Fordham Law Review

Volume 86 | Issue 4 Article 8

2018

Fixing Rule 702: The PCAST Report and Steps to Ensure the Reliability of Forensic Feature-Comparison Methods in the Criminal Courts

Eric S. Lander
Broad Institute of MIT and Harvard

Fixing Rule 702: The PCAST Report and Steps to Ensure the Reliability of Forensic Feature-Comparison Methods in the Criminal Courts

Erratum

Law; Criminal Law; Evidence; Courts; Judges

FIXING RULE 702: THE PCAST REPORT AND STEPS TO ENSURE THE RELIABILITY OF FORENSIC FEATURE-COMPARISON METHODS IN THE CRIMINAL COURTS

*Eric S. Lander**

INTRODUCTION

Within the Federal Rules of Evidence, Rule 702 marks the crossroads of law and science. For the most part, courts hear testimony about ordinary factual matters, which the triers of fact can evaluate based on common knowledge and experience (e.g., "the attacker had light brown hair"). But, the law recognizes that its search for truth may sometimes be aided by hearing the conclusions of experts with specialized scientific knowledge (e.g., "the hair found at the scene of the crime was microscopically indistinguishable from the defendant's hair with respect to seven specific parameters, and scientific studies show that this degree of similarity would be seen for only roughly 1 person in 10,000 in the population").

There is an obvious risk in permitting testimony from witnesses who come cloaked in the mantle of scientific authority, purporting to possess powerful knowledge that lies beyond the ken of ordinary people. Few jurors are equipped to assess the basis of an expert's reasoning—and cross-examination is a blunt instrument for probing complex scientific claims. As a result, expert conclusions must often be taken at face value. When the conclusions are wrong, they may be highly prejudicial, outweighing other evidence or the lack thereof.

Rule 702 therefore seeks to impose a strict limitation on the admissibility of expert testimony. Courts may not simply allow expert testimony that might be relevant and "let the jury decide." Instead, Rule 702 provides that judges may permit expert testimony only if they find that "the testimony is the product of reliable principles and methods" and "the expert has reliably applied the principles and methods to the facts of the case."[1]

---

* President and Founding Director, Broad Institute of MIT and Harvard. Former Co-Chair, President's Council of Advisors on Science and Technology. 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).

1. FED. R. EVID. 702(c)–(d).

In *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,[2] the U.S. Supreme Court held that Rule 702 requires courts to serve as "gatekeepers" who must assess the underlying "reliability" of proffered expert testimony. While recognizing that the inquiry should be "flexible" (that is, tailored to the type of scientific knowledge being proffered), the meaning of "reliability" must be based on actual "scientific validity": the trial judge must determine "whether the reasoning or methodology underlying the testimony is scientifically valid";[3] "[i]n a case involving scientific evidence, *evidentiary reliability* will be based on *scientific validity*";[4] and the "overarching subject [of a judge's inquiry under Rule 702] is the scientific validity—and thus the evidentiary relevance and reliability—of the principles that underlie a proposed submission."[5] Rule 702 (as well as cognate rules in many states) thus necessitates a dialog between law and science.

Many commentators agree that, in civil litigation, Rule 702 has largely fulfilled the intended goal. By preventing juries from hearing evidence that purports to be scientific but does not actually rest on scientifically valid methods, it has acted as a quality-control filter.

In contrast, Rule 702 has largely failed in criminal law—even though quality control should be *more* important when depriving individuals of liberty than of money.[6] Various explanations have been suggested for the failure, including that the vast majority of forensic science laboratories serve only one side, the prosecution; most defendants lack the resources to mount serious challenges; and judges are reluctant to question practices that have long been used and admitted in court. Whatever the reasons, it is clear that courts have historically admitted—and continue today to admit—some forensic-science methods that fail to meet the most basic requirements of scientific validity.[7]

The risks are not merely hypothetical. Starting in the 1990s, DNA analysis revealed that many individuals convicted of crimes were irrefutably innocent.[8] These discoveries have led so far to hundreds of exonerations, including of inmates on death row or who had spent decades in prison.[9] The true number of wrongful convictions must be considerably larger since evidence that could prove innocence is only rarely available and preserved.

Roughly half of these cases involved forensic-science evidence that was faulty—sometimes egregiously so. The problem could not simply be blamed

2. 509 U.S. 579 (1993).
3. *Id.* at 592–93.
4. *Id.* at 591 n.9.
5. *Id.* at 594–95.
6. *See generally* Paul C. Giannelli, *The Supreme Court's "Criminal"* Daubert *Cases*, 33 SETON HALL L. REV. 1071 (2003).
7. *See generally id.*
8. *See* 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 44 n.94 (2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf [https://perma.cc/R76Y-7VU].
9. *Id.*

on a few "bad apples" among forensic examiners. Rather, the failure was systemic in that some of the supposedly scientific methods had never been shown to be scientifically valid.

