affected conviction rates for participants in the voice comparison groups (conclusion type $\times$ error rate interaction, $p = 0.084$) for ID conclusion. Thus, when considering fingerprint evidence, instructions on error rates decrease guilty verdicts when the expert is making a simple identification conclusion; when the fingerprint expert is offering a likelihood ratio conclusion, error rate instructions do not
were presented ($p$ Model $v/Co$ Constant evidence case ($9 Conclusion type 9 Conclusion type $9 p$ at the standardized regression coefficient. Bolded predictors are statistically significant 897 $N$ (are significantly fewer guilty verdicts than fingerprint evidence decrease guilty verdicts. When considering voice evidence, there is an impact of condition on the ratings of the reliability of fingerprint evidence, and the error rate information exacerbated this. No other condition or interaction in the three-way model was significant, nor was there an impact of any condition on response to the reliability of voice evidence in general (all $p > 0.05$).

We also asked participants a question about risk aversion in the context of criminal cases generally, a question that relates to people’s preferences regarding the standard of proof in a criminal case. This question has been posed in several prior studies concerning lay evaluation of forensic evidence (5,12). We asked, “Which type of error is worse?” and 46.9% (421 of 897) answered “erroneously convicting an innocent person”; 6.9% (62 of 897) chose “failing to convict a guilty person”; and 46.2% (414 of 897) responded “the errors are equally bad.” We ran a series of one-way ANOVAs with their response to this as the independent variable and a variety of views on evidence and justice system as outcomes to see how this belief affected other relevant views about convictions. As displayed in Table 2 below, people who were more concerned about failing to convict a guilty person were more likely to vote guilty, believe the case was stronger, have a higher opinion of the reliability of voice and fingerprint evidence, and believe more people brought to criminal trial are guilty (all $p < 0.01$; see Table 2 for means).

2 × 3 Reanalysis

In addition to our primary 2 × 2 × 2 analyses above, we also conducted 2 × 3 analyses, at the urging of Reviewer 1. Specifically, because the three-way interaction here was particularly difficult to understand and had low power than other analyses, we also collapsed our design to exclude the Likelihood Ratio + Error Rate conditions into a 2 (evidence: fingerprint, voice) × 3 (Identification + No Error Rate, Identification + Error Rate, and Likelihood Ratio) design. In the interest of space, we summarize these analyses here, and the full analyses and output can be found in Appendix B. We note that this analysis was not our complete experimental design and is reanalyzing much of the same data as above; as such, we rely less on these results than the primary 2 × 2 × 2 analysis.

![FIG. 1—Evidence type and proportion of participants finding defendant guilty.](image)
Discussion

The type of forensic evidence matters when considering how to present forensic testimony. More specifically, and consistent with prior studies focusing on probabilistic DNA evidence, laypeople already have prior views regarding the reliability of forensic evidence. They may incorporate those views whether there are given information about a likelihood ratio or error rate, or not. However, their views may not be empirically based, but rather based on cultural perceptions. Thus, we found support for our first hypothesis that fingerprint evidence was given more weight than voice comparison evidence. Prior work has documented how laypeople place great weight on latent fingerprint comparison evidence; they assume that a fingerprint is “unique” and that comparisons are highly reliable (5). We expected participants would be fairly unfamiliar with voice comparison evidence and that they might be skeptical of its strength.

These results also suggest that a novel and unfamiliar type of forensic evidence may not provide as powerful evidence of guilt. This has potential implications for new techniques developed, using new technology that might be unfamiliar to jurors. The two forensic techniques studied here were presented as having the same error rates. In the future, new forensic techniques will no doubt be developed that are more reliable and have lower error rates than those used in the past. Future research should assess whether jurors undervalue the reliability of those new techniques, due to their novelty. In a domain that has come under question in the past, it is not unreasonable for jurors to be skeptical of voice comparison. If new methods are validated and have comparatively low error rates, it will be important to study whether participants give appropriate weight to these new methods.

