# HOW *DAUBERT* AND ITS PROGENY HAVE FAILED CRIMINALISTICS EVIDENCE AND A FEW THINGS THE JUDICIARY COULD DO ABOUT IT

*David H. Kaye**

## INTRODUCTION

Much of criminalistics concerns identification—associating traces such as fingerprints, fibers, glass fragments, paint chips, bullets, and DNA with their possible sources.[1] As this type of evidence became a staple of litigation, concerns over its accuracy surfaced.[2] With increasing urgency, observers called for greater regulation of crime laboratories[3] and better research into the validity of the scientific techniques.[4] Books and articles with titles such

* Distinguished Professor of Law, Penn State Law, University Park. 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). Timothy Lau and Jay Koehler provided helpful comments on a draft.

1. On the meaning of "criminalistics," see, for example, KEITH INMAN & NORAH RUDIN, PRINCIPLES AND PRACTICE OF CRIMINALISTICS: THE PROFESSION OF FORENSIC SCIENCE 10–12 (2001); CHARLES E. O'HARA & JAMES W. OSTERBURG, AN INTRODUCTION TO CRIMINALISTICS, at xii (1949) (defining criminalistics as "that science which applies the physical sciences in the investigation of crimes").

2. In 1979, for example, Professor Edward Imwinkelried described a government report on proficiency testing at some 240 crime laboratories as "alarming." Edward J. Imwinkelried, *The Constitutionality of Introducing Evaluative Laboratory Reports Against Criminal Defendants*, 30 HASTINGS L.J. 621, 636 (1979). Two years later, he deemed this "an understatement" and wrote that "'[s]hocking' would be more precise." Edward J. Imwinkelried, *A New Era in the Evolution of Scientific Evidence—A Primer on Evaluating the Weight of Scientific Evidence*, 23 WM. & MARY L. REV. 261, 268 (1981). However, the percentages in the report were not broken down into false-positive, false-negative, and falsely inconclusive findings, and they had other deficiencies. A more refined analysis suggested that the error rates on the early proficiency tests were inflated. Joseph L. Peterson & Penelope N. Markham, *Crime Laboratory Proficiency Testing Results, 1978–1991, II: Resolving Questions of Common Origin*, 40 J. FORENSIC SCI. 1009, 1009 (1995). Even so, traditional proficiency tests are not designed to measure the risk of error in actual casework and probably underestimate it in most fields. *See* Jonathan J. Koehler, *Proficiency Tests to Estimate Error Rates in the Forensic Sciences*, 12 L. PROBABILITY & RISK 89, 90–91 (2013).

3. *E.g.*, Randolph N. Jonakait, *Forensic Science: The Need for Regulation*, 4 HARV. J. L. & TECH. 109, 109 (1991).

4. *See* Paul C. Giannelli, Daubert *and Forensic Science: The Pitfalls of Law Enforcement Control of Scientific Research*, 2011 U. ILL. L. REV. 53, 64–65 (describing the history of research in the field).

as *Forensic Science Under Siege* and *Failed Forensics* followed.[5] With evidence of serious errors mounting in both high- and low-profile cases, Congress appropriated funds for "the National Academy of Sciences to create an independent Forensic Science Committee" to study and make recommendations to improve the practice of forensic science.[6] A report emerged in 2009.[7] It confirmed much of the earlier academic criticism. The seventeen-member committee pointedly wrote that "[i]n a number of forensic science disciplines, forensic science professionals have yet to establish either the validity of their approach or the accuracy of their conclusions, and the courts have been utterly ineffective in addressing this problem."[8] The committee also observed that "[f]ederal appellate courts have not with any consistency or clarity imposed standards ensuring the application of scientifically valid reasoning and reliable methodology in criminal cases involving *Daubert* questions."[9] This situation, it added, was "not really surprising" given that *Daubert* is so "flexible."[10]

The years that followed proved frustrating to those who had hoped that the courts would demand, as a condition for admissibility, the scientific proof of validity and accuracy that the committee found absent in some areas.[11] Of all the published opinions responding to challenges to unique identification via largely subjective comparisons of patterns and impressions, a grand total of two saw the sentence bemoaning the "utterly ineffective" judicial treatment of validity or accuracy as important enough to quote.[12] To be sure, many courts acknowledged the existence of the committee's calls for research to demonstrate these qualities, but they read them as not particularly relevant to the issue of admissibility. After all, these judges wrote, the recommendations for filling even the most gaping holes in foundational research were not directed at the courts,[13] and, even if the committee had

5. *See generally* KELLY M. PYREK, FORENSIC SCIENCE UNDER SIEGE (2007); Michael J. Saks & David L. Faigman, *Failed Forensics: How Forensic Science Lost Its Way and How It Might Yet Find It*, 4 ANN. REV. L. & SOC. SCI. 149 (2008).

6. S. REP. NO. 109-88, at 46 (2005).

7. NAT'L RESEARCH COUNCIL, STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES: A PATH FORWARD 1 (2009), https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf [https://perma.cc/CLW3-Y6VQ].

8. *Id.* at 53.

9. *Id.* at 96 (citing Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579 (1993)).

10. *Id.*

11. *E.g.*, David L. Faigman et al., *Preface* to 1 MODERN SCIENTIFIC EVIDENCE, at x (David L. Faigman et al. eds., 2016) (noting that "experience-based specialties have suffered . . . what was thought to be a terminal blow in 2009 with the publication of the [NRC] Report" but that "courts have largely ignored the virtually consensus opinion of mainstream academic scientists" and have responded to it with "indifference" and "intransigence").

12. Almeciga v. Ctr. for Investigative Reporting, Inc., 185 F. Supp. 3d 401, 415–16 n.8 (S.D.N.Y. 2016); State v. Hull, 788 N.W.2d 91, 104 n.4 (Minn. 2010).

13. *E.g.*, United States v. Aman, 748 F. Supp. 2d 531, 536 (E.D. Va. 2010) ("[T]he NRC Report does not recommend barring fire investigators from offering opinions . . . ."); Commonwealth v. Fulgiam, 73 N.E.3d 798, 820 n.26 (Mass. 2017) ("[T]he [NRC] Report does not draw the conclusion that fingerprint evidence lacks such reliability that courts should no longer deem it admissible . . . .").

opined on the admissibility of a given type of evidence, that recommendation could "not bind federal courts."[14]

In response to the marginalization of the NRC report and other critiques of some fields of criminalistics, the President's Council of Advisors on Science and Technology (PCAST)—a group of the nation's most eminent scientists and engineers that makes policy recommendations for the executive branch—issued a report late in 2016 on *Ensuring Scientific Validity of Feature-Comparison Methods*.[15] The report is far more direct in its approach to legal questions than was the 2009 report. It argues that under Federal Rule of Evidence 702, which applies to expert testimony generally and to scientific expert testimony as described in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,[16] admissibility requires scientific validity, that validity only can be established through empirical studies of how the methods work, and that for largely subjective conclusions of criminalists on the possible sources of trace evidence, performance studies must show a rate of false-positive identifications of no more than 5 percent.[17] The report concludes that some commonly used methods of identification have not been shown to satisfy these criteria. The ineluctable conclusion is that the courts cannot admit findings from these methods.[18]

This moment thus provides an opportunity for reflection on the principal rules governing the reception of criminalistics evidence in the courts. Should Federal Rule of Evidence 702 be rewritten to make it clear that criminalistics evidence requires certain kinds of validity studies before it can be considered admissible? Would "an Advisory Committee note, providing guidance to Federal judges concerning the admissibility under Rule 702 of expert testimony based on forensic feature-comparison methods" suffice?[19] Is more judicial education on *Daubert* and the nature and practice of science the solution?

