Fordham Law Review

Volume 86 | Issue 4 Article 6

2018

The Critical Role of Statistics in Demostrating the Reliability of Expert Evidence

Karen Kafadar
University of Virginia

The Critical Role of Statistics in Demostrating the Reliability of Expert Evidence

Erratum

Law; Criminal Law; Evidence; Courts; Judges

THE CRITICAL ROLE OF STATISTICS IN DEMONSTRATING THE RELIABILITY OF EXPERT EVIDENCE

*Karen Kafadar**

Federal Rule of Evidence 702, which covers testimony by expert witnesses, allows a witness to testify "in the form of an opinion or otherwise" if "the testimony is based on sufficient facts or data" and "is the product of reliable principles and methods" that have been "reliably applied." The determination of "sufficient" (facts or data) and whether the "reliable principles and methods" relate to the scientific question at hand involve more discrimination than the current Rule 702 may suggest. Using examples from latent fingerprint matching and trace evidence (bullet lead and glass), I offer some criteria that scientists often consider in assessing the "trustworthiness" of evidence to enable courts to better distinguish between "trustworthy" and "questionable" evidence. The codification of such criteria may ultimately strengthen the current Rule 702 so courts can better distinguish between demonstrably scientific sufficiency and "opinion" based on inadequate (or inappurtenant) methods.

* Commonwealth Professor and Chair, Department of Statistics, University of Virginia. The author thanks the Honorable Harry T. Edwards (U.S. Court of Appeals for the District of Columbia Circuit), Dr. Hari Iyer (NIST), Professor Brandon Garrett (University of Virginia), and Chief Brendan Max (Chicago Public Defender's Office) for their valuable comments on earlier versions of this Article. While this Article benefitted from their leadership, the statements made herein (apart from attributable citations) remain the sole responsibility of the author and do not necessarily represent the views of individual reviewers. The author is grateful to the Issac Newton Institute for Mathematical Sciences for its hospitality during its program *Probability and Statistics in Forensic Sciences* which was supported by EPSRC Grant Number EP/K032238/1. This Article was prepared in part with support from a cooperative agreement with the Special Programs Office, National Institute of Standards and Technology (NIST), via a subcontract to Iowa State University. This Article was prepared for the *Symposium on Forensic Expert Testimony,* Daubert*, and Rule 702*, held on October 27, 2017, at Boston College School of Law. The Symposium took place under the sponsorship of the Judicial Conference Advisory Committee on Evidence Rules. For an overview of the Symposium, see Daniel J. Capra, *Foreword: Symposium on Forensic Expert Testimony,* Daubert*, and Rule 702*, 86 FORDHAM L. REV. 1459 (2018).

INTRODUCTION

Federal Rule of Evidence 702 ("FRE 702") provides a list of five traits by which a witness may qualify as an "expert" and four conditions that the testimony must satisfy:

> A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:
>
> (a) the experts scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
>
> (b) the testimony is based on sufficient facts or data;
>
> (c) the testimony is the product of reliable principles and methods; and
>
> (d) the expert has reliably applied the principles and methods to the facts of the case.[1]

Although experience, education, and training can be documented, and knowledge and skill can be demonstrated via proficiency tests, none is required by Rule 702.[2]

The Advisory Committee's notes that accompany this statement provide guidance on how to apply or interpret the above four criteria. These notes include a statement about opinions: "The use of opinions is not abolished by the rule, however. It will continue to be permissible for the experts to take the further step of suggesting the inference which should be drawn from applying the specialized knowledge to the facts."[3]

The notes also provide the scope to which Rule 702 might apply: "The fields of knowledge which may be drawn upon are not limited merely to the 'scientific' and 'technical' but extend to all 'specialized' knowledge."[4] Thus, an expert may have only one of five traits (knowledge, skill, experience, training, or education), and, moreover, is allowed to draw inferences from facts, which in this case include statistical data.

---

1. FED. R. EVID. 702.
2. *Id.* (requiring "knowledge, skill, experience, training, *or* education" (emphasis added)).
3. *Id.* r. 702 advisory committee's notes on proposed rules.
4. *Id.*

The notes further refer to the case *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,[5] which "set forth a non-exclusive checklist for trial courts to use in assessing the reliability of scientific expert testimony."[6]

Rule 702 provides criteria that *allow* a person to serve as an "expert" but interestingly do not provide much guidance in limiting the *scope* of the testimony to *only* the areas of the person's expertise. This shortcoming is glaring when it comes to (1) data collection and presentation, (2) statistical methods and analysis, and (3) inferences and interpretations from data. *Merriam-Webster's Collegiate Dictionary* places these activities squarely under the discipline of statistics—"a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data."[7] But Rule 702 seems to allow *any* "expert" to draw inferences from data—which could arise from biased collections rather than from representative samples from the relevant population—even if that expert's knowledge of statistics is nonexistent. In fact, the notes accompanying Rule 702 cite *Kumho Tire Co. v. Carmichael*[8]: "[W]e conclude that the trial judge must have considerable leeway in deciding in a particular case how to go about determining whether particular expert testimony is reliable."[9] This statement endows the trial judge with the ability to recognize statistical arguments in the testimony of a forensic scientist.

Most judges will readily appreciate that statisticians cannot be allowed to testify as experts about matters of chemistry—yet fail to understand that chemists, forensic glass experts, latent print examiners, hair microscopists, and other forensic practitioners are routinely being allowed by Rule 702 "to take the further step of suggesting the inference which should be drawn from applying the specialized knowledge to the facts"[10]—in other words, to testify as a statistician. So, in addition to the failure of Rule 702 to appreciate the statistician as the appropriate expert for data collection, analysis, and inference, Rule 702 presumes that the trial judge has "considerable leeway in deciding . . . whether particular expert testimony is reliable"[11] and, hence, will recognize when the expert is testifying about statistical matters beyond the expert's expertise. It is the failure of most "gatekeepers" to distinguish statistical testimony from other scientific testimony that has led to many of the problems that forensic science has encountered, most of which could have been avoided if statisticians had been called into the problem earlier.

Surely, there have been cases where Rule 702 provided adequate guidance to ensure appropriate, useful, and proper testimony. But Rule 702 has clear shortcomings in cases where forensic testimony is presented. This Article discusses two types of forensic evidence that have been admitted under Rule

5. 509 U.S. 579 (1993).
6. FED. R. EVID. 702 advisory committee's notes to 2000 amendments.
7. *Statistics*, MERRIAM-WEBSTER'S COLLEGIATE DICTIONARY (11th ed. 2003).
8. 526 U.S. 137 (1999).
9. FED. R. EVID. 702 advisory committee's notes to 2000 amendments (quoting *Kumho*, 526 U.S. at 152).
10. *Id.* r. 702 advisory committee's notes on proposed rule.
11. *Id.* r. 702 advisory committee's notes to 2000 amendments (quoting *Kumho*, 526 U.S. at 152).

702 as having satisfied its conditions but were in fact less than reliable—or worse, incomplete and misleading. These examples provide opportunities to enhance Rule 702 with further and more specific conditions so that Rule 702 *will* be successful for its intended purpose: to ensure that reliable and useful information is conveyed to decision makers.