This research also has implications for efforts to regulate forensic testimony. Currently, the US Federal Rules of Evidence 702 already requires that “(b) the testimony is based on sufficient facts or data” and “(c) the testimony is the product of reliable principles and methods” (2). As noted, a proposed amendment to Rule 702 would permit judges to require that a forensic expert state “the probative value” of “any similarity of match” with an empirical basis (4). Scientists, including noted statisticians, have called for an increasing focus on presenting sound statistical information when experts testify, particularly in criminal cases (25). These results suggest that some such quantitative information can inform jurors, but that attention must also be paid to the preconceptions that jurors have about the reliability of evidence.

Our second hypothesis was that presenting an error rate would reduce the guilty vote rate in a trial based on forensic evidence. We found this to be true paired with a fingerprint identification but not with voice identification. This is consistent with the findings of prior experimental work which has found that error rate information can partially counter the strong weight that laypeople place on fingerprint evidence (5). For those hearing an identification with voice evidence, the guilty vote rates were similar across all instruction and conclusion conditions. Error rate information did appear to reduce guilty vote based on fingerprint evidence when hearing a simple identification conclusion.

This finding is also consistent with prior research on jury interpretation of quantitative DNA evidence, noting the role that lay priors regarding reliability plays when interpreting statistical conclusions (20). Thus, error rate information may not just induce skepticism. Indeed, it may be that for types of evidence that laypeople view with undue skepticism, error rate information may cause laypeople to adjust the weight that they place on the evidence upwards.

Our third hypothesis was not met: likelihood ratio information did not have a significant impact on guilty vote rates, either for

### TABLE 2—Error preferences and jury assessment of voice and fingerprint evidence.

<table>
<thead>
<tr>
<th>Which Type of Error They Believe is Worse</th>
<th>Convicting Innocent Person (n = 421)</th>
<th>Failing to Convict Guilty Person (n = 62)</th>
<th>ANOVA Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guilty vote rate</td>
<td>25% (26.77)a</td>
<td>58% (23.57)b</td>
<td>p &lt; 0.001*</td>
</tr>
<tr>
<td>Likelihood that defendant actually committed the crime</td>
<td>53% (27.57)a</td>
<td>68% (23.57)b</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Strength of prosecution’s case</td>
<td>2.45 (0.98)</td>
<td>3.18 (1.09)</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>General reliability of fingerprint evidence</td>
<td>3.51 (1.00)</td>
<td>3.92 (0.84)b</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>General reliability of voice comparison evidence</td>
<td>2.33 (0.92)a</td>
<td>2.82 (0.86)</td>
<td>p &lt; 0.001</td>
</tr>
</tbody>
</table>

Standard deviations in parentheses. Means that share superscripts of the same letter do not significantly differ from each other (p < 0.05 with Tukey’s adjustments) across a row.

*Guilty vote rate was done as a $\chi^2$ rather than ANOVA because it is a categorical outcome.

First, the reanalysis largely supported our findings from the $2 \times 2 \times 2$ analysis that error rate information decreased guilty votes for fingerprint evidence, but not voice comparison evidence. Further, likelihood ratio information did not affect guilty votes in any conditions. Specifically, we found that participants who heard fingerprint evidence were more likely to vote guilty than those that heard voice evidence analysis ($B = 1.96, p < 0.001, OR = 7.06, 95% CI [3.75, 13.31]$). We also found a significant interaction between Instruction/Conclusion type and Evidence type, such that the judge’s instructions concerning error rates caused guilty vote rates to drop (relative to the condition where the error rates were not presented), but only for fingerprint evidence and only when the fingerprint expert testified to an identification, rather than presenting a likelihood ratio ($B = -1.16, p = 0.007, OR = 0.32, 95% CI [0.14, 0.73]$). We did not find this same trend in participants who heard the voice comparison analysis, however.