At the risk of disappointing, this Article does not give firm answers to these questions. Its goal is less ambitious. It supplies information to assist judicial bodies concerned with possible rules changes—and courts applying the current rules—in improving their regulation of criminalistics identification evidence. Part I documents how courts have failed to faithfully apply *Daubert*'s criteria for scientific validity to this type of evidence. It describes how ambiguities and flaws in the terminology adopted in *Daubert* combined

14. *Aman*, 748 F. Supp. 2d at 536.

15. 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 1 (2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf [https://perma.cc/R76Y-7VU].

16. 509 U.S. 579 (1993).

17. *See infra* note 112 and accompanying text.

18. *See* David H. Kaye, *PCAST on "Foundational Validity," Evidentiary Reliability, and the Admissibility of "Firearms Analysis,"* FORENSIC SCI. STAT. & L. (Oct. 23, 2016, 2:21 PM), http://for-sci-law.blogspot.com/2016/10/pcast-on-foundational-validity.html [https://perma.cc/PCV5-AQQV].

19. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 15, at 145.

with the opaqueness of forensic-science publications and standards have been exploited to shield some test methods from critical judicial analysis. Simply desisting from these avoidance strategies would be an improvement.

Part II notes how part of the U.S. Supreme Court's opinion in *Kumho Tire Co. v. Carmichael*[20] has enabled courts to lower the bar for what is presented as scientific evidence by mistakenly maintaining that there is no difference between that evidence and other expert testimony that need not be scientifically validated. It suggests that a version of Rule 702 that explicitly insists on more rigorous validation of evidence that is promoted or understood as being "scientific" would be workable and more clearly compatible with the rule's common law roots.

Part III sketches various meanings of the terms "reliability" and "validity" in science and statistics, on the one hand, and in the rules and opinions on the admissibility of expert evidence, on the other. It discusses the two-part definition of "validity" in the PCAST report and the proposed criteria for demonstrating scientific validity of subjective pattern-matching testimony. It contends that if "validity" means that a procedure (even a highly subjective one) for making measurements and drawing inferences is fit for its intended use, then whether test results that have higher error rates than the ones selected in the report might nevertheless assist fact finders who are also appropriately informed of the evidence's probative value must be evaluated.

Finally, Part IV articulates two distinct approaches to informing judges or jurors of the import of similarities in features: the traditional one in which examiners opine on the truth and falsity of source hypotheses and a more finely grained one in which criminalists report only on the strength of the evidence. It suggests that the rules for admitting scientific evidence need to be flexible enough to accommodate the latter, likelihood-based testimony when it has a satisfactory empirically established basis.

## I. DODGING *DAUBERT*

*Daubert* resolved a conflict among the circuit courts as to whether "'general acceptance' [in the relevant scientific community was] the exclusive test for admitting expert scientific testimony" under Rule 702.[21] The Court held that it was not—that the trial court must employ a broader framework for evaluating "whether the reasoning or methodology underlying the testimony is scientifically valid."[22] Moreover, Justice Harry Blackmun's opinion for the Court supplied a nonexhaustive list of "pertinent consideration[s],"[23] namely, (1) "whether it can be (and has been) tested,"[24] (2) "whether the theory or technique has been subjected to peer review and publication,"[25] (3) "the known or potential rate of error,"[26] (4) "the existence

20. 526 U.S. 137 (1999).
21. *Daubert*, 509 U.S. at 589.
22. *Id.* at 592–93.
23. *Id.* at 593.
24. *Id.*
25. *Id.*
26. *Id.* at 594 (citing United States v. Smith, 869 F.2d 348, 353–54 (7th Cir. 1989)).

and maintenance of standards controlling the technique's operation,"[27] and (5) the "degree of acceptance within [a relevant scientific] community."[28] As applied in the lower courts, however, this list was not always used to structure a thoughtful inquiry into the "overarching subject [of] scientific validity."[29] Instead, and particularly with criminalistics evidence, they sometimes devolved into a superficial, if not pro forma, checklist. A brief sketch of this development with respect to each factor follows.[30]

### *A. Testability and Testing*

The Supreme Court began its explanation of scientific validity by observing that "a key question . . . in determining whether a theory or technique is scientific knowledge . . . will be whether it can be (and has been) tested."[31] Indeed, the Court added that "generating hypotheses and testing them to see if they can be falsified . . . is what distinguishes science from other fields of human inquiry."[32]

In applying this first factor, two problems have emerged. First, some courts have been impressed with testability rather than actual testing. In *Lee v. Martinez*,[33] the Supreme Court of New Mexico deemed the "testability" prong of *Daubert* satisfied merely because "the control question polygraph examination *can* be tested."[34] And in *United States v. Mitchell*,[35] Judge Edward Becker wrote for the Third Circuit that "the hypotheses that undergird the discipline of fingerprint identification are testable, if only to a lesser extent actually tested by experience, and so we find this factor to weigh in favor of admitting the evidence."[36]

Although theories that cannot be falsified—or at least tested to some degree—by experiments or observations are not part of science, the abstract possibility of testing adds almost nothing to a claim of scientific knowledge. Testability or falsifiability alone does not come close to satisfying the first *Daubert* factor. The mere possibility of systematically checking on the predictions of astrologers, for instance, would not "weigh in favor of admitting" such predictions.

---

27. *Id.* (citing United States v. Williams, 583 F.2d 1194, 1198 (2d Cir. 1978)).
28. *Id.* (quoting United States v. Downing, 753 F.2d 1224, 1238 (3d Cir. 1985)).
29. *Id.* at 594–95.
30. For a more complete treatment, see DAVID H. KAYE ET AL., THE NEW WIGMORE ON EVIDENCE: EXPERT EVIDENCE § 7.3.2. (2d ed. 2010).
31. *Daubert*, 509 U.S. at 593.
32. *Id.* (quoting Michael D. Green, *Expert Witnesses and Sufficiency of Evidence in Toxic Substances Litigation: The Legacy of* Agent Orange *and Bendectin Litigation*, 86 NW. U. L. REV. 643, 645 (1992)).
33. 96 P.3d 291 (N.M. 2004).
34. *Id.* at 299 (emphasis added).
35. 365 F.3d 215 (3d Cir. 2004).
36. *Id.* at 238; *see also* United States v. John, 597 F.3d 263, 275 (5th Cir. 2010) ("A number of circuits have determined that this 'sliding-scale' procedure [for deciding whether two fingerprints come from the same finger] is testable . . . ."); United States v. Love, No. 10cr2418–MMM, 2011 WL 2173644, at *3 (S.D. Cal. June 1, 2011) ("The fact that latent fingerprint analysis can be tested for reliability, without more, allows the first *Daubert* 'factor to weigh in support of admissibility.'" (quoting *Mitchell*, 365 F.3d at 238)).

Second, courts have been quite willing to find the more weighty "has been tested" facet fulfilled by nonscientific forms of testing. One finds statements such as "the reliability of the technique has been tested in the adversarial system for over a century"[37] and "unquestionably the technique has been subject to testing, albeit less rigorous than a scientific ideal, in the world of criminal investigation, court proceedings, and other practical applications . . . ."[38] This "adversarial testing"[39] may be a good thing, but it is no substitute for scientific testing.[40] The fortuitous and haphazard discovery of error in the justice system surely is not "what the Supreme Court meant when it discussed testing as an admissibility factor."[41]

### *B. Peer Review and Publication*

The second *Daubert* factor is "whether the theory or technique has been subjected to peer review and publication."[42] Again, two judicial practices often have drained the substance from this consideration. First, although the best reading of *Daubert* is that this factor refers only to publication in a rigorously refereed scientific journal,[43] a surprising number of courts have used "peer review" to mean a second opinion in a given case—such as the routine review by a laboratory supervisor or a second analyst.[44] This kind of "peer review" does not address the validity of a scientific theory or method. It merely shows that two individuals applying the same methodology can reach the same conclusion. It enhances confidence that the criminalist has

37. *John*, 597 F.3d at 275.