I. RELIABILITY, VALIDITY, AND REPRODUCIBILITY

Scientists judge research by many criteria, including how well it seems to work, whether the methods are described clearly so they can be reproduced (particularly on the data on which the methods were illustrated so other researchers can duplicate the findings), and whether the method has been demonstrated to be reliable and valid. Some of these criteria are stated explicitly in Rule 702, which requires "*reliable* principles and methods" that are "*reliably* applied."[12] The 2016 President's Council of Advisors on Science and Technology (PCAST) report states the requirements for demonstrating validity and reliability: "Scientific validity and reliability require that a method has been subjected to empirical testing, under conditions appropriate to its intended use, that provides valid estimates of how often the method reaches an incorrect conclusion."[13]

Reliability also carries with it the connotation of consistency, repeatability, and "trustworthiness." In other words, if the method were repeated on the same piece of evidence by another person and/or with other equipment, the results would be consistent (within some stated level of uncertainty).[14] A *valid* method is one that is founded on sound principles. In theoretical statistics, a *valid* hypothesis test is one that achieves its stated level of probabilities, but in common parlance, it usually refers to a method that is effective and accurate.[15] Finally, a *reproducible* method or procedure is one that is specified with enough detail that it can be repeated, presumably with similar answers.[16] Note that these concepts are not binary (e.g., reliable or not reliable); their definitions imply that some *threshold*s for "closeness" (accuracy) and "consistency" have been offered. Hence a method deemed "valid" and "reliable" for some purposes (e.g., your home scale) might be hopelessly inadequate for another purpose (e.g., National Institute of Standards and Technology's (NIST) measurements of a standard kilogram).

---

12. *Id.* r. 702 (emphasis added).

13. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., EXEC. OFFICE OF THE PRESIDENT, FORENSIC SCIENCE IN CRIMINAL COURTS: ENSURING SCIENTIFIC VALIDITY OF FEATURE-COMPARISON METHODS 143 (2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf [https://perma.cc/R76Y-7VU].

14. *See* EDWARD G. CARMINES & RICHARD A. ZELLER, RELIABILITY AND VALIDITY ASSESSMENT 11 (1979).

15. *See id.* at 12–13.

16. *See* BARRY N. TAYLOR & CHRIS E. KUYATT, NAT'L INST. OF STANDARDS & TECH., NIST TECHNICAL NOTE 1297: GUIDELINES FOR EVALUATING AND EXPRESSING THE UNCERTAINTY OF NIST MEASUREMENT RESULTS 14–15 (1994 ed.), https://www.nist.gov/sites/default/files/documents/2017/05/09/tn1297s.pdf [https://perma.cc/TYP5-V77H].

While these concepts can be defined, *demonstrating* that they hold for a given method or procedure can be more challenging. From a scientist's perspective, demonstration requires some way of quantitatively measuring a process. Because a process can involve many steps, from evidence collection at the crime scene to evidence processing and final examination, different metrics may be needed for different process steps, some of which may be more consistent and reproducible than others. A straightforward way to assess the entire process is to consider the two-by-two table of correct and incorrect conclusions:

	From Test Method	
	Claim "Same"	Claim "Different"
Same source/class	Correct (Sensitivity)	*False Negative*
Different source/class	*False Positive*	Correct (Specificity)

With multiple same-source (or same-class) pairs, the procedure should have a high probability of concluding "same source"; this probability is called *sensitivity*.[17] Likewise, with multiple different-source pairs, the procedure should have a high probability of concluding "different sources"; this probability is called *specificity*.[18] These probabilities, which provide measures of a procedure's performance in classifying evidence as being associated or not associated with a suspect (e.g., same source/class versus different source/class), can be estimated (with appropriately quantified uncertainties) via well-designed experiments. Moreover, a procedure's performance will be subject to many sources of variability, such as quality of the evidence, examiner skill and experience, or type of system or instrument being used. So the experiments to assess the method's performance in terms of reproducibility, sensitivity, and specificity are key components of characterizing the reliability and validity of a method or procedure.

In real life, of course, we do not know the truth (same or different source); all we have is the test result. For the test to be trustworthy, we want to have high confidence that the conclusions from the forensic examination (e.g., "same" or "different") are correct. Positive predictive value (PPV) is the probability that, for example, if the examiner states "same source," the evidence from the suspect and the crime scene really do come from the same source.[19] Negative predictive value (NPV) is the probability that, if the examiner states "different source," the evidence from the suspect and the

17. *See* WORKING GROUP 2, JOINT COMM. FOR GUIDES IN METROLOGY, INTERNATIONAL VOCABULARY OF METROLOGY—BASIC AND GENERAL CONCEPTS AND ASSOCIATED TERMS 40 (3rd ed. 2008), https://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2008.pdf [https://perma.cc/S57W-DQJR]; Karen Kafadar, *Statistical Issues in Assessing Forensic Evidence*, 83 INT'L STAT. REV. 111, 114 (2015).

18. *See* Kafadar, *supra* note 17, at 115.

19. *Id.* at 115.

crime scene really do come from different sources.[20] PPV and NPV are functions of sensitivity, specificity, and the size of the population from which the evidence might have come.[21]

The notes section of Rule 702 also refers to *Daubert*'s "non-exclusive checklist for trial courts to use in assessing the reliability of scientific expert testimony":

(1) "whether the expert's technique or theory can be or has been tested—that is, whether the expert's theory can be challenged in some objective sense, or whether it is instead simply a subjective, conclusory approach that cannot reasonably be assessed for reliability;"
(2) "whether the technique or theory has been subject to peer review and publication;"
(3) "the known or potential rate of error of the technique or theory when applied;"
(4) "the existence and maintenance of standards and controls; and"
(5) "whether the technique or theory has been generally accepted in the scientific community."[22]

"Objectivity" in point (1) is a goal in scientific procedures, even if the initial concept for the procedure might have arisen from intuition. For example, others may have speculated that fingerprints are unique but "Galton stressed that identification was accomplished precisely only through attention to the *minutia* of the prints—tiny islets and forks in the ridges."[23]

However, some aspects of these criteria are inappropriate. First, error rates are never "known," as stated in point (3);[24] at best they can be only estimated (with uncertainty). All data, measured or otherwise collected or recorded, are affected by many sources of variability (observation errors, recording errors, environmental influences on the measurements, etc.), and this variability translates into uncertainty in estimating error rates. Error rates should never be presented as "known"; at best, they are estimated with error, so they should be presented as intervals that have high probabilities of containing the "true" error rates (e.g., "95 percent confidence interval" for the true error rate). In fact, *all* estimates—of false-positive rates, of population means, or of specific proportions—need to be presented with appropriate confidence intervals so that regions of "plausible" and "implausible" values can be determined.

Second, "generally accepted in the scientific community" in point (5) is a rather low threshold. Indeed, Earth was believed to be both flat and the center of the universe for many centuries. Moreover, the phrase has been interpreted

20. *Id.* at 115–16.