Main effects of each conclusion type (likelihood ratio, and identification + error rates) were not significant ($p = 0.664$, and $p = 0.140$, respectively), and the Instruction/Conclusion type by Evidence type interaction for likelihood ratio conclusion was nonsignificant as well ($p = 0.129$). As such, we found no evidence that likelihood ratio conclusions lowered conviction rates for either types of evidence. Thus, as a caveat to the the $2 \times 2 \times 2$ analyses above, when likelihood conclusions with Error Rate information conditions are excluded, likelihood conclusion seem to do very little, if anything.
voice evidence or fingerprint evidence. It may be the case that, unlike error rate information, likelihood ratio results are not fully comprehended by laypersons. This may come as a surprise to some practitioners and scholars who have advocated for one approach toward presenting the limitations of forensic evidence over the other. Debates that have become quite heated, both in forensics and in statistics, regarding whether likelihood ratios are either always, or never appropriate for use in legal settings and what other information is needed to qualify them, if any (26,27).

We do not intervene here in those debates concerning which approach is more faithful to scientific and statistical assumptions; instead we focus on which is more impactful to lay jurors’ understanding—which, based on our results, appears to be error rate information.

Similarly, the two methods cannot be directly compared with each other; an error rate and a likelihood ratio express related concepts, but in conceptually different ways. What we find is that the two approaches may be differentially understood by laypeople and the subsequent impact of how this information relates to the case. Perhaps each type of information cues participants to fingerprint evidence having some room for interpretation or error—either that the methodology in general (error rates) or the specific analysis (likelihood ratios)—but here, we see only error rate information affects guilty decisions. Presenting participants with different error rates, independent of or in conjunction with, different likelihood ratios would likely tease apart how each type of information is affecting jurors. For example, given the closeness of the fingerprint guilty verdict rate with a likelihood ratio only (46.2%) and with an identification plus and error rate (42.4%), and how much lower they were than presenting an identification conclusion without either (59.2%), it is possible that a slightly lower likelihood ratio than the 1 in 3000 number given here may work similarly to an error rate in lowering guilty vote rates in another study. Further research on likelihood ratios is warranted however, as these conclusions are drawn primarily from the significant three-way interaction, which as we noted before is not the strongest evidence because of the unrealistic experimental conditions underlying a lower-powered analysis.

Our findings support the view that quantified suggestion of possible error can reduce the strength that jurors place on forensic evidence, but we did not see likelihood ratios having this effect. Further, this effect is not a one-way ratchet. Indeed, it may be particularly important if jurors have prior views that evidence is extremely reliable in general. However, there is more than one way to inform jurors of prior perceptions of the strength of forensic evidence. We are unable to determine exactly why error rate information was more impactful than a likelihood ratio. One possibility is that error rate information was more comprehensible than likelihood ratios. Similarly, it may be that jurors could more intuitively apply error rate information than likelihood ratio information to a specific case. Future research should consider why likelihood ratio information did not reduce verdicts—particularly if there is a desire for likelihood ratios to be presented to jurors. Further, we do not know the effect varying the strength of information provided by error rates and likelihood ratios on jurors—perhaps more extreme values than what we used here, either positive or negative, could have a differential impact on jurors.

No U.S. courts currently provide judicial instructions of the type that we included in this mock trial study, nor do they require that experts themselves express the limitations of their own disciplines, at least in a quantified manner or with respect to measured error rates. However, as discussed, proposals to encourage if not require that empirical evidence accompany conclusions regarding the probative value of expert evidence are increasingly being considered. These findings suggest such judicial instructions are worth consideration, but equally worth consideration are requirements that examiners themselves express the limitations of findings.

These findings also suggest that it is important to look at laypeople’s preconceived notions of evidence strength. As Thompson and Newman (18) put it: “The findings indicate that perceptions of forensic science evidence are shaped by prior beliefs and expectations as well as expert testimony and consequently that the best way to characterize and explain forensic evidence may vary across forensic disciplines.” Expressing limitations of the evidence may have induced more skepticism of fingerprint evidence, which laypeople place a great deal of weight on, but possibly induced less skepticism of the voice comparison evidence, which laypeople did not place as much weight on. Of course, other potential mechanisms exist as well, such as an experimental artifact of participants giving such little regard to voice analysis that they exhibited a floor effect. More research should be done to examine what relative weight laypeople place on a range of other types of commonly used forensic evidence, and how that affects efforts to express the limitations of those types of evidence.

