How Should Forensic Scientists Present Source Conclusions?

William C. Thompson[*]

I. INTRODUCTION 774
II. THE LOGIC OF SOURCE CONCLUSIONS 775
III. POSSIBLE FORMATS FOR REPORTING 776
A. Statements About the Balance of Likelihoods 778
1. Likelihood Ratios (LRs) 778
2. Strength-of-Support Statements 781
B. Statements Based on a Two-Step Analysis 782
1. Identification/Individualization Based on Uniqueness of Features 782
2. Match Frequencies/Random Match Probabilities (RMPs) 784
3. Likelihood of Observed Correspondence 786
4. Matches/Correspondence Without Frequencies or Probabilities 787
5. True and False Positive Rates 787
C. Statements Based, in Part, on Prior Odds of a Common Source 789
1. Statements About Source Probability 790
2. Identifications Without Uniqueness 792
IV. EVALUATING LAY REACTIONS TO FORENSIC EVIDENCE 793
A. Lay Perceptions of the Relative Strength of Various Reporting Statements 794
B. Assessing the Appropriateness of Lay Responses to Forensic Science Evidence 799
1. Sensitivity to the Strength of Evidence 800
2. Logical Coherence 802
i. Underutilization of Forensic Evidence 803
ii. Fallacious Reasoning 805

[*] Professor Emeritus, Department of Criminology, Law & Society, and School of Law, University of California, Irvine. The author's work on this article was supported by the Center for Statistics and Applications in Forensic Evidence (CSAFE), which was in turn supported by a grant from the National Institute of Standards and Technology (NIST).

V.SUMMARY AND CONCLUSIONS 809

I. INTRODUCTION

In a symposium celebrating the career and contributions of Professor D. Michael Risinger, it seems appropriate to discuss an issue he has considered and commented upon for many years—the challenge of communicating forensic science findings to ordinary human beings, such as those who serve on juries.[1] I will focus on *source conclusions*, which are the conclusions that forensic scientists reach after comparing items to evaluate whether they have, or might have, a common source. Examples include comparison of fingerprints, biological samples, tool marks, shoe prints, and handwriting.

The question I will address is how forensic scientists should communicate source conclusions in reports and testimony. The answer, I will argue, depends on two issues: (1) what conclusions can be justified logically and empirically;[2] and (2) what conclusions (among those that can be justified logically and empirically) are most likely to be understood and used appropriately. I will first review various possible ways that forensic scientists might report source conclusions, pointing out logical and empirical difficulties with some reporting methods. Then I will discuss what is currently known about lay understanding of such reports.

This analysis will, unfortunately, yield no ideal or preferred solution. It will instead suggest that the legal system faces trade-offs: the reporting formats that are easiest for lay people to understand are difficult to justify logically and empirically, while reporting formats that are easier to justify logically and empirically are more difficult for lay people to understand. To find the best solution, I will argue, we need careful consideration of the options and more empirical research.

[1] Examples of Risinger's commentary on this topic include D. Michael Risinger, *Reservations About Likelihood Ratios (and Some Other Aspects of Forensic 'Bayesianism,'* 12 LAW, PROBABILITY & RISK 63 (2012); D. Michael Risinger, *Against Symbolization*, 11 LAW, PROBABILITY & RISK 247 (2012).

[2] There is wide agreement (I hope) that forensic scientists should be limited to presenting conclusions that are scientifically valid. That is the essence of Rule 702 of the Federal Rules of Evidence and the *Daubert* standard. *See* Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579 (1993).

II. THE LOGIC OF SOURCE CONCLUSIONS

Let's begin by considering the logic of forensic comparison—that is, the logical steps by which a forensic examiner may go from observations about the physical characteristics of a pair of items to a conclusion about whether the items have a common source. I will use fingerprint comparison as an illustration because fingerprints are easy to understand and because the logic of drawing conclusions from a fingerprint comparison is essentially same as the logic of drawing conclusions from the comparison of other items, such as footwear impressions, tool marks, bite marks, handwriting, and the like.

The examiner's goal is generally to assess two alternative hypotheses about the origin of the prints being compared: (1) that the prints came from the same finger; and (2) that the prints came from different fingers. Examiners make this assessment based on the observed physical characteristics of the prints, focusing particularly on similarities and differences between the ridge patterns. The analysis is inherently probabilistic; the only logical way for an examiner to derive source conclusions is to consider the probability of the observed patterns under the alternative hypotheses about their origin.[3]

Suppose that the ridge patterns of the two fingerprints appear quite similar, but the examiner observes some small discrepancies. The examiner must consider how likely those discrepancies are if the prints were made by the same finger. This might involve consideration of the likelihood that slipping or torsion of the finger, or some other process, could have distorted one or both of the prints enough to produce the discrepancies. The examiner must also consider the likelihood that the patterns would be as similar if the prints were made by different fingers, which would require consideration of the rarity of the shared features.[4]

---

[3] Latent print examiners have only recently begun to recognize and acknowledge the probabilistic nature of their analyses. Heidi Eldridge, *The Shifting Landscape of Latent Print Testimony: An American Perspective*, 3. J. FORENSIC SCI. & MED. 72 (2017). In the past, it was common for them to claim they were simply determining whether the ridge patterns "match" or share unique features. Simon A. Cole, *Forensics Without Uniqueness, Conclusions Without Individualization: The New Epistemology of Forensic Identification*, 8 LAW, PROBABILITY & RISK 233 (2009); Simon A. Cole, *Individualization is Dead, Long Live Individualization! Reforms of Reporting Practices for Fingerprint Analysis in the United States*, 13 LAW, PROBABILITY & RISK 117 (2014). Whether they recognize it or not, this determination requires them to think about the probability that any discrepancies between the prints could have arisen if the prints came from the same finger, and the probability that the similarities between the prints could have arisen if the prints were from different fingers.

[4] For a detailed discussion and analysis of latent print examination, see William Thompson, John Black, Anil Jain & Joseph Kadane, *Forensic Science Assessments: A Quality and Gap Analysis—Latent Fingerprint Examination*, AM. ASS'N FOR THE ADVANCEMENT OF SCI. (Sept. 2017), https://mcmprodaaas.s3.amazonaws.com/s3fs-public/reports/Latent%20Fingerprint%20Report%20FINAL%209_14.pdf?i9xGS_EyMHnIPLG6INIUyZb66L5cLdlb

It is the balance between these two likelihoods that allows inferences to be drawn about whether the traces have a common source. The observed results support the hypothesis of a common source to the extent that the likelihood of the observed features is higher if the traces have a common source than if they have a different source. The observed results support the hypothesis of a different source to the extent that the likelihood of the observed features is higher if the traces have a different source than if they have a common source. That is the fundamental and inescapable logic of forensic comparison; it applies regardless of how examiners choose to report their conclusions, although some reporting methods do a better job than others of making this logic transparent.[5]

III. POSSIBLE FORMATS FOR REPORTING

There are (at least) three schools of thought on how examiners should report their conclusions. One approach requires the examiner to make statements reflecting the balance of likelihoods. The examiner either makes a statement about the relative likelihood of the observed findings under alternative hypotheses or makes a statement about the strength of the forensic evidence that is based on the balance of likelihoods.[6] I will discuss variants on this first approach to reporting in Part A.

A second approach, more common in the United States, requires a two-step analysis. First, the examiner compares the items, looking for distinguishing features that would rule out the hypothesis that the items have a common source.[7] When distinguishing features are found, the examiner

[hereinafter AAAS REPORT].

[5] More complete discussions of the logic of forensic inference can be found in BERNARD ROBERTSON, G.A. VIGNAUX & CHARLES E.H. BERGER, INTERPRETING EVIDENCE: EVALUATING FORENSIC SCIENCE IN THE COURTROOM (2d ed. 2016); COLIN AITKEN & FRANCO TARONI, STATISTICS AND THE EVALUATION OF EVIDENCE FOR FORENSIC SCIENTISTS (2d ed. 2004); COLIN AITKEN, PAUL ROBERTS & GRAHAM JACKSON, COMMUNICATING AND INTERPRETING STATISTICAL EVIDENCE IN THE ADMINISTRATION OF CRIMINAL JUSTICE: 1. FUNDAMENTALS OF PROBABILITY AND STATISTICAL EVIDENCE IN CRIMINAL PROCEEDINGS (2010), http://www.rss.org.uk/Images/PDF/influencing-change/rss-fundamentals-probability-statistical-evidence.pdf.

[6] *See* ROBERTSON, VIGNAUX & BERGER, *supra* note 5; John Buckleton, *A Framework For Interpreting Evidence*, *in* JOHN S. BUCKLETON, CHRIS M. TRIGGS & SIMON J. WALSH, FORENSIC DNA EVIDENCE INTERPRETATION 27–63 (2005); Geoffrey Stewart Morrison, *The Likelihood-Ratio Framework and Forensic Evidence in Court: A Response to* R v T., 15 INT'L J. OF EVID. & PROOF 1–29 (2012); Graham Jackson, *Understanding Forensic Science Opinions*, *in* HANDBOOK OF FORENSIC SCIENCE 419–45 (Jim Fraser & Robin Williams eds., 2009).

[7] Whether the hypothesis of common source can be ruled out is a decision made by the examiner based on an assessment of the probability that the observed discrepancies would occur if the items being compared have the same source. When the probability of the observed discrepancies is sufficiently low, the examiner decides, in effect, to reject the hypothesis of a common source and accept the hypothesis of a different source. There typically are no

reports that the items do not have a common source—which is often called "exclusion." When the items cannot be distinguished (i.e., the hypothesis of common source cannot be ruled out), then, as a second step, the examiner makes an assessment of the rarity or distinctiveness of the shared features. If the examiner believes the shared features are so distinctive as to be unique (one-of-a-kind), the examiner may conclude (and report) that the items have a common source—this conclusion is often called an individualization or identification.[8] If the examiner believes the shared features are not unique, then several options exist. The examiner may make a statement about the rarity of the matching features, or the probability that a random item of the same type would have such features. Alternatively, the examiner might simply report that the items are indistinguishable, or that they "match," without commenting on the rarity of the matching features. Finally, the examiner might report that the comparison was inconclusive. I will discuss variants on this two-step approach in Part B.

A third approach requires the examiner to draw conclusions about the probability that the items have a common source, which can be expressed either with numbers (e.g., "there is a 99% chance this bite mark was made by the suspect") or with words (e.g., "it is highly probable that these marks were made by the same tool"). These conclusions are sometimes called *source probabilities*. A distinctive feature of this third approach, which distinguishes it from both the first approach (balance of likelihoods) and the second approach (two-step analysis), is that it requires the examiner to take a position or make assumptions about the prior odds that the items being compared have a common source.[9] In other words, the examiner's conclusion necessarily rests on something more than an evaluation of the physical properties of the items being compared. I will discuss variants of the source probability approach in Part C.

objective standards either for the estimation of probability or for the threshold of decision; both are subjective judgments.

[8] Eldridge, *supra* note 3; AAAS report, *supra* note 4.

[9] *See* Jackson, *supra* note 6, at 426–27.

A. *Statements About the Balance of Likelihoods*

1. Likelihood Ratios (LRs)

In Europe, forensic examiners often describe their perception of the balance of likelihoods using numbers called likelihood ratios (LRs).[10] LRs represent the expert's view of the relative probability of the observed features under the alternative hypotheses about the source of the traces.[11]

LRs are commonly used in the United States to report the results of comparisons involving mixed DNA samples.[12] The analyst might report, for example, that the genetic characteristics found in a mixed specimen are "X times more likely" under one assumed hypothesis (e.g., "the specimen consists of DNA from the suspect and an unknown person") than under an alternative hypothesis (e.g., "the specimen consists of DNA from two unknown persons").[13] LRs have also been used to characterize the strength of forensic voice comparison evidence.[14]

Forensic DNA analysts and forensic voice comparison analysts can compute LRs based on databases and statistical models. In many fields of forensic science, however, the empirical foundation for such estimates is more limited. There are relatively few studies of the frequency of various features of fingerprints, tool marks, bite marks, handwriting, footwear impressions and the like.[15] Furthermore, it is often difficult to model the

---

[10] EUROPEAN NETWORK OF FORENSIC SCI. INSTS., ENFSI GUIDELINE FOR EVALUATIVE REPORTING IN FORENSIC SCIENCE: STRENGTHENING THE EVALUATION OF FORENSIC RESULTS ACROSS EUROPE (2015), http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf [hereinafter ENFSI REPORT]; Wim Kerkhoff, Reinoud Stoel, Erwin Mattijssen & Rob Hermsen, *The Likelihood Ratio Approach in Cartridge and Bullet Comparison*, 45 AFTE J. 284 (2013) (describing the adoption of the LR approach by the firearms section of the Netherlands Forensic Institute); Charles E.H. Berger et al., *Evidence Evaluation: A Response to the Court of Appeal Judgment in* R v. T, 51 SCI. & JUST. 43 (2011).