21. For further information about these concepts and how they are used in the forensic context, see *id.*

22. FED. R. EVID. 702 advisory committee's notes to 2000 amendments.

23. Stephen M. Stigler, *Galton and Identification by Fingerprints*, *in* 140 PERSPECTIVES ON GENETICS: ANECDOTAL, HISTORICAL AND CRITICAL COMMENTARIES ON GENETICS 857, 857 (James F. Crow & William F. Dove eds., 1995).

24. FED. R. EVID. 702 advisory committee's notes to 2000 amendments (noting "the known or potential rate of error of the technique or theory when applied").

by forensic practitioners to mean "generally accepted in the *relevant* scientific community"[25] so bite-mark evidence meets the criterion as the discipline is generally accepted in the *relevant* scientific community of forensic odontologists.[26]

Finally, due to the proliferation of journals, the existence of a peer-reviewed publication, as set forth in point (2), no longer carries the prestige that it once did. John P.A. Ioannidis and Muin J. Khoury note that "[t]housands of new journals publish work for a fee, regardless of the quality of the work."[27] Journals published by professional societies generally have careful review procedures and have historically had relatively low (10 to 25 percent) acceptance rates.[28] But even a researcher can have trouble distinguishing between respectable and questionable journals. It would seem that the *Daubert* criteria also are not effective in keeping "junk science" out of the courtroom.

II. THE ROLE OF MODELS

Comprehensive experiments to demonstrate a method's performance can be very costly to conduct, especially when they include factors that can influence performance (e.g., evidence quality, examiner skill level, or instrument manufacturer). Conveniently, some forensic evidence processes can rely on models. Two examples where inference relies on models are DNA and drug assessment.

Consider first DNA analysis, which is based in the combinations of two short tandem repeat (STR) alleles at twenty loci.[29] Each locus can have two alleles (one from each parent) selected from the six to more than twenty-one possible alleles, which translates to between twenty and more than 200

25. *See* Transcript of *Frye* Hearing at 266, New York v. Dean, No. 4555-2007 (N.Y. Sup. Ct. June 12, 2012) [hereinafter June Transcript of *Frye* Hearing]. The author provided the Innocence Project with pro bono testimony on basic scientific principles in a pretrial "*Frye* hearing" assessing the validity and reliability of bite-mark analysis, which the judge admitted based on forensic odontologist David Senn's testimony that bite-mark analysis is "generally accepted among forensic odontologists." *See* Transcript of *Frye* Hearing at 2–116, New York v. Dean, No. 4555-2007 (N.Y. Sup. Ct. Feb. 25, 2013).

26. *See* June Transcript of *Frye* Hearing, *supra* note 25, at 81.

27. John P.A. Ioannidis & Muin J. Khoury, *Assessing Value in Biomedical Research*, 312 J. AM. MED. ASSOC. 483, 483 (2014).

28. *See, e.g.*, Daniel W. Apley, Technometrics *2017 Editor's Report*, 59 TECHNOMETRICS 413, 415 (2017) (noting an acceptance rate of 21 percent for *Technometrics*); David Dunson & Piotr Fryzlewicz, *Report of the Editors—2017*, 80 J. ROYAL STAT. SOC'Y SERIES B 3, 3 (2018) (noting an acceptance rate of less than 10 percent for the *Journal of the Royal Statistical Society*); Diane Lambert et al., *Editors' Report for 1996*, 92 J. AM. STAT. ASSOC. 391, 391 (1997) (noting acceptance rates of 25 to 30 percent for the *Journal of the American Statistical Association*); Tilmann Gneiting, Annals of Applied Stat., Annual Report for 2016, at 1 (2017), http://imstat.org/officials/reports/AnnualReports2017.pdf [https://perma.cc/X4RD-WD4U] (noting acceptance rates of 13 to 22 percent for the *Annals of Applied Statistics*).

29. *Frequently Asked Questions on CODIS and NDIS*, FBI, https://www.fbi.gov/services/laboratory/biometric-analysis/codis/codis-and-ndis-fact-sheet [https://perma.cc/55B2-4F32] (last visited Feb. 14, 2018).

genotypes.[30] Each genotype (a pair of alleles) can be characterized by its frequency of occurrence in the specific population of interest. To assess the probative value of DNA evidence, we resort to a *multinomial distribution model* for the probabilities of two specimens that have the same two alleles at each locus.

Consider locus TH01, which has six alleles (of frequency at least 1 percent).[31] The sample could contain two copies of any one of the six alleles (six possibilities) or two different alleles (fifteen possibilities),[32] for a total of twenty-one possibilities. Because the Combined DNA Index System (CODIS) database is extremely large, the frequencies of occurrence for those twenty-one genotypes have been estimated from the profiles in the database. Now suppose we have a twenty-one-sided die whose faces have the same probabilities of appearing when the die is rolled. We can calculate the probability that the die will land on the face corresponding to the genotype in the sample (for example, if all twenty-one genotypes are equally likely, then the probability is 1/21). Now we move to the next locus, say TPOX, and address the same issue: TPOX has seven alleles (of frequency at least 1 percent), or twenty-eight genotypes.[33] If all are equally likely, the probability of having the same genotype at that locus is 1/28. Assuming die rolls are independent—that is, the genotype at locus TH01 gives no information at all about the genotype at locus TPOX—we can multiply the probabilities. Repeating with eighteen more loci, the chance that the sample has the same genotype at all twenty loci is "1" if the samples came from the same source, and very tiny otherwise. DNA genotypes may not be exactly like rolling multisided dice, but, for purposes of calculating random-match probabilities, the model serves us well.[34] Those with specialized education in statistical methods are best prepared to evaluate the appropriateness of proposed models.

Elemental concentrations can be measured via inductively coupled plasma mass spectrometry (ICP-MS), where the mass spectrometer measures the signal (as a peak in the spectrum) that is generated by the ion in proportion to its concentration.[35] A convenient model for the logarithm of this

30. There can be more; often only those alleles having frequencies of at least 1 percent in the population are noted.

31. Bruce Budowle et al., *Partial Matches in Heterogeneous Offender Databases Do Not Call into Question the Validity of Random Match Probability Calculations*, 123 INT'L J. LEGAL MED. 59, 62 (2008).

32. These fifteen possibilities are found by counting the number of different permutations that exist in the six options: one and two, or one and three, and so on, through five and six.

33. Budowle et al., *supra* note 31, at 62.

34. In practice, the risk of sample contamination is much greater than the random match probability, so models are needed to describe the entire DNA process, not just the perfect identification of peaks in the spectrum corresponding to the presence of alleles. For a discussion about rarity of profiles and effects of dependence among outcomes at the different loci, both of which can render the above model inadequate, see generally Cecelia Laurie & B.S. Weir, *Dependency Effects in Multi-Locus Match Probabilities*, 63 THEORETICAL POPULATION BIOLOGY 207 (2003); Bruce S. Weir, *The Rarity of DNA Profiles*, 1 ANNALS APPLIED STAT. 358 (2007).

