2018

**Forensic Science: *Daubert’s Failure***

Paul C. Giannelli

Follow this and additional works at: https://scholarlycommons.law.case.edu/caselrev

Part of the Law Commons

**Recommended Citation**


Available at: https://scholarlycommons.law.case.edu/caselrev/vol68/iss3/18

This Article is brought to you for free and open access by the Student Journals at Case Western Reserve University School of Law Scholarly Commons. It has been accepted for inclusion in Case Western Reserve Law Review by an authorized administrator of Case Western Reserve University School of Law Scholarly Commons.
FORENSIC SCIENCE: DAUBERT’S FAILURE

Paul C. Giannelli†

“The man who discovers a new scientific truth has previously had to smash to atoms almost everything he had learnt, and arrives at the new truth with hands bloodstained from the slaughter of a thousand platitudes.”

Contents

Introduction .............................................................................................................................. 870
  A. Daubert and Rule 702 ................................................................................................. 871

I. Discredited Techniques .................................................................................................. 876
  A. Bite Mark Comparisons .............................................................................................. 876
    1. Foundational Research ......................................................................................... 878
    2. Texas Forensic Science Commission (2016) ................................................... 880
  B. Microscopic Hair Analysis ......................................................................................... 883
    1. FBI Hair Review ..................................................................................................... 886
    2. White House PCAST Report (2016) .................................................................. 888
  C. Arson Investigations .................................................................................................... 889
    1. Willingham Case .................................................................................................... 889
    2. Han Tak Lee Case .................................................................................................. 896
    3. National Fire Protection Association Guidelines .............................................. 897
    4. Dog-Sniff Evidence ............................................................................................... 898
    5. Post-Daubert Cases .............................................................................................. 899
  D. Comparative Bullet Lead Analysis ............................................................................ 901
    2. Post-Report Developments .................................................................................. 905

II. Misleadingly Presented Techniques ............................................................................. 909
  A. Firearms & Toolmark Identifications ........................................................................ 909
    1. Post-Daubert Cases .............................................................................................. 910

† Distinguished University Professor Emeritus & Albert J. Weatherhead III and Richard W. Weatherhead Professor Emeritus, Case Western Reserve University. The author wishes to thank Jules Epstein and Julia Leighton for their insightful comments.

INTRODUCTION

In 2015, Judge Alex Kozinski of the Ninth Circuit Court of Appeals noted that “[m]any defendants have been convicted and spent countless years in prison based on evidence by arson experts who were later shown to be little better than witch doctors.” In the same year, Dr. Jo Handelsman, a White House science advisor, observed: “Suggesting that bite marks [should] still be a seriously used technology is not based on science, on measurement, on something that has standards, but more of a gut-level reaction.” According to Judge Catharine Easterly of the D.C. Court of Appeals, “[a]s matters currently stand, a certainty statement regarding toolmark pattern matching has the same probative value as the vision of a psychic.” A New York Times editorial echoed these sentiments:

[C]ourts have only made the problem worse by purporting to be scientifically literate, and allowing in all kinds of evidence that

2. Alex Kozinski, Preface: Criminal Law 2.0, 44 GEO. L.J. ANN. REV. CRIM. PROC. iii, v (2015); see also Almeciga v. Ctr. for Investigative Reporting, Inc., 185 F. Supp. 3d 401, 415 (S.D.N.Y. 2016) (Rakoff, J.) (“There have been too many pseudo-scientific disciplines that have since been exposed as profoundly flawed, unreliable, or baseless for any Court to take this [gatekeeping] role lightly.”).


would not make it within shouting distance of a peer-reviewed journal. Of the 329 exonerations based on DNA testing since 1989, more than one-quarter involved convictions based on “pattern” evidence—like hair samples, ballistics, tire tracks, and bite marks—testified to by so-called experts.\(^5\)

These criticisms are valid—which raises a puzzling and consequential question: Why didn’t the Supreme Court’s “junk science” decision, *Daubert v. Merrell Dow Pharmaceuticals, Inc.*\(^6\) prevent or restrict the admissibility of testimony based on flawed forensic techniques? After all, *Daubert* was decided in 1993, twenty-five years ago.

### A. Daubert and Rule 702

*Daubert* was considered a revolutionary decision.\(^7\) It “radically changed the standard for admissibility of scientific testimony”\(^8\) by sweeping away the *Frye v. United States*\(^9\) “general acceptance” test,\(^10\) which had been the majority rule in both federal and state cases.\(^11\) The *Frye* standard gave great deference to the views of forensic practitioners and not to empirical testing.\(^12\) *Daubert* promised to be different. The Supreme Court held that “[p]roposed testimony must be supported by

---


8. United States v. Barnette, 211 F.3d 803, 815 (4th Cir. 2000); see also United States v. Alatorre, 222 F.3d 1098, 1100 (9th Cir. 2000) (“*Daubert* has become ubiquitous in federal trial courts.”).