[11] There is a simple mathematical description of the LR that lawyers and judges may encounter when reviewing forensic evidence. Let E represent the observed features of two traces that a forensic scientist is asked to compare; let H_s represent the hypothesis that the items have the same source and H_d the hypothesis that they have a different source. The likelihood ratio is then $p(E|H_s)/p(E|H_d)$, which is read as "the probability of E given H_s over the probability of E given H_d."

[12] JOHN BUTLER, FORENSIC DNA TYPING: BIOLOGY, TECHNOLOGY, AND GENETICS OF STR MARKERS ch. 22 (2d ed. 2005).

[13] The LR for a particular comparison is the examiner's estimate of $p(E|H_s)/ p(E|H_d)$.

[14] Geoffrey Stewart Morrison & William C. Thompson, *Assessing the Admissibility of a New Generation of Forensic Voice Comparison Testimony*, 18 COLUM. SCI. & TECH. L. REV. 326 (2017), http://www.stlr.org/download/volumes/volume18/morrisonThompson.pdf.

[15] *See generally* NAT'L RES. COUNCIL OF THE NAT'L ACAD. OF SCIENCES, STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES: A PATH FORWARD 7 (2009), https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf [hereinafter NAS REPORT); PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., FORENSIC SCIENCE IN CRIMINAL COURTS: ENSURING SCIENTIFIC VALIDITY OF FEATURE-COMPARISON METHODS (2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf

probability of obtaining specific sets of features because the individual features are not necessarily statistically independent.[16] Examiners in these fields nevertheless make judgments about the probability of observing particular sets of features under alternative hypotheses. Rather than relying on empirical data and statistical modeling, however, they typically make a subjective evaluation based on their training and experience. In some instances, they can rely partly on empirical data and partly on training and experience.[17]

The European Network of Forensic Science Institutes (ENFSI) has recommended that forensic examiners always use likelihood ratios to evaluate and describe the strength of source conclusions, even if examiners must make subjective judgments about the relevant likelihoods.[18] While some commentators (most notably Professor Risinger) have expressed reservations about presenting LRs derived from subjective evaluations rather than empirical data,[19] the practice of presenting subjective likelihood ratios appears to have taken hold in many European countries.[20]

Advocates of subjective LRs point out that a forensic examiner must make subjective evaluations of likelihood in order to draw *any* conclusions about whether two items have a common source.[21] If the examiner is unable to assess the relevant likelihoods, then the examiner has no basis for evaluating the strength of the forensic evidence—and hence, should not be reporting source conclusions at all. Asking examiners to report LRs is simply asking them to use numbers to describe their subjective beliefs about the relevant likelihoods. By making these beliefs explicit, examiners increase the transparency of their inferential process, making it easier for those who rely on their conclusions to evaluate and appreciate potential weaknesses in the examiner's logic or in the scientific foundations for the

[hereinafter PCAST REPORT].

16 *See, e.g.*, AAAS REPORT, *supra* note 4, at 21–22 (citations omitted), which notes: "The probability of finding a set of genetic features in a DNA test is relatively easy to estimate because the features occur at rates that are statistically independent of one another. Statistical dependencies are more likely for fingerprint features and will make it far more difficult to estimate the frequency of combinations of features for fingerprints than for DNA profiles. Consequently, research of this type, while important, is unlikely to yield quick answers."

17 *See, e.g.*, Alex Biedermann, Franco Taroni & Christophe Champod, *How to Assign a Likelihood Ratio in a Footwear Mark Case: An Analysis and Discussion in Light of* R v. T. 11 LAW, PROBABILITY & RISK 259 (2012).

18 ENFSI REPORT, *supra* note 10.

19 *See* Risinger, *Reservations About LRs*, *supra* note 1 (calling subjective LRs "numbers from nowhere").

20 Berger et al., *supra* note 1010; ENFSI REPORT, *supra* note 10.

21 Marjan J. Sjerps & Charles E.H. Berger, *How Clear Is Transparent? Reporting Expert Reasoning in Legal Cases*, 11 LAW, PROBABILITY & RISK 317 (2012), https://academic.oup.com/lpr/article-pdf/11/4/317/2748435/mgs017.pdf.

examiner's conclusion.[22]

On the other hand, examiners may be reluctant to put specific numbers on their subjective beliefs about the relevant likelihoods, even if those beliefs are well-grounded. An examiner may justifiably believe that the observed results are far more probable if the items being compared have the same source than a different source, for example, without being able to say with any precision how much more probable. Forcing examiners to articulate numbers may lend a false air of precision to a subjective approximation.[23]

One way to deal with this problem is to allow examiners to express conclusions about the balance of likelihoods using words rather than numbers. In 2012, the Expert Working Group on Human Factors in Latent Print Analysis, sponsored by the National Institute of Standards and Technology (NIST), issued an important and carefully reasoned report that suggested that latent print examiners avoid claiming they can link a latent print to a single possible finger and instead make statements like the following: "[I]t is far more probable that this degree of similarity would occur when comparing the latent print with the defendant's fingers than with someone else's fingers.[24]" This is an imprecise verbal characterization of a likelihood ratio designed to convey the examiner's opinion that the balance of likelihoods strongly favors the hypothesis of same-source.

Statements of this type may be easier to defend than seemingly precise numerical LRs. An examiner who says the observed results are at least 10,000 times more likely if the prints were made by the same finger than different fingers is likely to face skeptical questions about the basis for that number, while a claim like "far more probable" may be more readily accepted. Of course the problem of unwarranted precision is evaded at the cost of imprecision, and perhaps obfuscation. Does "far more probable" mean 10 times more likely, 100 times more likely, 1000 times more likely, 10,000 times more likely? Without quantification, the meaning of such phrases can be unclear. Moreover, vague terminology may help examiners evade legitimate questions about how accurately they can judge the relevant probabilities, thereby obscuring the shaky scientific foundation for these

---

[22] *See* Sjerps & Berger, *supra* note 21; Biederman, Taroni & Champod, *supra* note 17; William C. Thompson, *Hard Cases Make Bad Law: Reactions to* R v. T., 11 LAW, PROBABILITY & RISK 347 (2012), https://academic.oup.com/lpr/article-pdf/11/4/347/2748692/mgs020.pdf.

[23] *See* Risinger, *Reservations*, *supra* note 1, at 72 ("[T]here is something about the generation of likelihood ratios with numbers from nowhere that tends to cover up the weakness of the ingredients.").

[24] EXPERT WORKING GRP. ON HUMAN FACTORS IN LATENT PRINT ANALYSIS, NAT'L INST. OF STANDARDS & TECH., LATENT PRINT EXAMINATION AND HUMAN FACTORS: IMPROVING THE PRACTICE THROUGH A SYSTEMS APPROACH 134 (2012) ws680.nist.gov/publication/get_pdf.cfm?pub_id=910745.

judgments.[25]

2. Strength-of-Support Statements

LRs suffer the disadvantage of being difficult for lay people to understand.[26] As discussed later in this Article, people often mistakenly believe LRs are a statement about the probability that a particular hypothesis is true, rather than about the strength of the evidence for supporting a particular hypothesis.[27]

One way to clarify the meaning of LRs is to translate them into a verbal statement about the strength of the evidence for supporting a particular hypothesis. An interesting example is the proposal of the United Kingdom-based Association of Forensic Science Providers (AFSP)[28] that forensic scientists use the "verbal expressions" shown in Table 1 as a means of explaining LRs in reports and testimony:

Table 1. Recommended Likelihood Ratio Terminology (AFSP, 2009)

Likelihood Ratio	Verbal Expression (Strength of Support)
1–10	Weak or limited support
10–100	Moderate support
100–1,000	Moderately strong support
1000–10,000	Strong support
10,000–1,000,000	Very strong support
>1,000,000	Extremely strong support

[25] What Professor Risinger, *supra* note 23, called "the weakness of the ingredients" may be covered up as effectively by vague terminology as by subjectively generated numbers.

[26] *See* Kristy A. Martire, Richard I. Kemp, Ian Watkins, Malindi A. Sayle & Ben R. Newell, *The Expression and Interpretation of Uncertain Forensic Science Evidence: Verbal Equivalence, Evidence Strength, and the Weak Evidence Effect*, 37 LAW & HUM. BEHAV. 197 (2013), http://www2.psy.unsw.edu.au/users/bnewell/MKWSN2013.pdf; William C. Thompson & Eryn J. Newman, *Lay Understanding of Forensic Statistics: Evaluation of Random Match Probabilities, Likelihood Ratios, and Verbal Equivalents*, 39 LAW & HUM. BEHAV. 332 (2015); Barbara A. Spellman, *Alternative Suggestions for Communicating Forensic Evidence*, 48 SETON HALL L. REV. 827. The question of how well lay people understand LRs is discussed in more detail later in this article.

[27] Thompson & Newman, *supra* note 26; William C. Thompson, Suzanne O. Kaasa & Tiamoyo Peterson, *Do Jurors Give Appropriate Weight to Forensic Identification Evidence?*, 10 J. EMPIRICAL LEGAL STUD. 359 (2013).

[28] Ass'n of Forensic Sci. Providers, *Standards for the Formulation of Evaluative Forensic Science Expert Opinion*, 49 SCI. & JUST. J. 161 (2009) [hereinafter AFSP Standards].

Table 1 shows a range of verbal expressions designed to be presented in place of (or along with) numerical LRs. For example, a forensic scientist who determines (by whatever means) that the results observed in a forensic comparison are 500 times more probable if the items have a common source than if they have a different source would report that the comparison provides "moderately strong" support for the conclusion that the items have a common source.

The *strength-of-support* statements recommended by the AFSP were not derived from empirical research; they simply reflect the best judgment of that association as to the kind of reporting statements that will be correctly understood by a lay audience. Later in this Article, I will discuss some recent empirical research that casts light on whether lay people perceive these statements to be as strong as the corresponding LRs.

B. *Statements Based on a Two-Step Analysis*

In a two-step analysis, the examiner first compares the items, looking for distinguishing features that rule out the hypothesis that the items have a common source. If no distinguishing features are found, the examiner then considers the rarity or distinctiveness of the shared features, which can lead to several different conclusions.

1. Identification/Individualization Based on Uniqueness of Features

When examiners determine that the features shared by the items are unique or one-of-a-kind, they may conclude, on that basis, that the two items must have a common source. This conclusion is often called an *identification* or *individualization*. In the United States latent print examiners have traditionally limited themselves to reporting one of three possible conclusions: that the prints being compared were made by the same finger (identification/individualization), or that they were made by different fingers (exclusion), or that the results of the comparison are inconclusive.[29] In other pattern-matching disciplines, examiners are allowed to reach a broader array of possible conclusions, but identification/individualization based on uniqueness of features is among the conclusions that examiners sometimes report.[30]

---

[29] *See supra* note 3.