We found that people who were most concerned with convicting an innocent person, reflecting the beyond-a-reasonable-doubt standard in criminal cases, were less likely to vote guilty, believed the case was weaker, and were less likely to view the evidence as reliable. In contrast, people who were most concerned with an error resulting in the release of a guilty person were more likely to vote guilty, believe the case was stronger, and have a higher opinion of the reliability of both voice and fingerprint evidence. The guilty vote rate among people who stated that freeing a guilty person was the worse type of error was 58%, compared with a guilty vote rate of 25% among people who stated that convicting the innocent is the worse type of error. These findings are similar to those observed in prior studies regarding lay assessment of forensic evidence, including Mitchell, Scurich, and Garrett (12), which focused on the role of risk aversions when laypersons evaluated proficiency information concerning forensic experts. The findings raise troubling questions regarding how effective the beyond-a-reasonable-doubt instruction, which was provided verbatim in this mock trial scenario, is in informing jurors of their obligations under the Constitution. The findings raise broader questions regarding whether more attention must be paid to the risk preferences of jurors in criminal cases and how they relate to their acceptance of ambiguous forensic evidence.

While these findings suggest that lay jurors can take into account new information about the reliability of forensic evidence, there are limitations of the methods used. We used a mock trial model. We did not ask jurors to deliberate in reaching a verdict; they were surveyed individually. However, studies have found that perceptions of the strength of the prosecution case are strong predictors of jury verdicts. Indeed, as Bornstein and Greene (28) have summarized, in reviewing the literature on the psychology of jury decision-making, that “[i]n approximately 90% of trials, the position favored by the majority at the beginning of deliberations becomes the jury verdict.” That said, future research should examine deliberation dynamics, with respect to scientific evidence specifically.

As noted, we did exclude participants that did not correctly answer a manipulation check. The written materials were fairly
short, to focus jurors on the presentation of the forensic evidence itself. They did not include cross-examination, opening or closing statements, or other features of a full criminal trial; our goal was not to assess verdicts in an entire trial presentation. Nor did participants see a defendant (and future studies could use an image of a mock-defendant), or otherwise feel pressed to make a real-life momentous decision whether to convict a person. Future studies can further examine whether presenting information limiting forensic conclusions affects outcomes, using more lengthy and videotaped trial materials. Such efforts will have to control for the confounding factors that arise in a full criminal trial, such as the appearance and manner of the forensic expert, lawyers, judge, and defendant. As described, our findings are consistent with prior mock trial studies, in person and online, finding that laypeople place great weight on forensic evidence and that many traditional judicial remedies are not effective in reducing that weight (29).

Conclusion

We find both reason for optimism and caution. We find that jurors have prior views regarding the reliability of forensic evidence and that in adjusting those priors, they can take into account new information about the reliability of that evidence. First, laypeople gave more weight to fingerprint evidence than voice comparison evidence. Second, presenting an error rate moderated weight of evidence only when paired with a fingerprint identification. This suggests that error rate information is particularly important for types of forensic evidence that people may assume are highly reliable. Third, likelihood ratio information did not significantly alter verdicts up or down, suggesting it is an ineffective way to communicate the limitations of forensic testing and reduce the guilty vote rate when compared to an identification presented without an error rate. If likelihood ratios are to be used, more research is needed to discover how they should be presented, and in what contexts, for them to effectively inform jurors.

Thus, we find that the same error rate information does not impact all forensic evidence types equally. Laypeople do not blindly assume that forensics are reliable and laypeople can take into account information about the strengths and limitations of forensics. In doing so, laypeople adjust their views about forensic evidence, and not necessarily downwards. Error rate information was neutral in its effect on persons evaluating voice comparison evidence. This has implications as new forensic technologies and techniques are introduced in court. More research should be done to examine what relative weight laypeople place on a range of other types of commonly used forensic evidence.