In 2005, Congress mandated that the National Research Council (NRC), the research arm of the U.S. National Academy of Sciences, undertake the first serious study of forensic science. Published in early 2009, the NRC's report found disturbing problems across many commonly used forensic methods, including a lack of rigorous and appropriate studies establishing their scientific validity.[10] In a scathing assessment, it found that "much forensic evidence—including, for example, bite marks, firearm, and toolmark identifications—is introduced in criminal trials without any meaningful scientific validation, determination of error rates, or reliability testing to explain the limits of the discipline."[11]

The NRC report made various recommendations, of which the most important was the establishment of a federal agency to promote the development of forensic science into a "mature field."[12] The report urged that the agency have "a culture that is strongly rooted in science" and "must not be part of a law enforcement agency," owing to the inherent conflict of interest between proponents and evaluators of forensic methods.[13]

The report triggered consternation among some in the forensic-science and law enforcement communities. While some forensic scientists sought to remedy the lack of evidence of scientific validity, many others disputed the NRC's assessment. For their part, prosecutors argued strenuously that, while the NRC report had identified room for improvement, its findings should have no bearing on the admissibility of commonly used forensic-science methods.[14]

The Obama administration responded to the report in several ways. While bowing to opposition by the U.S. Department of Justice (DOJ) against creating a forensic-science agency not tied to law enforcement, the administration took three actions. First, it increased overall funding for forensic-science research. Second, the DOJ, in collaboration with the National Institute of Standards and Technology (NIST), established the National Commission on Forensic Science (NCFS) to provide the Attorney General with guidance and policy recommendations on forensic science.[15]

---

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

11. *Id.* at 107–08 (footnotes omitted).

12. *Id.* at 81.

13. *Id.* at 80.

14. *See generally* Simon A. Cole & Gary Edmond, *Science Without Precedent: The Impact of the National Research Council Report on the Admissibility and Use of Forensic Science Evidence in the United States*, 4 BRIT. J. AM. LEGAL STUD. 585 (2015); Paul C. Giannelli, *The 2009 NAS Forensic Science Report: A Literature Review*, 48 CRIM. L. BULL. 378 (2012).

15. *National Commission on Forensic Science*, U.S. DEP'T JUST., https://www.justice.gov/archives/ncfs [https://perma.cc/Z8C9-3QQ4] (last visited Feb. 14. 2018).

Finally, the President tasked his President's Council of Advisors on Science and Technology (PCAST) to recommend additional actions that the federal government could take to ensure the scientific reliability of forensic evidence used in the nation's legal system.

PCAST is the leading scientific and technological advisory body to the executive branch, originally chartered by President Eisenhower in the weeks after the launch of Sputnik.[16] Together with White House science advisor John Holdren, I cochaired the council from 2009 to 2017. During this time, PCAST prepared thirty-nine reports (including two classified reports) making recommendations to the federal government on topics including cybersecurity, biological weapons, nanotechnology, spectrum policy, climate change, energy technologies, advanced manufacturing, ecosystems and the economy, antibiotic resistance, drug discovery and development, semiconductors, hearing aids, pandemic flu vaccines, health information technology, STEM education, agriculture, and big data and privacy.[17]

PCAST's report on forensic science, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*, was released in September 2016.[18] The unanimous report was the result of a year-long study, during which PCAST reviewed 2100 scientific papers, as well as hundreds of pages of input invited from the forensic-science community. Forensic-science experts and others at the FBI and NIST provided valuable and detailed assistance, including carefully reviewing multiple drafts of the report. PCAST also constituted a panel of senior advisors, which included ten current or former judges, a former U.S. Solicitor General, two law-school deans, and two statisticians.[19] As with all PCAST reports, the conclusions represent those of the presidential science advisors. The complete report included a 174-page main text, a 131-page appendix containing responses to PCAST's request for public input in 2015, a 98-page appendix listing the scientific papers consulted, and a 9-page addendum approved on January 6, 2017.[20]

Agreeing with the NRC's assessment that many forensic methods had long been used in courts despite the lack of meaningful evidence of scientific validity, PCAST focused considerable attention on the issue of the admissibility of forensic testimony under Rule 702. The report outlined the scientific meaning of "reliability" and "scientific validity" for a key class of forensic methods—including how these concepts specifically relate to 702(c)

16. *Celebrating the Contributions of the President's Council of Advisors on Science and Technology*, WHITE HOUSE (Jan. 9, 2017, 2:30 PM), https://obamawhitehouse.archives.gov/blog/2017/01/09/celebrating-contributions-presidents-council-advisors-science-and-technology [https://perma.cc/W4BZ-3B4P].

17. *PCAST Documents & Reports*, WHITE HOUSE, https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports [https://perma.cc/P8YV-KVYX] (last visited Feb. 14, 2018).

18. *See generally* PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 8.

19. *Id.* at viii–ix.

20. *See generally* PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 8.

and 702(d), which PCAST referred to respectively as "foundational validity" and "validity as applied."

The report made eight recommendations to the federal government, including both the executive and judicial branches. Among these, PCAST recommended that the Judicial Conference of the United States, through its Standing Advisory Committee on Evidence Rules, should provide guidance to the federal courts about the standards for admissibility under Rule 702 of expert testimony on certain forensic-science methods, through a new Advisory Committee note and a best-practices manual.[21]

In response to PCAST's recommendation, the Standing Advisory Committee on Evidence Rules convened a meeting on forensic expert testimony, *Daubert*, and Rule 702 on October 27, 2017, at Boston College Law School to inform itself about the issues.[22] The meeting included presentations by twenty-six speakers (including myself) and discussion among the attendees.

The purpose of this Article is to summarize aspects of the PCAST report relevant to its recommendation to the Standing Advisory Committee on Evidence Rules and to propose a path forward with respect to Rule 702.

I. FORENSIC FEATURE-COMPARISON METHODS

The PCAST report focused on a specific class of forensic methods, termed "forensic feature-comparison methods."[23] The category includes the analysis of DNA, hair, latent fingerprints, firearms and spent ammunition, toolmarks, shoe prints and tire tracks, bite marks, and handwriting.[24] In each method, examiners compare distinctive features (e.g., DNA fragment sizes, impressions, and so on) in two samples (e.g., from a crime scene and suspect) to determine whether they are likely to come from the same source.[25] Some of the methods are fully objective, while others involve examiners making subjective judgments.[26]

PCAST chose to focus on these methods for several reasons: the methods are widely used in criminal forensics, practitioners have historically claimed them to be highly accurate, the lay public largely regards them as highly accurate, wrongful convictions have occurred in cases involving this class of methods, and the methods all involve metrology—the science of measurement and comparison—which is a discipline with well-defined scientific standards.[27] In short, it is both *important* and *feasible* to ensure that these methods are reliable.

