VOLUME 130 MARCH 2017 NUMBER 5

HARVARD LAW REVIEW FORUM

FORENSIC COMMENTARY SERIES

HYPOTHESIS TESTING IN LAW AND FORENSIC SCIENCE: A MEMORANDUM

David H. Kaye[*]

This *Forum* Commentary series presents views on a letter from a group of lawyers and judges advising the Organization of Scientific Area Committees for Forensic Science (OSAC).[1] In response to calls for improving the practices of forensic science,[2] the National Institute of Standards and Technology (NIST) created the Scientific Area Committees in 2014 to promote and develop standards "that are fit-for-purpose and based on sound scientific principles."[3] The memorandum from the Legal Resource Committee (LRC)[4] responds to a question

[*] Associate Dean for Research and Distinguished Professor of Law, Penn State Law. This Introduction benefited from discussions with José Almirall, Karen Kafadar, and members of the Legal Resource Committee (LRC) of the Organization of Scientific Area Committees for Forensic Science (OSAC). The views expressed here are the author's. They should not be attributed to NIST, OSAC, the LRC, or any other individual or organization.

[1] *Se*e Memorandum from the Legal Res. Comm. to the Org. of Sci. Area Comms. for Forensic Sci., Nat'l Inst. of Standards & Tech., Question on Hypothesis Testing in ASTM 2926-13 and the Legal Principle that False Convictions Are Worse than False Acquittals 6 (rev ed. Oct. 7, 2016), *reprinted in* 130 HARV. L. REV. F. 137 (2017) [hereinafter LRC Memo].

[2] *See, e.g.*, COMM. ON IDENTIFYING THE NEEDS OF THE FORENSIC SCI. CMTY., NAT'L RESEARCH COUNCIL, STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES: A PATH FORWARD 53 (2009).

[3] *About OSAC,* OSAC NEWSLETTER (Nat'l Inst. of Standards & Tech., Gaithersburg, Md.), Oct. 2015, at 3, 3, https://www.nist.gov/sites/default/files/documents/forensics/OSACNewsLetterOctober2015.pdf [https://perma.cc/29QA-XRJC].

[4] When the letter was written, the LRC was composed of ten individuals from prosecutor and public defender offices, the judiciary, the Innocence Project, and law school faculty. The LRC comments on proposed standards and advises OSAC on legal issues. The scientific committees need not follow the LRC's recommendations or give effect to its opinions. OSAC-approved standards go to private standards-development organizations for possible adoption (perhaps with modifications). If OSAC approves of standards adopted by these external groups, NIST incorporates them into a registry of approved standards. For more information about the LRC, see *OSAC Roles and Responsibilities*, NAT'L INST. STANDARDS & TECH. (Jan. 26, 2016), https://www.nist.gov/forensics/osac-roles-and-responsibilities [https://perma.cc/3V99-CJ5L]. How much influence these standards will have on forensic laboratories, courts, or legislatures remains to be seen.

from a scientist on OSAC's governing board about whether the criminal law's concern with avoiding false convictions at the expense of false acquittals should affect the choice of a "significance level" for deciding whether pieces of glass match in their chemical composition (and hence might have a common origin). Must a criminalist favor the hypothesis that similarities are coincidental over the hypothesis that the fragments have a common origin? The underlying issue applies to many forms of identification evidence, including fingerprints, fibers, paint chips, bullets, and biological fluids. Indeed, arguments over the choice of a significance level arise for statistical evidence of all sorts, from econometrics to epidemiology.[5]

This Introduction is a preamble to the memorandum. Part I describes the technical standard that prompted the letter. Part II sketches the statistical ideas in the letter by using glass comparisons to illustrate the three main statistical approaches to reasoning about the implications of evidence. This Introduction is followed by the letter itself and two commentaries.

I. AN EXAMPLE OF STATISTICAL HYPOTHESIS TESTING: ASTM E2926-13

The standard that prompted the memorandum is ASTM E2926-13, promulgated in 2013 by ASTM International, a private standards-development organization.[6] It bears the forbidding title, "Standard Test Method for Forensic Comparison of Glass Using Micro X-ray Fluorescence (μ-XRF) Spectrometry."[7] Bombarding a material such as glass with high-energy X-rays produces a spectrum of radiation from the material. Each element contributes characteristic peaks,[8] permitting chemists to infer the elements that are present and their relative

[5] *See, e.g.*, DAVID H. KAYE ET AL., THE NEW WIGMORE: A TREATISE ON EVIDENCE: EXPERT EVIDENCE § 12.8.3, at 563–64 (2d ed. 2011).