These findings suggest we should study requirements that experts explain strengths and limitations of the evidence in testimony, or judicial instructions designed to inform jurors of strengths and limitations of forensic evidence. No U.S. courts currently impose such requirements or use such instructions. Proposals to introduce information concerning the limitations of forensic methods are worth further consideration. Nevertheless, our results suggest that if jurors have strong views concerning the reliability or unreliability of a type of evidence, the impact of disclosures or instructions may be blunted. In short, reliability matters to jurors, error rates can inform jurors of strengths and limitations of evidence, but their impact will depend on the prior views of jurors regarding reliability of forensics.

References

Appendix A

Mock Jury Materials

All participants were given the following information about a criminal trial:

A convenience store was robbed. The clerk took the stand and stated that she did not get a close enough look at the robber’s face to make an identification because the robber wore a ski mask. Next, the detective gives a statement about arriving at a convenience store. The clerk told the detective that a man wearing a black ski mask walked into the convenience store, pulled out a gun, and demanded that the cashier hand over the money in the cash register.

As the cashier was pulling out the money to hand to him, she pressed a hidden button that activated an alarm and called the police.

Startled by the alarm bells, the robber rushed out of the store and dropped his gun when his hand caught on the door.

The detective arrested a person, the Defendant in this case, who was found in the vicinity shortly after the robbery. The detective apprehended the Defendant based on the description but admitted that the description was generic and could have matched other men as well. He did not find proceeds of the crime on the Defendant.

Following that scenario, the participants were randomly assigned to eight different conditions, with over one hundred participants in each. Four of the conditions involved fingerprint evidence and four involved voice comparison evidence. The participants who heard about fingerprint evidence were told:

The police identified a fingerprint on the handle of the gun dropped at the crime scene. A fingerprint examiner compared fingerprints taken from the handle of the gun dropped at the crime scene to the known fingerprints taken from the Defendant on an inked card.

The fingerprint examiner’s report is then provided, and it states as follows:

I formed the following opinion and interpretation using accepted scientific and professional practices:

1. Processing and analysis of the gun recovered from the crime scene revealed one fingerprint suitable for comparison.
2. I compared this fingerprint to the known fingerprints taken from the defendant on an inked card.
3. I followed a process known as ACE-V, which stands for Analysis, Comparison, Evaluation, and Verification. In this process, the examiner first determines whether the print is suitable for comparison, based on whether there is adequate information in it. If so, then the examiner makes a side-by-side comparison between the latent print and a known print taken from a suspect.

Fingerprints are formed before birth, and they remain the same until after death, barring deep scarring. The underside of our fingers and hands and feet are covered with raised skin, called friction skin, which is usually covered with a thin film of perspiration or oil. When the finger or hand touches an item, a reproduction of those ridges is left by means of that perspiration or oil. That reproduction is called a latent print.

The examiner evaluates the degree of similarity between the latent print and the known print and can conclude that the prints are the same, different, or that the comparison is inconclusive. Finally, if the examiner decides that the prints could have come from the same person, the prints are passed to another examiner, to verify that conclusion.

In the first and second conditions, a simple conclusion of identification was presented. The examiner testified as follows:

4. Based on the information and analysis described, I made a conclusion that the that the two friction ridge skin impressions, the crime scene print and the Defendant’s print, originated from the same source. This conclusion is my decision that the observed friction ridge skin features are in sufficient correspondence such that I would not expect to see the same arrangement of features repeated in an impression that came from a different source and insufficient friction ridge skin features in disagreement to conclude that the impressions came from different sources.