---

21. *Id.* at 145.

22. *See generally* ADVISORY COMMITTEE ON RULES OF EVIDENCE OCTOBER 2017 AGENDA BOOK (2017), http://www.uscourts.gov/sites/default/files/a3_0.pdf [https://perma.cc/R6FB-APZB]; Daniel J. Capra, *Foreword: Symposium on Forensic Expert Testimony*, Daubert, *and Rule 702*, 86 FORDHAM L. REV. 1459 (2018).

23. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 8, at 44.

24. *Id.* at 23.

25. *Id.* at 146.

26. *Id.* at 47.

27. *Id.* at 44–46.

Forensic feature-comparison methods all involve answering two fundamental questions:

(1) Do two samples match? More precisely, are their features within a given degree of similarity?

(2) How meaningful is the match? More precisely, what is the probability that two samples from different sources would show features with the same degree of similarity?

Both questions must be answered before one can draw a conclusion about the likely origin of a sample. The finding of a match between two samples cannot be interpreted—in fact, it is meaningless—unless one knows how often *unrelated* samples show the observed degree of match. It is obviously crucial to know whether a method produces false-positive matches at a rate of 1 in 5000 or 1 in 5.

U.S. District Judge John Potter nicely expressed this point in his opinion in *United States v. Yee*,[28] an early case on the use of DNA analysis: "Without the probability assessment, the jury does not know what to make of the fact that the patterns match: the jury does not know whether the patterns are as common as pictures with two eyes, or as unique as the Mona Lisa."[29]

II. SCIENTIFIC VALIDITY OF FORENSIC FEATURE-COMPARISON METHODS

According to Rule 702 and *Daubert*, courts must consider a key question: What does it mean for a forensic feature-comparison method to be scientifically valid?

The basic answer is simple: *scientific validity requires empirical evidence of how well a method works in practice*. This is nothing more than a restatement of the foundational principle of science established 400 years ago—namely, that assertions about the world cannot be accepted based on authority but must be subjected to empirical testing.

The PCAST report emphasized that direct empirical testing was the *only* way to establish the scientific validity of a forensic feature-comparison method—that nothing else could substitute for it.[30] The report laid out for courts the two essential elements[31]:

(1) Reproducible procedure. The method must have a well-defined, reproducible procedure for identifying and comparing the features in two samples and for determining whether they share sufficient similarity (often called a matching rule). Without this, one does not even have a method to test.

(2) Estimation of false-positive rate. The method must be empirically tested, under conditions appropriate to the intended

28. 134 F.R.D. 161 (N.D. Ohio 1991), *aff'd sub nom.* United States v. Bonds, 12 F.3d 540 (6th Cir. 1993).

29. *Id.* at 181.

30. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 8, at 47.

31. *Id.* at 48.

> use, to determine its accuracy (i.e., how often samples from different sources are erroneously declared to match), which must be suitable for the intended use. Without this, the results cannot be interpreted.

PCAST noted that scientific validity does not require that a method be *perfect*.[32] But, it does require knowing the chances of falsely declaring a match between samples from different sources (e.g., 1 in 1 million, 1 in 600, 1 in 50, or 1 in 3).[33]

Feature-comparison methods can be classified as objective or subjective depending on whether they involved significant human judgment.[34] Subjective methods require special scrutiny because they effectively involve a "black box" in each examiner's head.[35] To assess their accuracy, one must therefore conduct so-called "black-box studies" in which one presents examiners with samples from the same or different sources and records how often examiners give the correct answer.[36] As discussed below, the FBI laboratory has done pioneering work using black-box studies to assess the reliability of latent fingerprint analysis.[37]

The PCAST report noted six scientifically self-evident criteria for any scientifically valid study to assess the accuracy of a method. Specifically, (1) the study should employ samples that are representative of the intended application and numerous enough to provide a meaningful estimate of accuracy, (2) examiners should not know the correct answers in advance nor should the study design allow them to make inferences about the correct answers, (3) the criteria for evaluating the study (especially what constitutes an error) should be specified in advance, not after seeing the results, (4) the study should be conducted or overseen by scientists with no stake in the outcome, (5) the results should be available for review by other scientists, and (6) the conclusions should be reproduced by a second group.[38]

Strikingly, PCAST's report produced diametrically opposed reactions. To scientists, the discussion of scientific validity seemed obvious. By contrast, many forensic practitioners and prosecutors objected to the idea that empirical testing was an absolute requirement. Instead, they insisted, forensic methods could be considered "reliable" even without direct empirical testing to assess their accuracy. To grasp this response, it is necessary to understand the history of forensic science.

---

32. *Id.*
33. *Id.*
34. *Id.* at 47.
35. *Id.* at 49.
36. *Id.*
37. *See infra* notes 69–72 and accompanying text. Once the accuracy of a method has been established, one can also use "white-box studies" to try to shed light on factors that affect examiners' accuracy. Although not necessary for admissibility, such studies can be valuable for improving a method. For a brief description of a white-box study the FBI has conducted, see PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 8, at 99–100.
38. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 8, at 52–53.