[30] See the reporting standards for latent print examination, *Guideline for the Articulation of the Decision Making Process for Individualization in Friction Ridge Examination*, SCI. WORKING GRP. ON FRICTION RIDGE ANALYSIS, STUDY & TECH. (Mar. 13, 2013), clpex.com/swgfast/documents/articulation/130427_Articulation_1.0.pdf; tool mark examination, Ass'n of Firearms & Tool Mark Examiners, *Theory of Identification*, AFTE J. (1992),

The scientific basis for identification/individualization is weak. A number of commentators have questioned whether forensic scientists can accurately determine whether the features they examine are unique.[31] In 2009, the National Academy of Sciences (NAS) declared in an authoritative report about the state of forensic science: "With the exception of nuclear DNA analysis, however, no forensic method has been rigorously shown to have the capacity to consistently and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source."[32] Ironically, DNA analysts typically report findings using likelihood ratios or match probabilities; "identifications" are most likely to be made in the very disciplines for which NAS found insufficient proof of a capacity to connect evidence to a specific individual or source. The NAS report called on forensic scientists to stop claiming that they can uniquely identify the source of an item, saying "the concept of 'uniquely associated with' must be replaced with a probabilistic association . . . ."[33]

Even for fingerprints, which are sometimes thought to be unique, it is problematic for forensic examiners to claim they can identify individuals on that basis. A 2017 American Association for the Advancement of Science (AAAS) report on latent print examination explained the matter as follows:

> Even if the ridge detail of every finger were unique and unchangeable, it does not follow that every impression made by every finger will always be distinguishable from every impression made by any other finger, particularly if the impressions are of poor quality (e.g., limited detail, smudged, distorted, overlaid on another impression). By analogy, it may be that every human face is unique, but we can still mistake one person for another, particularly when comparing poor-quality photos.[34]

---

https://afte.org/about-us/what-is-afte/afte-theory-of-identification; Ronald G. Nichols, *Defending the Scientific Foundations of the Firearms and Tool Mark Identification Discipline: Responding to Recent Challenges*, 52 J. FORENSIC SCI. 586 (2007); document examination, *SWGDOC Standard Terminology for Expressing Conclusions of Forensic Document Examiners*, SCI. WORKING GRP. FOR FORENSIC DOCUMENT EXAMINATION (Jan. 2015), http://www.swgdoc.org/index.php/standards/published-standards; footwear examination, SCI. WORKING GRP. FOR SHOEPRINT & TIRE TREAD EVID., *Range of Conclusions Standard for Footwear and Tire Impression Examinations* (Mar. 2013), https://www.swgtread.org/images/documents/standards/published/swgtread_10_conclusions_range_201303.pdf.

[31] *See* Michael J. Saks & Jonathan J. Koehler, *The Coming Paradigm Shift in Forensic Identification Science*, 309 SCI. 892 (2005); Cole (2009); Cole (2014), *supra* note 3; Jonathan J. Koehler & Michael J. Saks, *Individualization Claims in Forensic Science: Still Unwarranted*, 75 BROOK. L. REV. 1187 (2010).

[32] NAS REPORT, *supra* note 15, at 7.

[33] *Id.* at 184.

[34] AAAS REPORT, *supra* note 4, at 13.

Based on a comprehensive review of the scientific literature on latent print examination, the AAAS's report concluded that a sufficient scientific basis did not exist for the claim that latent print examiners can "identify" or "individualize" a latent print:

> Examiners may well be able to exclude the preponderance of the human population as possible sources of a latent print, but there is no scientific basis for estimating the number of people who could not be excluded and there are no scientific criteria for determining when the pool of possible sources is limited to a single person.[35]

The same problem arises in other forensic disciplines where examiners claim to be able to identify the source of an item based on the uniqueness of its features. Indeed, most of the pattern-matching disciplines have a weaker scientific foundation than latent print examination; less is known about the rarity of the features sets evaluated by examiners, and less is known about the accuracy of examiners' conclusions.[36]

2. Match Frequencies/Random Match Probabilities (RMPs)

After comparing items and finding no distinguishing features, examiners following a two-step analysis sometimes report that the items are indistinguishable, and then, to explain the probative value of this finding, provide estimates of how frequently indistinguishable items would be found in a reference population. This occurs most commonly in forensic DNA analysis, where genetic databases provide an empirical basis for estimating the frequency of DNA profiles in various human populations. Forensic DNA analysts sometimes present these estimates as *match frequencies*—e.g., "The blood stain at the crime scene and the reference blood sample from the suspect have the same DNA profile. This DNA profile is estimated to occur in 1 in 10 million people among Caucasian-Americans." Alternatively, they may present these estimates as *random match probabilities (RMPs)*—e.g., "The defendant could not be excluded as a possible contributor to the DNA in the crime scene sample. The probability that a random Caucasian-American would fail to be excluded as a possible contributor is 0.0000001 or 1 in 10 million."[37] While experts could in principle use numbers to report their subjective beliefs about the match probability, I am not aware of any cases in which forensic examiners have done that. When numbers are

[35] *Id.* at 5. The 2012 report of the NIST Expert Working Group, *supra* note 24, reached the same conclusion.

[36] *See generally* PCAST REPORT, *supra* note 15.

[37] *See* Butler, *supra* note 12.

reported, they are always derived from databases and statistical models.

Estimates based on empirical data and statistical models are far easier to defend on scientific grounds than assertions about the uniqueness of features or subjective estimates of the match probability. Questions can, of course, be raised about whether the statistical models underlying the estimates are appropriate and whether the databases represent relevant reference populations.[38] When relevant data are available, however, empirically-based estimates of match frequency are undoubtedly preferable to subjective evaluations based on experts' training and experience. The major barrier to wider use of this reporting method, as noted earlier, is that most forensic disciplines do not have adequate databases from which to make such estimates, and that statistical modeling may be complicated by the lack of statistical independence of critical features observed when making comparisons.

One limitation of this reporting method is that RMPs do not always provide a complete account of the probative value of a forensic comparison. RMPs do not take account of discrepancies between the items being compared that may be too minor to justify exclusion but may nevertheless undermine the probative value of the comparison, such as discrepancies between fingerprints that may have arisen from distortion, or discrepancies between DNA profiles that could have arisen from degradation or allelic drop-out.[39] In such instances, it may be better to report results using LRs because match frequencies/RMPs provide an incomplete and potentially

[38] For a general discussion of uncertainty and assumptions in modeling, see Steven P. Lund & Hari Iyer, *Likelihood Ratio as Weight of Forensic Evidence: A Closer Look*, 122 J. RES. NATL. INST. STAND. & TECH. 1 (2017); https://doi.org/10.6028/jres.122.027. For a discussion of modeling assumptions and relevant reference populations in the field of forensic voice comparison, see Morrison & Thompson, *supra* note 14; for a discussion of similar issues in forensic DNA analysis, see DAVID H. KAYE, THE DOUBLE HELIX AND THE LAW OF EVIDENCE (2010); William C. Thompson, *Evaluating the Admissibility of New Genetic Identification Tests: Lessons From the "DNA War,"* 84 J. CRIM. LAW & CRIMINOLOGY 22 (1993).

[39] The problem can be understood most easily by comparing a match frequency with a LR. The LR takes into account two probabilities: (1) the probability of the observed findings if the items have the same source; and (2) the probability of the observed findings if the items have a different source. Match frequencies/RMPs are statements only about the second probability (2). In some cases, the second probability is all that needs to be considered. If the observed findings were certain to be found if the items have a common source, i.e., $p(E|H_s)$=1.0, then all one need consider (to evaluate the strength of the evidence) is the probability of the observed findings if the items have a different source. In such cases, the match frequency or RMP is the complement of the LR and conveys the same information. For example, a random match probability of 1 in 100 would be equivalent in strength to an LR of 100. Problems arise, however, when the features observed when making a particular comparison are ones that are not certain to be observed if the items have the same source, *i.e.*, $p(E|H_s)$<1.0. This can occur when there are discrepancies between the items that make the observed findings unlikely, but not impossible, under the hypothesis that there is a common source.

misleading account of the strength of the forensic comparison.[40]

3. Likelihood of Observed Correspondence

Some laboratories have opted to make qualitative statements about the likelihood that the observed correspondence between the items could arise randomly if the items do not have a common source. In other words, they make verbal statements about the match frequency/RMPs, which may rest in whole or in part on an examiner's subjective judgment.[41] For example, in 2015, the Defense Forensic Science Center (DFSC) of the Department of the Army adopted the following reporting statement for positive latent print comparisons: "The latent print on Exhibit ## and the record finger/palm prints bearing the name XXXX have corresponding ridge detail. The likelihood of observing this amount of correspondence when two impressions are made by different sources is considered extremely low."[42]

This statement is similar to the verbal statement about relative likelihood recommended by the NIST Expert Working Group (as discussed previously). The difference is that the Army statement addresses only the likelihood of the observed pattern under the hypothesis of different source—that is, it speaks only to the denominator of the likelihood ratio. In cases where the likelihood of the observed evidence under the same-source hypothesis is high, there may be little conceptual difference between the Army statement and the NIST Expert Working Group statement, but the Army approach could be misleading in cases where the numerator of the LRs may be significantly less than one.[43] This could arise where, for example, two prints appear to share a distinctive set of features but also have discrepancies that are difficult to explain under the same-source hypothesis.[44]

---

[40] *See* James M. Curran & John Buckleton, *Inclusion Probabilities and Dropout*, 55 J. FORENSIC SCI. 1171–73 (2010) (discussing this issue in connection with forensic DNA evidence); Morrison & Thompson, *supra* note 14, at 358–60 (discussing this issue in connection with forensic voice comparison evidence).

[41] These can be viewed as verbal statements of the examiner's beliefs about the likelihood $p(E|H_d)$.

[42] DEPT. OF THE ARMY DEFENSE FORENSIC SCI. CTR., INFORMATION PAPER RE: USE OF THE TERM "IDENTIFICATION" IN LATENT PRINT TECHNICAL REPORTS 1 (2015), http://onin.com/fp/DFSC_LP_Information_Paper_Nov_2015.pdf.

[43] *See supra* note 39 and accompanying text.

[44] I offer further comments on the Army reporting statement in the concluding section of this Article.

4. Matches/Correspondence Without Frequencies or Probabilities

Examiners sometimes report that they have compared items and found them to be indistinguishable without providing information about the rarity of the matching features or the probability of a random match. The examiner might simply report that the items are "indistinguishable," that they "match," or "correspond," that one "cannot be excluded" from having the same source as another, or similar language.[45] Reports of this type are typically offered when the examiner is uncertain about the rarity of the shared features, or thinks the shared features are not particularly rare.

The problem with this approach is that it provides no meaningful information about the probative value of the forensic evidence, and may imply more than the examiner intends:

> . . . recipients of an opinion expressed as a '*match*' may translate that into meaning that the two '*matching*' samples share the same origin. This would be different from the meaning that the scientist would want to convey, namely that the samples share the same attributes. So, even when scientists and laypeople use the same word, the meaning to these two sets of people can be quite different.[46]

Furthermore, even if the examiner clarifies that *match* means *same attributes* rather than *same source*, the meaning of the "match" remains unclear in the absence of information about the likelihood of observing those attributes under alternative propositions about whether the items have a common source. The probative value of the evidence is a matter about which recipients must guess based on whatever knowledge or preconceptions they have about the evidence in question.[47]

5. True and False Positive Rates

In 2016, the President's Council of Advisors on Science and Technology (PCAST) issued an important report titled "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods."[48] The PCAST report sets forth in considerable detail what is required to establish the validity of a method for assessing whether two items, such as fingerprints or tool marks, have a common source based on a

---

[45] See Jackson, *supra* note 6, for a more complete account of terminology of this type.

[46] *Id.* at 422.

[47] Lay reactions to such testimony are discussed later in this Article. *See infra* note 78, and accompanying text.

[48] PCAST REPORT, *supra* note 15.

comparison of their features. It emphasized the need for empirical testing to establish the accuracy of each method: "Without appropriate estimates of accuracy, an examiner's statement that two samples are similar—or even indistinguishable—is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact. Nothing—not training, personal experience nor professional practices—can substitute for adequate empirical demonstration of accuracy."[49]

PCAST recommended that forensic examiners disclose the error rates observed in black-box validation studies on the method in question when they report and testify about forensic comparisons.[50] Under this approach, error rate data become an integral part of forensic science reporting, allowing the trier of fact better insight into the probative value of the expert's conclusion. When reporting an identification or exclusion, the examiner also reports the rate of false identifications and false exclusion that have occurred in validation studies of the method in question. This changes the nature of the examiner's report from an unqualified statement that the items have the same source (or a different source) to a statement that the examiner has made a determination that is related probabilistically to whether the items have the same source. This shift in reporting makes the examiner's conclusion easier to justify logically and empirically, but may make the report more difficult for recipients to understand and evaluate.