38. United States v. Baines, 573 F.3d 979, 990 (10th Cir. 2009); *see also* United States v. Crisp, 324 F.3d 261, 266 (4th Cir. 2003) ("[F]ingerprint analysis has been tested and proven to be a reliable science over decades of use for judicial purposes." (quoting United States v. Joseph, No. CR. A. 99-238, 2001 WL 515213, at *1 (E.D. La. May 14, 2001))). For discussion of why "longstanding use establishes something, [but] it establishes less than its advocates suggest," see Jennifer L. Mnookin et al., *The Need for a Research Culture in the Forensic Sciences*, 58 UCLA L. REV. 725, 748 (2011).

39. United States v. Havvard, 260 F.3d 597, 599 (7th Cir. 2001) (quoting United States v. Havvard, 117 F. Supp. 2d 848, 854 (S.D. Ind. 2000)).

40. Practical applications are relevant to general acceptance, but what could be more practical than the treatment of life and death diseases—a practice that is littered with the bodies of therapies shown by controlled experiments to have been worthless or unnecessary? *See* KAYE ET AL., *supra* note 30, § 8.7.2. On the absence of valid expertise in a variety of practical domains in which expert advice is commonly sought, see generally DAVID H. FREEDMAN, WRONG: WHY EXPERTS* KEEP FAILING US—AND HOW TO KNOW WHEN NOT TO TRUST THEM (2010).

41. United States v. Sullivan, 246 F. Supp. 2d 700, 704 (E.D. Ky. 2003) (quoting United States v. Llera Plaza, No. 98-362-10, 2002 WL 27305, at *10 (E.D. Pa. Jan. 7, 2002)). In an opinion vacating the one quoted above, Judge Louis Pollak remained unimpressed with the government's arguments about "adversarial testing." In that opinion, he again expressed his disagreement "with [the] contention[] that . . . a century of litigation has been a form of 'adversarial' testing that meets *Daubert*'s criteria" and "concluded . . . that *Daubert*'s testing factor was not met . . . ." United States v. Llera Plaza, 188 F. Supp. 2d 549, 564 (E.D. Pa. 2002).

42. Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 593 (1993).

43. *See* KAYE ET AL., *supra* note 30, § 7.3.2(b)(2).

44. For opinions of the Third, Fourth, Seventh, and Ninth Circuits substituting this quality-control measure for the "peer review" preferred in *Daubert*, see *id.* § 7.3.2(b)(4); *id.* § 7.6.3(b) (2d ed. Supp. 2016).

performed the assigned task carefully, but the review of a forensic test by a second examiner does not satisfy the concern with the validity of a theory or method that motivated the phrase "peer review" in *Daubert*.[45] The point of the discussion of peer review in every opinion in *Daubert*, starting with the district court and culminating in the Supreme Court, was to ensure that scientific theories and methods are scrutinized in the scientific community before they are used in the courtroom. The number of forensic examiners participating in a particular procedure is irrelevant to this concern.

Second, there is a tendency to count publications of all stripes as indicia of scientific knowledge. Publications in the *Journal of Clinical Ecology* with an editorial board of believers in this fringe theory would or should be given little credence in a toxic tort case.[46] The same result should occur for forensic-science publications that are not readily accessible to research scientists and whose editors and referees lack broad expertise in statistics and empirical research methods. Otherwise, the publications become comparable to talk within congregations of true believers and bear little resemblance to the desired scientific practice of critical review and debate mentioned in *Daubert*.[47] Yet, assurances that methods have been discussed in practitioner journals have been accepted without further inquiry.[48]

### *C. Controlling Standards*

The *Daubert* list of factors also includes "the existence and maintenance of standards controlling the technique's operation."[49] Here too, courts have been overinclusive in applying the indicator of validity. *Daubert* referred only to standards for making measurements or drawing inferences. It cited to *United States v. Williams*[50] as "noting [a] professional organization's standard governing spectrographic analysis."[51] In *Williams*, the Second Circuit referred to a rule that ten matching features must be found in voice spectra "before a positive identification can be made."[52] Rules like these, which control analyst discretion, enhance reliability within and across examiners. Such repeatability and reproducibility, as these two types of

45. United States v. Baines, 573 F.3d 979, 991 (10th Cir. 2009); *Sullivan*, 246 F. Supp. 2d at 703; *Llera Plaza*, 2002 WL 27305, at *10, *vacated*, 188 F. Supp. 2d 549 (E.D. Pa. 2002).

46. *See* Sterling v. Velsicol Chem. Corp., 855 F. 2d 1188, 1208–09 (6th Cir. 1988); Bert Black, *A Unified Theory of Scientific Evidence*, 56 FORDHAM L. REV. 595, 689 (1988).

47. Thus, one of the first documents approved by the National Commission on Forensic Science was an expression of views on what can be deemed part of the requirements of "scientific literature." Nat'l Comm'n on Forensic Sci., *Scientific Literature in Support of Forensic Science and Practice*, U.S. DEP'T JUST. (Jan. 30, 2015), http://www.justice.gov/sites/default/files/ncfs/pages/attachments/2015/02/25/scientific_literature_views_document_as_adopted_1_30_15.pdf [https://perma.cc/GJQ7-WJ8M].

48. *E.g.*, United States v. Ashburn, 88 F. Supp. 3d 239, 246 (E.D.N.Y. 2015) (relying on articles in the *AFTE Journal*). For discussion of the nature of the journal, see KAYE ET AL., *supra* note 30, § 7.6.3(b) (2d ed. Supp. 2016); Mnookin et al., *supra* note 38, at 754–58.

49. Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 594 (1993).

50. 583 F.2d 1194 (2d Cir. 1978).

51. *Daubert*, 509 U.S. at 594.

52. *Williams*, 583 F.2d at 1198.

reliability are sometimes called,[53] affect the validity of a procedure by making the outcomes less erratic.[54]

Many of the identification methods in common use are devoid of such controlling standards. Instead, published standards contain circular or vacuous statements about the extent to which two samples must display similarities for a criminalist to conclude that they are (or simply could be) from the same source. An example is the *Standard Guide for Forensic Paint Analysis and Comparison* that "describes methods to develop discriminatory information."[55] To discriminate is "to distinguish between two samples based on significant differences."[56] A "*significant difference*" is "a difference between two samples that indicates that the two samples do not have a common origin."[57] Round and round we go. A controlling standard would prescribe when a difference is "significant" and how "significant" it is in including or excluding possible sources.

Some courts seem to recognize that some "standards" do nothing to confine discretion,[58] but others are impressed with such unedifying directives as "7.12.5 Evaluate the similarities, differences, and limitations. Determine their significance individually and in combination" and "7.13 Form a conclusion based on results of the above analyses, comparisons, and evaluations."[59] Indeed, the Fourth Circuit perceived controlling standards in the fact that examiners look at the same features (to make highly discretionary judgments) and undergo proficiency tests of these judgments.[60]

Such practices are desirable, but they do not constitute "controlling standards" for the evaluation of similarities and differences within the

---

53. *E.g.*, JOINT COMM. FOR GUIDES IN METROLOGY, INTERNATIONAL VOCABULARY OF METROLOGY—BASIC AND GENERAL CONCEPTS AND ASSOCIATED TERMS 38–42, 98, 104–05 (3d ed. 2008), https://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2008.pdf [https://perma.cc/S57W-DQJR].