III. THE LONG AND UNFINISHED PATH FROM FORENSICS TO FORENSIC SCIENCE

Forensic feature-comparison methods (with the notable exception of DNA analysis) did not emerge from scientific laboratories but rather were developed by police departments as rough-and-ready tools to aid in criminal investigations.[39] As a result, practitioners of these methods devoted much effort to practical issues, such as characterizing features and refining laboratory techniques, but paid virtually no attention to the foundational issue of accuracy.[40]

PCAST surveyed the troubling history for five non-DNA-based feature-comparison disciplines: latent fingerprints, firearms, hair, bite marks, and footwear.[41] In each case, the disciplines were admitted in court based on extraordinary claims unsupported by empirical evidence. Only slowly are these claims being subjected to empirical testing—revealing that they were grossly inaccurate, often by many orders of magnitude.[42]

The history might be characterized as having three successive stages: (1) data-free theories, (2) spurious estimates, and (3) meaningful empirical testing. We discuss these stages in turn, with an overview provided in Table 1.

A. Stage 1: Data-Free Theories

In this stage, various types of arguments are made about why a method should, in principle, be extremely accurate—without actually testing the method empirically. Much attention, for example, has been devoted to "uniqueness studies" aimed at proving that no two objects give identical patterns (e.g., fingerprints, shoe prints), with the implication that feature-comparison analysis will thus never yield false positives. For example, in a 2012 paper on shoe prints, the author studied thirty-nine Adidas Supernova Classic size-twelve running shoes worn by a single runner over eight years, by applying black shoe polish to the soles and having the owner carefully produce tread marks by walking on sheets of legal paper on a hardwood floor. The author reported that small identifying differences could be found between different pairs of shoes.[43]

The PCAST report noted that:

> uniqueness studies miss the fundamental point. The issue is not whether *objects* or *features* differ; they surely do if one looks at a fine enough level. The issue is how well and under what circumstances *examiners* applying a

39. *Id.* at 32.
40. *See id.* at 32–33.
41. *Id.* at 83–122.
42. *See id.* at 76 (describing an instance where the prosecutor told the jury that the chance of a false positive was 1 in 1 billion when the actual probability could have been as low as 1 in 2); *see also infra* tbl.1.
43. *See id*. at 62 (citing Hilary D. Wilson, *Comparison of the Individual Characteristics in the Outsoles of Thirty-Nine Pairs of Adidas Supernova Classic Shoes*, 62 J. FORENSIC IDENTIFICATION 194 (2012)).

> given metrological method can reliably *detect* relevant differences in features to reliably identify whether they share a common source. Uniqueness studies, which focus on the properties of features themselves, can therefore never establish whether a particular *method* for measuring and comparing features is foundationally valid. Only empirical studies can do so.[44]

Another popular approach has been to invoke mathematical calculations. In such studies, authors consider various types of features that might make up a pattern and calculate the number of patterns that might theoretically arise. Given enough features assumed to occur independently and be detected perfectly, the potential number of patterns is guaranteed to be astronomical. A widely cited 1984 paper measured twelve parameters in roughly 400 bite marks carefully made in wax wafers and calculated that the chance that two different sources would produce matching bite marks is less than one in six trillion.[45] The paper was entirely theoretical: it did not even undertake any actual comparisons.[46] Similarly, a 2006 paper on footwear examination, cited by the FBI, used a mathematical model to assert that the chance that two shoe prints from different sources would share three characteristics was less than 1 in 683 billion.[47] Again, the study analyzed no actual shoe prints.[48]

A third solution was simply to *declare* that methods are perfect. The DOJ took this approach in its 1984 publication *The Science of Fingerprints*, which asserted that, "Of all the methods of identification, fingerprinting alone has proved to be both infallible and feasible."[49] At the time, no empirical studies of accuracy had been undertaken.[50] The DOJ conceded in a 2016 draft guidance document about appropriate language for testimony and reports that this earlier assertion was unjustified.[51]

The Association of Firearms and Tool Mark Examiners (AFTE) provides an example of data-free circular reasoning in its *Theory of Identification as It Relates to Toolmarks*.[52] The "theory" (1) declares that an examiner is

---

44. *Id.*

45. *See* PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 8, at 84. *See generally* Raymond D. Rawson et al., *Statistical Evidence for the Individuality of the Human Dentition*, 29 J. FORENSIC SCI. 245 (1984).

46. *See generally* Rawson et al., *supra* note 45. As discussed below, recent empirical studies of bite-mark examiners have found stunningly high error rates.

47. *See generally* Rocky S. Stone, *Footwear Examinations: Mathematical Probabilities of Theoretical Individual Characteristics*, 56 J. FORENSIC SCI. 577 (2006).

48. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 8, at 115.

49. FBI, U.S. DEP'T OF JUSTICE, THE SCIENCE OF FINGERPRINTS: CLASSIFICATION AND USES, at iv (1984).

50. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 8, at 87 (citing FBI, *supra* note 49).

51. *Id.* (citing U.S. DEP'T OF JUSTICE, SUPPORTING DOCUMENTATION FOR DEPARTMENT OF JUSTICE PROPOSED UNIFORM LANGUAGE FOR TESTIMONY AND REPORTS FOR THE FORENSIC LATENT PRINT DISCIPLINE 15 (2016)).

52. *See generally* Comm. for the Advancement of the Sci. of Firearm & Tool Mark Identification, *Theory of Identification as It Relates to Toolmarks: Revised*, 43 AFTE J. 287 (2011).

justified in concluding that two toolmarks have a common origin if they are in "sufficient agreement" and (2) defines "sufficient agreement" as meaning that the agreement between the two toolmarks is such that it is a "practical impossibility" that they have different origins.[53] AFTE still contends that its document constitutes a meaningful scientific theory.

