INTUITIVE ERROR RATE ESTIMATES FOR THE FORENSIC SCIENCES

Jonathan J. Koehler*

ABSTRACT: There is a risk of error associated with all forensic science conclusions. However, no proper studies have been conducted in any of the forensic sciences to estimate the rate at which forensic science conclusions are wrong. As a result, jurors and other consumers of forensic science information are left to speculate about what those error rates might be. The present paper provides data from an online study (n = 210) about what jury-eligible people (“jurors”) think the false positive error rates are for five types of forensic science evidence (DNA, fingerprints, bite marks, microscopic hair, and handwriting). Jurors’ median estimates range from a high of 1 in 100,000 (handwriting) to a low of 1 in 1,000,000,000 (DNA). The significance of these estimated false positive error rates is discussed.


I. DEARTH OF DATA ON FORENSIC SCIENCE ERROR RATES

Nobody knows the rate at which forensic science examiners produce false match reports1 or otherwise reach the wrong conclusion. Unlike in medicine,2 there is no systematic research program in the forensic sciences that seeks to identify the rates at which errors are made in practice.

---

1. A false match report (or false positive error) occurs when an examiner falsely concludes that two forensic items (e.g., hairs, DNA samples, fingerprints, tire tracks) come from a single source when, in fact, they were produced from two different sources. See COLIN AITKEN ET AL., PRACTITIONER GUIDE NO 1: FUNDAMENTALS OF PROBABILITY AND STATISTICAL EVIDENCE IN CRIMINAL PROCEEDINGS 92 (2010), http://www.rss.org.uk/Images/PDF/influencing-change/rss-fundamentals-probability-statistical-evidence.pdf (“False Match: a match is declared but the identification is false. This could arise for a variety of reasons, including: (i) faulty criteria for declaring a match; (ii) misapplication of those criteria in practice, e.g., a fingerprint examiner erroneously judges two characteristics to be similar when they are dissimilar; (iii) confusion, contamination, or degradation of samples; or (iv) the crime sample and the control sample genuinely do match, but the accused is not in fact the source of the crime sample.”).

2. See e.g., COMM. ON QUALITY OF HEALTH CARE IN AM., INST. OF MED., TO ERR IS HUMAN: BUILDING A SAFER HEALTH SYSTEM, at xi (Linda T. Kohn et al. eds., 1999) (reviewing hundreds of studies on medical errors and error rates).
Koehler

This is not to say that forensic examiners are never tested. But the tests that examiners take are generally so easy, unrealistic, and otherwise unlike casework, that even the test manufacturers have said that error rates on these “proficiency tests” should not be used to estimate casework error rates.4 Elsewhere, I have distinguished sharply between two types of proficiency tests: the type that forensic examiners routinely take that are designed to help them improve their skill set and to identify areas in need of improvement (Type I proficiency tests), and the type that forensic examiners almost never take that are designed to identify the rate at which they reach accurate conclusions (Type II proficiency tests).5 The routine administration of Type I rather than Type II proficiency tests is problematic for two key reasons. First, it prevents us from learning how often forensic scientists reach erroneous conclusions in practice. Second, it invites us to confuse and substitute the results of Type I tests with those of Type II tests. Thus, attorneys, judges, jurors, and even the experts themselves frequently rely on the results Type I tests (or what they hear others say about those test results) when considering the risk of error for a forensic science method. This is a dangerous practice: because Type I proficiency tests are notoriously easy,6 those

3. Elsewhere, I have defined a proficiency test as "an assessment of the performance of laboratory personnel." Jonathan J. Koehler, Proficiency Tests to Estimate Error Rates in the Forensic Sciences, LAW, PROBABILITY & RISK 89, 89 (2013). Similarly, one of the forensic science professional organizations defines these tests as an "evaluation of practitioner performance against pre-established criteria." SCI WORKING GRP. FOR FORENSIC ANTHROPOLOGY, PROFICIENCY TESTING 1 (Aug. 2, 2012), http://swghanth.startlogic.com/Proficiency%20Testing%20Rev0.pdf.

4. COLLABORATIVE TESTING SERVS., INC., TOOLMARKS EXAMINATION: TEST NO. 15-528 SUMMARY REPORT 1 (2015), http://www.ctsforensics.com/assets/news/3528_Web-Revised.pdf (Tests provided by this agency are "not intended to be an overview of the quality of work performed in the profession and cannot be interpreted as such."). See also COLLABORATIVE TESTING SERVS., INC., CTS STATEMENT ON THE USE OF PROFICIENCY TESTING DATA FOR ERROR RATE DETERMINATIONS 3 (2010), http://www.ctsforensics.com/assets/news/CTSErrorRateStatement.pdf ("The design of an error rate study would differ considerably from the design of a proficiency test. Therefore, the results found in CTS' Summary Reports should not be used to determine forensic science discipline error rates."). In a 2015 question and answer session with the National Commission on Forensic Science, John Butler and Christopher Caryca argued that DNA proficiency tests are similar to math tests in which a student is asked if they know how to compute $2 + 2$. In contrast, casework is more like doing calculus. See Nat’l Comm’n on Forensic Sci., NCFS Meeting 7, Part 2, NAT’L INST. OF STANDARDS & TECH. (Aug. 31, 2015), https://www.nist.gov/video/ncfs-meeting-7-part-2-(beginning-at-1:10:20). John Butler used slides from an earlier presentation to answer the question. See John M. Butler, Oral Presentation at the International Forensics Symposium, To Err is Human, but How Might We Measure Error Rates in Forensic DNA Testing and What Would These Error Rates Really Mean? (July 23, 2015), https://www.nist.gov/sites/default/files/documents/2016/11/22/error_management_dna_error_butter.legalfact.pdf.