The PCAST report operated on the assumption that forensic scientists will report categorically—that is, the examiner, after comparing items, will report one of a limited number of conclusions, e.g., "identification," "exclusion," or "inconclusive." It is fairly easy to establish hit rates and false positive rates when experts use a limited number of reporting categories. Error rates are more difficult to evaluate when experts use continuous scales (e.g., LRs or source probabilities), although with enough data it is possible to establish the relationship between the numbers expert's report and ground truth.[51] One might calculate, for example, how much more likely experts are

[49] *Id.* at 46.

[50] Black-box studies are designed to test examiners' accuracy when comparing items known (to researchers, but not examiners) to have the same source, or to have a different source. The goal of these studies is to determine how often examiners correctly and incorrectly determine the source of the items they are asked to compare. Ideally these studies require examiners to evaluate items that are comparable to the items encountered in casework following procedures that are as similar as possible to those used in casework. *Id.* at 66. *See also* AAAS REPORT, *supra* note 4, at 43–51 (discussing ways to conduct empirical studies on the accuracy of latent print examination).

[51] For further discussion of the validation of likelihood ratios, see Geoffrey Stewart Morrison, *Measuring the Validity and Reliability of Forensic Likelihood-Ratio Systems*, 51 SCI. & JUST. 91 (2011), and Didier Meuwly, Daniel Ramos & Rudolf Haraksim, *A Guideline For the Validation of Likelihood Ratio Methods Used For Forensic Evidence Evaluation*, 276 FORENSIC SCI. INT'L 142 (2017), https://www.sciencedirect.com/science/article/pii/S0379073

to report a LR in a given range (e.g., 10–100) when evaluating same-source samples than different-source samples. Research of this type would be helpful for determining whether LR estimates are well-calibrated and for identifying weaknesses in LR-based methods, although results of this type will undoubtedly be more difficult to explain to juries than hit rates and false positive rates for a categorical reporting system.

C. *Statements Based, in Part, on Prior Odds of a Common Source*

One virtue of the reporting methods discussed thus far is that examiners can reach the reported conclusions by examining and comparing the items in question without giving any consideration to other evidence in the case. For example, a latent print examiner can reach conclusions based on an evaluation of the ridge patterns and other features of the prints being compared. The examiner need not consider any other evidence (e.g., DNA test results; investigative facts; witness statements; suspect's alibi) that bears on the claim that the prints have a common source. These approaches are therefore consistent with a recommendation of the National Commission on Forensic Science that forensic scientists avoid being influenced by "task-irrelevant" information.[52] According to the National Commission, forensic scientists should draw source conclusions by considering the physical properties of the items being compared and any information needed to evaluate likelihood of observing those properties under relevant hypotheses about the source of the items. The Commission offered persuasive reasons why forensic scientists should not consider other information about the case, even if that information can be used to draw inferences about whether the items have a common source. For example, information that a suspect confessed to touching an item might support an inference that a fingerprint found on the item is his, but a latent print examiner should not consider that information because the examiner is supposed to draw conclusions from an examination of the prints, not from other evidence in the case.[53]

816301359.

[52] Nat'l Comm. on Forensic Sci., *Views of the Commission: Ensuring That Forensic Analysis is Based Upon Task-Relevant Information* (Dec. 8, 2015), https://www.justice.gov/archives/ncfs/file/818196/download [hereinafter National Commission, 2015].

[53] For additional discussion of the concept of task-relevance, and of reasons forensic scientists should avoid being influenced by task-irrelevant information, see William C. Thompson, *Determining the Proper Evidentiary Basis for an Expert Opinion: What Do Experts Need to Know and When Do They Know Too Much?*, *in* BLINDING AS A SOLUTION TO BIAS IN BIOMEDICAL SCIENCE AND THE COURTS: A MULTIDISCIPLINARY APPROACH 133–49 (Christopher T. Robertson & Aaron S. Kesselheim eds., 2016) [hereafter Thompson, 2016]; William C. Thompson, *What Role Should Investigative Facts Play in the Evaluation of Scientific Evidence?*, 43 AUSTL. J. FORENSIC. SCI. 123 (2011); D. Michael Risinger, Michael J. Saks, William C. Thompson & Robert Rosental, *The Daubert/Kumho Implications of Observer Effects in Forensic Science: Hidden Problems of Expectation and Suggestion*, 90

Despite the Commission's recommendation, forensic scientists sometimes report source conclusions in a way that requires them to evaluate or make assumptions about the *prior odds* that the items being compared have a common source.[54] These reporting statements require the examiner, either implicitly or explicitly, to make assumptions or draw conclusions about the strength of the other evidence for supporting the hypotheses of a common source. Because this method of reaching source conclusions differs in this important way from both the first approach (balance of likelihoods) and the second approach (two-step analysis), I discuss it here as a third major approach to reporting. Using this third approach, examiners can reach two kinds of conclusions that are not logically possible under the first two approaches.

1. Statements About Source Probability

Forensic examiners sometimes offer opinions on the probability that two items have a common source. Opinions of this type can be expressed quantitatively, using probabilities or percentages. For example, a forensic scientist might say there is a 99% chance that two items have a common source. It is more common, however, for examiners to express such conclusions with words rather than numbers. For example, the forensic scientist might say it is "moderately probable;" or "highly probable;" or "practically certain" that two items have a common source.[55]

A number of commentators have criticized this reporting method on grounds that it requires forensic scientists to look beyond the forensic evidence and consider (or take positions) on matters beyond their expertise, matters that have traditionally been left to the trier-of-fact in criminal cases rather than evaluated by experts.[56] The simplest way to illustrate this point

CAL. L. REV. 1 (2002).

[54] I borrow the term *prior odds* from the field of Bayesian statistical inference where it is used to refer to an evaluator's perception of the odds that a hypothesis is true before considering a relevant item of evidence. For example, in a latent print analysis, the prior odds would be the examiner's perception of the odds that the prints in question have a common source *before* conducting the examination of the prints. Bayesian statistical analysis concerns that manner in which an evaluator's prior odds should be updated in light of new evidence. *See* ROBERTSON, VAGNAUX & BERGER, *supra* note 5.

[55] *See* Jackson, *supra* note 6.

[56] After comparing two items, a forensic examiner may be able to estimate the likelihood of the observed results under the alternative hypotheses: $p(E|H_s)$ and $p(E|H_d)$. But these likelihoods are not the same as source probabilities; source probabilities are the inverse of these conditionals—i.e., $p(Hs|E)$ and $p(Hd|E)$. To infer source probabilities from the likelihoods, the examiner must take into account the prior probability that the items have the same source, $p(H_s)$, or different source, $p(H_d)$. According to Bayes' rule, $p(H_s|E)/p(H_d|E) = p(H_s)/p(H_d) \times p(E|H_s)/p(E|H_d)$. This means that conclusions about source probability cannot rest solely on what the examiner observes when making the comparison but must also depend on assumptions or conclusions about the *a priori* probability the items have the same source.

is with an example. Suppose that a DNA analysis shows that a bloodstain found at a crime scene has the same DNA profile as a particular suspect. How rare would that profile need to be in order for an examiner to conclude that it probably came from the suspect? Suppose, for example, that the DNA profile in question would be found in only one person in 1 million in the general population. Would that be rare enough to justify a conclusion that the profile *probably* came from the suspect?

A moment's reflection should make it clear that the forensic scientist can draw no conclusion about the source probability based solely on the information given so far. To assess source probabilities, the forensic scientist must also consider (or make assumptions about) other evidence in the case. Consider that in a nation as large as the United States there are likely to be over 300 people who share the one-in-a-million DNA profile found in the bloodstain. If there was no other evidence of the suspect's guilt, beyond the DNA match, then it is not necessarily likely that he, rather someone else with the matching profile, was the source of the bloodstain.[57] Indeed, if the suspect has a solid alibi, it might be far more likely that the bloodstain came from someone else, notwithstanding the DNA match. On the other hand, if the suspect already appeared likely, based on other evidence, to be the source of the bloodstain, then finding he also shares the one-in-a-million DNA profile with the bloodstain might well support a conclusion that he is likely to be the source of the stain. My point is that the source probability cannot be inferred from the DNA evidence alone; it also depends on how strongly other evidence points toward or away from the suspect.

The same problem arises when forensic examiners attempt to infer source probabilities from any type of forensic evidence. This is an inherent problem of source probabilities. Even if the examiner does not realize it, a statement that a suspect is "highly likely" to be the source of a bloodstain, for example, necessarily rests, in part, on an implicit assessment or assumption about the prior odds that the suspect is the source. Making those assessments or assumptions takes forensic scientists beyond their scientific expertise in ways that arguably usurp the role of legal fact-finders. Experts are rarely in a good position to evaluate the prior odds that the items they are comparing have a common source, and arguably have no business doing so. Consequently, many commentators have suggested that forensic experts

Consequently, examiners must necessarily consider or make assumptions about matters beyond forensic science in order to reach source conclusions. *See*, I.W. Evett, *Toward a Uniform Framework For Reporting Opinions in Forensic Science Casework*, 38 SCI. & JUST. 198 (1998); Buckleton, *supra* note 6; Morrison, *supra* note 6; ROBERTON, VIGNAUX & BERGER, *supra* note 5; Thompson (2012), *supra* note 22.

[57] There might be little or no evidence of a suspect's guilt other than a DNA match in a case in which the suspect is identified through a "cold-hit" in a search of a large database.

avoid presenting source probabilities.[58] As Professor Redmayne explained: . . . the expert should not testify in terms such as . . . : the blood probably came from the defendant", because one can only reach conclusions of this sort by making assumptions about the strength of other evidence against the defendant.[59]

Forensic scientists who report in this manner arguably have a special obligation to explain in their reports and testimony that their conclusions are not based solely on a comparison of the items in question but also rest partly on their assumptions about the strength of other evidence: "Unless the receiver of the opinion understands the prior, non-scientific information that influenced the scientist, there can be no assessment of the reliability and fairness of the opinion."[60]

2. Identifications Without Uniqueness

The same problem sometimes arises when forensic scientists claim to have "identified" or "individualized" the source of an item. In recent years, forensic scientists in pattern matching disciplines have begun to acknowledge that they do not have a scientific basis for determining whether the features of a particular item (e.g., a latent print) are unique in the entire world. For example, some latent print examiners no longer claim to be able to associate a particular latent print to a single finger to the exclusion of all other fingers in the world. They nevertheless claim to know (or be able to determine) whether the features of a latent print are rare enough that they are unlikely to be duplicated among the smaller group of individuals who are possible sources of a particular latent print. They reason that the possible sources of a latent print at a typical crime scene is likely to be limited to people with access to the scene, and this group (which is sometimes called "the relevant population") may be much, much smaller than the entire human population. Consequently, to identify the suspect as the source of the latent print in question, they do not need to know whether the features shared by the latent print and his print are unique in the entire world; they just need to know that those shared features are sufficiently rare that it is unlikely another member of the relevant population will have those same features. Thus it is possible to have identification without uniqueness.[61]

---

[58] Evett, *supra* note 56; Buckleton, *supra* note 6; Morrison, *supra* note 6; ROBERTON, VIGNAUX & BERGER, *supra* note 5; Thompson (2012), *supra* note 22.

[59] MIKE REDMAYNE, EXPERT EVIDENCE AND CRIMINAL JUSTICE 46 (2001).

[60] Jackson, *supra* note 6, at 426.

[61] Another way to explain this approach is to say the examiner's assessment of the "uniqueness" of the observed features is made with respect to a limited, localized sub-population, rather than for the entire population of such items.

While the logic of this approach seems sound, it clearly requires the examiner to consider matters beyond the characteristics of the items being compared. The examiner must determine that the suspect is a member of the relevant population—e.g., a possible source of the latent print, which presumably entails either an evaluation of, or assumptions about, the strength of any alibi the suspect might advance. It also requires the examiner to know enough about the case to judge the size of the relevant population—i.e., how many people other than the suspect might have been the source of the item in question. That, in turn, requires the examiner to consider, or make assumptions about, a number of matters that are likely to be contested if the case is tried. So, like examiners who present source probabilities, examiners who make identifications in this manner must make assumptions about the strength of other evidence against the suspect.