From a scientific standpoint, such efforts to justify forensic feature-comparison methods as scientifically valid would be amusing—except that the arguments were accepted by courts in criminal cases.

B. Stage 2: Spurious Estimates

In this stage, estimates of accuracy are made based on contrived situations that do not correspond to the method's use in practice. Expert testimony in 2009 by a former head of the FBI's fingerprint unit provides an example of how *not* to estimate accuracy from empirical data. He told the court that the FBI's latent fingerprint analysis had "an error rate of one per every 11 million cases."[54] He had arrived at that estimate, he explained, because among 11 million fingerprint identifications performed by the agency, he was aware of only one error.[55]

In a classic study of microscopic hair analysis in the 1970s (and quoted approvingly by the DOJ in 2016), all pairwise comparisons were performed between hairs from different sources and showed a remarkably low false-positive rate of 1 in 40,000.[56] Unfortunately, the result is meaningless because the examiner knew that every comparison involved hairs from *different* sources![57] With no risk of missing true matches, they could safely focus on finding differences—whether real or imagined.[58] As noted below, a rigorous evaluation of hair analysis found a dramatically higher false-positive rate.[59]

Finally, firearms analysis presents a subtler issue. Starting about two decades ago, forensic scientists undertook studies in which they presented examiners with samples of spent ammunition and asked them to identify a match within a set of samples fired from a collection of known guns.[60] The examiners performed well, with a false-positive rate of roughly 1 in 5000.[61] However, these studies had serious flaws. In contrast to casework, many involved "closed set" comparisons, where examiners knew or could infer that

53. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 8, at 59 (citing Comm. for the Advancement of the Sci. of Firearm & Toolmark Identification, *supra* note 52).

54. United States v. Baines, 573 F.3d 979, 990–91 (10th Cir. 2009).

55. *Id.* at 989.

56. B.D. Gaudette & E.S. Keeping, *An Attempt at Determining Probabilities in Human Scalp Hair Comparisons*, 19 J. FORENSIC SCI. 599, 599 (1974); *see also* PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 8, at 28.

57. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 8, at 118–19 (citing Gaudette & Keeping, *supra* note 56).

58. *See id.*

59. *See infra* notes 79–81 and accompanying text.

60. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 8, at 106–09.

61. *Id.* at 111.

a correct answer was always present in the known set.[62] Such knowledge provides a big leg up: examiners can safely match an unknown sample to the closest matching known (rather than worrying that there may not be a match).[63] In some studies, they could use results from some samples to narrow the options for other samples.[64] The Director of the Defense Forensic Science Center analogized such studies to solving a "Sudoku" puzzle, where initial answers can be used to help fill in subsequent answers.[65]

In fairness, the scientists who designed the studies likely did not recognize the problem. However, recent studies that employed "open-set" designs (where examiners have no ancillary information bearing on whether any pair of samples matches) have yielded error rates closer to 1 in 50—that is, one hundredfold higher than the earlier closed-set designs.[66] PCAST rejected the earlier studies as providing overly optimistic estimates of accuracy.[67]

C. Stage 3: Meaningful Empirical Testing

In this final stage, forensic scientists obtain a scientifically valid measure of accuracy by conducting black-box studies that directly measure examiners' accuracy in a setting that resembles the method's use in practice but in which the evaluators know the right answers. Notably, the first black-box studies for subjective feature-comparison methods were only undertaken after the NRC report called attention to the lack of evidence for scientific validity for most forensic methods.[68]

The first properly designed black-box study on latent fingerprint analysis was reported in 2011.[69] In a paper by FBI scientists and their collaborators published in the prestigious journal *Proceedings of the National Academy of Sciences*, they asked each of 169 examiners to analyze 100 pairs of fingerprints.[70] The paper found a false identification rate of roughly 1 in 600 (with a confidence interval ranging up to 1 in 300).[71] A subsequent black-box study conducted by the Miami-Dade Police Department Forensic Services Bureau, with funding from the National Institute of Justice, found a higher error rate of 1 in 137 (if one excludes false positives that the authors suggest are likely to represent clerical errors) or 1 in 24 (if one includes these errors, as one would in a clinical trial).[72] These error rates are a far cry from

62. *Id.* at 108–09.
63. *Id.*
64. *Id.* at 106.
65. *Id.*
66. *Id.* at 109–11.
67. *Id.* at 111.
68. *Id.* at 9, 11.
69. *See generally* Bradford T. Ulery et al., *Accuracy and Reliability of Forensic Latent Fingerprint Decisions*, 108 PROC. NAT'L ACAD. SCI. 7733 (2011).
70. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 8, at 94 (citing Ulery et al., *supra* note 69).
71. *Id.* (citing Ulery et al., *supra* note 69).
72. *Id.* at 94–95 (citing IGOR PACHECO ET AL., MIAMI-DADE RESEARCH STUDY FOR THE RELIABILITY OF THE ACE-V PROCESS: ACCURACY AND PRECISION IN LATENT FINGERPRINT

the DOJ's original claim of infallibility, but they are perfectly serviceable estimates of reliability that would allow a jury to weigh fingerprint testimony relative to other evidence in a criminal case.

The first black-box study of firearms analysis was reported by the Ames Laboratory in 2014; the work was stimulated and funded by the Defense Forensic Science Center, whose director had criticized the Sudoku-like nature of previous closed-set studies.[73] Similar in its basic design to the FBI's latent-fingerprint study, the authors evaluated the performance of 218 examiners on fifteen separate comparison problems.[74] They reported an error rate of 1 in 66 (with a confidence interval ranging to 1 in 46).[75] As noted above, the error rate is approximately one hundredfold higher than the closed-set studies.[76] With only a single well-designed study estimating accuracy, PCAST judged that firearms analysis fell just short of the criteria for scientific validity, which requires reproducibility.[77] A second study would solve this problem.