5. See Jonathan J. Koehler, Forensics or Fauxrensics? Ascertaining Accuracy in the Forensic Sciences, 49 ARIZ. ST. L.J. (forthcoming 2018) (manuscript at 7, 27–28), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2773255. There is no relationship between Type I and Type II proficiency tests and Type I and Type II errors in statistical hypothesis testing. In the present context, Type I and Type II refer to two types of proficiency tests that have different purposes and therefore are designed and conducted in different ways. For further commentary on the use of Type I and Type II nomenclature in other fields. See id. (manuscript at 27, n.137).

6. Forensic science proficiency tests have often been criticized for being too easy. See e.g., United States v. Crisp, 324 F.3d 261, 274 (4th Cir. 2003) (Michael, J., dissenting) ("Proficiency testing is typically based on a study of prints that are far superior to those usually retrieved from a crime scene."); Simon A. Cole, More than Zero: Accounting for Error in Latent Fingerprint
who rely on Type I test results as a proxy for casework error rates will likely underestimate the risk of error.

Many people, including judges and scientists, believe that the story is different for DNA analyses. They assume that because DNA evidence has become so important to our criminal justice system over the past three decades that scientists have studied the accuracy of DNA analysis. But the truth is that we do not know any more about human or laboratory error rates in DNA cases than we know about human or laboratory error rates in cases that employ other forensic science techniques. Although many DNA examiners regularly participate in proficiency tests, these tests are of the Type I variety referred to above. Consequently these test results cannot serve as a reliable proxy for casework error rates. At best, the results of Type I proficiency tests offer an indication of a lower bound casework error rate.8

It should go without saying that it is important to know what the error rates are for forensic science methods and conclusions. Without error rate information, judges cannot assess the reliability of proffered forensic methods, and jurors cannot assess the probative value of the evidence produced from those methods. Information pertaining to the training, knowledge, and experience of the forensic scientist who performed the requisite tests will not suffice. Unsupported opinions about “low” rates of error from forensic scientists will not suffice. References to prior admissibility rulings will not suffice. Forensic science must be a science and, as such, must embrace a rigorous empiricism pertaining to its own capabilities. Reliable error rate data are absolutely crucial to our understanding of the value of forensic science evidence. Yet those data do not exist.

Error Rate Estimates


8. If Type I proficiency tests are easier than casework (see supra note 6), then it stands to reason that casework error rates will be at least as high as those observed in the Type I proficiency tests. In this manner, Type I proficiency test results may serve as lower bound error rate estimates. This point has been made previously. See Cole, supra note 6, at 1027 (“Given the acknowledged weaknesses in the studies that generated these false positive rates, these should be regarded as lower bounds of the actual error rate.”); Michael J. Saks & Jonathan J. Koehler, The Coming Paradigm Shift in Forensic Identification Science, 309 SCIENCE 892, 895 (2005) (“Forensic science proficiency test[] . . . data are probably best regarded as lower-bound estimates of error rates. Because the tests are relatively easy (according to test participants), and because participants know that mistakes will be identified and punished, test error rates (particularly the false-positive error rate) probably are lower than those in everyday casework.”).

9. See Jennifer Mnookin et al., The Need for a Research Culture in the Forensic Sciences, 58 UCLA L. REV. 725, 742 (2011) (“Claims, both about a field and about particulars, should be expected as a matter of course to be data-driven.”).
An implication of the lack of reliable data from which to estimate forensic error rates is that legal decision makers are left to their own devices when it comes to estimating the chance that a reported match is erroneous. What do jurors think that chance is? Do their error rate estimates vary across the different types of forensic sciences? Is there reason to believe that their estimates are too high or too low? I address these questions in an online study with 210 participants. But first, I review prior research that bears on people’s intuitions about error rates in the forensic sciences.

### II. INTUITIVE ERROR RATE ESTIMATES: PRIOR RESEARCH

Previous research provides some indication about how people perceive the accuracy and reliability of forensic science evidence. In general, it appears that people think the accuracy and reliability of forensic science evidence is high. Lieberman and his coauthors (in their 2008 study) surveyed 233 people who had completed jury duty in Las Vegas, and asked them how they would rate various types of forensic science evidence for accuracy using a scale that ranged from 0% to 100%.\(^{10}\) They reported mean accuracy ratings of 94.9% for DNA, 91.4% for fingerprints, and 89.2% for hair and fiber.\(^{11}\) Garrett and Mitchell recently described an online study of 251 laypeople in which about 3/4 of respondents indicated that they thought fingerprint evidence was either “very reliable” (25.9%) or “reliable” (50.6%), and the rest thought that it was either “somewhat reliable” (19.5%) or “somewhat unreliable” (4.0%).\(^{12}\) The spread in Garrett and Mitchell’s data set suggests that whereas most people were likely to accept a reported fingerprint match as true, a significant minority were more skeptical.