35. *See generally* AM. SOC'Y FOR TESTING & MATERIALS, STANDARD E2330-12: STANDARD TEST METHOD FOR DETERMINATION OF CONCENTRATIONS OF ELEMENTS IN GLASS

concentration is the Gaussian (normal) distribution.[36] That assumption allows certain useful characterizations about the precision of the mean measured concentration. One such characterization is that the true concentration lies within a calculated interval with high probability. But if the one who analyzes the data fails to recognize nonnormality or the various sources of variation that affect them, then inferences are based on an improper model. As above, education in statistical methods is essential for evaluating the appropriateness of proposed models.

In neither case does the model guarantee that the model is correct. The model may not be at all relevant, or it may be plausible but wrong (e.g., distribution is not normal but has heavier tails). It merely provides a framework for calculating probabilities. The Gaussian model for characterizing the distribution of measurements is a familiar one—but, like all models, inferences from it (e.g., "99.7 percent of the population lies within three standard deviations of the mean" and then *estimating* that mean and standard deviation with small samples) can be badly misleading if the data are not consistent with the Gaussian model. Those who use models should be extremely familiar with the errors that arise in using them when the models are not appropriate. Unfortunately, most people who use statistical methods did not have to learn the underlying mathematical theory that dictates the consequences of using statistical methods when assumptions do not hold—and, hence, the inferences that they draw can be highly misleading.

Statisticians routinely use methods to assess the appropriateness of models for a given set of data and know well that inappropriate inferences can arise if they fail to check model assumptions. It also is important to examine whether other models adequately fit the data (and, if so, offer the conclusions that those other models admit). Furthermore, it is important to acknowledge whether other models were *not* considered. The fact that an alternate model was not considered does not mean that the alternative model does not also adequately fit the data.

III. LATENT PRINT EXAMINATIONS

Expert testimony on latent prints appears to satisfy all four conditions of Rule 702. Per Rule 702(a), a latent print examiner (LPE) with specialized experience or knowledge can explain the evidence and can reliably apply the analysis, comparison, evaluation, and verification (ACE-V) process.[37] But the ACE-V process itself involves many subjective aspects which examiners cannot quantify. Presumably the full spatial assortment of all features (ridge

SAMPLES USING INDUCTIVELY COUPLED PLASMA MASS SPECTROMETRY (ICP-MS) FOR FORENSIC COMPARISONS (2017).

36. *See* GEORGE W. SNEDECOR & WILLIAM G. COCHRAN, STATISTICAL METHODS 38–64 (8th ed. 1989).

37. For an overview of the ACE-V process, see EXPERT WORKING GRP. ON HUMAN FACTORS IN LATENT PRINT ANALYSIS, LATENT PRINT EXAMINATION AND HUMAN FACTORS: IMPROVING THE PRACTICE THROUGH A SYSTEMS APPROACH 1–20 (2012).

endings, bifurcations, etc.) on a fingerprint is unique to the individual.[38] But whether the LPE's selection of a *subset* of these features is unique, much less consistent across examiners, is far less certain.[39] Even the estimation of the frequencies of single features in the population, let alone pairs or triples of features in combination, remains vague, so it is impossible to characterize the accuracy, validity, or reliability of the method in terms of sensitivity and specificity. Ideally, one would have catalogued the types of features in a latent print, and, from thousands of prints, obtained some estimates of their frequencies of occurrence. In that way, a latent print examiner could state, based on data, estimates of how "rare" or "common" such combinations of features in a print might be. But such a catalogue has not been developed. Moreover, the effects of making multiple comparisons of features induce higher error rates.[40] Experts have noted several additional reasons why fingerprint evidence has been receiving increased scrutiny.[41]

Consequently, one resorts to "black-box" studies to assess fingerprint accuracy. In these studies, an LPE is given many test pairs of prints; the test administrator knows which pairs "match" (are mated) and which do not and tries to estimate accuracy of LPEs' calls based on the study's specific collection of prints.[42] The level of difficulty in these collections likely varies from study to study: some pairs may be easy, others challenging, and still others very difficult. The largest sources of variability in latent print accuracy are likely to be "level of difficulty" and "examiner"—but even this statement has not been fully assessed in any study (in part because an objective measure of latent print quality has yet to be created).[43] Furthermore, the study should be conducted as "double blind" so that neither the LPE nor the test administrator who assigns the case knows that it is a test (people are much more careful when they know they are being tested). To date, no "blinded," much less double-blinded, studies have been conducted.[44]

---

38. For some history on Sir Francis Galton's basis for believing that fingerprint patterns are unique, see generally Stigler, *supra* note 23.

39. Various studies have shown that LPEs do not all select exactly the same subsets of features on a print. *See, e.g.*, Simon A. Cole, *Grandfathering Evidence: Fingerprint Admissibility Rulings from* Jennings *to* Llera Plaza *and Back Again*, 41 AM. CRIM. L. REV. 1189, 1226–31 (2004); Itiel E. Dror & David Charlton, *Why Experts Make Errors*, 56 J. FORENSIC IDENTIFICATION 600, 608–09 (2006); Itiel E. Dror et al., *Cognitive Issues in Fingerprint Analysis: Inter- and Intra-Expert Consistency and the Effect of a 'Target' Comparison*, 208 FORENSIC SCI. INT'L 10, 16 (2011); *see also* Itiel E. Dror, *A Hierarchy of Expert Performance*, 5 J. APPLIED RES. MEMORY COGNITION 121, 122 (2016).

40. *See generally* Yoav Benjamini & Yosef Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*, 57 J. ROYAL STAT. SOC'Y SERIES B 289 (1995).

41. *See, e.g.*, Sandy L. Zabell, *Fingerprint Evidence*, 13 J. L. & POL'Y, 143, 143–44 (2005).

42. *See generally* Bradford T. Ulery et al., *Accuracy and Reliability of Forensic Latent Fingerprint Decisions*, 108 PROC. NAT'L ACAD. SCI. 7733 (2011).

43. *See* Adele P. Peskin & Karen Kafadar, A New Measurement for the Quality of Individual Minutiae in Latent Fingerprints 2 (2016) (unpublished manuscript) (on file with author).

44. Many forensic practitioners claim that "double blind" is impossible—much the way the medical profession objected many decades ago. Today, double-blind testing is standard in

The most comprehensive study of LPE accuracy was conducted in 2011 with 169 LPEs, each of whom volunteered to examine 100 print pairs from a collection of 356 latent and 484 exemplar pairs (520 mated and 224 nonmated).[45] The print pairs were selected by "subject matter experts . . . from a much larger pool of images to include a broad range of attributes and quality" intended to be representative of real casework.[46] The study estimated a false-positive rate of 0.15 percent (6/4083); the upper 95 percent confidence limit for this error rate would be about 1 in 345 (i.e., if the study were repeated under exactly the same conditions, one would not expect to see a false-positive error rate any higher than 0.29 percent, or 1/34). However, as the study relied on LPEs who agreed to participate *and who knew that they were being tested*, this estimated false-positive error rate is likely to be a lower bound. Moreover, a subsequent study showed that LPEs changed their decisions for about 10 percent of the cases.[47] Further studies under more realistic conditions should be conducted. Such a study could include a representative sample from the population of LPEs who do not know that the case is part of a study and a documented range of print quality levels.