54. Of course, the standardized outcomes could all be wrong, in which case the procedure would be invalid. Consistency is not validity, but it is a necessary precondition for validity. For a report questioning the validity of using voice spectrograms to identify speakers, see COMM. ON EVALUATION OF SOUND SPECTROGRAMS, NAT'L RESEARCH COUNCIL, ON THE THEORY AND PRACTICE OF VOICE IDENTIFICATION 49 (1979).

55. AM. SOC'Y FOR TESTING & MATERIALS, STANDARD E1610-17: STANDARD GUIDE FOR FORENSIC PAINT ANALYSIS AND COMPARISON § 1.2 (2017).

56. *Id.* § 3.2.4.

57. *Id.* § 3.2.10.

58. *E.g.*, United States v. Mitchell, 365 F.3d 215, 241 (3d Cir. 2004); United States v. Johnsted, 30 F. Supp. 3d 814, 819 (W.D. Wisc. 2013); *cf.* United States v. Baines, 573 F.3d 979, 991 (10th Cir. 2009) ("[S]earching this record for evidence of standards that guide and limit the analyst in exercise of these subjective judgments, we find very little.").

59. *See* AM. SOC'Y FOR TESTING & MATERIALS, STANDARD E2290-07A: STANDARD GUIDE FOR EXAMINATION OF HANDWRITTEN ITEMS (2007); *see also* Pettus v. United States, 37 A.3d 213, 224–25 (D.C. 2012) (relying on the fact that "FBI document examiners . . . are trained according to and employ national standards recommended by ASTM International . . . and at each step look for multiple handwriting characteristics that conform to standards recognized by ASTM International and published in recognized questioned document texts."). ASTM withdrew Standard E2290-07a in 2016. *See ASTM E2290-07a*, ASTM INT'L, https://www.astm.org/Standards/E2290.htm [https://perma.cc/6TFG-LUCB] (last visited Feb. 14, 2018).

60. United States v. Crisp, 324 F.3d 261, 269 (4th Cir. 2003).

meaning of *Daubert*. Nevertheless, it is tempting for courts to refer to the mere existence of documents from standards-development organizations on different matters and quality-assurance measures as a basis for finding that the "controlling standards" factor argues for admissibility.

### *D. Error Rates*

Along with "standards controlling the technique's operation,"[61] the *Daubert* Court spoke of "the known or potential rate of error,"[62] again using spectrographic voice identification as an example. Lower courts had relied on experiments with voice exemplars (with little to no analysis of the comprehensiveness and design of the experiments) in ruling on the admissibility of the technique.[63] A scientifically rigorous way to validate claims of accuracy (under experimental conditions) is to compare analysts' judgments of the origin of pairs of exemplars as coming from the same speaker, or instead from different speakers, when the experimenters (but not the analysts) know the true state of affairs.

Many post-*Daubert* opinions do not adhere to this type of validity study in discussing error rates for identification tests. Some courts accepted the meaningless claim of "a potential error rate of zero for the method [because] any error is attributable to examiners."[64] Some opined that "the known error rate remains impressively low"[65] because the examiners do not know of many mistakes that they have made in their casework[66] and they make no mistakes on training or later proficiency tests—even though these tests "are not shown to be accurate facsimiles of the tasks undertaken by fingerprint analysts in actual cases."[67] This kind of information is encouraging, but it is far removed from the error-rate statistics cited to in *Daubert*.

---

61. Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 594 (1993) (citing United States v. Williams, 583 F.2d 1194, 1198 (2d Cir. 1978)).

62. *Id.* (citing United States v. Smith, 869 F.2d 348, 353–54 (7th Cir. 1989)).

63. The *Smith* court referred to experiments finding false-positive rates of 2.4 percent and 0.31 percent, false-negative rates of 6 percent and .53 percent, and "no errors whatsoever." *Smith*, 869 F.2d at 354. The court also noted two studies with far higher rates of 62.7 percent and 83.33 percent (presumably for false positives). *Id.* Unsurprisingly, the experiments found better performance on "closed" sets of exemplars (those in which the analyst knew that the questioned sample came from a small number of possible speakers) than in "open" sets. COMM. ON EVALUATION OF SOUND SPECTROGRAMS, *supra* note 54, at 24.

64. United States v. Mahone, 453 F.3d 68, 72 (1st Cir. 2006). The claim is meaningless because there is no inherent rate of error for the "method" that can be separated from the performance of the human analyst. Because it is a claim that cannot be falsified by any conceivable study, it is outside the realm of science.

65. United States v. Baines, 573 F.3d 979, 991 (10th Cir. 2009).

66. The court of appeals reached this conclusion on the basis of an FBI supervisor's testimony that he knew of only one error per eleven million cases, although it allowed that this estimate might be on the low side. *Id.* at 990–91.

67. *Id.* at 990.

### *E. Degree of Acceptance*

The final factor articulated in *Daubert* is general acceptance. Under the previously leading case of *Frye v. United States*,[68] this consideration was determinative. Under *Daubert*, "explicit identification of a relevant scientific community and an express determination of a particular degree of acceptance within that community"[69] remain important. Theories and methods that are generally accepted in the scientific community are more likely to be valid than those that are not.

Clearly, if one limits the "scientific community" to individuals who produce the challenged evidence or write textbooks and standards on how to generate such evidence, the methodology will be generally accepted. But "forensic science service provider[s]" or "forensic science practitioner[s]"[70] are not coterminous with a scientific community. The National Commission on Forensic Science regarded "forensic science" as encompassing either "scientific *or* technical practices."[71] In its view, an "individual who . . . applies scientific or technical practices to the recognition, collection, analysis, or interpretation of evidence" can be a forensic-science practitioner.[72] Traditionally, this scientific and technical community has been affected by, but not imbued with, the kind of research culture associated with other fields of science.[73] As the theories and claims of criminalists have come under scrutiny from a wider range of research scientists, it has become harder for courts to discern general agreement on these matters. The recent PCAST report is an extreme example of discordant voices in the scientific community.

In the face of disagreements about the scientific status of some methods, a number of courts have substituted general acceptance within "the expert community,"[74] "the forensic identification community,"[75] the "*Daubert* community,"[76] and "the courts"[77] for acceptance in the scientific community. The Third Circuit reasoned that the fact that "fingerprint identification is generally accepted within the forensic identification community"[78] was enough to place a checkmark for *Daubert*'s general acceptance factor on the government's scorecard.[79] Likewise, the Massachusetts Supreme Judicial Court categorically asserted that "[a] technical community, or a community

68. 293 F. 1013 (D.C. Cir. 1923).
69. Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 594 (1992).
70. These are terms that the National Commission on Forensic Science introduced. *See* Nat'l Comm'n on Forensic Sci., *Views of the Commission Defining Forensic Science and Related Terms*, U.S. DEP'T JUST. 1 (May 1, 2016), https://www.justice.gov/archives/ncfs/file/786571/download [https://perma.cc/4VVG-9E7F].
71. *Id.* (emphasis added).
72. *Id.*
73. *See generally* Mnookin et al., *supra* note 38.
74. United States v. Crisp, 324 F.3d 261, 268 (4th Cir. 2003).
75. United States v. Mitchell, 365 F.3d 215, 241 (3d Cir. 2004).
76. Commonwealth v. Patterson, 840 N.E.2d 12, 25 (Mass. 2005).
77. *Crisp*, 324 F.3d at 268.
78. *Mitchell*, 365 F.3d at 241.
79. *Id.*

of experts who have some other specialized knowledge, can qualify as a relevant *Daubert* community in the same way a scientific community can."[80]

The problem with such statements is that not every "*Daubert* community" is fungible.[81] General acceptance of relativity theory among physicists is one thing; acceptance among arson investigators of "crazed glass" as an indicator of accelerants is another.[82] Because the standards of acceptance within the expert community are crucial to gauging the significance of general acceptance, courts that apply the general-acceptance factor mechanically are missing the meaning of *Daubert*.[83]