9. 293 F. 1013 (D.C. Cir. 1923).

10. Id. at 1014 (stating that a technique “must be sufficiently established to have gained general acceptance in the particular field in which it belongs”).


12. See Michael J. Saks, *Merlin and Solomon: Lessons from the Law’s Formative Encounters with Forensic Identification Science*, 49 Hastings L.J. 1069, 1138 (1998) ("*Frye* does not work because its measure of validity is the judgment of ‘the field,’ and the field may consist of nonsense. For example, the *Frye* doctrine cannot exclude astrology.").
appropriate validation—i.e., ‘good grounds,’ based on what is known. In short, the requirement that an expert’s testimony pertain to ‘scientific knowledge’ establishes a standard of evidentiary reliability.\footnote{13} In making this reliability determination, the Daubert Court highlighted five factors: (1) empirical testing, (2) peer review and publication, (3) error rate, (4) maintenance of standards, and (5) general acceptance.\footnote{14} The first and most important factor is empirical testing. The other factors are supplementary.\footnote{15} Peer review and publication are designed to expose defects in testing. Acceptance of a technique within the scientific community is achieved through the publication of valid test results. Similarly, both error rates and standards are derived from testing.

Daubert was followed in 1999 by Kumho Tire Co. v. Carmichael,\footnote{16} which held that Daubert’s reliability standard applied to all expert testimony, not only scientific evidence.\footnote{17} By 2000, the Supreme Court was describing Daubert as establishing an “exacting” standard.\footnote{18} In the same year, Federal Rule of Evidence 702 was amended to incorporate the Daubert-Kumho standard.\footnote{19} Although a handful of jurisdictions continue to apply the Frye test, about forty jurisdictions have adopted the Daubert standard in one form or another.\footnote{20}

During this time, there was no shortage of commentary on the lack of empirical research in forensic science.\footnote{21} For example, shortly after

\begin{footnotes}
\item[14] Id. at 593–94.
\item[15] Id.
\item[17] Id. at 141.
\item[21] A few perceptive scholars had noted the lack of empirical testing prior to Daubert. See Randolph N. Jonakait, Forensic Science: The Need for Regulation, 4 HARV. J.L. & TECH. 109, 137 (1991) (“Forensic science is supported by almost no research. The laboratory practices are based on intuitions and deductions, not on empirical proof.”); D. Michael Risinger et al., Exorcism of Ignorance as a Proxy for Rational Knowledge: The Lessons of Handwriting Identification “Expertise,” 137 U. PA. L. REV. 731, 738 (1989) (“Our literature search for empirical evaluation of handwriting identification turned up one primitive and flawed validity study from nearly 50 years ago, one 1973 paper that raises the issue of consistency among examiners but that presents only uncontrolled impressionistic and anecdotal information not qualifying as data in any rigorous sense, and a summary of
\end{footnotes}
Daubert was decided, Professor Margaret Berger wrote: “Considerable forensic evidence made its way into the courtroom without empirical validation of the underlying theory and/or its particular application.”\textsuperscript{22} After Kumho, two commentators—citing bite mark, hair, and firearm analysis—observed that “little rigorous, systematic research has been done to validate the discipline’s basic premises and techniques, and in each area there was no evident reason why such research would be infeasible.”\textsuperscript{23}

Notwithstanding Daubert’s promise, scholars soon discerned its uneven application in civil and criminal cases: “[T]he heightened standards of dependability imposed on expertise proffered in civil cases has continued to expand, but . . . expertise proffered by the prosecution in criminal cases has been largely insulated from any change in pre-Daubert standards or approach.”\textsuperscript{24} The title of a 2005 article summed up the state of the law—“The (Near) Irrelevance of Daubert to Criminal Justice.”\textsuperscript{25} In short, in the criminal context, courts applied Daubert-lite.

\textbf{B. National Academy of Sciences Forensic Report (2009)}

In 2006 Congress entered the picture by authorizing the National Academy of Sciences (“NAS”) to conduct a study of forensic science. After a three-year investigation, NAS issued a landmark report. One of