Schklar and Diamond (in their 1999 study) examined how intentional tampering, unintentional error, coincidental matching, and the expectations that people have about each of these risks affect the weight that mock jurors assign to DNA evidence in a hypothetical sexual assault case.\(^{13}\) The participants in this study were jury-eligible undergraduate psychology students, and a summary of the case was presented using written stimulus materials.\(^{14}\) In one condition, 31 participants were provided with a laboratory error rate for the DNA evidence (either 2 in 100 or 1 in 1 billion) but no other statistics pertaining to the value of the reported DNA match.\(^{15}\) The median estimate participants provided for the

---

11. Id. at 35.
14. Id. at 168–69.
15. Id. at 176.
risks of intentional tampering, coincidental match, and laboratory error were 1 in 5,000, 1 in 1,000,000, and 1 in 1,000 respectively. In an online study, Thompson and Newman (in their 2015 study) asked 541 jury-eligible adults to estimate the chance that “an innocent man in a case like this one” would be falsely incriminated by a frame up, laboratory error, or coincidence in a hypothetical sexual assault case that included a reported forensic match. The forensic evidence was either DNA recovered from a faucet, or a shoeprint recovered from a freshly waxed floor. In one set of conditions, participants were provided with a random match probability (RMP) for the DNA or shoeprint match (1 in 1,000,000 or 1 in 100). Participants then estimated the chance that the match was due to a frame up (which is similar to Schklar and Diamond’s “intentional tampering” condition), coincidence, or laboratory error. When the DNA RMP was 1 in 1,000,000, the median estimates were 1 in 100,000 for a frame up, 1 in 1,000,000 for a coincidental match, and 1 in 10,000 for laboratory error. When the DNA RMP was 1 in 100, the median estimates for the three risks were higher: 1 in 1,000 for a frame up, 1 in 100 for coincidental matching, and 1 in 1,000 for laboratory error. Estimates for the risks associated with a reported shoeprint match revealed a similar pattern. When the shoeprint RMP was 1 in 1,000,000, the median estimates were 1 in 55,000 for a frame up, 1 in 100,000 for a coincidental match, and 1 in 10,000 for laboratory error. When the shoeprint RMP was 1 in 100, the median estimates were 1 in 1,000 for a frame up, 1 in 100 for coincidental matching, and 1 in 1,000 for laboratory error.

Though similar, the studies by Schklar and Diamond and Thompson and Newman differed in some important ways. Whereas participants in Schklar and Diamond estimated the chance of a coincidental match after receiving laboratory error rate information, this information was not provided to participants in Thompson and Newman. Whereas Schklar and Diamond provided an RMP of either 2 in 100 or 1 in 1 billion before asking participants to estimate the lab

16. Id. at 176–77.
18. Id. at 336.
19. See generally id. Thompson and Newman examined two other presentation modes as well: likelihood ratios and verbal equivalent of likelihood ratios. For simplicity and ease of comparison with other studies, data produced from these other presentation modes are not considered here.
20. The RMP is the probability that a randomly selected person in a given reference population will share the target profile. It also identifies the probability that a randomly selected person from this reference population (such as an innocent suspect) will match by coincidence. The RMP is approximately equivalent to the risk that an innocent suspect will match the target profile by sheer coincidence. See id. at 333.
21. Id. at 338.
22. Id. at 341–42 figs. 5, 6, & 7.
23. See id.
24. See id. at 342 fig. 7 (data interpolated from the figure).
25. See id.
error rate, Thompson and Newman provided their participants with an RMP of either 1 in 100 or 1 in 1 million.27 Whereas jurors in Schklar and Diamond wrote down their relevant probability estimates, jurors in Thompson and Newman expressed their answers as odds or used a 17-point force-choice log scale.28

Despite these and other differences between the two studies, together they hint that laboratory error is the most significant perceived risk to the accuracy of a reported forensic match. Participants across conditions in both studies generally indicated that they believed that a laboratory error that falsely incriminated the defendant was more likely than a false incrimination caused by either a coincidental forensic match or fraudulent behavior by examiners. This result suggests that the chance of laboratory error should impose a lower bound on the perceived risk of an incriminating forensic error of some sort in jurors’ minds. In other words, if a juror thinks that the chance that a reported DNA match is caused by a laboratory error is 1 in 1,000, then this juror should (as a matter of logic) believe that the overall risk of a false positive error of some sort is at least 1 in 1,000 and probably larger.29

Importantly, participants in these two studies did not estimate the overall risk of a false positive error. Instead, they estimated each of the central components—fraud, coincidental match, and laboratory error—that comprise the overall false positive error risk. This is an important distinction to acknowledge because research indicates that people have a poor understanding of the relationship between the overall false positive error rate and the individual risks that feed into it. In the forensic context, people may be confused about the significance of the RMP (which is often an extremely small number in the DNA context) for estimating the overall false positive error risk, as well as for estimating the other risks that comprise the overall risk. Thompson and Newman found that RMPs affected their jurors’ estimates about the risk of both laboratory error and a frame up for shoeprint and DNA evidence.30 Jurors in this study thought that the laboratory error and frame up risks were lower when the RMP was very