## II. ADMITTING CRIMINALISTICS EVIDENCE FOR WHAT IT IS: OF BABIES AND BATHWATER

I have tried to lay bare how courts have deviated from *Daubert* by altering or misapplying the five factors it provided as a framework for judging scientific validity. Repeatedly, they have shied away from scrutinizing criminalistics evidence of identity as *Daubert* originally seemed to require. Beneath this doctrinal dissection lies the question of causation. Why have courts applied a weakened or mutated form of *Daubert* to this type of evidence? This is essentially a question of psychology, sociology, and political science. Commentators have pointed to such psychological and institutional factors as disparities in resources between prosecutors and defendants, deficiencies of defense counsel, a lack of knowledge or scientific competence, strong prior beliefs about validity, proprosecution attitudes, conservatism, and a conviction that evidence is useful even if the claims of "scientific knowledge" under Rule 702 are not fully validated.[84]

Understandably, a constant refrain in the opinions rejecting *Daubert* challenges to criminalistics is that even if the evidence is too imperfectly validated to wear the mantle of science, it is still valuable, and "wholesale exclusion" would be too "drastic,"[85] would impose "an extremely high degree of intellectual purity,"[86] and would "make the best the enemy of the good."[87] The doctrinal route to admitting expert evidence that does not quite

80. *Patterson*, 840 N.E.2d at 25.

81. *See* KAYE ET AL., *supra* note 30, § 6.3.3(b).

82. *See* JOHN J. LENTINI, SCIENTIFIC PROTOCOLS FOR FIRE INVESTIGATION 472 (2d ed. 2013) (discussing such "old firemen's tales" and the lack of "natural scientific skepticism" among arson investigators).

83. They also are distorting *Kumho Tire*, which allows trial courts to apply some or all of the *Daubert* factors to nonscientific expert testimony. *See* Kumho Tire Co. v. Carmichael, 526 U.S. 137, 141 (1999). *Kumho Tire* does not mean that general acceptance is equally supportive of admitting the evidence regardless of the nature of the community that accepts it.

84. *See* Stephanie L. Damon-Moore, *Trial Judges and the Forensic Science Problem*, 92 N.Y.U. L. REV. 1532, 1556 (2017); Michael J. Saks, *Explaining the Tension Between the Supreme Court's Embrace of Validity as the Touchstone of Admissibility of Expert Testimony and Lower Courts' (Seeming) Rejection of Same*, 5 EPISTEME 329, 339 (2008); Joseph Sanders, *"Utterly Ineffective": Do Courts Have a Role in Improving the Quality of Forensic Expert Testimony?*, 38 FORDHAM URB. L.J. 547, 558–59 (2010).

85. United States v. Crisp, 324 F.3d 261, 268 (4th Cir. 2003).

86. United States v. Baines, 573 F.3d 979, 989 (10th Cir. 2009).

87. United States v. Llera Plaza, 188 F. Supp. 2d 549, 572 (E.D. Pa. 2002).

meet the requirements expected of scientific evidence is to call it nonscientific. With that label, the evidence need not constitute the "scientific knowledge" sought in *Daubert*. Instead, "the court may admit the testimony as non-scientific expert testimony under Rule 702 and *Kumho Tire*."[88] Thus, District Judge Louis Pollak, who famously ruled that latent print examiners could not make source attributions[89]—and then reversed his ruling[90]—ultimately concluded that examiners were not engaged in "a science"[91] but were "like accountants, vocational experts, accident-reconstruction experts, appraisers of land or of art, [or] experts in tire failure analysis."[92] Similarly, state courts have short-circuited the special scrutiny normally given to scientific evidence by characterizing some pattern comparisons as nonscientific or within the jury's grasp.[93]

If criminalistics evidence is to pass muster on these grounds, the theory must be followed to its logical conclusion. The trial court must ensure that experts do not "wrap themselves in a scientist cloak."[94] This will be quite difficult if the witnesses are called forensic scientists, if they have employed esoteric scientific devices in their analyses, or if the court has designated them as experts in front of the jury.[95] They (or the court) would have to call attention to the parts of their testimony that cannot be said to rest on adequate validity studies, cautioning jurors that those parts do not constitute scientific knowledge.

---

88. Restivo v. Hessemann, 846 F.3d 547, 577 (2d Cir. 2017). *Kumho Tire* involved an engineer's poorly validated testimony about the cause of a fatal tire failure based on a visual inspection. *Kumho Tire*, 526 U.S. at 145–47. There, the Supreme Court explained that the overarching requirement of extra "reliability" read into Rule 702 in *Daubert* (and later codified in an amendment to the Rule) applies to all expert testimony, and it held that trial courts may rely on whatever *Daubert* factors would be helpful in gauging the "reliability" of that testimony. *See id.* at 148–49.

89. *See* United States v. Llera Plaza, No. 98-362-10, 2002 WL 27305, at *10 (E.D. Pa. Jan. 7, 2002), *vacated*, 188 F. Supp. 2d 549 (E.D. Pa. 2002).

90. *Llera Plaza*, 188 F. Supp. 2d at 549.

91. *Id.* at 560.

92. *Id.* at 563. Judge Posner's opinion in *United States v. Herrera* is similar in deeming latent print examinations to be admissible as nonscientific "expert evidence . . . on the style of a particular artist would be [admissible as] the expert's opinion, based on comparison with other paintings, of the genuineness of the painting alleged to be a forgery." United States v. Herrera, 704 F.3d 480, 486 (7th Cir. 2013).

93. *See* KAYE ET AL., *supra* note 30, § 8.6 (questioning this reasoning as applied to examinations of bite marks, hair, handwriting, and toolmarks); *id.* § 8.9.4 (2d ed. Supp. 2016) (questioning this reasoning as applied to adhesive tape tears).

94. Sanders, *supra* note 84, at 557. An example of such cloaking is the Scientific Working Group for Firearms and Toolmarks's (SWGGUN) reliance on "Richard Feynman's writings on the character of physical laws for guidance in articulating the critical scientific method elements used in the discovery, origination, and evolution of firearm and toolmark identification." *See* SCI. WORKING GRP. FOR FIREARMS & TOOLMARKS, THE FOUNDATIONS OF FIREARM AND TOOLMARK IDENTIFICATION 1 (2013) (footnote omitted), https://www.nist.gov/sites/default/files/documents/2016/11/28/swggun_foundational_report.pdf [https://perma.cc/R9BZ-FVH4].

95. The last practice is inadvisable. Nat'l Comm'n on Forensic Sci., *Views of the Commission: Judicial Vouching*, U.S. DEP'T JUST. (June 21, 2016), https://www.justice.gov/archives/ncfs/file/880246/download [https://perma.cc/X8KS-YTAF].

It is worth noting that the concern here is not connected to any abstract philosophical analysis of the distinction between science and other forms of knowledge. Even if there is no epistemologically distinctive "scientific method,"[96] invoking a "scientific" basis for a conclusion has special rhetorical and persuasive power.[97] Consequently, "it is the Court's role to ensure that a given discipline does not falsely lay claim to the mantle of science, cloaking itself with the aura of unassailability that the imprimatur of 'science' confers and thereby distorting the truth-finding process."[98] Attending to this concern enables courts to accomplish what the Supreme Court in *Kumho Tire* cursorily dismissed as impractical and unnecessary—namely, distinguishing "between 'scientific' knowledge and 'technical' or 'other specialized' knowledge."[99]

For much of the last century, courts distinguished between scientific and other forms of expert testimony, demanding more of the former than the latter. Generally, they were successful in demarcating the type of expert testimony that needed heightened scrutiny. A version of Rule 702 that explicitly insists on more rigorous validation of evidence that is promoted or understood as being "scientific" would be workable and more clearly compatible with the rule's common law roots.