\begin{itemize}
  \item one study in a 1978 government report. Beyond this, nothing.”); Michael J. Saks & Jonathan J. Koehler, \textit{What DNA “Fingerprinting” Can Teach the Law About the Rest of Forensic Science}, 13 \textit{Cardozo L. Rev.} 361, 372 (1991) (“[F]orensic scientists, like scientists in all other fields, should subject their claims to methodologically rigorous empirical tests. The results of these tests should be published and debated. Until such steps are taken, the strong claims of forensic scientists must be regarded with far more caution than they traditionally have been.”).
  \item Margaret A. Berger, \textit{Procedural Paradigms for Applying the Daubert Test}, 78 \textit{Minn. L. Rev.} 1345, 1354 (1994) (“Courts never required some of the most venerable branches of forensic science—such as fingerprinting, ballistics, and handwriting—to demonstrate their ability to make unique identifications.”).
  \item Paul Giannelli & Edward Imwinkelried, \textit{Scientific Evidence: The Fallout from Supreme Court’s Decision in Kumho Tires}, \textit{Crim. Just.}, Winter 2000, at 12, 40. For an insightful analysis of how identification science was accepted by the courts, see Saks, \textit{supra} note 12.
\end{itemize}
its most riveting passages concluded: “Among existing forensic methods, only nuclear DNA analysis has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between an evidentiary sample and a specific individual or source.”26 The report went on to state that “some forensic science disciplines are supported by little rigorous systematic research to validate the discipline’s basic premises and techniques.”27 Such common forensic techniques as fingerprint examinations,28 firearm (“ballistics”) and toolmark identifications,29 handwriting examinations,30 microscopic hair analysis,31 and bite mark comparisons32 fell into this category.

Not only did the NAS report highlight flaws in forensic science, it sharply criticized the judiciary for failing to demand the validation that Daubert required: “The bottom line is simple: In 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.”33 In a later passage, the report declared that “Daubert has done little to improve the use of forensic science evidence in criminal cases.”34 The disparate treatment of civil actions and criminal prosecutions was also noted. After finding that “trial judges rarely exclude or restrict expert testimony offered by prosecutors,” the report commented: “ironically, the appellate courts appear to be more willing to second-guess


27. Id. at 22. At another point, the Report stated: “The simple reality is that the interpretation of forensic evidence is not always based on scientific studies to determine its validity. This is a serious problem.” Id. at 8; see also id. at 6 (“Often there are no standard protocols governing forensic practice in a given discipline. And, even when protocols are in place . . . , they often are vague and not enforced in any meaningful way.”).

28. Id. at 144 (noting that research is needed “[t]o properly underpin the process of friction ridge [fingerprint] identification”).

29. Id. at 154 (“Sufficient studies [on firearms identification] have not been done to understand the reliability and repeatability of the methods.”).

30. Id. at 166 (“The scientific basis for handwriting comparisons needs to be strengthened.”).

31. Id. at 161 (“[T]estimony linking microscopic hair analysis with particular defendants is highly unreliable.”).

32. Id. at 174 (“No thorough study has been conducted of large populations to establish the uniqueness of bite marks . . . .”).

33. Id. at 53 (emphasis added).

34. Id. at 106.
trial court judgments on the admissibility of purported scientific evidence in civil cases than in criminal cases.”

Despite the NAS report, courts continued to admit the same evidence. Only a handful of courts applied the “exacting” standard that the Supreme Court said Daubert demanded.

This Article examines the justice system’s failure by reviewing the status of two categories of forensic techniques. The first category involves discredited techniques: (1) bite mark analysis, (2) microscopic hair comparisons, (3) arson evidence, and (4) comparative bullet lead analysis. The second category involves techniques that have been misleadingly presented, including firearm and toolmark identifications and fingerprint examinations. Both categories present Daubert issues. The Article argues that the system’s failure can be traced back to its inability to demand and properly evaluate foundational research—i.e., Daubert’s first factor, empirical testing. Indeed, the justice system may be institutionally incapable of applying Daubert in criminal cases.

A different paradigm is needed, one that assigns an independent agency the responsibility of evaluating foundational research. As discussed in Part IV, this approach was recently recommended by the National Commission on Forensic Science and the President’s Council of Advisors on Science and Technology (“PCAST”).

Both recommended that the National Institute of Standards and Technology (“NIST”) evaluate all forensic disciplines on a continuing basis, thereby injecting much needed scientific expertise into the criminal justice system. The recent reports on latent fingerprints and arson

---

35. Id. at 11.

36. See, e.g., Weisgram v. Marley Co., 528 U.S. 440, 455 (2000). A former federal district judge, Nancy Gertner, noted, “a busy trial judge can rely on the decades of case law to legitimize decisions rejecting a hearing or motions in limine. And the trial judge can count on the Court of Appeals likely concluding that rejecting the challenge was not an abuse of the judge’s discretion.” Nancy Gertner, Commentary on The Need for a Research Culture in the Forensic Sciences, 58 U.C.L.A. L. Rev. 789, 790 (2011).


investigations,\textsuperscript{40} which were published by the American Association for the Advancement of Science ("AAAS"), buttress the need for independent scientific evaluations.