28. See id. at 334 fig.1.
29. When discussing false positive errors, it is important to define the evidence and hypothesis terms carefully. See Jonathan J. Koehler, If the Shoe Fits They Might Acquit: The Value of Forensic Science Testimony, 8 J. EMPIRICAL L. STUD. 21, 21 (2011). Here, the evidence is “the defendant reportedly matches some forensic item” (e.g., DNA or shoeprint). The hypothesis is “the defendant is guilty of the crime” (e.g., a sexual assault). A false positive error in this context occurs when the defendant is not guilty of the sexual assault, and yet there is a forensic match. Such an error could occur if the match is merely coincidental, if a laboratory error occurs, if there is some sort of frame up or evidence tampering, or if the forensic evidence was deposited innocently by the defendant (i.e., not during an assault). From the standpoint of basic logic and probability theory, the overall false positive error risk is at least as large as the risk associated with each of its individual components. This is why I state above that the perceived laboratory error rate should impose a lower bound on the overall perceived false positive error rate. But, as is well-known, human judgment is not always so logical nor is it constrained in practice by probability axioms. See Daniel Ellsberg, Risk, Ambiguity, and the Savage Axioms, 75 Q.J. Econ. 643, 646, 659–60, 669 (1961); see also DANIEL KAHNEMAN, THINKING FAST AND SLOW 6–9 (2011); Amos Tversky & Itamar Simonson, Context-Dependent Preferences, 39 MGMT. SCI. 1179 (1993).
small. This is an intriguing finding because there is no logical relationship among these risks: the frequency of a DNA profile in a population (which is what the RMP estimates) has no bearing on the risks that laboratory personnel will mix up two samples or that investigators will choose to frame an innocent person. So why did people think otherwise? One possible explanation is that some people mistakenly believe either that the RMP is the false positive error rate. A related explanation is that some people believe that the RMP somehow incorporates all of the various risks, including lab error and fraud. In short, the observed dependence between the RMP associated with a reported match and estimates about other false positive error risks may arise because people think the RMP is a more important indicator of the value of the evidence than it actually is.

Some research supports the hypothesis that, in the context of DNA evidence, people assign too much importance to the RMP statistic. In our 1995 study, Audrey Chia, Samuel Lindsey, and I reported that small RMPs caused mock jurors to behave as if a much larger laboratory error rate was irrelevant to the probative value of a reported match. In this study, mock jurors who were provided with DNA evidence in a hypothetical criminal case were relatively unwilling to convict the defendant when told that the laboratory error rate was .001 or .02. However, when those error rates were supplemented with a very small, but diagnostically irrelevant RMP of 1 in one billion, the conviction rate doubled. As the study explained, from a normative standpoint, the RMP is the component that is largely irrelevant to an estimate of the probative value of a reported match when it is paired with a much larger laboratory error rate, not the other way around. Scurich and John (in their 2013 study) reported a similar confusion among mock jurors in the context of DNA database match evidence. Nance and Morris (in their 2005 study) and Schklar and Diamond likewise found that people combined RMPs and laboratory error rates in inappropriate ways, even after receiving instruction.

To summarize, prior research indicates that mock jurors generally think that (a) the accuracy and reliability of forensic science evidence in general—and DNA evidence in particular—is high, (b) the risks of coincidental match and

31. Id.
32. Id. at 346.
34. Id. at 212.
35. Id. at 214.
36. Id. at 210–11; see also Jonathan J. Koehler & John B. Meixner, Decision Making and the Law: Truth Barriers, in THE WILEY-BLACKWELL HANDBOOK OF JUDGMENT AND DECISION MAKING 749, 756 (Gideon Keren & George Wu eds., 2016) (arguing that the RMP contributes little to jurors’ assessment of the probative value of a forensic match beyond that which is given by the false positive error rate).
fraud are low, and (c) the risk of lab error is also low, but not quite as low as the risks of coincidental match and fraud. Prior research also indicates that mock jurors do not necessarily have a good sense of how to combine the various risks to identify the probative value of a reported forensic match. Some data suggest that some people use the RMP (when available) as a kind of one-stop proxy for identifying the overall false positive error risk. The problem with this approach is that when the laboratory error risk is much larger than the other false positive error risks (as it often is), then people who use one of those other risks—for example, the RMP—as a proxy for the overall false positive error rate will underestimate the overall risk of error.

The present study asks and answers a few important exploratory questions related to perceived forensic error rates that have yet to be tackled. First, what do people think the overall false positive error rates are for the various types of forensic evidence? Because there are no reliable data in any forensic science area that address casework error rates, it is important to know what people think the false positive error rates are because this rate is so essential to identifying the probative value of a reported forensic match. Although no study has collected these estimates across the forensic sciences, Thompson, Kaasa, and Peterson (in their 2013 study, Experiment 2) did ask 61 members of the Orange County California jury pool who were assigned to the control condition in their experiment on how jurors weigh DNA evidence to estimate a “false report probability” for DNA evidence in a sexual assault case. The median estimate was 1 in 1,000,000. This value, which is surely much lower than the actual rate of false DNA match reports, further suggests that many people confuse the low DNA RMPs that they have no doubt heard about with the risk of false match reports. As a secondary matter, the present study will explore this possibility a bit further. It will ask what role, if any, presentation of an extremely small (but realistic) RMP plays in juror’s estimates about the risk of false positive error (in the aggregate). Based on the findings in Thompson and Newman (in their 2015 study) and Thompson and his coauthors (in their 2013 study, Experiment 2), I predict that introduction of such an RMP will drive jurors’ estimates of the overall false positive error risk downward in violation of a normative basis for doing so. Because RMPs are only in widespread use at the present time for DNA evidence, I confine the inquiry on this second matter to DNA evidence.

---

41. Id. at 386.
III. ONLINE STUDY

A. Method

1. Participants

The participants were 210 jury-eligible42 people ("jurors") drawn from a panel of Internet users affiliated with Survey Sampling International (SSI). SSI is a third party vendor that uses opt-in e-mail-based recruiting methods. E-mail invitations were sent to an age-stratified random sample of panel members with the goal of approximating the jury-eligible population. The invitations were spread over different times of day during a one-week period in May 2015. Jurors were compensated by SSI in accordance with its policy of providing members with small guaranteed cash incentives and entry in drawings for additional cash prizes. Jurors covered a broad cross-section of the jury-eligible population on age (23% younger than 30, 22% older than 60), educational level (16% high school graduate or less, 15% graduate degrees), and gender (65% women).