## III. DEFINING RELIABILITY AND VALIDITY

The PCAST report has reinvigorated debate on the extent to which certain fields of criminalistics are based on scientific knowledge as opposed to less impressive foundations.[100] Unlike the 2009 NRC committee, which avoided overt legal conclusions, PCAST presents an analysis of the implications of its findings for the admissibility of several types of evidence and creates a somewhat neoteric vocabulary (compared to conventional usage in statistics) to map its criteria for "validity" onto Rule 702.

Because any revision or advice on Rule 702 should carefully consider which terms to use, it may be helpful to note the specialized meanings of the

---

96. *See, e.g.*, SUSAN HAACK, DEFENDING SCIENCE—WITHIN REASON: BETWEEN SCIENTISM AND CYNICISM 94–95 (2003); *cf.* SCI. WORKING GRP. FOR FIREARMS & TOOLMARKS, *supra* note 95, at 1–2 (noting that "1) observation of a phenomenon, 2) developing a premise, forming a hypothesis, 3) develop a testing model, 4) using reliable methodology, and 5) forming a theory" constitute the "scientific method elements consistent to those rudiments described by Feynman").

97. HAACK, *supra* note 96, at 18. According to Haack,

> "Scientific" has become an all-purpose term of epistemic praise meaning "strong, reliable, *good*." No wonder, then, that psychologists and sociologists and economists are sometimes so zealous in insisting on their right to the title. No wonder, either, that practitioners in other areas—"Management Science," "Library Science," "Military Science," even "Mortuary Science"—are so keen to claim it.

*Id.*

98. Almeciga v. Ctr. for Investigative Reporting, Inc., 185 F. Supp. 3d 401, 415 (S.D.N.Y. 2016); *see also* KAYE ET AL., *supra* note 30, § 6.2, at 247 n.8 (referring to the many cases recognizing this concern).

99. Kuhmo Tire Co. v. Carmichael, 526 U.S. 137, 148 (1999). For an extended analysis and a proposed solution to this "boundary problem," see KAYE ET AL., *supra* note 30, ch. 7.

100. *See* KAYE ET AL., *supra* note 30, § 15.7.5 (2d ed. Supp. 2018). Parts of this discussion are drawn from this work.

terms "validity" and "reliability" in science and law. Confusion can arise because in science "reliability" and "validity" are not synonyms, and "reliability" in the law of evidence does not mean scientific or statistical reliability.[101]

### *A. Legal Reliability*

For better or worse, the *Daubert* Court chose to use the word "reliability" to mean "trustworthiness." Justice Blackmun explained that:

> [S]cientists typically distinguish between "validity" (does the principle support what it purports to show?) and "reliability" (does application of the principle produce consistent results?). . . . [O]ur reference here is to *evidentiary* reliability—that is, trustworthiness. In a case involving scientific evidence, *evidentiary reliability* will be based upon *scientific validity*.[102]

Federal Rule of Evidence Rule 702 goes beyond the validity spoken of in *Daubert* for principles or methods and uses the term "reliable" in its legal sense to encompass the application of those principles or methods in a particular case. It requires not only that "the testimony is the product of reliable principles and methods"—as required in *Daubert*—but also that "the expert has reliably applied the principles and methods to the facts of the case." This amended version of the original Rule codifies the dubious conflation of method and conclusion adopted in *General Electric Co. v. Joiner*.[103]

### *B. Scientific Reliability of a Measurement Procedure*

In most scientific fields, "reliability," as Justice Blackmun acknowledged, pertains to consistency. A rigid ruler is a reliable measuring device (when used properly). It gives the same measurements for the length of a straight line when used repeatedly. An elastic ruler would produce more variable measurements of the same line. In *Hall v. Florida*,[104] the Supreme Court used "reliability" in this statistical sense in its discussion of the reliability of IQ scores.[105] Test developers use clever methods to measure reliability,[106] and no standardized test would be marketed without an estimate of its reliability.

---

101. *See id.* § 12.7 (2d ed. 2010); *id.* § 15.7.5 (2d ed. Supp. 2018).

102. Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 590 n.9 (1993) (using "trustworthiness" in the sense of "reliable sources of information" (quoting FED. R. EVID. 602 advisory committee's notes on proposed rules)).

103. 522 U.S. 136 (1997). For a criticism of *Joiner*, see KAYE ET AL., *supra* note 30, ch. 9.

104. 134 S. Ct. 1986 (2014).

105. *Id.* at 1994.

106. *See generally* David H. Kaye, *Deadly Statistics: Quantifying an 'Unacceptable Risk' in Capital Punishment*, 16 L. PROBABILITY & RISK 7, 7–8 (2017).

### *C. Scientific Validity of a Measurement Procedure*

Reliability is not validity. If the markings on the rigid ruler were too wide apart, measurements made with it would consistently understate true length. Its use would be reliable but not valid. The highly elastic ruler would be neither reliable nor valid for measuring length. The Law School Admissions Test has been validated as a predictor of first-year law school grades. It is not perfect, but it is considerably better than guessing (or predicting that everyone will receive the average grade). It is less valid as a predictor of success in law practice, but the scores are the same (and equally reliable) for either use. The polygraph reliably and validly measures physiological variables such as respiration rate. It is less valid (but no less reliable) for measuring conscious deception. As these examples indicate, scientific validity of a measurement procedure involves its accuracy for a specified use.

The PCAST report redefines scientific validity to fit both parts of the legal mold of Rule 702.[107] For PCAST (unlike *Daubert*), there are at least "two types of scientific validity"[108]—"foundational validity" corresponding to the legal requirement in Rule 702(c) of "reliable principles and methods"[109] and "*validity as applied* mean[ing] that the method has been [correctly] applied *in practice*."[110] Translating these legal-scientific terms back into the standard scientific ones, we could say that "foundational validity" refers to the validity of a measurement procedure, and "validity as applied" refers to a flawed application of the valid measuring system. For example, if an instrument is shown to validly (i.e., accurately) measure breath ethanol concentration when properly calibrated, it has "foundational validity" for the purpose of measuring breath alcohol. It can do what it is supposed to do. If it is not correctly calibrated, however, it cannot be trusted to do the job—at least, not as accurately as expected.[111]

The PCAST report's requirements for showing "foundational validity" of largely subjective feature-comparison methods rest on the assumption that examiners will provide categorical conclusions about the true source. In response to such testimony, PCAST promulgated the following categorical, quantitative rule for validity:

> Methods with a high FPR (false positive rate) are scientifically unreliable for making important judgments in court about the source of a sample. To be considered reliable, the FPR should certainly be less than 5 percent and

107. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 15, at 142.

108. *Id.* at 4. Another type of scientific validity—"external validity"—also is part of *Daubert* and Rule 702's "reliability." *See* KAYE ET AL., *supra* note 30, § 12.5.4. This type of validity refers to generalizability or projectability of findings from specific studies (such as the "black box" experiments discussed in the PCAST report) to actual casework. The report cautions that experiments must involve "known and representative samples from each relevant population." PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 15, at 152.

109. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 15, at 4–5.

110. *Id.* at 5.

111. It is not obvious what is gained by denominating the proper application of a valid system for measuring something as a form of "validity."