In the third and fourth conditions, a likelihood ratio was presented. The examiner testified as follows:

5. I used a statistic called the likelihood ratio to measure the strength of my findings. A likelihood ratio indicates how much more likely it would be to observe the friction ridge skin features found in the crime scene print if the culprit was the Defendant than if the culprit was some other person from the relevant population. On the basis of my comparisons, it would be 3000 times more likely to get the fingerprint details observed on the latent fingerprint had they been produced by the Defendant than had they been produced by some other person from the relevant population.

The jurors were then provided with judicial instructions, which in the basic conditions read as follows:

You have heard the testimony of a fingerprint examiner, who claims special qualification in the field of fingerprint comparison, including the classification of association between a latent print (from the Defendant) and a crime scene print.

Just because a witness is allowed to offer opinion testimony does not mean that you must accept his or her opinion. When you decide whether you believe the expert witness’ opinion, you may consider the method or technique or the witness used, and whether there is a strong enough scientific basis for drawing conclusions from that method or technique. You may want to consider how much of the method is objective, and how much is subjective in making the comparison.

All forensic analysis methods, like any human endeavor, are subject to error and have some error rate greater than zero. Not even highly automated tests have a zero error rate. The error rate of a fingerprint examiner’s method cannot be inferred from that examiner’s case work alone, nor the examiner’s expression of confidence about his opinion or about the accuracy of the field. Instead, a fingerprint examination method’s error rate can only be determined from conducting scientific experiments that actually test how often such examiners get the right answer. These studies have not yet been conducted in this field but you can assume the error rate is not zero.

In the error rate condition, an additional portion of the judicial instructions were provided (replacing the last paragraph above stating that there is some error rate but none that has been measured) with the text as follows:
To determine whether a fingerprint examination method is scientifically valid, you must also consider its error rate, or likelihood of producing an inaccurate conclusion. All forensic analysis methods, like any human endeavor, are subject to error and have some error rate greater than zero. The error rate of a fingerprint examiner’s method cannot be inferred from that examiner’s case work alone, nor the examiner’s expression of confidence about his opinion or about the accuracy of the field. Instead, a fingerprint examination method’s error rate can only be determined from conducting scientific experiments that actually test how often such examiners get the right answer.

In this field, there have been two studies of something called an “upper bound” on the “false-positive rate” in fingerprint comparisons based on scientific testing, which involves giving certified examiners samples with known sources to judge and looking at what conclusions are drawn. One study’s estimate was given as “1 in 306,” and the other study’s estimate was given as “1 in 18.”

Let me explain in more detail what those numbers would mean, if you choose to accept them. A “false-positive rate” is one type of error rate estimating how often prints that actually came from two different people are incorrectly judged to be from the same person by examiners who are given two prints of known sources to compare. A false-positive rate of 1 in 306 means that for every 600 print comparisons that actually come from different people, we expect examiners to incorrectly classify them as likely coming from the same person about 2 times. A 1 in 18 false-positive rate would mean that there would be about 30 prints classified as likely coming from the same person out of 600 print comparisons that actually come from different people.

Four additional conditions involved voice comparison evidence. In those conditions, the participants were told:

The police obtained a recording of the event captured on a security camera within the store that took audio and video of the incident. A voice comparison expert compared a recording taken of the Defendant’s voice to the audio recording of the robber at the convenience store.

The voice comparison examiner issued a report stating:

1. Processing and analysis of the recording from the crime scene revealed a voice sample suitable for comparison.
2. I compared this voice sample to a known voice recording taken from the defendant.
3. I followed a process to make acoustic measurements, using measurements called MFCC’s or mel-frequency cepstral coefficients. They are a standard measurement used in speech processing and speaker recognition systems. They measure the frequency components of speech. These acoustic properties vary among individuals for several reasons. For example, individuals vary physiologically in the size of their vocal tracts. People also vary in the way they have learned to speak. People with different accents, for example, will sound different. Voices are not necessarily unique. Voices of different people may be highly similar due to similarity in their physiology and how they speak. It is necessary to calculate probabilistically how similar two voices are and how typical they are in the relevant population. The examiner evaluates the degree of similarity between the two voices and generates a result. Finally, if the examiner decides that the voices could have come from the same person, the recordings are passed to another examiner, to verify that conclusion.