[6] Originally known as the American Society for Testing Materials, ASTM International issues voluntary standards for everything "from the toy in a child's hand to the aircraft overhead." *About ASTM International*, ASTM INT'L, https://www.astm.org/ABOUT/overview.html [https://perma.cc/BNS4-RNM7].

[7] ASTM INT'L, ASTM E2926-13 STANDARD TEST METHOD FOR FORENSIC COMPARISON OF GLASS USING MICRO X-RAY FLUORESCENCE (μ-XRF) SPECTOMETRY (2013) [hereinafter E2926-13].

[8] Roughly speaking, electrons orbiting the nucleus of an atom do so at discrete distances that depend on the composition of the nucleus. The electrons that are farther away are at higher energy levels. A high-energy X-ray can knock an inner electron from an atom, creating what is effectively a hole in the shell of electrons. An outer electron will drop into the hole, emitting a photon whose energy is the difference between the higher and lower energy levels of the electrons. Measuring the energies (the spectrum) of the emerging photons (the fluorescence) reveals peaks in the spectrum that are characteristic of specific kinds of atoms (elements). *See* EUGENE P. BERTIN, PRINCIPLES AND PRACTICE OF X-RAY SPECTROMETRIC ANALYSIS 21 (2d ed. 1975).

concentrations. For present purposes, I shall refer to this information as an elemental profile.

The ASTM standard discusses how to ascertain and use elemental profiles for forensic inference. Suppose a burglar broke a window to enter a building and glass fragments are found on a defendant's coat. The analyst has "known" or "reference" specimens — the pieces from the broken window — and "unknown" or "questioned" specimens — the fragments from the coat. The question of legal interest is whether both sets of fragments come from the same window. The profiles themselves do not answer this question. One must first gauge how similar the profiles are and then evaluate what this degree of similarity reveals about their origin.

One method in the standard is to take replicate measurements, compute the variance of these measurements, construct a 99.7% confidence interval of "±3s" for the true value of each elemental ratio in the known window,[9] and then declare that a profile from a questioned coat fragment is "distinguishable" if any elemental ratio in the questioned profile falls outside the corresponding interval. If that happens, "it may be concluded that the specimens are not from the same source."[10] On the other hand, "[i]f the specimens are indistinguishable in all of these observed and measured properties, the possibility that they originated from the same source of glass cannot be eliminated."[11]

In other words, the criminalist performs a hypothesis test in which differences thought to be extreme prompt the categorical conclusion that the questioned and known specimens do not have a common origin; lesser differences generate the conclusion that they have the same true elemental profile (and hence might be from the same source). The same-profile conclusion relates to the statistical hypothesis — the true values of the relative concentrations are the same in the two sets of specimens — while the same-source pertains to the legal hypothesis — that the specimens have a common origin. The legal

[9] Presumably, "s" is the standard error as estimated from the repeated measurements. If this estimate is exactly equal to the true standard error, then the interval of ±3 standard errors achieves 99.7% "confidence" for "a normally distributed population." E2629-13 § 10.7.3.2. In general, the "confidence" for an interval is not the probability that the true value lies within the interval. It is a statement about how frequently the intervals will cover the unknown, true value when the statistical model of measurement error is correct. *See, e.g.*, KAYE ET AL., *supra* note 5, § 12.6.4, at 546–47. The standard also approves of a simplistic "Elemental Ratio Range Overlap" method. E2629-13, *supra* note 7, § 10.7.3.1.

[10] E2629-13, *supra* note 7, § 10.7.3.1.

[11] *Id.* introductory cmt. On its face, most statisticians would find the details of this method puzzling. However, concerns over such matters as the choice of the test statistic are outside the scope of the memorandum. For an overview arguing that the ASTM's decision rule does not assure 99.7% confidence, see David H. Kaye, *Broken Glass, Mangled Statistics*, FORENSIC SCI., STAT. & L. (Feb. 3, 2016, 3:37 PM), http://for-sci-law.blogspot.com/2016/02/broken-glass-mangled-statistics.html [https://perma.cc/4KC9-C3SC].

hypothesis depends on the elemental profiles of all pieces of glass in the world that could have ended up on the defendant's coat. Only if an elemental profile is unique does the statistical hypothesis that two true profiles are the same deductively imply that the legal hypothesis of a common source is true.[12]

Although many forensic statisticians believe that this "match/no-match" framework discards information and is ill adapted to the legal setting,[13] when it is beyond dispute that the questioned specimens could not realistically have come from the same window, it is reasonable to report that the two sets of fragments must have come from different sources. But what of specimens that are not so radically different? When the hypothesis of a common source is not summarily dismissed, how strongly does the observed degree of similarity point to identity? How well does it warrant the conclusion that the trace and the known specimens have the same elemental profile? That if they do have the same profile, they have a common origin? How should criminalists present their findings to enable factfinders to draw well-informed inferences about these hypotheses? On these legally crucial matters, ASTM E2926-13 is silent.