> it may be appropriate that it be considerably lower, depending on the intended application.[112]

The meaning of "scientifically unreliable" is slippery at best. The conventional choice of the 0.05 level for a statistical test may or may not be appropriate here, but reliability and validity are not binary quantities. As the passage (and the earlier examples) indicate, reliability and validity come in degrees.[113] Demanding a low rate of false positives will increase the rate of false negatives. At first blush, it might seem that the threshold for validity (and hence admissibility) should be very high in a criminal case since the legal system regards false convictions as worse than false acquittals. But the law certainly does not require that each piece of evidence prove guilt beyond a reasonable doubt. That standard applies to the totality of the evidence. If the scientific brick in the wall of evidence adds some structural integrity, it normally can be inserted.[114]

Thus, one can deny that a false-positive probability below 0.05 is essential to validity.[115] A method that generates probative evidence can still be scientifically valid for its intended use.[116] The use is to inform the factfinder of a possible association to the trace material so that the judge or jury will use the information to make a better decision.[117] And if that is to happen, the fact finder must know about something else emphasized in the 2009 NRC report and the newer PCAST report—the uncertainty in the findings. Whether dictated by the scientific-validity requirement of Rule 702,[118] by

---

112. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 15, at 151–52.

113. *Cf.* Dale A. Nance, *Two Concepts of Reliability*, 5 J. PHIL. SCI. & L. 1, 4 (2005) (making this point about "evidentiary reliability").

114. *See* United States v. Herrera, 704 F.3d 480, 486 (7th Cir. 2013); Nance, *supra* note 113, at 4 ("'Sufficiently reliable *to be considered*' is not the same as 'sufficiently reliable *to warrant a verdict*.'"). I am barely scratching the surface of these issues. For more analysis of the PCAST 0.05 level for validity and false-positive probabilities in statistical hypothesis testing, see generally David H. Kaye, *Hypothesis Testing in Law and Forensic Science: A Memorandum*, 130 HARV. L. REV. F. 127 (2017); David H. Kaye, *The Source and Soundness of PCAST's 5% Rule*, FORENSIC SCI. STAT. & L. (July 23, 2017, 9:58 PM), http://for-sci-law.blogspot.com/2017/07/the-source-and-soundness-of-pcasts-5.html [https://perma.cc/YU8T-HGQC].

115. KAYE ET AL., *supra* note 30, § 15.7.5(c) (2d ed. Supp. 2018). As noted elsewhere,

> If judgments can be shown to be somewhat informative, and the modest degree to which they are discriminating can be explained, does it follow that they are inadmissible on grounds of scientific validity—or is this a legal judgment under Rule 403? PCAST seems to regard validity as a purely scientific question under Rule 702—scientists inform the courts of scientific validity. But a case can be made for regarding slightly probative results as satisfying Rule 702(c)'s "reliability" requirement, and then considering whether the somewhat discriminating evidence can be presented and used for what it is worth.

*Id.*

116. *See, e.g.*, David H. Kaye, *Ultracrepidarianism in Forensic Science: The Hair Evidence Debacle*, 72 WASH. & LEE L. REV. ONLINE 227, 252–53 (2015).

117. *See Herrera*, 704 F.3d at 486–87.

118. The report refers to exaggerated expressions for either probative value or the probability of a conclusion from measurements that come from a valid-as-applied and foundationally valid method as "not scientifically valid." PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 15, at 145.

Rule 403, which warrants exclusion to avoid unfairly prejudicial evidence,[119] or by scientific and prosecutorial ethical norms, purportedly scientific evidence should be conveyed in a manner that reduces the risk of its being grossly overvalued. The next Part therefore sketches two different ways to express uncertainty in forensic-science findings.

## IV. EXPRESSING UNCERTAINTY

Empirical science—indeed, all ampliative reasoning—can achieve only degrees of certainty, and "probability is the logic of uncertainty."[120] Broadly speaking, two approaches to presenting the results of comparisons of traces to known samples are in use.[121] Traditionally (and overwhelmingly in the United States), criminalists speak to the probability of possible conclusions about the source of the trace. An alternative approach that dominates the academic literature on forensic inference[122] requires criminalists to openly and transparently address the probability of the measured or observed similarities under different theories of the origin of the samples.[123] No revision to Rule 702 should foreclose the second method of interpreting the data.

### *A. Conclusions About Hypotheses*

The traditional mode of testimony supplies opinions as to the probability that a hypothesis about the source of a trace is true. For example, the Association of Firearm and Tool Mark Examiners (AFTE) encourages "opinions of common origin to be made when the unique surface contours of two toolmarks are in 'sufficient agreement.'"[124] "Sufficient agreement" is a "subjective" judgment "based on the examiner's training and experience" that the toolmarks are more similar than ones the examiner remembers as having "been produced by different tools."[125] It "means that . . . the likelihood another tool could have made the mark is so remote as to be considered a practical impossibility."[126] If this explanation of a reported association between a cartridge case and a known gun is presented in court, the examiner is stating (1) that the patterns being compared are remarkably similar and (2) that the probability that any other tool made the mark is close

119. *See* FED. R. EVID. 403.

120. JOSEPH K. BLITZSTEIN & JESSICA HWANG, INTRODUCTION TO PROBABILITY 1 (2015).

121. *See, e.g.*, KAYE ET AL., *supra* note 30, chs. 13–15. *See generally* BERNARD ROBERTSON ET AL., INTERPRETING EVIDENCE: EVALUATING FORENSIC SCIENCE IN THE COURTROOM (2d ed. 2016).

122. INMAN & RUDIN, *supra* note 1, at 170; Geoffrey S. Morrison et al., *A Comment on the PCAST Report: Skip the "Match"/"Non-Match" Stage*, 272 FORENSIC SCI. INT'L, at e7, e7–e8 (2017).

123. *See, e.g.*, I.W. Evett et al., *Finding the Way Forward for Forensic Science in the U.S.—A Commentary on the PCAST Report*, 278 FORENSIC SCI. INT'L 16, 23 (2017).

124. *AFTE Theory of Identification as It Relates to Toolmarks*, ASS'N FIREARM & TOOL MARK EXAMINERS, https://afte.org/about-us/what-is-afte/afte-theory-of-identification [https://perma.cc/UVP3-N555] (last visited Feb. 14, 2018).

125. *Id.*

126. *Id.* This is said to occur when the examiner does not recall encountering the same degree of similarity in marks known to have come from the same source. *Id.*

to 0. Statement (1) is an opinion about the features, while (2) is a statement about the probability of the hypothesis that a particular gun is the source of the mark. That probability, in the examiner's mind, is practically 1. Most courts would allow the examiner to testify to the categorical conclusion that bullets came from the same gun.[127]

PCAST maintains that although one well-designed experiment shows a sufficiently small rate of errors for subjective same-source classifications for bullets from one make of gun, that experiment (together with studies of other, less rigorous designs) is not enough to meet the criteria for scientific validity. Perhaps recognizing that not all courts will follow those demanding criteria, however, the report urges courts allowing source attributions to require that they be accompanied by statements of the upper limit of the one-sided 95 percent confidence interval on the false-positive error rate from the study.[128] The details of the proposal to use the false-positive error rate from controlled experiments to estimate errors are debatable. Moreover, there are better statistical measures of probative value. Even so, if the current practice of making source attributions is to continue, supplying jurors with some objectively determined estimate of the accuracy achieved by examiners as tested in cases like the one at bar is vital.[129]

### *B. Evaluations of Support for Competing Hypotheses*

Opining on the truth or falsity of a source hypothesis is not the only way to present subjective findings from feature comparisons. Instead of somehow judging the probability of a source attribution given the similarity in the features, the criminalist can describe the degree to which the comparison supports the source attribution as opposed to the extent to which it supports an inference to some other source.[130] He or she can do this by estimating the probability of observing the measured similarities when the source hypothesis is true and when it is false. If the probabilities are the same, the evidence has no probative value. If the perceived degree of similarity occurs as often when one item is the source as when another item is, then the similarities do not help us choose between these possibilities. But if the degree of similarity is more common when examining traces from the same

127. *E.g.*, Hinton v. Alabama, 134 S. Ct. 1081, 1089–90 (2014). A small number of federal district courts have limited the degree of certainty that the examiner can express in the truth of the hypothesis. KAYE ET AL., *supra* note 30, § 15.2.4. *See generally* David H. Kaye, *Firearm-Mark Evidence: Looking Back and Looking Ahead*, 68 CASE W. RES. L. REV. (forthcoming 2018).