As with the fingerprint testimony, in the likelihood ratio conditions, the voice examiner explained:

4. I used a statistic called the likelihood ratio to measure the strength of my findings. A likelihood ratio indicates how much more likely it would be to observe the acoustical properties found in the crime scene recording if the culprit was the Defendant than if the culprit was some other speaker from the relevant population. On the basis of these comparisons, it would be 3000 times more likely to get the acoustical properties on the offender recording had they been produced by the Defendant than had they been produced by some other speaker from the relevant population.

In the basic identification condition, however, the expert simply “concluded that the voice on the recording was identified as the voice of the defendant.” Similarly, the judicial instruction either generally instructed, with the same language as with fingerprinting, that jurors are to consider the opinions of experts in context. In the error rate conditions, the jurors were told that two studies had been done of false positives and false negatives in the field of voice comparison, and that “One study’s estimate was given as ‘1 in 306,’” and the other study’s estimate was given as ‘1 in 18.”

In all eight conditions, the written materials concluded as follows: “The Defendant maintains that he has nothing to do with the robbery and was just walking from his home to the grocery store at the time of the robbery.”
Appendix B

TABLE A1—Logistic Regression results for “guilty” decisions for a 2 (evidence type: voice reference group, fingerprint) × 3 (conclusion type: identification + no error reference group, identification + error, likelihood ratio) analysis

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>AIC</th>
<th>BIC</th>
<th>df</th>
<th>( \chi^2 )</th>
<th>( p )</th>
<th>McFadden ( R^2 )</th>
<th>Nagelkerke ( R^2 )</th>
<th>Tjur ( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0 )</td>
<td>882.1</td>
<td>884.106</td>
<td>888.627</td>
<td>678</td>
<td></td>
<td></td>
<td>70.873</td>
<td>&lt;0.001</td>
<td>0.080</td>
</tr>
<tr>
<td>( H_1 )</td>
<td>811.2</td>
<td>823.234</td>
<td>850.357</td>
<td>673</td>
<td>70.873</td>
<td>&lt;0.001</td>
<td></td>
<td>0.080</td>
<td>0.136</td>
</tr>
</tbody>
</table>

Coefficients

<table>
<thead>
<tr>
<th>Estimate</th>
<th>SE</th>
<th>Odds Ratio</th>
<th>( z )</th>
<th>( p )</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.587</td>
<td>0.259</td>
<td>0.205</td>
<td>-6.135</td>
<td>&lt;0.001</td>
<td>0.123</td>
</tr>
<tr>
<td>Evidence (Fingerprint)</td>
<td>1.955</td>
<td>0.323</td>
<td>7.062</td>
<td>6.046</td>
<td>&lt;0.001</td>
<td>3.747</td>
</tr>
<tr>
<td>Conclusion (Categorical + Error)</td>
<td>0.500</td>
<td>0.339</td>
<td>1.649</td>
<td>1.477</td>
<td>0.140</td>
<td>0.849</td>
</tr>
<tr>
<td>Conclusion (Likelihood Ratio)</td>
<td>0.154</td>
<td>0.355</td>
<td>1.167</td>
<td>0.434</td>
<td>0.664</td>
<td>0.582</td>
</tr>
<tr>
<td>Conclusion (Cat + Error) * Evidence (Fingerprint)</td>
<td>-1.156</td>
<td>0.430</td>
<td>0.315</td>
<td>2.689</td>
<td>0.007</td>
<td>0.136</td>
</tr>
<tr>
<td>Conclusion (LR + NoError) * Evidence (Fingerprint)</td>
<td>-0.676</td>
<td>0.445</td>
<td>0.509</td>
<td>-1.520</td>
<td>0.129</td>
<td>0.213</td>
</tr>
</tbody>
</table>

Verdict level “1” (guilty) coded as class 1. Bolded variables are significant predictors of the outcome variable at \( p < 0.05 \) level.