The noted scientist Sir Ronald Fisher, who contributed vast research to the field of statistics, was reported to have said that he would be more inclined to trust a result that had shown moderate significance (0.05) in ten studies than a result that had shown strong significance (0.005) in only one study.[48] Using that philosophy, LPE accuracy should be investigated further, ideally in a double-blind (or at least blind) fashion.

IV. COMPOSITIONAL ANALYSIS OF BULLET LEAD

From the 1960s until 2005, the FBI performed compositional analysis of bullet lead (CABL), a forensic technique that compared the trace elemental compositions in bullets found at a crime scene to those in bullets found in a suspect's possession.[49] CABL was used when no gun could be recovered or

clinical trials. Dr. Peter Stout, director of the Houston Forensic Science Center, is willing to work with the author in creating double-blind tests.

45. Ulery et al., *supra* note 42, at 7734.

46. *Id.* Recently, the Defense Forensic Science Center has proposed a more objective measure of "similarity" between two fingerprint images. The method relies on examiner-selected features for comparison, so its relevance to real-world error rates has not been demonstrated. Henry Swofford, Def. Forensic Sci. Ctr., Remarks at the SAMSI Forensics Transition Workshop, Development and Evaluation of a Model to Quantify the Weight of Fingerprint Evidence (May 9, 2016), https://www.samsi.info/wp-content/uploads/2016/03/SAMSI-2016-Swofford-DFIQI-A_and_C-Combined_HJS_REVISED.ppt [https://perma.cc/RK5Y-MGBT].

47. Bradford T. Ulery et al., *Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners*, PLOS ONE, Mar. 2012, at 1, 1 ("Examiners repeated 89.1% of their individualization decisions, and 90.1% of their exclusion decisions; most of the changed decisions resulted in inconclusive decisions.").

48. The medical literature is replete with single studies whose results have later been shown to be less than reliable.

49. *See* Steve Pierson & Karen Kafadar, *Statisticians and Forensic Science: A Perfect Match*, CHANCE, Feb. 2016, at 4, 6.

when bullets were too small or fragmented to compare striations on the casings with those on the gun barrel.

FBI chemists designed a "suite" of seven trace elements,[50] whose concentrations, measured via inductively coupled plasma optical emission spectrometry (ICP-OES), were believed to provide unique "signatures" for all bullets in the box.[51] Thus, they measured these concentrations in bullets found at a crime scene; if the concentrations were "close" to those in bullets found in the suspect's possession, they were deemed "analytically indistinguishable" and, hence, implied evidence of "guilt."[52] FBI chemists might then be called to testify in court that the two sets of bullets "matched."[53]

Presumably, the technique met the Rule 702 requirements. The FBI bullet-lead examiner was well trained in matters of chemistry, the measurement technique was reliable, the testimony was based on data, and the chemist applied ICP-MS reliably. Nonetheless, the "FBI laboratory announce[d] discontinuation of bullet lead examinations" in September 2005.[54] Why did the FBI discontinue the method?

The FBI had asked the National Research Council to convene a committee in the National Academy of Sciences to evaluate "the scientific method, the data analysis, and the interpretation of the results" from bullet-lead examinations.[55] Accordingly, the committee investigated the reliability and validity of CABL, both in terms of the analytical chemistry *and its method of inference from the data*. The committee's report, *Forensic Analysis: Weighing Bullet Lead Evidence*, included discussions on the effects of the manufacturing process on the validity of the comparisons, the precision and accuracy of the chemical measurement technique, and the statistical methodology used to compare two bullets and to test for a "match."[56] Briefly, the committee found that the chemical analysis (ICP-OES) was sound and that the selection of the seven elements for comparison was sensible.[57]

---

50. These elements were antimony (Sb), arsenic (As), bismuth (Bi), cadmium (Cd), copper (Cu), silver (Ag), and tin (Sn). In the 1960s, when the FBI began conducting CABL, it did so with only antimony, copper, and arsenic concentrations measured via neutron activation analysis (NAA). NAT'L RESEARCH COUNCIL, *supra* note 51, at 39. Over time, however, additional elements were added and the final suite of seven elemental concentrations, measured via inductively coupled plasma optical emission spectroscopy, was completed with the formal addition of cadmium in 1995. *Id.* at 19.

51. *See* NAT'L RESEARCH COUNCIL, FORENSIC ANALYSIS: WEIGHING BULLET LEAD EVIDENCE 1–2 (2004). *See generally* Paul C. Giannelli, C*omparative Bullet Lead Analysis: A Retrospective*, 47 CRIM. L. BULL. 306 (2011).

52. *Id.*

53. *See* K.D. Pan & K. Kafadar, *Statistical Analysis of Forensic Glass*, 12 ANNALS APPLIED STAT. (forthcoming June 2018) (manuscript at 3).

54. Press Release, FBI, FBI Laboratory Announces Discontinuation of Bullet Lead Examinations (Sept. 1, 2005), https://archives.fbi.gov/archives/news/pressrel/press-releases/fbi-laboratory-announces-discontinuation-of-bullet-lead-examinations [https://perma.cc/6T7Y-EYUT].

55. *Id.*

56. *See generally* NAT'L RESEARCH COUNCIL, *supra* note 51.

57. *Id.* at 3.

However, at least five problems should be noted. First, the consistency of the manufacturing process for making bullets was high, resulting in perhaps thousands of bullets that could have nearly identical seven-element "signatures."[58] Second, the "match rule" for determining when two bullets were "analytically indistinguishable" was too generous, leading to an uncomfortably high false-positive rate (claiming a "match" when in fact the elemental concentrations were quite different).[59] Third, the concentrations were not independent as was believed (antimony and copper were noticeably correlated).[60] Four, the collection of 1837 bullets on which the FBI tested their "match rule" was useful for some purposes (e.g., for providing information on approximate ranges in levels of concentrations of seven trace elements that might be observed in a production batch of bullet lead, existence of recording errors, etc.) but not for estimating the false-positive error rate. The bullets in the collection were not a random sample of bullets but rather were "selected" to be different[61] ("one specimen from each combination of bullet caliber, style, and nominal alloy class was selected and that data was placed into the test sample set").[62] Fifth, and consequently, the FBI's stated false-positive error rate of 0.04 percent (about 1 in 2500)[63] was not valid.

The chemist had specialized knowledge, the measurement technique was sound and was properly applied, and much bullet-lead data had been collected over the years. So in what ways did Rule 702 "fail" in CABL? Quite simply, *chemists* were permitted to testify, not only about their chemical measurement technique *but also, unjustifiably, about statistical methodology, the data set being used for "validating" error rates, and data analysis*. Chemists may have some statistics training, just as many statisticians have taken college courses in chemistry. But allowing a chemist to testify about the inferences drawn from data (which is badly biased in favor of unrealistically low false-positive rates) makes about as much sense as allowing a statistician to testify about ICP-OES or ICP-MS. No one would even think of it. Yet, for some reason, chemists have been allowed to testify repeatedly about the inferences from very limited sample sizes (three measurements per element) based on a method that had little statistical grounding and a "validation" method on a biased data set.