128. The report's treatment of confidence intervals is problematic in relatively small ways. David H. Kaye, *PCAST's Sampling Errors*, FORENSIC SCI. STAT. & L. (Oct. 24, 2016), http://for-sci-law.blogspot.com/2016/10/pcasts-sampling-errors.html [https://perma.cc/25QQ-SA2J]; David H. Kaye, *PCAST's Sampling Errors (Part II: Getting More Technical)*, FORENSIC SCI. STAT. & L. (Dec. 11, 2016, 3:15 PM), http://for-sci-law.blogspot.com/2016/12/pcasts-sampling-errors-part-ii-getting.html [https://perma.cc/6F4J-KX44].

129. *See* Kaye, *supra* note 127.

130. In speaking of "some other source," I am glossing over important subtleties in the interest of simplicity and brevity.

source than when encountering traces from different sources, then the similarity is probative evidence.[131]

Thus, what determines relative support is the ratio of the probability of the similarity given that the trace came from the same source to the probability given that it came from a different source. To be more concrete, suppose that the observed level of similarity in the bullet cartridges that might have come from the defendant's gun occurs ten times more often for same-gun bullets than for different-gun bullets. Then even if the false-positive error probability exceeds 0.05, the observed similarity is somewhat probative of the cartridge recovered from the scene of the shooting having been in the defendant's gun.[132]

The ratio I have described is known as a likelihood ratio. It is a more complete measure of probative value than is a false-positive probability. But there are questions about whether a lay fact finder will correctly use either a categorical conclusion accompanied by a false-positive probability or a likelihood ratio to give the evidence the weight it deserves.[133] For example, counterintuitively—and contrary to what some courts have written—a false-positive probability is *not* the probability that the positive report on the source is false.[134] Similarly, the likelihood ratio cannot generally be equated to the odds in favor of the source hypothesis.[135] Nonetheless, the state of research into misuse of expressions of uncertainty is still primitive, and there may be presentations that would reduce the risk of misunderstanding. At this point, it would be premature to assume that it is better to exclude moderately probative criminalistics evidence because it might be misunderstood.

Of course, the fact that the likelihood ratio is a more conceptually complete measure of the probative value of evidence does not remove the need to show that criminalists relying on heavily subjective impressions[136] will provide accurate statements of the magnitude of evidentiary support.[137] If the

131. *E.g.*, KAYE ET AL., *supra* note 30, § 14.2; David H. Kaye, *Digging into the Foundations of Evidence Law*, 115 MICH. L. REV. 915, 923–25 (2017).

132. A Bayesian statistician would say that it changes the odds in favor of the defendant's gun from whatever they were before considering the evidence to ten times those odds. *See, e.g.*, KAYE ET AL., *supra* note 30, § 14.3.1; Kaye, *supra* note 131, at 925–27.

133. In addition, the witness who presents it is not forced into making a somewhat arbitrary match/no-match declaration. Morrison et al., *supra* note 122, at e7–e8. For an example of such testimony, see David H. Kaye, *Likelihoodism, Bayesianism, and a Pair of Shoes*, 53 JURIMETRICS J. 1, 1–3 (2012).

134. David H. Kaye, *The False-Positive Fallacy in the First Opinion to Discuss the PCAST Report*, FORENSIC SCI. STAT. & L. (Nov. 3, 2016, 11:10 AM), http://for-sci-law.blogspot.com/2016/11/the-false-positive-fallacy-in-first.html [https://perma.cc/U2C4-NVT9].

135. KAYE ET AL., *supra* note 30, § 14.2.2; *id.* § 14.5.2(a) (2d ed. Supp. 2016).

136. The impressions will be grounded in training and experience, but the concern remains that they are not informed by objective, applicable, and quantified data on the variability of the features in samples from a common source as opposed to those from two different sources. In time, for some evidence types, we can expect criminalists testifying in the likelihood-ratio format (whether qualitative or quantitative) to be assisted by computer programs and databases and to move beyond "the expression of personal, egocentric and generally badly articulated opinions." Christophe Champod, *Fingerprint Identification: Advances Since the 2009 National Research Council Report*, PHIL. TRANSACTIONS B, Aug. 5, 2015, at 1, 7.

137. *Cf.* David V. Budescu & Timothy R. Johnson, *A Model-Based Approach for the Analysis of the Calibration of Probability Judgments*, 6 JUDGMENT & DECISION MAKING 857,

probabilities that determine the likelihood ratio are impressionistic rather than data driven, will examiners actually report that the similarities are the kind that arise more often for same-source traces when they are in fact from the same source—but not when they are from different sources? Just using likelihood ratios to express probative value does not eliminate the concern about unvalidated, subjective judgments. As with testing examiners who make subjective source attributions and exclusions, performance studies of examiners who present highly subjective ratios are critically important.[138]

## CONCLUSION

Federal Rule of Evidence 702 has not performed well in regulating the admission of putatively scientific identification methods for associating traces with their possible sources. The original rule was just a shell that referred to expert knowledge of various types. The Supreme Court stepped in repeatedly to infuse the rule with more content, but for several fields of criminalistics the lower courts dodged the bullet that *Daubert* might have been.

Forensic science has grown stronger over the years, partly in response to criticism from scientists and lawyers alike.[139] But much remains to be done, and prominent reports from the National Academies and the President's Council of Advisors on Science and Technology have focused attention on the largely subjective source conclusions of criminalists. Most of these efforts to identify traces are not "junk science"—they can produce at least modestly helpful evidence.[140] But all have been oversold in the courtroom, and the current mode of source attribution, as opposed to expressions of evidentiary support, is not optimal.

Modifications to or authoritative advice on Rule 702 could be useful to encourage courts to look more critically at claims like the following: "it could be falsified," "there are many publications," "a second examiner is a form of peer review," "there are many standards," and "acceptance in a '*Daubert* community' is good enough." A revised rule could explicitly insist on more rigorous validation of evidence that is promoted as "scientific" and try to ensure that evidence deemed admissible as other "specialized knowledge" is descientized. It could sharpen the definitions of terms like "valid" and "reliable." It could remain flexible enough to allow the admission of scientific or experiential evidence that has probative value when

857 (2011) ("A specific property of the probability judgments—their *calibration*—has been accepted as the 'common standard of validity' in the empirical literature."). *But see generally* Franco Taroni et al., *Dismissal of the Illusion of Uncertainty in the Assessment of a Likelihood Ratio*, 15 L. PROBABILITY & RISK 1 (2016).

138. I have taken some liberties with the terms "objective" and "subjective" as applied to probabilities. Some statisticians maintain that all probabilities are ultimately subjective or personal. *E.g.*, DENNIS V. LINDLEY, UNDERSTANDING UNCERTAINTY 37–38 (2007). But some personal probabilities, and hence some personal likelihood ratios, are more reasonable than others. Judgments informed by systematically collected data, in a manner that can be clearly articulated, have a stronger claim to be interpersonally acceptable.

139. *See generally* DAVID H. KAYE, THE DOUBLE HELIX AND THE LAW OF EVIDENCE (2010).

140. *See* Kaye, *supra* note 116, at 235 n.31.

that value is reasonably estimated and presented to the fact finder. It could replace source attributions with expert evaluations of whether and how much support the observations provide for those attributions. No doubt, courts *could* apply Rule 702 (and Rule 403) as currently phrased to do all these things. But they have not done so, and judicial inertia is hard to overcome.