Fifty-eight additional potential jurors were excluded from the study because they failed to complete the study, failed one or more screening questions (designed to eliminate people who were not reading carefully), or were not jury-eligible.

2. Materials and Procedure

Jurors were invited to participate in a short study on decision making involving forensic science evidence. Each juror was assigned at random to one of three groups.

Jurors in Group 1 (n = 70) estimated the chance that a "qualified, experienced forensic scientist" in each of five different forensic areas—microscopic hair evidence, bite marks, DNA, fingerprint, and document examination—would commit a false positive error when examining a pair of samples in a case. Jurors were told that a false positive error occurs when an examiner concludes that two hairs (or DNA samples, etc.) share a common source when, in fact, they do not. Jurors provided false positive error estimates using the logarithmic scale first published in Thompson, Kaasa, and Peterson’s 2013 study, where it was also used as a scale for capturing forensic error rate assessments. The scale included 14 levels that ranged from “1 time in 2” to “1 time in a quadrillion.” See the Appendix for a sample question and the 14-level response scale that was used.

Jurors in Group 2 (n = 70) received a one-paragraph summary of a hypothetical criminal case that included evidence of a DNA match against the defendant. Jurors were told that a qualified, experienced DNA examiner testified

42. U.S. citizens aged 18 or older who have not been convicted of a felony.
43. See Thompson et al., supra note 40, at 362 (explaining that use of a logarithmic response measure “makes it easier [for participants] to express high and low probability estimates”).
that a DNA profile extracted from semen at the crime scene matched the defendant’s DNA profile. They were also told that the expert offered his opinion that the defendant was the source of the DNA recovered from the semen. The expert did not provide any quantitative information related to the rareness of the matching DNA profile. Jurors in this group estimated the false positive error risk for the DNA match report in this case using the same logarithmic scale described above. Unlike jurors in Group 1, jurors in Group 2 did not estimate the overall false positive error rate for DNA evidence or for any other forensic science evidence.

Jurors in Group 3 (n = 70) received a similar one-paragraph summary of the hypothetical criminal case that was provided to jurors in Group 2. However, whereas jurors in Group 2 were not provided with quantitative information related to the rareness of the matching DNA profile, jurors in Group 3 were told that the expert testified that “the ‘random match probability’ associated with that DNA profile is approximately 1 in 987,000,000,000 (one in 987 billion). He further explained that the random match probability identifies the chance that a person selected at random from the general population would match the DNA from the crime scene. So, in this case, the chance is about one in 987 billion (1 in 987,000,000,000) that a random person would happen to have the DNA profile that was found in the semen at the crime scene, and the defendant has that DNA profile. Finally, the DNA examiner offers his opinion that the defendant is the source of the DNA recovered from the semen.” After estimating the false positive error risk for the case, jurors in Group 3 indicated their degree of agreement or disagreement with a claim that the 1 in 987 billion RMP provided “a reasonable estimate of the chance that the DNA examiner made a critical mistake in his testing procedure.” This purpose of this question was to learn whether some people confuse the RMP with the risk of human error.

B. Results

Jurors’ median false positive error rate estimates were all very low (see Table 1). The median estimates for jurors in Group 1 (who estimated the false positive error rate for each of five forensic sciences) ranged from a high of 1 in 100,000 for handwriting to a low of 1 in 10,000,000 for DNA. Median false positive error rate estimates for microscopic hair, bite marks, and fingerprints were 1 in 1,000,000, 1 in 1,000,000, and 1 in 5.5 million, respectively. A repeated measures MANOVA on the Group 1 data revealed that error rate estimates differed significantly as a function of forensic science type (F(4, 66) = 4.96, p = .001). Post-test contrasts indicated that jurors thought that handwriting had a higher error rate than each of the other four disciplines (p values range from .02 (bite marks) to < .001 (DNA and fingerprints)). Jurors also thought that DNA had a lower error rate than bite marks and hair (p = .002 and .03 respectively). Jurors also thought that fingerprint evidence had a lower error rate than bite mark evidence (p = .02).

Jurors in Groups 2 and 3 (who received a summary of a hypothetical DNA case), estimated the risk of a false positive DNA error in the target case to be extremely low. The median estimate was 1 in 1,000,000 for Group 2 (no RMP
Error Rate Estimates

provided) and 1 in 1,000,000,000 for Group 3 (RMP of 1 in 987 billion provided). This difference was statistically significant \((Z = 2.91, p = .004, \text{ two-tailed})\). Error rate estimates for DNA evidence for Groups 1 and 2 were significantly different \((Z = 2.28, p = .022, \text{ two-tailed})\), but not significantly different between Groups 1 and 3 \((Z = 1.11, p = .267, \text{ not significant})\). Finally, a large proportion of jurors who received the DNA RMP probability value (Group 3) misunderstood its meaning. Fifty-three percent of jurors in Group 3 agreed or strongly agreed with the statement that the RMP “provides a reasonable estimate of the chance that the DNA examiner made a critical mistake in his testing procedure.”

Regarding demographics, there were no differences between males and females on any of the key dependent measures \((ps\text{ range from } .26 \text{ to } .99)\). There was a difference between younger and older jurors on fingerprint error rate estimates. Whereas younger jurors thought that the risk of error for fingerprints was about 1 in 1,000,000, older jurors thought that risk was even smaller, about 1 in 100,000,000 \((t(68) = 2.36, p = .02)\). There were no age differences on any of the other key dependent measures \((ps\text{ range from } .18 \text{ to } .76)\). Finally, there was a significant difference between more and less educated jurors on microscopic hair error rate estimates. Whereas relatively less educated jurors thought that the risk of error for microscopic hair analysis was about 1 in 1,000,000, relatively more educated jurors (college or graduate degree) thought that risk was even smaller, about 1 in 100,000,000 \((t(68) = 2.64, p = .01)\). There were no educational level differences on any of the other key dependent measures \((ps\text{ range from } .33 \text{ to } .98)\).