Further, chemists have been allowed to state far more definitive conclusions than could be justified. Any undergraduate major in statistics would recognize the bias in the data set and would know the proper statistical

58. *See* Clifford H. Spiegelman & Karen Kafadar, *Data Integrity and the Scientific Method: The Case of Bullet Lead Data as Forensic Evidence*, 19 CHANCE 17, 17–18 (2006); *see also* Pan & Kafadar, *supra* note 53 (manuscript at 3–4).

59. *See* Spiegelman & Kafadar, *supra* note 58, at 22; Pan & Kafadar, *supra* note 53 (manuscript at 2).

60. *See* Pierson & Kafadar, *supra* note 49, at 7.

61. *See* Spiegelman & Kafadar, *supra* note 58, at 21–22.

62. NAT'L RESEARCH COUNCIL, *supra* note 51, at 175; *see also* Robert D. Koons & JoAnn Busaglia, *Forensic Significance of Bullet Lead Compositions*, 50 J. FORENSIC SCI. 341, 343 (2005).

63. NAT'L RESEARCH COUNCIL, *supra* note 51, at 193.

technique for comparing univariate means (one element at a time: Student's t) and for comparing multivariate means (seven elements simultaneously when those elements are correlated, as they were here: Hotelling's T^2). The chemists' "2-SD-overlap" technique, whereby means and standard deviations from only three measurements on each element yielded an interval constructed from the mean plus or minus two standard deviations and checking to see if all seven sets of intervals overlapped, led to claiming "analytically indistinguishable" (which jurors hear as "match") for an uncomfortably high proportion of bullets that actually differed considerably in their concentrations.[64] Moreover, the data set used for "validating" the false-positive rate consisted of *selected* samples and hence clearly was not "unbiased."[65] Most statisticians are very well versed in recognizing biased data sets.[66]

To clarify, chemists certainly should be allowed to testify about the method that was used to measure the concentration, the number of measurements that were taken, and even the mean and standard deviation of those measurements. But the further steps of calculating "match intervals" and drawing inferences from them fall outside their domain of expertise. A statistician need not be called in to every court case involving trace evidence, but a statistician should have been consulted in the development of the inference procedure from those data and in the proper estimation of error rates that could arise from that procedure.

Sadly, this ground is likely to be covered again, this time with forensic glass comparisons. The Organization of Scientific Area Committees (OSAC) just approved the posting on the OSAC Registry of Standards the present American Society for Testing and Materials Standard E2926-17, *Standard Test Method for Forensic Comparison of Glass Using Micro X-ray Fluorescence (μ-XRF) Spectrometry*.[67] This standard is for glass concentrations measured via the μ-XRF technique;[68] two other standards use other measurement techniques (E2927-16 uses Laser Ablation ICP-MS[69] and E2330-12 uses ICP-MS[70]). All three standards include section 10 (section 11 in E2927-16): "Calculation and Interpretation of Results." The basic

64. *See* Pierson & Kafadar, *supra* note 49, at 6.

65. Koons & Busaglia, *supra* note 62, at 341.

66. For an excellent nontechnical discussion of key principles of sampling to ensure that sample data are representative of the relevant populations, see generally William G. Cochran et al., *Principles of Sampling*, 49 J. AM. STAT. ASS'N 13 (1954).

67. *See* Org. of Sci. Area Cmtys. for Forensic Sci., *OSAC Newsletter*, NIST (Mar. 15, 2017), https://www.nist.gov/topics/forensic-science/osac-newsletter-march-2017 [https://perma.cc/CB8Q-TFYL].

68. *See generally* AM. SOC'Y FOR TESTING & MATERIALS, STANDARD E2926-17: STANDARD TEST METHOD FOR FORENSIC COMPARISON OF GLASS USING MICRO X-RAY FLUORESCENCE (μ-XRF) SPECTROMETRY (2013).

69. *See generally* AM. SOC'Y FOR TESTING & MATERIALS, STANDARD E2927-16: STANDARD TEST METHOD FOR DETERMINATION OF TRACE ELEMENTS IN SODA-LIME GLASS SAMPLES USING LASER ABLATION INDUCTIVELY COUPLED PLASMA MASS SPECTROMETRY FOR FORENSIC COMPARISONS (2017).

70. *See generally* AM. SOC'Y FOR TESTING & MATERIALS, *supra* note 35.

components of the test method are the same; for simplicity, they are described below with specific reference to E2330-12, ICP-MS:

- Several trace element concentrations are measured (at least) three times in each fragment of glass from the two sources (the crime scene, or "recovered," and suspect, or "known");[71]
- Calculate the mean and standard deviation from the ≥ 3 measurements on each element;[72]
- Calculate a "match interval" of mean ±4 standard deviations using the data from the "known" fragment;[73]
- "If the mean concentration of one (or more) element(s) in the Recovered fragment falls outside the match interval for the corresponding element in the Known fragments, the element(s) does not 'match' and the glass samples are considered distinguishable."[74]

The concerns with the inferences in "Calculation and Interpretation of Results" are the same:

- The "match interval" procedure (for comparing many mean concentrations from two samples) fails to acknowledge multiple sources of variability in the measurements.
- The use of only three measurements for estimating a standard deviation is highly unstable. Many people do not realize the large samples needed to have 95 percent confidence in just one digit of accuracy in estimating the standard deviation from an idealized Gaussian distribution: it can be as small as 31 or as large as 600.[75]
- An "optimal" statistical procedure can be derived *if the data have a specified distribution*. If not, then a more robust procedure that demonstrates good performance across a range of assumptions, especially when the measurements on the concentrations are correlated (some very highly so, such as zirconium and hafnium), is needed.
- The data sets on which false-positive error rates have been estimated are likewise biased toward including a diverse set of samples—as diverse as possible. The fact that one finds even a single false match is surprising given that the samples are included to represent the diversity of glass. Only two papers using LA-ICP-MS appear to have measured the same fragment multiple times and multiple fragments from the same pane of glass, thereby providing some tentative estimates of within-fragment variability and within-pane variability. One hopes both of them are small compared to between-pane variability, but they may not be if manufacturers produce highly consistent batches of glass—glass panes themselves exhibit different elemental concentrations in different parts of the pane.

71. *Id.* at 3.
72. *Id.*
73. *Id.*
74. *Id.*
75. More are needed if that digit is nine, fewer if that digit is one.

- In reality, the variation in the measurements is likely to be much larger than the variation represented by "±4 standard deviations" based on only three measurements on each element. ROC curves[76] can be constructed using many sets of data using the ±4 standard deviations compared with ±2 SD or ±6 SD, and so on.[77]
- If *all* of the recovered fragment means fall within the corresponding known fragment "match intervals" then the impression of a juror may well be that that the glass samples cannot be "considered distinguishable"—that is, they are "indistinguishable," leading one to assume "guilt."[78]

Based on preliminary results (to be submitted for publication), I fear that the trajectory for forensic glass evidence will follow that of CABL. In the meantime, chemists should not be permitted to present expert testimony about statistical inferences from data, unless those inferences were developed using sound statistical methods with error rates estimated properly using unbiased, representative data sets. Statistical inference and data analysis require considerably more statistical expertise than the mere calculation of sample means and standard deviations.