As with source probabilities, examiners who "identify" the source of items in this manner arguably incur a special obligation to disclose their underlying assumptions. Suppose, for example, that an examiner "identified" an individual as the source of a particular item based, in part, on the assumption that the suspect was one of a small number of people who could have been the source. Without knowing that the examiner's conclusion depended on this assumption, recipients of this information have no way to assess whether the conclusion is reasonable and fair.[62] Even if the assumption is disclosed, it might be difficult for recipients to judge the probative value of this evidence, particularly if they had reason to question or disagree with the examiner's assumption.[63]

IV. EVALUATING LAY REACTIONS TO FORENSIC EVIDENCE

Having discussed various ways source conclusions might be presented in reports and testimony, I will now turn to the question of how the presentation format affects lay people's perceptions of this evidence. This requires discussion of a growing body of research on lay reactions to forensic science evidence.[64]

---

[62] *See* Jackson, *supra* note 6, at 426.

[63] To avoid "double-counting" of evidence, a fact-finder would need to assess the "incremental probative value" of the expert's conclusion, which is the value added by the examiner's opinion beyond the value provided by other evidence in the case that the fact-finder has already considered. *See* Thompson, 2016, *supra* note 53, for explanation of this point.

[64] For recent reviews of this literature, see Graham Jackson, David H. Kaye, Cedric Neumann, Anjali Ranadive & Valerie F. Reyna, *Communicating the Results of Forensic Science Examinations: Final Technical Report for NIST Award 70NANB12H014* (Penn State Law Research Paper No. 22-2015, 2015), https://papers.ssrn.com/sol3/papers.cfm?abstract id=2690899; Dawn McQuiston-Surrett & Michael J. Saks, *Communicating Opinion Evidence in the Forensic Identification Sciences: Accuracy and Impact*, 59 HASTINGS L.J. 1159 (2008). For reviews of earlier studies, *see* D.H. Kaye & Jonathan J. Koehler, *Can Jurors*

The most common method for studying lay reactions is a trial simulation study in which participants are asked to evaluate hypothetical criminal cases while the researchers experimentally vary the nature of the evidence. Studies relevant to forensic evidence have experimentally varied whether the evidence against the defendant includes (or does not include) testimony of a forensic examiner; these studies have also varied the type and strength of the forensic science evidence and the presentation format.[65] The studies differ in how elaborately they present the hypothetical case, ranging from simple written summaries of evidence to elaborate simulations of trial testimony. Participants in these studies also vary, ranging from undergraduates, to participants in online research panels, to people recruited from jury pools.

A. *Lay Perceptions of the Relative Strength of Various Reporting Statements*

Studies that have tested the effects of presentation format on people's reactions to forensic science evidence have often found that the format makes a difference. Presentations that should logically be given the same weight are sometimes treated differently. For example, Thompson and Schumann[66] found that participants in a trial simulation study gave more weight to statements about conditional probability (e.g., "a two percent chance the defendant's hair would be indistinguishable . . . if he were innocent") than statements that focused on the percentage and number ("2% of people would be indistinguishable; that would be 20,000 in a city of 1 million"). Goodman[67] and Lindsey et al.[68] found that people gave more weight to RMPs when stated as percentages (e.g., 1/10th of 1%) than frequencies (1 in 1000).

Koehler and his colleagues have also shown that minor (and logically irrelevant) variations in the way statistics are presented in connection with a forensic match can have striking effects on the weight people give the

Understand Probabilistic Evidence?, 154 J. ROYAL STAT. SOC. 75 (1991); William C. Thompson, *Are Juries Competent to Evaluate Statistical Evidence?*, 52 LAW & CONTEMP. PROBS. 9 (1989).

[65] Some of the studies employ between-subject designs in which different versions of the case are presented to different groups of participants. Some of the studies employ within-subject designs in which participants evaluate the case before and after receiving forensic evidence.

[66] William C. Thompson & Edward L. Schumann, *Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy*, 11 LAW & HUM. BEHAV. 167 (1987).

[67] Jane Goodman, *Jurors' Comprehension and Assessment of Probabilistic Evidence*, 16 AM. J. TRIAL ADVOC. 361 (1992).

[68] Samuel Lindsey, Ralph Hertwig & Gerd Gigerenzer, *Communicating Statistical DNA Evidence*, 43 JURIMETRICS 147 (2003).

evidence. For example, participants in a trial simulation study gave much more weight to a DNA match when the random match probability was expressed as a percentage using language that focuses on the suspect ("the probability the suspect would match the blood drops if he is not their source is 0.1%") than when it was expressed as a frequency within a broader reference group ("One in 1000 people in Houston would also match the blood drops").[69] This effect probably arises from people's tendency to evaluate the strength of a forensic match according to the ease with which they can imagine that other people, besides the suspect, might match. According to "exemplar cueing theory," formats that draw attention to the number of people who could match in a large population make it easier to imagine false matches occurring, which reduces the impact of the forensic evidence.[70]

Several studies have compared perceptions of LRs with perceptions of comparable match frequencies (or RMPs), although the results have been mixed. Koehler found that jury-eligible students judging a DNA case gave more weight to LRs (100 times more likely) than comparable match frequencies (1 person out of every 100).[71] Nance and Morris conducted two important studies with members of a jury pool.[72] In both studies they found that participants gave significantly more weight to LRs than RMPs, but only when the LRs were accompanied by a chart explaining how to update a prior probability in light of a LR. Without the chart, the weight given LRs and RMPs did not significantly differ. Thompson & Newman[73] found that participants recruited from an online labor pool did not differ in the weight they gave LRs and RMPs when evaluating DNA evidence; but found participants gave more weight to RMPs than LRs when evaluating shoe print evidence.

With some colleagues at the University of California, Irvine I have recently been studying lay perceptions of the relative strength of various conclusions that a forensic scientist might present about whether two items (fingerprints; biological samples) have a common source.[74] We adopted a

---

[69] Jonathan J. Koehler, *When Are People Persuaded by DNA Match Statistics?*, 25 LAW & HUM. BEHAV. 493 (2001).

[70] Jonathan J. Koehler & Laura Maachi, *Thinking About Low Probability Events: An Exemplar-Cuing Theory*, 15 PSYCH. SCI. 540 (2004).

[71] Jonathan J. Koehler, *On Conveying the Probative Value of DNA Evidence: Frequencies, Likelihood Ratios and Error Rates*, 67 U. COLO. L. REV. 859 (1996).

[72] Dale A. Nance & Scott B. Morris, *An Empirical Assessment of Presentation Formats for Trace Evidence With a Relatively Large and Quantifiable Random Match Probability*, 42 JURIMETRICS J. 403 (2002); Dale A. Nance & Scott B. Morris, *Juror Understanding of DNA Evidence: An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Small Random Match Probability*, 34 J. LEGAL STUD. 395 (2005).

[73] *Supra* note 26.

[74] The materials reported here are drawn from a working paper: William C. Thompson,,

method known as paired comparison that was originally used in the field of psychometrics to study perceptions of the strength of physical stimuli (e.g., the brightness of a light, the intensity of a sound).[75] In a series of studies, we presented statements to participants[76] in pairs and asked participants to judge which statement indicated the evidence was stronger for proving the items have a common source. For example, in one study participants were asked to evaluate the relative strength of two statements about the result of a fingerprint comparison: (1) the examiner's claim to have "identified" the prints as coming from the same finger, or (2) the examiner's claim that "the likelihood of observing this amount of corresponding ridge detail when two fingerprints are made by different people is less than 1 in 100,000." Various statements were paired randomly and each participant was asked to evaluate the relative strength of about sixteen random pairs. By combining data across participants, we were able to generate rankings of the perceived strength of the statements, relative to one another. By fitting the data to statistical models we were able to assess the statistical significance of differences in rankings.[77]

Our results suggest that statements involving large numbers—either RMPs and LRs—are perceived as very powerful. For a fingerprint comparison, a RMP of 1 in 100,000 was as strong as the categorical statement that the examiner had "identified" or "individualized" the print; the RMP of 1 in 10 million was even stronger. For DNA evidence, a LR of 100,000 was as strong as the categorical statement that the evidence proved the suspect "was the source;" and a LR of 10 million was stronger still.

Rebecca Grady, Eric Lai & Hal S. Stern, Perceived Strength of Forensic Scientists' Reporting Statements About Source Conclusions (Dec. 10, 2017) (unpublished draft) (on file with authors). A version of this paper has been accepted for publication in the journal *Law, Probability and Risk*.

[75] People have difficulty providing meaningful evaluations of the strength of such stimuli on rating scales (e.g., "How loud is this sound on a scale of 1-10?"). Responses tend to be unreliable, poorly calibrated, and affected by contextual factors, such as the volume of previously heard sounds, and by the nature of the rating scale. People do better, however, when reporting which of two sounds is louder than when rating the loudness of various sounds on a scale. In 1927, L.L. Thurstone demonstrated that pair-wise comparison can be used to order multiple items in a scale of strength or magnitude. L.L. Thurstone, *A Law of Comparative Judgment*, 34 PSYCHOL. REV. 273 (1927). We used this method to estimate the perceived strength of forensic science reporting statements.

[76] Participants were jury-eligible U.S. adults from mTurk, an online labor pool.

[77] The data were fitted to Thurstone-Mosteller paired comparison models to obtain rank-ordered lists of the various statements and an indication of the perceived differences among them. *See*, Thurstone, *supra* note 75; Fredrick Mosteller, *Remarks on the Method of Paired Comparisons: I. The Least Squares Solution Assuming Equal Standard Deviations and Equal Correlations*, 16 PSYCHOMETRIKA 3 (1951); Hal Stern, *A Continuum of Paired Comparisons Models*, 77 BIOMETRIKA 265 (1990).

Some non-numerical statements about strength of support were also perceived to be powerful. Saying that a fingerprint comparison provides "extremely strong support" for the theory that the suspect made the print was seen as roughly equivalent to saying that it was "a practical certainty that the suspect was the source;" saying the comparison provides "extremely strong support" for the theory that the suspect made the print was perceived to be stronger than saying that it is "highly probable" the suspect made the print. Saying a DNA match provides "extremely strong support" for the theory of a common source was seen as roughly equivalent to saying that the RMP is 1 in 100,000 or the LR is 100,000. So our results suggest that it is possible to make a strong statement about the probative value of a forensic comparison by talking about strength of support rather than relying on the traditional claim of identification/individualization.

Interestingly, simply saying that the ridge patterns of two fingerprints "match" was also perceived to be an extremely strong statement about the probative value of a fingerprint comparison.[78] This finding raises concerns because, as noted earlier, forensic examiners use the term "match" merely to convey that the items share the same attributes, which does not necessarily imply that the forensic comparison is highly probative.[79] It is not clear whether the term "match" will be equally powerful with other forms of forensic evidence, but our findings suggest that forensic scientists should use this term cautiously, if at all, when reporting their conclusions, particularly when there is uncertainty about the probative value of the matching features for proving the items have a common source.

The statement on likelihood of correspondence proposed by the Defense Forensic Science Center of the United States Army was perceived to be significantly weaker, when used to characterize a fingerprint comparison, than reporting "identification" or "individualization." It was also weaker than saying the results provide "extremely strong support" for the theory of a common source Relative to the RMP statements, the army statement fell between "1 in 100,000" and "1 in 1000," differing significantly from both.

With regard to the verbal expressions recommended by the Association of Forensic Science Providers (*see* Table 1), our findings suggest that two of these expressions—"weak support" and "moderate support" are indeed perceived in the manner intended. In other words, participants perceived the verbal expression as being roughly equal in strength to the corresponding

[78] People also give a great deal of weight to testimony that two hairs "match." *See* Dawn McQuiston_Surrett & Michael J. Saks, *The Testimony of Forensic Identification Science: What Experts Say and What Factfinders Hear*, 33 LAW & HUM. BEHAV. 436 (2009); McQuiston-Surrett & Saks, *supra* note 64.

[79] *See supra* note 46 and accompanying text.

LR. On the other hand, our participants thought the expression "extremely strong support" was weaker than the corresponding LR. AFSP recommended using the term "extremely strong support" when the LR exceeds 1 million, but our participants found "extremely strong support" to be equivalent or weaker than a RMP of 1 in 100,000. Our findings suggest that forensic scientists who are seeking a verbal statement comparable in strength to a RMP of 1 in 1 million, or a LR of 1 million, more may need to find something stronger than "extremely strong support."