<table>
<thead>
<tr>
<th>Forensic Science</th>
<th>Group 1 ((n = 70))</th>
<th>Group 2 ((n = 70))</th>
<th>Group 3 ((n = 70))</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>1 in 10,000,000*</td>
<td>1 in 1,000,000</td>
<td>1 in 1,000,000,000**</td>
</tr>
<tr>
<td>Fingerprints</td>
<td>1 in 5,500,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bite Marks</td>
<td>1 in 1,000,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Microscopic Hair</td>
<td>1 in 1,000,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Handwriting</td>
<td>1 in 100,000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Table 1 reports participants’ median error rate estimates on a 14-level logarithmic scale. Group 1 gave error rate estimates for each of five forensic sciences without case specific information. Groups 2 and 3 received case specific DNA information and gave error rate estimates only for DNA. Group 3 also received a random match probability value of 1 in 987 billion. *Indicates a statistically significant difference between the 1 in 10,000,000 DNA median error rate estimate for Group 1 and the 1 in 1,000,000 DNA median error rate estimate for Group 2 \((Z = 2.28, p = .022, \text{ 2-tailed})\). **Indicates a statistically significant difference between the 1 in 1,000,000,000 DNA median error rate estimate for Group 3 and the 1 in 1,000,000 DNA median error rate estimate for Group 2 \((Z = 2.91, p = .004, \text{ 2-tailed})\).
C. Discussion

This study asked people to estimate the overall false positive error rates for five types of forensic science evidence (DNA, fingerprints, bite marks, hair, and handwriting). It also considered what role, if any, presentation of an extremely small RMP plays in jurors’ estimates about the risk of a false positive DNA error.

The results indicate that people think that the risk of a false positive error rates in all of the examined forensic sciences is extremely low. People think that DNA evidence has the lowest error rate (median estimates ranged from one out of one million to one out of one billion, and that handwriting analysis has the highest error rate (one out of one hundred thousand). The extremely low error rate estimates identified here for DNA evidence are in line with the Thompson, Kaasa, and Peterson study, which reported that the median estimated false report probability for DNA evidence among 61 members of the Orange County jury pool was 1 in one million.

At the same time, the median false positive error rate estimates in the present study are generally smaller than those reported for the individual components that comprise the false positive error rate as reported in Thompson and Newman. This might seem odd because, from a mathematical point of view, the overall false positive error rate must be at least as large as any of the individual components that feed into it. But if people believe that the RMP is not merely a single way in which a false positive error might occur but is instead either synonymous with false positive error or somehow incorporates the other false positive error risks (e.g., fraud and laboratory error), then we would expect the false positive error rate estimates to be similar to RMP values as they were for our participants in Group 2. As for why the median error rate estimates in the present study are generally smaller than those in Thompson and Newman rather than the same order of magnitude, the likely reason is a difference in RMP values. Whereas the smallest RMP used by Thompson and Newman was 1 in one million, the RMP in the present study was much smaller (1 in 987 billion).

Regarding the significant difference found between young and old jurors on fingerprint error rate estimates and between more and less educated jurors on microscopic hair error rate estimates, these are probably spurious differences. Whenever multiple comparisons are made (as they were here), a few significant differences can emerge by chance alone.

The fact that we know little about the rate at which forensic methods err is scandalous. Simply put, there is no way anyone can accurately assess the probative value of any item of forensic evidence offered at trial without this knowledge. Yet various types of forensic science are proffered and routinely

44. Thompson & Newman, supra note 17, at 336.
deemed admissible at trial. As a result, all of the relevant players (judges, jurors, attorneys, and the expert witnesses themselves) are essentially flying blind. Worse still, the absence of error rate data are likely to severely limit, or at least make ineffectual, any courtroom conversation about the risk of error.46

In the unlikely case that jurors give much thought to the risk of error, they will have to speculate about what that risk is, based on nothing that approaches hard evidence or expert summaries of what the hard evidence shows. The present study provides a window into what jurors think the risk of error is for some forensic sciences. The results suggest that people generally think that the risk of false positive error for DNA, fingerprints, bite marks, microscopic hair, and handwriting is one in a million or less. Consistent with prior empirical studies reporting that people trust and rely upon DNA evidence more than any other type of forensic science evidence,47 our jurors indicated that they thought the false positive error rate for DNA was the lowest of the five forensic sciences considered here. Because reliable data on false positive error rates do not exist,48 one cannot say just how far off the estimated false positive error rates for DNA or other forensic sciences are. But based on the lower bounds for error rate set by Type I proficiency test data49 and dozens of confirmed false positive errors,50 it seems safe to suggest that our jurors’ error rate estimates are too low.