V. A NOTE ABOUT LIKELIHOOD RATIOS

Several forensic scientists (notably in Europe) have argued that Bayesian inference is the only "logical" framework for presenting evidence.[79] This framework suggests that the forensic expert should not draw conclusions about "match" or "nonmatch" but rather state only the ratio of probabilities about the evidence under two different hypotheses. Using glass evidence as an illustration, these two probabilities (P{ }) are

(1) P{Evidence | Hypothesis: "Same source of glass"}

(2) P{Evidence | Hypothesis: "Different sources of glass"}

The ratio of these two probabilities, (1)/(2), is the likelihood ratio (LR). If we focus on a specific test method that generated the "evidence," then, referring back to the table in Part II, (1) can be viewed as the *sensitivity* of the test method and (2) is the *false negative rate*, or (1 – *specificity*).

As emphasized in Part II, the error rates (false-positive rate, or 1– *sensitivity*, and false negative rate, or 1 – *specificity*) must be *estimated* from real data. Sometimes *sensitivity* and *specificity* depend on other factors, such as experience or the number of replicates used in the test method. These estimates have *uncertainty*. For example, 100 tests of true mated pairs with seven false negatives yields a false negative rate of 0.07, and another 100

76. *See, e.g.*, PETER ARMITAGE ET AL., STATISTICAL METHODS IN MEDICAL RESEARCH 496, 697 (1971).

77. *See* Pan & Kafadar, *supra* note 53 (manuscript at 20).

78. Jessica Gabel-Cino, Presentation to the Second Annual Conference of the National Center for Forensic Science, Expert Witnesses and Lawyers: Can We All Get Along? (Oct. 17, 2017), https://ncfs.ucf.edu/ffsc_info/ [https://perma.cc/5SDQ-F4XP].

79. *See generally* C. Neumann et al., *Quantifying the Weight of Evidence from a Forensic Fingerprint Comparison: A New Paradigm*, 175 J. ROYAL STAT. SOC'Y SERIES A 371 (2012).

tests of true nonmated pairs with two false positives yields an estimated *sensitivity* of 0.98, suggesting a likelihood ratio of fourteen. If these 200 tests were repeated in a similar fashion, the LR is not likely to be fourteen again. The LR is reasonably likely to fall between eight and forty-eight, but it is unrealistic to believe that it will be exactly fourteen again.

As noted above, the LR does *not* give the probability that the person is guilty. It does *not* provide a list of sources from which the evidence might have come, nor the probabilities associated with each source. It merely provides a ratio of probabilities of seeing the evidence if the two sources are the same versus not the same. When the LR is multiplied by the *prior odds*

$$p/(1 - p) \text{ where } p = P\{\text{Same source of glass}\},$$

the product gives the *posterior odds*

$$P\{\text{Same source} \mid \text{Evidence}\}/ P\{\text{Different sources} \mid \text{Evidence}\}.$$

A likelihood ratio of 1 suggests that the evidence is equally consistent with the hypothesis that the suspect is innocent or that the suspect is guilty. A posterior odds ratio of 1 suggests that, in view of the evidence, the probability that the person is guilty equals the probability that the person is innocent. Figure 1 might help to explain the connection between the value of the likelihood ratio (horizontal axis) and the posterior odds of being guilty (versus not guilty) in light of the evidence for different values of one's prior probability p that the suspect is guilty (where p is one in ten, one in twenty-five, one in fifty, one in 100, and so on through 100,000). The highest line corresponds to a prior belief that the suspect might be one of only ten people who could have committed the crime; the lowest line corresponds to a prior belief that as many as 100,000 others might have been guilty. The plot shows that one needs very large likelihood ratios to believe that the suspect is more likely than not to be guilty.[80]

Some experts emphasize that the LR approach to decision-making is based on the expert's assessment of the probabilities in formulas (1) and (2), which is subjective (and hence may not be based on "sufficient facts or data") or may be based on many assumptions.[81] Failure to detect violations from assumptions does *not* imply that no other model can be better; the data may admit several plausible models, with possibly a range of conclusions. For example, the likelihood ratio could be 1000 if one assumes that the concentrations are normally distributed, but it could be ten if they are lognormally distributed. Moreover, the definitions of "same" and "different" depend on how "close" or "far" the concentrations are deemed to be in order to be judged as "same" and "different." Finally, a legitimate concern has been raised that a juror is likely to interpret the LR as the posterior odds, which it surely is not. An LR of 1000 would yield a posterior odds ratio of 100 to 1 (very persuasive) if the prior odds ratio is 0.1 (roughly one in ten panes of glass) but only 0.01 if the prior odds ratio is 0.00001 (1 in 10,000

80. *See infra* Figure 1.

81. *See generally* Steven P. Lund & Hariharan K. Iyer, *Likelihood Ratio as Weight of Forensic Evidence: A Closer Look*, J. RES. NAT'L INST. STANDARDS TECH., Oct. 2017, at 1.

panes of glass). Quoting from John Tukey, "At least until the literature has many more examples of how to think about choosing priors, then, I shall have my doubts of the wisdom of trying to formalize the whole process."[82]

VI. COURTROOM TESTIMONY

Rule 702 provides criteria that aim to ensure valid and well-founded scientific testimony and to eliminate unqualified experts, unfounded scientific claims, and inadequately demonstrated science. Admission of forensic evidence such as CABL, hair analysis, and bite marks illustrates that Rule 702 has not always succeeded.

Brendan Max, chief of the forensic science division at the Chicago Public Defender's Office, reports numerous instances of forensic examiners who are not required to answer questions about their knowledge of recent research in the latent print field.[83] The demonstration of such knowledge presumably is necessary for establishing "expertise" in the field. For example, LPEs now recognize the inappropriateness of stating "zero error." Yet "experts" are permitted to evade such questions during pretrial discovery[84] or even during trial.[85] Quoting from Chief Max:

> Continuing with forensic fingerprints as an example, significant benchmark literature now exists in the field. . . . Examiners who are unfamiliar with the fundamental research and its implications on casework

---

82. 3 JOHN W. TUKEY, *Foreword to the Philosophy Volumes* of THE COLLECTED WORKS OF JOHN W. TUKEY: PHILOSOPHY AND PRINCIPLES OF DATA ANALYSIS: 1949–1964 xxxix, xli (Lyle V. Jones ed., 2017).

83. Brendan Max, Chief, Forensic Sci. Div., Chi. Pub. Def. Office, Remarks at the International Conference on Forensic Inference and Statistics, Reforming Forensics: What Are the Odds We Do It and Get It Right? (Sept. 6, 2017).