II. STATISTICAL ERRORS AND LEGAL VALUES

In the ongoing review of ASTM E2926-13 and related standards, some OSAC members have noted that whereas the law is more concerned with avoiding false inclusions than false exclusions, the statistical procedures lean heavily in the opposite direction. An OSAC member asked the LRC for an opinion on whether "we [must] set up our statistical tests to favor making one error (allowing a guilty defendant go free) over the possibility of making the error that we will wrongly convict."[14] The LRC's answer was that "the widespread aversion to false convictions does not make reporting the result of a frequentist statistical test with small Type I and Type II error probabilities inadmissible per se in court."[15] At the same time, the Committee cautioned that "the choice of equal elemental composition for a null hypothesis makes the application of the usual statistical terminology of Type I

[12] *See, e.g.*, David H. Kaye, *Reflections on Glass Standards: Statistical Tests and Legal Hypotheses*, 27 STATISTICA APPLICATA 173, 177 (2015).

[13] Geoffrey Stewart Morrison et al., Letter to the Editor, *A Comment on the PCAST Report: Skip the "Match"/"Non-Match" Stage*, 272 FORENSIC SCI. INT'L e7 (2017); *see also* Tacha Hicks et al., *A Framework for Interpreting Evidence*, *in* FORENSIC DNA EVIDENCE INTERPRETATION 37 (John S. Buckleton et al. eds., 2d ed. 2016); F. Taroni et al., *Statistical Hypothesis Testing and Common Misinterpretations: Should We Abandon* P*-value in Forensic Science Applications?*, 259 FORENSIC SCI. INT'L e32, e32–e36 (2016).

[14] LRC Memo, *supra* note 1, at 1.

[15] *Id.* at 7.

and Type II errors potentially confusing," and noted "that a report of a match without more information about the probability of a match to other glass in the relevant population would not fulfill the expert's role of impartially and adequately educating the trier of fact about what the scientific measurements establish."[16] In addition, it concluded that "the principle that false convictions are more serious than false acquittals does not . . . prevent statistical equivalence testing or the use of a likelihood ratio or Bayes factor to inform the judge or jury of the probative value of the spectroscopic results."[17]

Although directed at the glass-comparison standard, the letter is concerned with the general nature of the statistical reasoning and its role in legal factfinding. To see this, we can use a stylized and simplified example for statistical comparisons of glass. Imagine that all glass in an isolated community comes from a single manufacturing process that produces glass with some normally distributed property X. This bell-shaped curve is centered at a mean of 50, and a standard deviation (σ_x) of 10 indicates its width. The instrumentation that measures X does so imperfectly. The measurements are correct on average, but they scatter around the true value with a standard error (σ_e) of 1.[18] Someone broke a window to gain entry, and glass fragments were recovered from the defendant's coat. For simplicity, imagine further that we make a huge number of measurements for the reference specimen (the window). Sure enough, these are normally distributed, and their mean is, say, 33.[19] Because the sample of measurements is so large, we will proceed as if the true mean for the reference sample is exactly 33.[20] We measure the questioned specimen only once and find that $x = 30$.

What have we learned? Are the specimens statistically indistinguishable or are they significantly different? If they are deemed indistinguishable, what does this reveal about the same-source and different-source hypotheses? Is there a better way to investigate the implications of the measurements? These basic questions prompted

[16] *Id.*

[17] *Id.* at 2 (emphasis omitted).

[18] For a gentle introduction to the same model in the context of IQ testing, see David H. Kaye, *Deadly Statistics: Quantifying an "Unacceptable Risk" in Capital Punishment*, 16 LAW, PROBABILITY & RISK (forthcoming 2017), https://academic.oup.com/lpr/article/doi/10.1093/lpr/mgw012/2975605/Deadly-statistics-quantifying-an-unacceptable-risk [https://perma.cc/BHE5-XACY].