Our data also support the claim that, in cases involving DNA evidence, people conflate the RMP (i.e., the frequency with which the DNA profile is deemed admissible at trial. As a result, all of the relevant players (judges, jurors, attorneys, and the expert witnesses themselves) are essentially flying blind. Worse still, the absence of error rate data are likely to severely limit, or at least make ineffectual, any courtroom conversation about the risk of error.46

In the unlikely case that jurors give much thought to the risk of error, they will have to speculate about what that risk is, based on nothing that approaches hard evidence or expert summaries of what the hard evidence shows. The present study provides a window into what jurors think the risk of error is for some forensic sciences. The results suggest that people generally think that the risk of false positive error for DNA, fingerprints, bite marks, microscopic hair, and handwriting is one in a million or less. Consistent with prior empirical studies reporting that people trust and rely upon DNA evidence more than any other type of forensic science evidence,47 our jurors indicated that they thought the false positive error rate for DNA was the lowest of the five forensic sciences considered here. Because reliable data on false positive error rates do not exist,48 one cannot say just how far off the estimated false positive error rates for DNA or other forensic sciences are. But based on the lower bounds for error rate set by Type I proficiency test data49 and dozens of confirmed false positive errors,50 it seems safe to suggest that our jurors’ error rate estimates are too low.

Our data also support the claim that, in cases involving DNA evidence, people conflate the RMP (i.e., the frequency with which the DNA profile is

46. Koehler, supra note 29, at 41–42. A mock jury study that found, among other things, that a forensic examiner’s concessions on cross-exam about the risk of error had relatively little impact on jurors’ beliefs about what the evidence showed. Id.

47. Lieberman et al., supra note 10, at 37 (empirical study that found undergraduate college students rated the accuracy and persuasiveness of DNA evidence higher than fingerprints, hair fibers, eyewitness testimony, or victim testimony); Craig M. Cooley, Forensic Individualization Sciences and the Capital Jury: Are Witherspoon Jurors More Deferential to Suspect Science Than Non-Witherspoon Jurors? 28 S. ILL. U. L.J. 273, 317–21 (2004) (empirical study that found law students rated the trustworthiness and persuasiveness of DNA evidence higher than fingerprints, bullet identification, tool marks, lip prints, handwriting, boot prints, and bite marks).

48. See Koehler, supra note 5 (manuscript at 3–4). It should also be noted that, for some complex forensic technologies such as DNA typing, it may not make sense to test for and refer to a single false positive error rate. For example, DNA samples that contain more than one DNA profile, are degraded, or that have a very limited quantity (i.e., low copy number DNA samples) are more challenging to analyze than larger, pristine, single-source samples. As a result, the former may have a higher false positive error rate. See e.g., JOHN M. BUTLER, ADVANCED TOPICS IN FORENSIC DNA TYING 332 (2011) (“In the world of forensic casework, DNA degradation or PCR inhibition may be a complicating factor along with the low levels of DNA template being recovered from an[ ] evidentiary item. When mixtures are observed at low DNA amounts, the individual components to the mixture will be even lower in amount and the stochastic effects become worse making it extremely challenging to recover the correct profile of the original contributors to the mixture.”).

49. Based on all reported Type I external proficiency test data from 1983 to 2004, the false positive error rate for fingerprints is nearly 1% (373 false positive errors out of 43,973 comparisons). See Cole, supra note 6, at 1030, 1072 (“Overall, the comparison false positive rate, aggregated over the entire test-taking period is around 0.8%.”).

50. Simon A. Cole has documented about two-dozen casework fingerprint errors alone. Cole also explains that because unusual circumstances are required to be in a position where a false positive fingerprint error could be exposed, the actual number of such errors is much greater than several dozen. See id. at 1017–28; see also Simon A. Cole, The Prevalence and Potential Causes of Wrongful Conviction by Fingerprint Evidence, 37 GOLDEN GATE U. L. REV. 39, 60–62 (2006).
found in a reference population) with the false positive error rate. Consistent with Thompson and Newman, participants in the present study who received an extremely low RMP (1 in 987 billion) estimated the overall risk of false positive error to be quite low. As indicated previously, this is problematic because, as a normative matter, the RMP typically has only a small effect on the overall false positive error rate, whereas the risk of laboratory error exerts a much larger effect on the overall error rate. Unfortunately, it may be the case that the RMP provides people with an anchor (albeit a largely irrelevant one) for estimating the false positive error rate.

To the extent our data indicate that people underestimate the risk of false positive error in the forensic sciences, it suggests an affirmative role for the courts, government, and scientific community. Specifically, courts should take seriously the possibility that jurors will overweight various types of forensic science evidence because they mistakenly believe that the risk of error is infinitesimal. To mitigate this risk, courts should carefully monitor and restrict the claims made by forensic scientists and attorneys at trial, and take steps to provide jurors with information about what we do and do not know about forensic science error rates based on rigorous scientific studies.

Regarding the government, relevant agencies such as the National Institute of Justice, the National Science Foundation, and the National Institute of Standards and Technology must encourage and support scientific research designed to identify accuracy rates in the forensic sciences, including identifying the conditions under which critical errors are more and less likely to arise. The scientific community must also study how to present the resultant accuracy data to judges, jurors, and others to convey their meaning and to minimize the risk of fallacious statistical thinking. In addition, the scientific community must take the lead in

51. See Thompson & Newman, supra note 17, at 341–42 figs. 5, 6, & 7.

52. Koehler et al., supra note 33, at 203; Thompson et al., supra note 39, at 50.

53. Several other studies have provided mock jurors with very small RMPs and found that jurors report low error rate estimates. See Thompson & Newman, supra note 17, at 336 (providing some mock jurors with an RMP of 1 in one million); Thompson et al., supra note 40, at 366 (providing all mock jurors with an RMP of 1 in one trillion); Schklar & Diamond, supra note 13, at 162 (providing some mock jurors with an RMP of 1 in one billion). Elsewhere, I have suggested that, as a policy matter, it might be best to simply provide jurors with error rate data rather than RMP data because the latter contribute nothing beyond the error rate to an understanding of the probative value of the reported match. See e.g., Jonathan J. Koehler, Linguistic Confusion in the Court: Evidence From the Forensic Sciences, 21 J.L. & POL’Y 515, 533 (2013) (“There is no need to provide the RMP [when the RMP = 1 in 1,000,000 and the false positive error rate = 1 in 500] . . . because it does not contribute anything beyond the false positive error rate in terms of helping jurors understand a fact in evidence. What should jurors be told in cases like the one described above? They should be told something like this: The suspect reportedly matches the DNA evidence found at the crime scene. The chance that we would report such a match on nonmatching samples, either because of a coincidence or because of an error, is approximately one in 500.”).