Of course, it is important to consider whether such strong statements are warranted when describing the strength of forensic source comparisons in disciplines other than DNA analysis. If it would be an exaggeration to report a LR of 100,000 or higher when explaining the strength of a latent print, tool mark, or footwear comparison, then arguably it is also an exaggeration to say that the comparison provides "extremely strong support" for the theory of a common source (given that "extremely strong support" is viewed as equivalent in strength to the LR of 100,000). In this regard it is noteworthy that the false identification rate of latent print examiners in the largest black-box study of examiner accuracy[80] was approximately 0.17%, or one false identification for every 588 comparisons of prints from different people for which examiners were able to reach a source conclusion.[81] Even higher error rates have been observed in some studies.[82] The error rates observed in these studies suggest that the LR describing the strength of a latent print identification may well be closer to 1000 than to 100,000. Hence, reporting that a latent print comparison provides "extremely strong" support for a common source may be more than adequate to convey an accurate impression of the strength of this evidence. In fact, if one were seeking a verbal expression that is equivalent to reporting a LR of 1000, then the statement proposed by the United States Army's Defense Forensic Science

[80] Bradford T. Ulery, R. Austin Hicklin, Joann Buscaglia & Maria Antonia Roberts, *Accuracy and Reliability of Forensic Latent Fingerprint Decisions*, 108 PROC. NAT'L. ACAD. SCI. 7733 (2011).

[81] For an analysis of the error rates in this study, *see* AAAS REPORT, *supra* note 4, at 55 ("A total of 4083 different-source pairs were deemed of value for identification, and examiners were able to make conclusive calls on 3638 of those pairs. Six of those calls were erroneous identifications (0.17%)"). The PCAST report, *supra* note 15, also analyzed this study and reported that the upper 95% confidence bound of the false identification rate was 0.33%, which corresponds to 1false identification for every 306 different-source comparisons that led to a source determination.

[82] *See, e.g.*, Igor Pacheco, Brian Cerchiai & Stephanie Stoiloff, *Miami-Dade Research Study for the Reliability of the ACE-V Process: Accuracy & Precision in Latent Fingerprint Examinations*, MIAMI DADE POLICE DEP'T FORENSIC SERVS. BUREAU (Dec. 2014), https://www.ncjrs.gov/pdffiles1/nij/grants/248534.pdf. For detailed discussion of error rates in latent print examinations, see AAAS REPORT, *supra* note 4; Simon A. Cole, *More than Zero: Accounting for Error in Latent Fingerprint Identification*, 95 J. CRIM. L & CRIMINOLOGY 985 (2005).

Center[83] would be a more reasonable choice.

One must be cautious about arguments of this kind, however, because they rely on research that compares the perceived strength of various possible reporting statements *relative to one another*. It seems reasonable to infer that people will give roughly equal weight to a latent print comparison when the examiner reports a LR of 100,000 as when the examiner reports that the comparison provides "extremely strong support" for a common source. But the research discussed thus far does not tell us whether the weight given to the evidence when presented in either manner will be appropriate. People may see the statements as roughly equivalent in strength but may give both statements more weight or less weight than they deserve.

Readers should also bear in mind that participants in these studies were responding to short written summaries of the examiners' conclusions about fingerprint and DNA comparisons, similar to what might be found in a written report. The same statements may well be viewed differently when presented in connection with other types of forensic evidence. It is also possible that people will respond differently to such conclusions when examiners have the opportunity to explain and elaborate upon them during testimony. In order to resolve these questions, additional research is needed on how widely these finding generalize across evidence types and presentation modalities.

B. *Assessing the Appropriateness of Lay Responses to Forensic Science Evidence*

What is an "appropriate" response to a forensic scientist's testimony? How do we determine whether a particular juror's conclusions are based on clear-eyed understanding of the expert's testimony, rather than a biased or incorrect distortion thereof? Suppose, for example, that a shift in reporting format—from "individualization" to LRs, for example—produces a change in lay reactions to forensic science evidence. How do we judge whether that change is harmful or beneficial? In the sections that follow, I will propose two criteria for assessing appropriateness: whether people's responses are sensitive to the strength of the forensic evidence, and whether people's responses are logically consistent with the forensic evidence.[84]

---

[83] *See supra* note 42, and accompanying text.

[84] Thompson & Newman, *supra* note 26, proposed three criteria for evaluating the appropriateness of reactions to forensic science evidence, which they called sensitivity, logical coherence, and susceptibility to fallacy. In the analysis presented here, I combine the second and third of these categories.

1. Sensitivity to the Strength of Evidence

If people are responding to forensic evidence appropriately, their judgments should be sensitive to the strength of the forensic evidence. They should give more weight to forensic evidence when it is strong and therefore deserves more weight; they should give less weight to forensic evidence when it is weak and therefore deserves less weight. Consequently, we should prefer presentation formats that promote sensitivity to the strength of evidence, and avoid presentation formats that render people insensitive to the strength of evidence.

Several studies have examined lay people's sensitivity to variations in RMPs when evaluating forensic science evidence in hypothetical criminal cases. The majority of these studies found that people give more weight to the forensic evidence when the RMP is low than when it is higher, as they should.[85] There were two exceptions in which variations in RMP did have a statistically significant effect on the weight given to forensic evidence,[86] but the literature as a whole suggests that people understand and respond appropriately to this variable. This conclusion is bolstered by the recent research my colleagues and I conducted on perceptions of the relative strength of reporting statements. When we asked people to compare two statements about RMPs for DNA or fingerprint evidence, most people correctly perceived the statement with the lower RMP to be stronger.[87]

Whether people are also sensitive to variations in LRs and Strength of Support Statements is less clear. Martire and her colleagues have reported that people "were only weakly sensitive to large differences" in LRs and Strength-of-Support statements when evaluating shoeprint evidence.[88] Thompson and Newman[89] also found that people were insensitive to LRs and Strength-of-Support statements when evaluating shoeprint evidence, but found that people were sensitive to these variables when evaluating DNA

[85] David L. Faigman & A.J. Baglioni, Jr., *Bayes' Theorem in the Trial Process: Instructing Jurors on the Value of Statistical Evidence*, 12 LAW & HUM. BEHAV. 1 (1988); Goodman, *supra* note 67 (Study 2); Brian C. Smith, Steven D. Penrod, Amy L. Otto & Roger C. Park, *Jurors' Use of Probabilistic Evidence*, 20 LAW & HUM. BEHAV. 49 (1996); Thompson & Newman, *supra* note 26.

[86] Goodman, *supra* note 67 (Study 1) (reporting that varying match frequencies of 10%, 5%, 1% and 0.1% produced only slight changes in participants' reactions to blood group evidence); Koehler, *supra* note 71 (finding no significant differences in reactions to a match frequency of 1 in 100 versus 1 in 1000 in a DNA case).

[87] Thompson, Grady, Lai & Stern, *supra* note 74. A minority of participants in these studies initially said the statement with the higher RMP was stronger, but most of them realized their error when given additional explanation.

[88] Martire et al., 2013, *supra* note 26; *see also* Kristy A. Martire et al., *On the Interpretation of Likelihood Ratios in Forensic Science Evidence: Presentation Formats and the Weak Evidence Effect*, 240 FORENSIC SCI. INT'L 61 (2014).

[89] *Supra* note 26.

evidence. Thompson and Newman speculated that people's preconceptions about forensic evidence may have made them skeptical when experts presented large LRs and strong Strength-of-Support statements in connection with shoeprint evidence, but more accepting of such statements when offered in connection with DNA evidence.[90] Additional research to follow up on these findings is clearly warranted.

The studies by Martire et al. and by Thompson and Newman required participants to evaluate written summaries of evidence. In a more recent set of studies, my colleagues and I have asked participants to view videos of simulated testimony by a forensic voice comparison expert who used likelihood ratios to describe the strength of his findings, and offered a clear explanation of what a LR is. These studies found that participants' judgments were sensitive to the strength of the reported LR—they gave significantly more weight to the voice print evidence when the reported LR was 3000 than when it was thirty.[91] While these findings are reassuring with respect to LRs, much more research is needed to test people's understanding of LRs and Strength of Support statements.

There is has been relative little research on lay reactions to the kind of error rate data recommended by PCAST, although the reported studies indicate people can be sensitive to error rate data. Thompson, Kaasa & Peterson[92] found that undergraduates were sensitive to the probability of a false match when evaluating the strength of DNA evidence. In a second study they found that members of a jury pool did not necessarily accept an expert's estimate of the probability of a false DNA match due to laboratory error, although their assessments of the strength of the DNA evidence varied appropriately in accordance with their own estimates of the probability of a false match.

In an earlier study, Kaasa et al. found that undergraduate participants in a jury simulation study gave more weight to bullet lead evidence when they were provided with statistical data suggesting it had strong "diagnostic value" than when the data suggested the evidence was worthless (non-diagnostic), or when no data were presented.[93] Interestingly, the study found

[90] Interestingly, Thompson and Newman also found that judgments about shoeprint evidence were sensitive to variations in RMPs. People tended to give more weight to shoeprint evidence when the RMP was low (1 in 1 million) than when it was more moderate (1 in 1000), but a similar variation in LRs produced no effect.

[91] William C. Thompson, Rebecca Grady & Eryn J. Newman, Lay Understanding of Likelihood Ratios Presented by a Forensic Voice Comparison Expert (unpublished manuscript) (on file with author).

[92] *Supra* note 27.

[93] Suzanne O. Kaasa, Tiamovo Peterson, Erin K. Morris & William C. Thompson, *Statistical Inference and Forensic Evidence: Evaluating a Bullet Lead Match*, 31 LAW & HUM. BEHAV. 433 (2007). Participants were asked to judge the value of bullet lead evidence for proving a bullet found at a crime scene came from a stock of bullets owned by a defendant.

that group deliberation improved participants' ability to draw conclusions from the statistical data. It was only after group deliberation that participants appreciated that the bullet lead evidence deserved no weight in the condition where statistical data indicated the evidence was non-diagnostic and was therefore worthless. The study also found variation across participants in their sensitivity to variations in the statistical data. Participants who expressed more confidence in their ability to understand and use numerical data were more likely to respond to the statistical data in the appropriate manner (by giving weight to the evidence only when the data indicated it had diagnostic value), while those who expressed less confidence were insensitive to the variation in statistical data.

A recent study on people's reaction to negative evidence—the failure to detect gunshot residue on an individual suspected of firing a gun—found that participants recruited from an online database were sensitive to statistical data about the probability of detection, although the authors note that "jurors may evaluate negative evidence according to a fairly crude metric—-giving it no weight if the probability of detection is zero, a great deal of weight if the probability of detection is 100%, and moderate weight if the probability of detection is somewhere in between." [94]

2. Logical Coherence

A second criterion for evaluating whether people are responding to forensic science evidence appropriately is whether their responses follow logically from the evidence. If a particular presentation format causes people to respond to the evidence in a manner that is illogical, or logically incoherent,[95] that is cause for concern. Researchers have often used Bayes' rule as a normative standard for determining whether people's responses to forensic evidence are logical.[96]

Statistics were provided on the probability of finding bullets matching the chemical profile of a crime scene bullet among the defendant's stock of bullets and the probability of finding such bullets among a sample of bullets found in the broader community were the crime occurred. When the former probability greatly exceeded the latter, the evidence was deemed diagnostic (meaning it has strong probative value); when the two probabilities were equal, the evidence was non-diagnostic, meaning it had no probative value for determining whether the crime scene bullet came from the defendant or someone else.

[94] William C. Thompson, Nicholas Scurich, Rachel Dioso-Villa & Brenda Velazquez, *Evaluating Negative Forensic Evidence: When Do Jurors Treat Absence of Evidence as Evidence of Absence?*, 14 J. EMPIRICAL LEGAL STUD. 569, 569 (2017).

[95] Logical incoherence consists of holding beliefs that are logically inconsistent with one another. *See* Thompson & Newman, *supra* note 26. An example of incoherence would be a person who believes that the chances of a false DNA match are extraordinarily low (which implies that the DNA evidence deserves a lot of weight), but who updates his beliefs about the source of an item relatively little after receiving the DNA evidence (which implies that the DNA evidence deserves little weight).