Although black-box studies have not yet been conducted for other disciplines, PCAST summarized the limited scientific studies undertaken for hair and bite-mark analysis.[78] The papers are notable because they debunk past claims about the accuracy of these disciplines.

A 2002 paper by FBI scientists revealed a stunningly high false-identification rate for hair analysis.[79] The study did not present examiners with test problems but rather used DNA analysis to reexamine hair samples from actual criminal cases that FBI examiners had declared were microscopically indistinguishable.[80] In contrast to earlier work claiming that hairs from different sources could be distinguished with an error rate of only 1 in 40,000 comparisons, DNA analysis of casework revealed that 11 percent of hairs (that is, 1 in 9) reported as microscopically indistinguishable actually came from different sources.[81]

Only a few small empirical studies have been reported on the accuracy of bite-mark examiners. The results have been consistently appalling. In a 2010 paper, for example, twenty-nine examiners were asked to inspect

---

EXAMINATIONS (2014), https://www.ncjrs.gov/pdffiles1/nij/grants/248534.pdf [https://perma.cc/4FZ6-F9N7]).

73. *See generally* DAVID P. BALDWIN ET AL., A STUDY OF FALSE-POSITIVE AND FALSE-NEGATIVE ERROR RATES IN CARTRIDGE CASE COMPARISONS (2014), https://afte.org/uploads/documents/swggun-false-postive-false-negative-usdoe.pdf [https://perma.cc/3JXK-6FS4].

74. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 8, at 109–10 (citing BALDWIN ET AL., *supra* note 73).

75. *Id.* at 110–11 (citing BALDWIN ET AL., *supra* note 73).

76. *See supra* note 66 and accompanying text.

77. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 8, at 111.

78. *Id.* at 117.

79. *See generally* Max M. Houck & Bruce Budowle, *Correlation of Microscopic and Mitochondrial DNA Hair Comparisons*, 47 J. FORENSIC SCI. 964 (2002).

80. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 8, at 28 (citing Houck & Budowle, *supra* note 79).

81. *Id.* (citing Houck & Budowle, *supra* note 79).

photographs of bite marks (produced in pig flesh by a mechanical biting machine using human dentition) and decide whether they came from individuals A, B, C, or none of the above.[82] When the correct answer was "none of the above," the examiners nonetheless attributed the bite marks to one of the three known sources for 17 percent of samples (that is, 1 in 6).[83] Other studies showed that bite-mark examiners performed poorly even in a closed-set design, when the correct source was always provided (error rates of 1 in 9).[84]

Finally, PCAST could find few relevant papers on footwear analysis and none that even came close to providing a serious evaluation of scientific validity.[85]

D. Current Status of Forensic Methods

In summary, there has been some progress since the NRC's report in 2009. Empirical studies have now provided scientifically valid estimates of the accuracy of latent fingerprint analysis, and firearms analysis is coming close to achieving the standard for scientific validity. With respect to hair analysis, little has been done to address the poor ability to distinguish different-source samples in casework revealed by the FBI's study. At the least, juries should be told that 1 in 9 identifications in casework proved to come from different sources. As for bite-mark analysis, the field does not appear to be salvageable; it should be abandoned. Finally, footwear analysis has yet to be subjected to empirical testing—although it continues to be used in court.

In light of the historical (and in some cases continuing) lack of empirical evidence, what has given forensic practitioners confidence that their methods were reliable? The answer is that they have had faith in their processes. Specifically, they point to (1) examiners' extensive "experience" and "judgment" in the course of casework and (2) good professional practices, such as the existence of professional societies, certification programs, peer-reviewed articles, proficiency testing, and codes of ethics.

There is a gaping hole in this logic. Extensive experience and good professional practices are clearly important, and forensic practitioners should be commended for their attention to these matters. But, experience and professional practices can never establish whether a method itself is reliable—for the simple reason that neither assesses a method's accuracy. Experience in casework provides no information about accuracy because the right answer is not known in casework. And, professional practices concern process not results.

To grasp the importance of this point, one need only note that practitioners of pseudoscience—such as psychics—can make the same claims about their fields. Psychics can claim extensive experience in mindreading and soothsaying, and they too have professional societies, certification programs,

82. *Id.* at 86.
83. *Id.*
84. *Id.*
85. *Id.* at 117.

peer-reviewed journals, proficiency testing, and codes of ethics. Despite these trappings of science, psychics' claims are not accepted as scientifically valid—and are not admissible under Rule 702—because their methods have not withstood appropriate empirical testing to determine their accuracy.

For forensic methods to be accepted as reliable and scientifically valid, there is simply no substitute for actual empirical testing of accuracy.

IV. THE DOJ'S RESISTANCE TO ADDRESSING THE ISSUES OF SCIENTIFIC VALIDITY

For its part, the DOJ has resisted the necessity of empirical testing. The resistance is understandable: acknowledging the need for empirical testing might lead to calls to revisit past convictions or jeopardize ongoing cases involving evidence based on forensic-science methods that had not been empirically shown to be reliable. The DOJ has thus sought to block or blunt recommendations from the scientific community.

When the NRC recommended the creation of a federal office separate from law enforcement to ensure the quality of forensic science,[86] the DOJ successfully lobbied for a weaker solution: an outside advisory committee that would make recommendations to the Attorney General. The National Commission on Forensic Sciences was established in 2013 but soon ran into trouble when the DOJ's efforts to limit the body's scope caused a federal judge who served on the commission to resign in protest.[87] The DOJ reversed course, and the commissioner returned.[88] The Commission made various recommendations, but only a few were implemented by the Attorney General.