[19] A further simplifying assumption is that the window is homogeneous — the true value of X is the same at any point in the window. The only differences between the measured values and the true value of 33 come from instrumental error.

[20] This mimics the ASTM standard's idea of using a 99.7% confidence interval around the mean of nine or more measurements of the reference specimen (without accounting for the uncertainty in the use of the small sample standard deviation as an estimate of the standard error of measurement). *See* E2629-13, *supra* note 7, § 10.7.3.2.

the Committee to refer to "frequentist, likelihood, and Bayesian" statistical approaches.[21]

A. The Frequentist "Match/No-match" Test

The frequentist hypothesis tester calls the prosecution's same-true-value hypothesis the "null hypothesis" (H_0) and demands a "significant difference" between $x = 30$ and the reference specimen's value of 33 to conclude that the different-true-value hypothesis (H_1) is true.[22] The values of x for which the difference is deemed great enough to reject the null hypothesis form the "rejection region."[23] For example, following ASTM E2926-13, we might use the following decision rule: reject H_0 (which incriminates the defendant) and accept H_1 (which favors the defense) if and only if the measurement on the questioned specimen is more than plus-or-minus three standard errors from the reference specimen's value. In fewer words, reject H_0 if and only if x is outside of the confidence interval (33 ± 3).[24] These extreme values ($x < 30$ and $x > 36$) form our rejection region, and $30 < x < 36$ is the acceptance region.

This statistical test procedure has well-defined "operating characteristics."[25] Suppose that H_0 is true (the true values of the questioned and reference specimens are the same). Then, using the rejection region given above would result in false rejections about 0.27% of the time.[26] These are the "Type I errors" mentioned in the memorandum.[27] Now suppose that H_1 is true — the true value of X for the questioned specimen is not 33. Maybe it is 29. On that assumption, measurements in the acceptance region of 33 ± 3 will occur about 16% of the time.[28] These are false acceptances — we retain the null hypothesis and say that the specimens are indistinguishable when H_1 is true (when they have different true values of X). These are also called "Type II errors."[29]

[21] LRC Memo, *supra* note 1, at 3.

[22] *Id.* at 4.

[23] *Id.*

[24] *Id.*

[25] *Id.* at 6.

[26] X is a normal random variable with mean 33 (for the null hypothesis) and standard deviation 1. The rejection region is all $x < 30$ as well as all $x > 36$. The area under the normal curve in this region is 0.0027 (for the 99.7% confidence interval mentioned in ASTM E2926-13).

[27] LRC Memo, *supra* note 1, at 4.

[28] When the mean is 29 and the standard deviation is 1, the area under the curve between 30 and 36 is 0.1587.

[29] The Type II error changes with the difference in the true values. If the true difference is less, the ±3 test will err more often. For instance, if the true value of the questioned specimen is 32, the specimens are harder to distinguish, and the probability that the test, with the same precision of measurement, will fail to distinguish them is a whopping 97.72%.

It should be clear that different rejection regions will produce different conditional error probabilities. The smaller the rejection region, the more demanding the statistical test — in the sense that it will be harder to reject the null hypothesis that the specimens have the same true value of X. The Supreme Court once remarked in a case of discrimination against Latinos in the selection of grand juries that for a normally distributed random variable, "if the difference between the expected value and the observed number is greater than two or three standard deviations, then the [null] hypothesis . . . would be suspect to a social scientist."[30] Particle physicists are even more demanding when it comes to announcing the discovery of a new elementary particle. Their rule of thumb is that a difference of "5σ" is necessary.[31] The more the scientist wants to avoid a Type I error (and tolerate a greater risk of a Type II error), the smaller the rejection region and the higher the significance level will be. Thus, higher significance levels make it harder to reject the null hypothesis, and that hypothesis enjoys a kind of advantage.

To avoid advantaging the null hypothesis of no-difference, one can switch the hypotheses around. In testing whether a new, brand-name drug is therapeutically comparable to an existing drug that has received regulatory approval, for example, if the null hypothesis is that the two drugs are equivalent, the new drug has an advantage. It takes strong evidence to disprove the claim of equivalence (at a conventional significance level). Instead, one can frame the null hypothesis as the claim that the two drugs are substantially different in therapeutic effect.[32] Now the evidence must dislodge the hypothesis of the specified difference in therapeutic effect to infer that the new drug is indeed comparable to the approved one. Although the exact methodology for such equivalence testing is not trivial, adapting it to test for the identity of trace evidence could make it more difficult to conclude that there is statistically significant evidence that the items being compared are the same.[33]

Recognizing these variations in scientific conventions for statistical significance and in the choice of a null hypothesis, the memorandum concludes that testimony of "no significant difference," "indistinguishable" specimens, or a "match[]" under a traditional null hypothesis of

[30] Castaneda v. Partida, 430 U.S. 482, 497 n.17 (1977).