84. Memorandum from Brendan Max, Chief, Forensic Sci. Div., Chi. Pub. Def. Office (Oct. 14, 2017) (quoting Letter from Leo Schmitz, Dir., Ill. State Police, to Brendan Max, Forensic Sci. Div., Chi. Pub. Def. Office (Nov. 22, 2016)) (on file with author). The letter states:

> Dear Chief Max:
>
> I received a copy of your correspondence addressed to the Illinois State Police (ISP) Forensic Science Center at Chicago. . . . To make forensic scientists available for extended meetings to discuss scientific foundation and peer review articles is outside the scope of these meetings and impedes the work done on the bench to ensure timely forensic analysis. We feel it is more appropriate to establish scientific foundation during voir dire at trial.

Id.

85. *Id.* The following exchange is taken from the examination of a Chicago Police Department fingerprint examiner:

> Q: And while you say that, you cannot say that you can exclude all other people in the world as a possible source of a latent impression, correct?
>
> A: Yes, I can.
>
> Q: I'm sorry, you said yes?
>
> A: Yes.
>
> Q: Yes, I'm correct or yes, you can exclude all others in the world?
>
> A: Yes, to the exclusion of all others, yes.

Id.

> are arguably not qualified to testify and will mislead the trier of fact if permitted to do so.
>
> Unfortunately, the message of reform has trickled down very unevenly across the many forensic labs in the U.S. . . .
>
> In one underfunded local law enforcement fingerprint lab, we have questioned examiners about the foundational literature in their field as a means of assessing qualifications, and we have identified unqualified examiners—they are unaware of the fundamental literature, don't understand the important concepts referenced in the literature, and can't correctly identify how the current research in the field effects their methods and conclusions. And yet, try as we may, we can't get their testimony excluded pursuant to Rule 702, and can't even get judges to conduct pre-trial hearings to assess qualifications.[86]

To address these problems, Chief Max recommends the following changes to Rule 702: (1) "Require pre-trial qualification evidentiary hearings upon written motion of a litigant," (2) "[r]equire any expert who is the subject of a pre-trial qualification hearing to submit to a compulsory deposition, and" (3) "[r]equire that experts disclose all the facts and data that support their proffered opinions (such as all features in a fingerprint case that support an association between a latent print and a suspect)."[87]

VII. FINAL COMMENTS

This Article provides some illustrations of the shortcomings in Rule 702 in the context of assuring expertise in forensic disciplines. Experts in forensic science should be required to disclose the basis for their expertise, either via comprehensive blinded proficiency tests or via pretrial discovery, as well as their methods for obtaining data and why those methods of data collection are scientifically valid and reliable. But they should *not* be allowed to testify about *inferences from those data*, especially when they rely on their (often inadequate) understandings about statistical methods, particularly because they often are (understandably) unaware of inherent statistical issues that should be considered when evaluating forensic evidence. Well-characterized, objective metrics (with appropriate intervals of uncertainty, such as 95 percent confidence intervals) need to be developed for each type of evidence. The studies to evaluate its performance on realistic cases need to be designed and conducted to be truly representative of the population of interest (unbiased data sets) and account for sources of variability that can affect the results. Such (ideally blind) studies will lead not only to the identification of conditions under which the evidence is valuable but also to issues which can be addressed and ultimately strengthen the value of the evidence.

Whether Rule 702 remains as it is or is strengthened to address the shortcomings noted in this Article, it is important for judges to understand the criteria involved in assessing reliability and validity. Judges should also

86. *Id.*
87. *Id.*

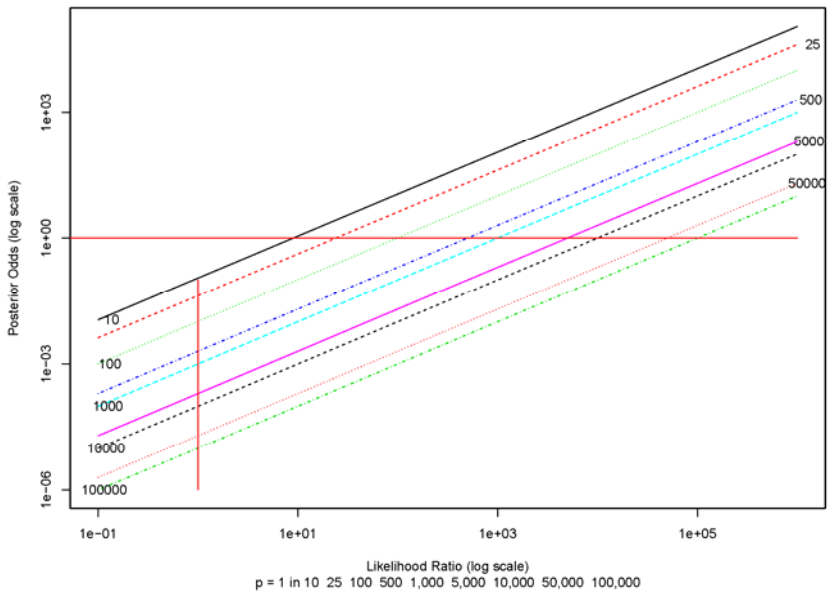
understand that, even if an expert is deemed to have met all Rule 702 criteria and is allowed to offer testimony, the appropriateness of the testimony should be assessed to ensure that the expert is not presenting opinions as if they were facts—especially when such opinions are based on methods that fall outside the expert's domain of expertise. The National Academy of Sciences report did *not* state how courts should treat the admissibility of forensic evidence. Presumably, admissibility of forensic evidence should not differ from that of any other kind of evidence that claims to be scientific. In the meantime, courts must continue to hear cases. Judge Edwards stated in his testimony to the Senate Judiciary Committee:

> It will be no surprise if the report is cited authoritatively for its findings about the current status of the scientific foundation of particular areas of forensic science. And it is certainly possible that the courts will take the findings of the committee regarding the scientific foundation of particular types of forensic science evidence into account when considering the admissibility of such evidence in a particular case. However, *each case in the criminal justice system must be decided on the record before the court pursuant to the applicable law, controlling precedent, and governing rules of evidence*. The question whether forensic evidence in a particular case is admissible under applicable law is not coterminous with the question whether there are studies confirming the scientific validity and reliability of a forensic science discipline.[88]

Until then, strengthening Rule 702 will benefit all of the sciences, including forensic science.

88. *The Need to Strengthen Forensic Science in the United States: The National Academy of Sciences' Report on a Path Forward: Hearing Before the S. Comm. on the Judiciary*, 111th Cong. 15 (2009) (statement of Harry T. Edwards, Senior Circuit J., U.S. Court of Appeals for the D.C. Circuit, and Co-Chair, Committee on Identifying the Needs of the Forensic Science Community, National Research Council of the National Academies) (emphasis added).

Figure 1: Visualizing Likelihood Ratio and Posterior Odds[89]



89. This figure shows the connection between likelihood ratio (x-axis) and posterior odds (y-axis) for different levels of prior probability of guilt (1 in 10, 1 in 25, and so on, through 1 in 100,000). The vertical line corresponds to a likelihood ratio of 1 (that is, the evidence is equally consistent with the hypothesis that the suspect is innocent or that the suspect is guilty). The horizontal line corresponds to a posterior odds ratio of 1 (that is, in view of the evidence, the probability that the person is guilty equals the probability that the person is innocent).