When PCAST briefed the DOJ on its preliminary conclusions at a meeting that I attended in late May 2016, DOJ officials acknowledged the lack of empirical studies establishing reliability for some disciplines but expressed concerns that the report could affect past convictions and ongoing cases. The DOJ proposed that PCAST delay its report until December 2016 and declare that its findings should not be applied retroactively.

While understanding the reasons for the DOJ's concern, PCAST declined these suggestions. In particular, it saw no scientific basis for distinguishing between past and present applications of forensic science. However, consistent with its past practices, PCAST invited the DOJ to provide comments on the draft report and identify any relevant material that PCAST might have missed. PCAST revised the report in response to several rounds

86. NAT'L RESEARCH COUNCIL, *supra* note 10, at 19–22.

87. Spencer S. Hsu, *U.S. Judge Quits Commission to Protest Justice Department Forensic Science Policy*, WASH. POST (Jan. 29, 2015), https://www.washingtonpost.com/local/crime/us-judge-quits-commission-to-protest-justice-department-forensic-science-policy/2015/01/29/cbed0a84-a7bb-11e4-a2b2-776095f393b2_story.html [http://perma.cc/MA55-QR99].

88. Spencer S. Hsu, *Judge Rakoff Returns to Forensic Panel After Justice Department Backs Off Decision*, WASH. POST (Jan. 30, 2015), https://www.washingtonpost.com/local/crime/in-reversal-doj-lets-forensic-panel-suggest-trial-rule-changes-after-us-judge-protests/2015/01/30/2f031d9e-a89c-11e4-a2b2-776095f393b2_story.html [https://perma.cc/6SBC-Q7UW].

of comments from the DOJ, including many helpful suggestions from the FBI laboratory.

In the end, however, the DOJ insisted, in written communications with PCAST, that the implications for Rule 702 should be deleted from the report. In its judgment, the President's science advisory council had no business opining on the meaning of scientific validity as it pertains to the admissibility of expert scientific testimony. Moreover, the DOJ asserted, the references in *Daubert* to evidentiary reliability being based on scientific validity were merely dicta. The DOJ asked the White House to block the release of the PCAST report, but the White House declined to do so.

When PCAST released its report on September 20, the Attorney General thanked PCAST for its work but stated that the agency would not accept the council's recommendations.[89] The statement also claimed PCAST had failed to mention "numerous published research studies" and that this "omission discredits the PCAST report as a thorough evaluation of scientific validity."[90] In response to a request from PCAST to identify relevant omissions, the DOJ eventually concluded in December 2016 that it could find none.[91]

Following the presidential transition in January 2017, the Attorney General decided to terminate the NCFS by allowing its charter to expire in April 2017.[92] The DOJ instead chose to rely solely on its own internal Senior Advisor on Forensics.[93] Whereas the previous incumbent had been a forensic scientist, the DOJ in August 2017 tapped as its new advisor a prosecutor without scientific training who had served as a law enforcement representative on the NCFS.[94] The new advisor has employed tactics often used to resist scientific consensus, such as characterizing basic scientific statements as extreme and alleging substantial disagreement within the scientific community. For example, at the symposium organized by the

89. Gary Fields, *White House Advisory Council Report Is Critical of Forensics Used in Criminal Trials*, WALL. ST. J. (Sept. 20, 2016, 4:25 PM), https://www.wsj.com/articles/white-house-advisory-council-releases-report-critical-of-forensics-used-in-criminal-trials-1474394743 [https://perma.cc/W84T-WNDA].

90. *Comments on: President's Council of Advisors on Science and Technology Report to the President*, FBI (Sept. 20, 2016), https://www.fbi.gov/file-repository/fbi-pcast-response.pdf [https://perma.cc/W9UQ-JEU4].

91. *See* President's Council of Advisors on Sci. & Tech., *An Addendum to the PCAST Report on Forensic Science in Criminal Courts*, WHITE HOUSE 5 (Jan. 6, 2017), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_addendum_finalv2.pdf [https://perma.cc/8A4D-57HX].

92. Press Release, U.S. Dep't of Justice, Attorney General Jeff Sessions Announces New Initiatives to Advance Forensic Science and Help Counter the Rise in Violent Crime (Apr. 10, 2017), https://www.justice.gov/opa/pr/attorney-general-jeff-sessions-announces-new-initiatives-advance-forensic-science-and-help [https://perma.cc/4JGZ-VPEG].

93. *See id.*

94. Press Release, U.S. Dep't of Justice, Justice Department Announces Plans to Advance Forensic Science (Aug. 7, 2017), https://www.justice.gov/opa/pr/justice-department-announces-plans-advance-forensic-science [https://perma.cc/F7Y3-FNPB]; *see also* Radley Balko, *Deputy AG Announces New Forensic Science Working Group but Still Doesn't Grasp the Extent of Problem*, WASH. POST (Aug. 7, 2017), https://www.washingtonpost.com/news/the-watch/wp/2017/08/07/deputy-ag-announces-new-forensic-science-working-group-but-still-doesnt-grasp-the-extent-of-problem [https://perma.cc/G7CY-EPWT].

Standing Advisory Committee on Evidence Rules, the DOJ took the position that (1) PCAST's list of criteria for reliability studies was a radical "nonseverable six-part test" (without actually identifying any criteria that were not correct)[95] and (2) a recent report by the American Association for the Advancement of Science (AAAS) supposedly held that empirical testing, as described in the PCAST report, was not the only way to establish the reliability of forensic feature-comparison methods.[96] The AAAS swiftly rejected the claim, issuing a statement that the PCAST and AAAS reports were in complete agreement on the issue.[97] In summary, it is not realistic to count on law enforcement to drive progress.