[31] David A. van Dyk, *The Role of Statistics in the Discovery of a Higgs Boson*, 1 ANN. REV. STAT. & ITS APPLICATION 41, 52–53 (2014).

[32] *See, e.g.*, Esteban Walker & Amy S. Nowacki, *Understanding Equivalence and Noninferiority Testing*, 26 J. GEN. INTERNAL MED. 192, 192 (2011).

[33] There are many potential applications of this variant to testing hypotheses of legal interest. A plaintiff alleging that a competitor's claim that its less expensive product is equivalent to the plaintiff's is false (and thus violates the Lanham Act, 15 U.S.C. § 1125 (2012)) might find it appealing.

absolutely no difference could be admissible — but only when accompanied by information on the probative value of such a match for the legally relevant source hypothesis.[34] Forensic statisticians have called the latter evaluation the "second stage" in the analysis of associative evidence.[35] In our example, it would mean reporting how often "indistinguishable" glass fragments arise in the larger population.[36] This step is required because the probability of Type I and Type II errors in deciding that specimens have the same true values must not be confused with the diagnosticity or specificity of these values. In the match/no-match paradigm, the criminalist first declares a match (with a risk of error), then uses other statistical information to quantify the distinct risk of a Type I error with respect to the legally important same-source hypothesis. (It may help to denote the same-source hypothesis as S_1 to distinguish it from the same-true-value hypothesis H_0.) In our example, the population distribution of property X is known, and the expert could testify that the questioned specimen could have come from many other sources of glass in the town — about 1 in 17 pieces of glass also would be indistinguishable under the 33 ± 3 rule.[37]

B. The Likelihood and Bayes Factor Analyses

The frequentist approach exemplified above treats the question for the expert witness as a decision problem for the witness followed by an inference problem for the factfinder (who then returns a verdict based on these and other inferences). The witness first decides whether there is a statistically significant match and, if there is, then uses a statistic such as the relative frequency or random match probability to inform the factfinder of the probative value of a match (H_0) for the legally relevant but distinct source hypothesis (S_1).

Other schools of statistical thought, however, hold that a better measure of probative value is available and that there is no fundamental reason to make an inherently arbitrary match/no-match decision.[38]

[34] LRC Memo, *supra* note 1, at 6.

[35] *See, e.g.*, COLIN AITKEN & FRANCO TARONI, STATISTICS AND THE EVALUATION OF EVIDENCE FOR FORENSIC SCIENTISTS 11 (2d ed. 2004); Kaye, *supra* note 12, at 182.

[36] LRC Memo, *supra* note 1, at 2.

[37] All questioned specimens with measured values between 30 and 36 would match. This match window encompasses 5.8% of the area under a normal curve with mean 50 and standard deviation 10. In other words, about 1 in 17 pieces of glass in the town would match.

[38] David H. Kaye, *Digging into the Foundations of Evidence Law*, 115 MICH. L. REV. (forthcoming 2017) (reviewing MICHAEL J. SAKS & BARBARA A. SPELLMAN, THE PSYCHOLOGICAL FOUNDATIONS OF EVIDENCE LAW (2016)). The match/no-match step treats a degree of similarity just shy of the match cutoff as radically different from the value just above the cutoff. It also treats all values above the cutoff as equivalent (the specimens are "distinguishable"), and all those below as equivalent ("indistinguishable"). LRC Memo, *supra* note 1, at 2; *see infra* note 42.

Instead of using a statistical test for exclusion and inclusion, the expert can present "likelihoods" for the latter hypotheses in light of the observed values of the property *X*.[39] The likelihood for a given hypothesis is the probability of the data when that hypothesis is true.[40] For the same-source hypothesis (S_1), 33 must be the true value for both specimens (as H_0 asserts). That, in turn, implies that the probability that x = 30 is 0.44%.[41] For the different-source hypothesis (S_0), the expert can argue that the fragment is essentially a random draw from glass in the town. The resulting probability for x = 30 is 0.54%. So the expert could report that the data are just about as likely to be seen regardless of whether the specimens come from the same source or instead come from different sources.[42] Another way to say this is that the data give approximately equal support to both source hypotheses.