[96] *See infra* note 97. Bayes' rule is a formal description of how a rational actor should

i. Underutilization of Forensic Evidence

A number of researchers have concluded that people "underutilize" forensic science, which means that they give less weight to the evidence than would be logical, given their beliefs about it.[97] Based on this conclusion, some scholars have argued that we need not fear that forensic science evidence will be overvalued and prejudicial, and that forensic scientists should favor presentation formats that maximize the impact of forensic science evidence in order to combat the tendency toward underutilization. Several commentators have suggested, for example, that jurors be instructed on Bayesian updating in order to help them appreciate the strength of forensic evidence and overcome their tendency to give it too little weight.[98]

I am skeptical of this conclusion. While the research shows patterns of judgment that deviate from Bayesian norms, "underutilization" of forensic evidence is not the only possible explanation. Many of the studies in this literature have methodological limitations that may have created a false appearance of "underutilization." Some of the studies used incomplete Bayesian models that failed to account for all possible sources of uncertainty. Participants in these studies may have updated their beliefs less than the Bayesian model specified because they were legitimately skeptical about the evidence for reasons not taken into account by the Bayesian models, rather than because their judgments were illogical.[99] A false appearance of

update beliefs about a particular hypothesis (or about a pair of alternative hypotheses) in light of new evidence. ROBERTSON, VIGNAUX & BERGER, *supra* note 5. Researchers can use Bayesian analysis to assess whether forensic science evidence causes people to change their beliefs about the source of an item in a manner that is logically consistent with their beliefs about the forensic evidence. For example, several researchers have used Bayesian models to determine how DNA evidence should affect a rational actor's belief about the probability that a defendant was the source of a biological sample, given the actor's beliefs about the probability that the DNA evidence could falsely implicate the defendant through such mechanisms as a coincidental match, examiner error, or a frame up. The models are, in effect, statements about the logical consistency of various beliefs a person might hold.

[97] Thompson & Schumann, *supra* note 66; Faigman & Baglioni, *supra* note 86; Goodman, *supra* note 67; Brian C. Smith, Steven D. Penrod, Amy L. Otto & Roger C. Park, *Jurors' Use of Probabilistic Evidence*, 20 LAW & HUM. BEHAV. 49 (1996); Jason Schklar & Shari Seidman Diamond, *Juror Reactions to DNA Evidence: Errors and Expectancies*, 23 LAW & HUM. BEHAV. 159 (1999); Nance & Morris (2002), *supra* note 72; Nance & Morris (2005), *supra* note 72; Martire et al. *supra* note 26, *but see* Thompson, Kaasa & Peterson, *supra* note 27 (finding both underutilization and overutilization of forensic DNA evidence).

[98] This suggestion was first put forth in 1970. *See* Michael O. Finkelstein and William B. Fairley, *A Bayesian Approach to Identification Evidence*, 83 HARV. L. REV. 489 (1970). It drew a famous response from Laurence Tribe, who objected to what he regarded as an unwarranted intrusion of mathematics into the trial process. Laurence H. Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329 (1971). But suggestion that jurors be instructed on Bayesian updating, as a means of mitigating "underutilization," have continued. *See, e.g.*, Faigman & Baglioni, *supra* note 85; Nance & Morris (2005), *supra* note 72.

[99] For details on this issue, see Thompson, Kaasa & Peterson, *supra* note 27, at 361;

underutilization might also have arisen from the use of probability elicitation methods that artificially restricted the range of possible responses, and hence the degree to which people could express changes in their beliefs after receiving forensic evidence. Ceiling effects—the well known tendency of people to avoid using the extreme ends of a response scale—may have created a false appearance of "underutilization" in some studies by making it difficult for some participants to indicate how much their beliefs had changed.[100]

Beyond the methodological issues, there are theoretical reasons to question the conclusion that people underutilize forensic evidence. Psychologists have long recognized that ordinary people do not think like Bayesians.[101] According to contemporary psychological theory, people employ a variety of heuristic strategies for evaluating evidence and updating beliefs—strategies that generally work well but can produce counter-normative judgments in specific situations.[102] Jonathan Koehler has extended these theories to encompass forensic evidence, showing how variations in reporting format can produce predictable variation in peoples' evaluations of the strength of forensic comparisons.[103] While these theories predict that people's judgments will sometimes deviate from Bayesian norms, it is difficult to see why the underlying cognitive processes posited by the theories would cause a systematic tendency to underutilize forensic evidence, rather than errors that could lead either to underutilization or overutilization depending on the specific circumstances.

My colleagues and I have recently re-visited the question of people's consistency with Bayesian norms using more complete Bayesian models and improved probability elicitation methods designed to avoid ceiling effects. Our studies present a somewhat different picture than previous studies, at least for DNA evidence. Thompson, Kaasa and Peterson found that people evaluating DNA evidence (participants included actual jurors) generally responded to it in a manner consistent with Bayesian norms; under some circumstances DNA evidence caused people's judgments to shift more than Bayesian norms would dictate—suggesting that they may have given too

Thompson & Newman, *supra* note 26, at 334.

100 The potential for ceiling effects in this research is discussed in Smith et al., *supra* note 97; Thompson, Kaasa & Peterson, *supra* note 27; Thompson & Newman, *supra* note 26.

101 *See* Spellman, *supra* note 26; Daniel Kahneman & Amos Tversky, *Subjective Probability: A Judgment of Representativeness*, 3 COGNITIVE PSYCHOL. 430, 450 (1972) (". . . man is apparently not a conservative Bayesian: he is not Bayesian at all").

102 S*ee generally* Amos Tversky & Daniel Kahneman, *Judgment Under Uncertainty: Heuristics and Biases*, 185 SCI. 1124 (1974), http://psiexp.ss.uci.edu/research/teaching/Tversky_Kahneman_1974.pdf; JUDGMENT UNDER UNCERTAINTIES: HEURISTICS AND BIASES 3 (Daniel Kahneman, Paul Slovic & Amos Tversky eds., 1982).

103 *See* Koehler (2001), *supra* note 69; Koehler & Maachi, *supra* note 70.

much weight to the DNA evidence.[104]

Thompson and Newman also found that people's responses to DNA evidence were consistent with Bayesian norms, but found people violated Bayesian norms when they evaluated shoeprint evidence.[105] People gave less weight to shoeprint evidence than the Bayesian model said they should even after taking into account their beliefs about the probability of a coincidental match, lab error and frame-up. People apparently perceived shoeprint evidence to be weak (relative to DNA evidence) in ways that were not captured by the Bayesian model. Thompson and Newman suggest that people's reactions to forensic evidence are complex and are influenced by a number of variables. Impressions about the value of evidence and the credibility of experts that have been shaped by popular culture may be as important as the words experts speak in determining the weight people give to forensic evidence.

While there are reasons to be skeptical of claims that people inevitably, or systematically, underutilize forensic science evidence, there clearly is much we do not know about factors affecting the weight people give this evidence. Additional research is needed on how people weigh various factors that affect the probative value of forensic evidence, such as the probability of a coincidental match, a false incrimination due to laboratory error, or a frame up. It would also be useful to explore whether people's beliefs about error rates, match probabilities, and related matters are reasonable (i.e., consistent with the best evidence), and how those beliefs are affected by the way forensic examiners present their conclusions. If misconceptions are distorting their views of the value of forensic evidence, that issue will need to be addressed as well.

ii. Fallacious Reasoning

Another kind of logical incoherence can be caused by people's tendency to evaluate forensic science using illogical strategies that arise from a fundamental misunderstanding of probabilistic evidence. One error, often called the prosecutor's fallacy, arises from mistakenly assuming that a random match probability equates to the probability the defendant is innocent.[106] As Thompson and Newman have explained:

---

104 Thompson, Kaasa & Peterson, *supra* note 27. The evidence for "overutilization" of DNA evidence occurred when random match probability was extremely low (1 in 1 trillion) but the probability of a false match due to cross-contamination in the laboratory was much higher (1 in 100). Participants may have had difficulty aggregating data on the probability of a coincidental match with data on the probability of a false match due to lab error, causing them to discount insufficiently for the latter source of error.

105 Thompson & Newman, *supra* note 26.

106 *See* Thompson & Schumann, *supra* note 66; Nance and Morris (2002), *supra* note 72; Nance and Morris (2005) *supra* note 72; David H. Kaye, Valerie P. Hans, B. Michael Dann,

> . . . people sometimes mistakenly assume that they can infer the probability that matching items have (or do not have) a common source from the random match probability (RMP). If an expert reports that a defendant matches a DNA sample and that the probability a random person would match is 1 in 1 million, for example, people sometimes assume that this necessarily means there is one chance in a million that the DNA sample came from someone other than the defendant—a mistake of logic that has been called the "source probability error" and the "fallacy of the transposed conditional." . . . In cases where the defendant's identity as the perpetrator is the sole issue . . . people sometimes mistakenly equate the RMP with the probability the defendant is innocent—an error known as "the prosecutor's fallacy" . . . .[107]

Thompson and Newman found that people are susceptible to the same error when evaluating LRs:

> The same erroneous logic (arising from transposition of conditional probabilities) might also lead to fallacious interpretation of likelihood ratios. If an expert says the DNA evidence is one million times more likely if the defendant, rather than a random person, is the source of a sample, for example, then people might mistakenly assume that this means it is one million times more likely that the defendant, rather than a random person, is the source of the sample.[108]

In either case, the fallacy may cause a serious misinterpretation of the evidence. ". . . The danger of this fallacy is that it leads people to think they can determine the probability the defendant is (or is not) the source from the forensic evidence alone, without considering the other evidence."[109]

The key issues for researchers are how often people fall victim to fallacious reasoning, how much these logical errors influence their judgments, and whether people's susceptibility to error is influenced by the presentation format used by the forensic scientist.

Researchers have provided a wide range of estimates of how frequently people fall victim to the prosecutor's fallacy. Differences across studies are

Erin Farley & Stephanie Albertson, *Statistics in the Jury Box: How Jurors Respond to Mitochondrial DNA Match Probabilities*, 4 J. EMPIRICAL LEGAL STUD. 797 (2007); Thompson, Kaasa & Peterson, *supra* note 27.

107 *See supra* note 26 at 335.

108 *Id.*

109 *Id.*

due in part to ways in which researchers judged whether people were making fallacious judgments. A common approach is to assume participants are committing the prosecutor's fallacy when they judge the probability of the defendant's guilt to be the exact complement of the random match probability. Thompson and Schumann[110] found that 13.2% of participants judged the probability of the defendant's guilt to be exactly 98% (which was higher than Bayesian norms) in a case in which the defendant was linked to the crime by serological evidence and the random match probability was 2%. These participants defended their conclusion with fallacious arguments, suggesting, for example, that because there was only a 2% chance the defendant would have "matched" if he was innocent, the evidence of the match implies a 98% chance that he is guilty. However, the percentage of participants who made judgments consistent with the prosecutor's fallacy dropped to only 3% in a second study[111] in which participants read lawyers' arguments about the strength of the forensic evidence. In subsequent studies, the percentage making judgments consistent with the prosecutor's fallacy has varied but it was generally low.[112] The frequency of judgments consistent with the prosecutor's fallacy undoubtedly depends, in part, on how the random match probability is presented and explained to the jury and on how easy it is for participants to express such judgments.[113]

The studies cited so far judged whether people committed the prosecutor's fallacy by seeing whether their estimate of the probability of guilt was the exact complement of the random match probability. This is a rather exacting requirement and may underestimate the number of people who are confused by fallacious reasoning. Thompson and Newman[114] took a different approach. After participants had reviewed the trial evidence, they presented the participants with a series of statements regarding the implications of the forensic science evidence and asked participants which

110 *Supra* note 66 (Study 1).

111 *Id.* (Study 2).

112 2% in Goodman, *supra* note 67; 4% in Smith et al., *supra* note 97; 6% in Nance and Morris (2002), *supra* note 72; 0.3% in Nance & Morris (2005), *supra* note 72; but an estimated 16% in Kaye et al., *supra* note 106.

113 One reason for the low percentage (0.3%) in Nance & Morris (2005), *supra* note 72, for example, may have been that the random match probability was 1 in 40,000 (or 0.000025), as a consequence of which the researchers coded participants' judgments as consistent with the prosecutor's fallacy only if they said the probability of guilt was exactly 0.999975 (99.9975%). As the researchers acknowledged, some participants who fell victim to fallacious reasoning may have been missed in this assessment because they miscalculated the percentage, or simply rounded to 99%. In a case involving mtDNA evidence where the RMP was less than 1%, Kaye et al., *supra* note 106, found that 48% of mock jurors agreed with the proposition that "the mtDNA evidence in this case shows that there is about a 1% chance that someone else besides the defendant committed the crime." But the researchers argued that jurors could have arrived at this conclusion without committing the prosecutor's fallacy.