V. FIXING RULE 702

With respect to forensic science, Rule 702 has clearly failed to accomplish its goal of ensuring that expert testimony must be based on reliable methods. Courts routinely admit testimony about feature-comparison methods that claim to be able to identify the source of a sample with high accuracy—even when the reliability of the methods have never been tested or when the methods have been tested and found to be unreliable.

The leading scientific advisory groups chartered by the legislative and executive branches—the National Academy of Sciences and the President's Council of Advisors on Science and Technology—have now weighed in. They have unanimously agreed that methods have historically lacked meaningful scientific validation, that their accuracy has been seriously overstated, and that their misuse has led to wrongful convictions. Moreover, they agree that requiring empirical testing is feasible and would increase the quality of forensic science—with benefits for prosecutors, defendants, and the public.

To fix Rule 702, it is important to understand some reasons for its failure. First, many judges simply do not know how to apply the concepts of reliability and scientific validity to any given scientific discipline. In the absence of a clear definition, they are often willing to accept the trappings of reliability (examiners' experience and professional practices) rather than insist on actual reliability. Second, many judges are also reluctant to challenge longstanding precedents concerning the admissibility of forensic methods, even when they were established long before current problems became apparent.

How, then, to restore the role of courts, articulated in *Daubert*, as gatekeepers ensuring quality control? The appellate process is not well suited to the task. Even if an appeals court wished to do so, it would be hampered by the high standard (abuse of discretion) for overturning admissibility decisions. And, because the vast majority of criminal cases occur in the state

95. *See Symposium on Forensic Expert Testimony,* Daubert, *and Rule 702*, 86 FORDHAM L. REV. 1463, 1520 (2018) (statement of Ted R. Hunt).

96. *Id.*

97. William C. Thompson, *AAAS, PCAST and Validation: Questions and Answers*, AM. ASS'N ADVANCEMENT SCI. (2017), https://mcmprodaaas.s3.amazonaws.com/s3fs-public/QA%20AAAS%20and%20PCAST%20Reports.pdf [https://perma.cc/UKS2-RS4Z].

courts, establishing a coherent jurisprudence would require parallel progress on the many cognate versions of Rule 702.

Instead, PCAST recommended that the most effective solution would be for the Judicial Conference of the United States to clarify the meaning of "reliable methods" for forensic feature-comparison methods. PCAST proposed that the Standing Advisory Committee on Evidence issue a new advisory committee note and a best-practices manual to provide clear guidance for courts. Alternatively or additionally, the committee could propose a revision to the Federal Rules of Evidence.[98]

Whatever the mechanism, the key message should be roughly the following:

> An expert witness may provide testimony based on a forensic examination conducted to determine whether an evidentiary sample is similar or identical to a source sample if (in addition to satisfying existing requirements under Rule 702):
>
> > (i) the witness's method is sufficiently repeatable, reproducible, and accurate for its intended use, as shown by empirical studies conducted under conditions appropriate to the intended use;
> >
> > (ii) the witness is capable of applying the method reliably and actually did so; and
> >
> > (iii) the witness accurately states the probative value of the meaning of any similarity or match between the samples.

With respect to the third point, it is useful to give a specific example of appropriate testimony. Suppose that two proper black-box studies have been performed and published. The data in each study allows empirical estimates to be made of a method's error rate. Courts should require a witness to describe in clear and simple terms what is known about accuracy and error rate based on these studies. An appropriate statement would be:

> Examiners sometimes make mistakes in associating a sample with a particular individual. Studies have therefore been done to see approximately how often such errors occur in situations similar to this one. In one study, examiners made false associations at a rate of 1 in every 300 comparisons performed. Given the number of tests carried out, the true error rate in this study might be somewhat higher—possibly 1 in 150. In a second study, examiners made false associations at a rate of 1 in every 75 comparisons; given the number of tests carried out, the true error rate in this study might be 1 in 40. In short, the method usually gives the correct answer, but errors do occur.

One might be tempted to try to craft a general rule that would provide guidance not just for forensic feature-comparison methods, but for all forensic-science testimony in criminal cases. But, such a course would be

98. Under current practices, new advisory notes are not issued without a change to the rules themselves. It might be worth loosening this stricture in appropriate circumstances.

problematic. As Justice Oliver Wendell Holmes cautioned, "the life of the law . . . is experience."[99]

We now have two decades of experience illuminating the problems and solutions for forensic feature-comparison methods. When adequate experience arises for other areas, they may be addressed in turn. It may not be necessary to repeat this exercise many times. Even a few examples may suffice to signal to courts that they should engage more generally in the essential dialog, contemplated in *Daubert*, between law and science.

99. O.W. HOLMES, JR., THE COMMON LAW 1 (1881).

Table 1: Summary of Claims About the Accuracy of Forensic Feature-Comparison Methods

Method	**Stage 1: Data-Free Theories**	**Stage 2: Spurious Estimates**	**Stage 3: Meaningful Empirical Testing**
Fingerprints	"Infallible"	1 in 11 million	1 in 600
Firearms	n/a	1 in 5000	1 in 50
Hair Analysis	n/a	1 in 40,000	1 in 9^{100}
Bite marks	1 in 6 trillion	n/a	1 in 6^{101}
Footwear	1 in 683 billion	n/a	None

100. Well-designed black-box studies have not been performed for this discipline, but other studies showed extremely high error rates. *See supra* note 81 and accompanying text.

101. Well-designed black-box studies have not been performed for this discipline, but other studies showed extremely high error rates. *See supra* notes 82–83 and accompanying text.