54. Thompson & Newman, supra note 17, at 347. The authors found that about two-thirds of participants in a mock jury study committed the “source probability error” in which they mistakenly equated the RMP with the probability that the matching defendant was not the evidentiary source. Id. They concluded that, “fallacious interpretation of forensic science evidence may play a significant role in the decision to convict.” Id. See generally Laurens Walker & John Monahan, Social Frameworks: A New Use of Social Science in Law, 73 VA. L. REV. 559, 568–69, n. 23 (1987) (discussing how and why such general testimony at trial may be useful to a jury).
challenging the claim that we already know that error rates across the forensic sciences are extremely low based on data from existing laboratory proficiency tests. As discussed elsewhere, existing proficiency tests are a poor vehicle for estimating casework error rates. A different type of proficiency test—a Type II proficiency test—must be developed and implemented across the forensic sciences for the express purpose of providing legal decision makers with an empirical basis for estimating the rate at which significant errors occur in forensic casework.

Some readers may wonder whether the anticipated benefits of a Type II proficiency testing program are large enough to justify its likely expense. Regarding those benefits, some may argue that even if Type II proficiency testing reveals error rates that are substantially higher than the intuitive medians reported here, jurors will not benefit from receiving this information. According to this argument, the error rate for a given technique will merely be used as a threshold value: if the error rate is low enough, the evidence will be admitted and jurors will give full weight to any identification; if the error rate is too high, the evidence will not be admitted or, if it is admitted, it will not be given any credence by jurors. What little empirical data exist that speak to the issue of how jurors respond to error rate data is mixed: some studies suggest that jurors will be insensitive to error rate variations, but other studies suggest the opposite. Whether people intuitively understand the importance of error rate variation or not, researchers must study how best to sensitize jurors to the importance of such variations. This work should proceed concurrently with Type II proficiency testing work.

Finally, a few words of caution about interpreting the present results. As in many behavioral studies, the precise wording of the questions asked may have influenced the answers jurors provided. Wording may be especially important when it comes to error rate estimates due to the ease with which linguistic confusion can affect the dissemination and interpretation of probabilistic forensic

55. See Koehler, supra note 5 (manuscript at 27).
56. See id. (manuscript at 27).
57. The cost of a comprehensive Type II proficiency testing program across the various forensic sciences will likely be at least several million dollars per year. See, e.g., Joseph L. Peterson et al., The Feasibility of External Blind DNA Proficiency Testing. I. Background and Findings, 48 J. FORENSIC SCI. 1, 8–9 (2003) (estimating the cost, as of 2003, of a government-run proficiency testing program that includes 1–2 tests per year for each of 150 DNA laboratories in the United States at $535,000–$814,000 per year).
58. At least one well-known academic forensic scientist has made this argument. Christophe Champod, Research Focused Mainly on Bias Will Paralyse Forensic Science, 54 SCI. & JUST. 107, 108 (2014).
59. See e.g., Koehler et al., supra note 33, at 213 n.49 (“No statistically significant differences were detected between the 2 percent and .1 percent laboratory error rates” in a mock juror study that involved a reported DNA match.); Scurich & John, supra note 37, at 428 (Mock jurors did not respond to DNA evidence as a function of whether the laboratory error rate was 1 in 10 or 1 in 1,000.).
60. See Schklar & Diamond, supra note 13, at 175 (Mock jurors responded differently to DNA evidence as a function of whether the laboratory error rate was presented as 2 in 100 or 1 in 1 billion.).
Similarly, to the extent that participants in this study confused the RMP with error rate, it is possible that some sort of admonition not to confuse the two entities would have reduced the number of extremely small error rate estimates. Caution should also be exercised when generalizing from the reported median error rate estimates. Although there are technical advantages to using a logarithmic response scale such as the one used here, it is possible that the available choices subtly encouraged jurors to select low values. Because people are often attracted to options that tend toward the middle of response scales, and because the scale had many small values in the middle range (see Appendix), some jurors may have assigned lower values than they would have assigned if they were using a different scale. Whatever influence the scale may have exerted on jurors’ estimates, it seems safe to conclude that jurors in our study thought that the error rates for various forensic sciences are extremely low.

**APPENDIX. Sample Question: Group 1**

Based on what you know about the science of forensics from all sources, estimate the frequency that a qualified, experienced forensic scientist will mistakenly conclude that two DNA samples were contributed by the same person when, in truth, the samples were contributed by different people. Such an error is likely to occur about 1 time in:

- 2
- 10
- 100
- 1,000
- 10,000
- 100,000
- 1,000,000 (one million)
- 10,000,000
- 100,000,000
- 1,000,000,000 (one billion)
- 1,000,000,000,000 (one trillion)
- 1,000,000,000,000,000,000 (one quadrillion)
- Less than 1 time in one quadrillion

Such an error is impossible

---


63. Roger Tourangeau et al., *The Psychology of Survey Response* 248 (2004) (noting a “response contraction bias” in which survey respondents tend to select options that are in the middle of the option set).