114 *Supra* note 26.

statements were a correct interpretation of what the expert had said. Some of the statements correctly indicated that the expert had testified about the probability of the observed results given the hypotheses.[115] Other statements incorrectly transposed the conditional and indicated that the expert had testified about the probability that the same-source hypothesis is true given the observed results. Nearly two-thirds of participants who heard testimony about RMPs or LRs thought the "correct interpretation" was a statement that transposed the conditional—suggesting a high level of misunderstanding. Furthermore, participants who chose the fallacious statement as a "correct interpretation" were more likely to have voted in favor of convicting the defendant—suggesting that misinterpretations of the expert's statistical statements can have important consequences.

In a subsequent study, Thompson, Grady and Newman[116] asked participants to evaluate simulated testimony by a voice comparison expert, who presented and explained likelihood ratios. About 80% of these participants thought the "correct interpretation" of the expert's testimony was a statement that transposed the conditional probabilities.[117] Those who agreed with the fallacious interpretation also gave more weight to the voice comparison evidence when assessing the defendant's guilt.

When evaluating forensic science evidence, people sometimes fall victim to another logical error called the "defense attorney's fallacy."[118] Victims of this fallacy mistakenly assume that a forensic match has little or no probative value for incriminating the defendant if someone other than the defendant could also have matched. This error may cause people to underutilize forensic evidence, or ignore it entirely, when evaluating a case.[119]

Finally, researchers regularly observe a third kind of logical error—some people respond to forensic evidence by changing their beliefs *in the wrong direction*. Generally this happens when people respond to forensic evidence that is reported to be weakly incriminating by revising downward their belief in the probability of the defendant's guilt, hence it has been called

---

115 Experts who presented RMPs were testifying about the probability of the observed results given the hypothesis that the items being compared had a different source; experts who presented LRs were testifying about the relative probability of the observed results under the two alternative hypotheses—i.e., same-source or different-source.

116 *Supra* note 91.

117 The expert had testified that the evidence observed in the forensic comparison of voices was either 30 or 3000 times more likely if the defendant was the speaker on a contested recording than if the speaker was another individual from the relevant population. The statement that 80% of participants viewed as the "correct interpretation" said the expert had testified that it was 30 or 3000 times more likely that the defendant was the speaker given the evidence observed in the forensic comparison of voices.

118 Thompson & Schumann, *supra* note 66; Thompson, Kaasa & Peterson, *supra* note 27.

119 Thompson & Newman, *supra* note 26.

a "weak evidence effect."[120] When Martire and her colleagues asked people to evaluate shoeprint evidence and fingerprint evidence that incriminated a defendant, they found that a majority of participants revised their belief in the defendant's guilt downward when the expert explained the strength of the evidence by saying it provided "weak or limited support" for the hypothesis that the defendant was the source of an incriminating item at the crime scene.[121] One possible explanation for this finding is that participants misunderstood the expert's statement about the strength of the evidence for supporting the hypotheses as a statement about the probability that the hypothesis is true. In other words, this phenomenon may be another manifestation of confusion over conditional probabilities.

V. SUMMARY AND CONCLUSIONS

I began this Article by suggesting two criteria for determining the way forensic scientists should report source conclusions: (1) whether the reported conclusions can be justified logically and empirically; and (2) whether the reported conclusions will be understood and used appropriately. In the discussion that followed, I suggested that some of the most commonly used forensic reporting formats fail (or might fail) the first criterion.

There is legitimate skepticism in the broader scientific community about claims that examiners in pattern-matching disciplines can link an item to a single possible source by discerning that the item has unique features. Consequently, forensic examiners are on shaky ground when they claim to have "identified" or "individualized" the source of a trace or impression found at a crime scene. Concern about this problem has prompted discussion among forensic scientists about the possibility of reporting source conclusions in different ways. One of my goals in writing this Article is to provide a helpful perspective to forensic scientists regarding possible options.

There is also reason to be skeptical about reporting source probabilities (e.g., claiming a "99% chance" or a "high probability" that items have a common source) because forensic examiners must consider or make assumptions about matters beyond their scientific expertise to draw conclusions about source probabilities. Source probabilities are hybrid conclusions, based partly on the examiner's analysis of the physical characteristics of the items being compared, and partly on the examiner's assumptions or conclusions about the strength of other evidence that bears on whether the items have a common source.[122] Whether hybrid conclusions

120 Martire et al. (2013), *supra* note 26; Martire et al. (2014), *supra* note 88.

121 Martire et al. (2013), *supra* note 26.

122 The same problems arise when examiners "identify" items as having the same source based, in part, on consideration of the prior odds (a process that I called "identification without

of this kind should even be admissible is a question evidence scholars should consider, given the danger that testimony of this type will invade the province of the trier-of-fact and the potential for confusion and double-counting that arises when the expert's conclusion rests (to an extent that may be unclear) on evidence the trier-of-fact will also consider, and may view differently.[123] If forensic scientists continue to present source probabilities, they should develop standards for disclosure of the examiner's underlying assumptions or evaluations about the prior odds, so that recipients will have a clearer understanding of the basis for these conclusions.

There are many possible alternatives to reporting "identification" or source probabilities, some of which require presenting numbers (LRs and RMPs) and some of which involve qualitative statements (about strength-of-support or likelihood of correspondence). Experts can reach these conclusions based solely on evaluation of the physical features of the items being compared, without making assumptions about the prior odds of a common source, so these approaches avoid the foundational problems of source probabilities. Of course, questions can and should be raised about the validity of these conclusions. When these conclusions rest on databases and statistical models, the appropriateness of the databases and the validity of the models must be considered. When these conclusions rest on subjective assessment of the relevant likelihoods, serious questions should be asked about how accurately examiners can make the necessary subjective assessments. We need thoughtful analysis of such questions in order to decide what forensic scientists should be allowed to say about source conclusions in reports and testimony.

The PCAST report suggested that forensic scientists validate subjective methods by conducting black-box studies to determine their accuracy.[124] Examiners could then explain the probative value of their conclusions by presenting data on true and false positive rates. For example, a latent print examiner could report "identification" of a particular suspect as the source of a latent print, but would then say, for example, that the rate of false identifications could be as high as 1 in 306 based on error rate data from black-box studies.[125] The introduction of error rate data changes the nature

uniqueness).

[123] I am aware of no cases in which litigants have raised foundational challenges to such testimony, but that may reflect a failure of lawyers and judges to appreciate that such conclusions require examiners to make assumptions or take positions on matters beyond their scientific expertise. Acceptance of such testimony may well diminish once lawyers and judges come to appreciate the problematic nature of the underlying logic.

[124] According to the PCAST report, latent print examination is the only forensic discipline for which appropriate black-box studies currently exist.

[125] There will inevitably be controversy about the exact numbers examiners use to report error rates. A false identification rate of 1 in 306 is PCAST's calculation of the upper 95%

of examiner's report from a straightforward, unqualified statement that the same-source hypothesis is true, to a statement that the examiner has made a determination that is diagnostic, in a probabilistic manner, of whether the same-source hypothesis is true. Reports in this form would better meet the first criterion of being justified logically and empirically.

The second criterion for determining how forensic scientists should present source conclusions is whether the reported conclusions will be understood and used appropriately. As forensic scientists contemplate the possibility of shifting from traditional reporting formats (e.g., "identification," source probabilities) to alternatives (e.g., LRs, strength-of-support statements), questions will inevitably arise about lay reaction to alternative reporting formats. The research reviewed here should allay concerns that people will dismiss or give little weight to statements about LRs, strength of support, match frequency or likelihood of correspondence. The alternative formats can be influential. Whether people will give such statements the correct weight, however, is a much harder question about which there remains much uncertainty.

Over the past few years, it has become increasingly clear that fallacious interpretation of forensic science testimony can have important consequences. Hence, it is important to look for ways to avoid these errors. For example, it is worth considering Professor Friedman's suggestion that judges provide explicit, detailed instructions on this issue.[126] In some cases there may also be a need for expert testimony to explain to the jury what would and would not be a proper inference from forensic science evidence. While expert testimony on this topic might be resisted as an effort to tell jurors how to think, the social science evidence suggests that on this issue they may need the help.

If efforts to educate jurors (and lawyers) about these errors prove fruitless, as Professor Spellman fears,[127] we will then need to consider whether it makes sense to avoid presenting forensic science findings in ways that are conducive to these errors. That might mean avoiding quantitative likelihood ratios and random match probabilities, although careful thought should be given to whether the disease is sufficiently serious to warrant such

confidence bound of the false identification rate in the largest black-box study of latent print examiners. *See supra* note 81. For additional commentary on error rate estimation in forensic science, see PCAST REPORT, *supra* note 15; Nat'l Comm. on Forensic Sci., *Views of the Commission: Facilitating Research on Laboratory Performance*, U.S. DEP'T OF JUSTICE & NIST (Sept. 13, 2016), https://www.justice.gov/archives/ncfs/page/file/909311/download; Jonathan J. Koehler, *Forensics or Fauxrensics? Ascertaining Accuracy in the Forensic Sciences*, 49 ARIZ. ST. L.J. 1369 (2018).

126 Richard D. Friedman, *Controlling the Jury-Teaching Function*, 48 SETON HALL L. REV. 815 (2018).

127 Spellman, *supra* note 26.

a drastic cure. That will depend, of course, on whether there are alternative presentation methods that work better—which we can only assess if we have yet more research on lay reactions to forensic evidence. For researchers interested in these issues, there is much more work to be done.

In an ideal world, forensic examiners would always present their findings in a manner that causes people to respond to the evidence in a manner commensurate with its probative value. If those receiving the evidence gave it too much, or too little weight, the reporting format could be tweaked to correct the problem. Forensic scientists would report their findings in the format that assured the best calibration between the probative value of the evidence and lay reactions to the evidence.

We are a long way from that ideal world, but I hope this Article provided some glimpses of what that world might look like and how we might move toward it. First, we need much better information about the probative value of the forensic identification evidence produced in each forensic science discipline. It would be useful to develop and apply a common metric for measuring the probative value of forensic evidence, such as a likelihood ratio. The probative value of a particular conclusion could be evaluated by assessing (through research studies) how frequently examiners reach that conclusion when comparing items known to be from the same source, relative to how frequently they reach that conclusion when comparing items known to be from different sources. The ratio of the two frequencies (which is essentially a LR) describes the strength of the evidence for proving the items have a common source. The PCAST report[128] and the AAAS report on latent print identification[129] each include extensive discussion of how such studies might be designed and carried out.

Second, we need more and better research on whether people respond to forensic evidence in a manner commensurate with its value, and on how their responses vary for different presentation formats. By combining assessment of the probative value of the evidence with assessment of people's reactions to it, we can help assure that responses are properly calibrated so that people give the evidence the weight it deserves.

Consider, for example, how we might evaluate the suitability of the reporting statement for latent print evidence proposed by the Defense Forensic Science Center of the United States Army.[130] We first must understand the probative value of this evidence. That requires data on the relative probability that an Army latent print examiner will issue such a report after comparing prints made by the same finger, and after comparing

128 *Supra* note 15.
129 *Supra* note 15.
130 *See supra* note 42 and accompanying text.

prints made by different fingers. The ratio of these two probabilities is a likelihood ratio (LR) that can be used to describe the probative value of the conclusion for proving the prints come from the same finger. We need to know roughly the magnitude of this LR—is it 100, 1000, 10,000 or more? We cannot determine whether people are responding appropriately to the Army's latent print evidence, after hearing the proposed reporting statement, until we know how much weight this evidence deserves.

The second step is to assess the weight that people give to the evidence, using the kind of research discussed in this Article. If we determine that a LR of 1000 describes the probative value of this evidence, then we can assess whether people think the reporting statement is as strong as reports about other evidence with a LR of 1000, and whether they update their beliefs after hearing this evidence in a manner logically commensurate with evidence of that value.

We are not yet able to make the kind of evaluation just described, but we have already taken some substantial steps toward that imagined future. I hope this Article will help guide us closer to that goal.