In Bayesian inference, the ratio of the likelihoods involving two mutually exclusive and collectively exhaustive statistical hypotheses is known as the Bayes factor.[43] It expresses how much the measurements shift the odds on the same-source hypothesis from whatever value they had prior to considering those measurements. In our example, the Bayes factor is .44/.54 = 0.82. If the prior odds that the glass from the defendant's coat came from the window were, say, 2-to-1, the similarity in the property *X* would imply that the 2 should be revised downward, to the slightly smaller figure of .82 × 2 = 1.64, or about 5-to-3. The memorandum notes that there are cases in which such posterior odds or probabilities have been presented.[44]

[39] Kaye, *supra* note 38.

[40] Strictly speaking, the likelihood is the conditional probability multiplied by a constant that is the same for all the hypotheses under consideration. Julia Mortera & A. Philip Dawid, *Forensic Identification Then and Now*, 27 STATISTICA APPLICATA 145, 154 (2015).

[41] This is the height, at x = 30, of the normal curve with mean 33 and standard deviation 1. Because the normal distribution is a function of a continuous variable, its height is a probability density. Consequently, it would be more accurate to say that the probability is proportional to the height.

[42] The expert who reports the components of a likelihood ratio offers more complete information than one who declares a match and gives the operating characteristics of the statistical test for a match. Suppose that the measured value of *X* for the questioned specimen was 29.99, just enough to declare a match in a test that has a false positive error rate of only 0.27%. The probative value of the evidence has not changed — the likelihood ratio for the same-source hypothesis is still about 1, yet the jury hears that scientific testing has shown that the fragments match — they are statistically indistinguishable. With a known specimen closer in value to the population mean of 50 and a questioned specimen with a difference of exactly +3 or −3, the likelihood for the same-source hypothesis (S_1) would be the same, but the likelihood for the different-source hypothesis (S_0) would be smaller. The measurements would more strongly support that hypothesis. For example, if the true value for the reference specimen were 50 and the measured value for the questioned sample were 47, the likelihood ratio for the different-source hypothesis would be 0.50/0.0044 = 114 — clearly tending to exculpate the defendant.

[43] *See generally* Robert E. Kass & Adrian E. Raftery, *Bayes Factors*, 90 J. AM. STAT. ASS'N 773 (1995); Kaye, *supra* note 38.

[44] LRC Memo, *supra* note 1, at 3 n.7.

Historically, the frequentist philosophy of matching dominated thinking about trace evidence, but European laboratories and academic writing on forensic inference and statistics have moved to likelihood-based evaluations.[45] The LRC memorandum takes no position on the relative merit of the two schools of thought. It maintains that although U.S. law clearly recognizes that Type I errors with respect to the final decision of guilt (false convictions) are generally more serious than Type II errors (false acquittals), this asymmetry does not require the use of one approach instead of the other.[46] With respect to the frequentist hypothesis-testing framework, the memorandum suggests that, in principle, the apparent unfairness in taking the no-difference hypothesis as the one to disprove can be avoided if the expert completes the second stage of the statistical evaluation.[47] If experts are to report that there is no detectable difference between two specimens, they also must have data to indicate the extent to which this degree of similarity proves that the specimens come from the same source.

[45] *See, e.g.*, EUR. NETWORK OF FORENSIC SCI. INSTS., ENFSI GUIDELINE FOR EVALUATIVE REPORTING IN FORENSIC SCIENCE 6–18 (2015); Morrison et al., *supra* note 13; Cedric Neumann et al., *Presenting Quantitative and Qualitative Information on Forensic Science Evidence in the Courtroom*, 29 CHANCE 37, 37 (2016) (referring to "the abundant literature published over the past 30 years advocating . . . [that] forensic scientists should report the relative support that forensic evidence provides to each side of the legal argument using a Bayes factor (also sometimes referred to as a likelihood ratio . . .)"); S.J. Walsh, *Significance*, *in* FORENSIC BIOLOGY 141, 142 (Max M. Houck ed., 2015) ("The use of Bayes' theorem in the context of forensic evidence interpretation is widely accepted in the international forensic community.").

[46] LRC Memo, *supra* note 1, at 5–7.

[47] In this stage, the expert estimates the relative frequency of a match (or a related quantity) to address the legally crucial hypothesis that the specimens have a common origin. *Id.*