I. Discredited Techniques

A. Bite Mark Comparisons

For decades, bite-mark evidence has been admitted in hundreds of trials,\textsuperscript{41} many of which were capital prosecutions.\textsuperscript{42} No reported American case has rejected bite mark testimony. Moreover, it is not uncommon for courts to speak of bite mark comparisons as a "science"\textsuperscript{43}—even an "exact science."\textsuperscript{44} Acceptance of the technique is so deeply entrenched that some courts have taken judicial notice of its validity,\textsuperscript{45} which means its reliability is indisputable.\textsuperscript{46}

Distinctive characteristics of a person’s dentition were first used to identify skeletonized remains and individuals in mass disasters such as...
plane crashes. Courts assumed that these distinctive characteristics can be transferred to another person’s skin during a violent crime—e.g., homicides, rapes, and child abuse—an assumption that overlooked some obvious problems. First, bite marks typically involve no more than the edges of six to eight front teeth, not thirty-two teeth with five anatomical surfaces that can be used when comparing a deceased person’s dentition with X-rays. Second, bite marks do not reveal artifacts such as fillings, crowns, etc., all of which assist in associating human remains with a person’s dental records. Moreover, human skin is extremely malleable and thus subject to various types of distortion. In addition, bite mark analysis is a subjective technique with no agreed-upon methodology.

47. 1 GIANNELLI ET AL., supra note 20, § 13.02 (discussing the admissibility of dental identifications).

48. See People v. Milone, 356 N.E.2d 1350, 1358 (Ill. App. Ct. 1976) (“The concept of identifying a suspect by matching his dentition to a bite mark found at the scene of a crime is a logical extension of the accepted principle that each person’s dentition is unique.”); People v. Smith, 443 N.Y.S.2d 551, 556–57 (Cty. Ct. 1981) (“The basic premise is the unique nature of individual dentition and the virtually infinite number of individual bite configurations.”) (citations omitted).

49. “Restorations alone, with varying shapes, sizes, and restorative materials, may offer numerous points for comparison. In addition to restorations, the number of teeth, prostheses, decay, malposition, malrotation, peculiar shapes, root canal therapy, bone patterns, bite relationship, and oral pathology may all provide identifying characteristics.” 1 GIANNELLI ET AL., supra note 20, § 13.02.

50. See I.A. Pretty & D. Sweet, The Scientific Basis for Human Bitemark Analyses—A Critical Review, 41 SCI. & JUST. 85, 87 (2001) (“Skin is a poor registration material since it is highly variable in terms of anatomical location, underlying musculature or fat, curvature, and looseness or adherence to underlying tissues. Skin is highly visco-elastic, which allows stretching to occur during either the biting process or when evidence is collected.”).
1. Foundational Research

Despite overwhelming judicial approval, bite mark evidence is not supported by foundational research.\(^{51}\) Indeed, the only rigorous studies are recent—and undercut the technique’s validity.\(^{52}\) The 2009 NAS forensic report concluded that “the scientific basis is insufficient to conclude that bite mark comparisons can result in a conclusive match.”\(^{53}\) Despite the NAS Report, courts continued to permit expert testimony on the subject. For example, in *State v. Prade*,\(^{54}\) decided in 2014, the expert testified that “bite mark evidence is generally accepted within the scientific community . . . .”\(^{55}\) Similarly, in *Coronado v. State*,\(^{56}\) a different expert stated that he did not “agree with the NAS Report’s conclusion that bite mark analysis cannot result in a conclusive match,” adding, “you do not have to be a ‘rocket scientist’ to see that, in some cases, there is a unique and distinct pattern of teeth that can be identified.”\(^{57}\) In addition, these experts rejected the valid research

---

51. See Saks, *supra* note 12, at 1120 (“[R]ather than the field convincing the courts of the sufficiency of its knowledge and skills, admission by the courts apparently convinced the forensic odontology community that, despite their doubts, they really were able to perform bite mark identifications.”).

52. Dr. Mary Bush and her colleagues at the Laboratory for Forensic Odontology, State University of New York at Buffalo, have published over a dozen studies that have undermined the assumptions underpinning bite mark evidence. *E.g.*, Mary A. Bush et al., *Statistical Evidence for the Similarity of the Human Dentition*, 56 J. FORENSIC SCI. 118, 122 (2011) (“Our results show that given our measurement parameters, statements concerning dental uniqueness with respect to bitemark analysis in an open population are unsupportable. . . . Confidence in the notion of dental uniqueness in bitemark analysis has been based on anecdotal knowledge, the use of inappropriate statistics, and precedence of admission in the courtroom.”). Bush and her team also reported the results of a study where twenty-three bites were made in cadaver skin with the same dentition using an instrumented-biting machine. The cadavers were moved and re-photographed in different positions. Subsequent measurements showed differences between all bite marks. In addition, postural distortion was significant. *See* Mary A. Bush et al., *Biomechanical Factors in Human Dermal Bitemarks in a Cadaver Model*, 54 J. FORENSIC SCI. 167, 169–170, 174 (2009). One survey of fifteen odontologists involved their opinions of six images of supposed bite marks. The “practitioner agreement was at best fair, with wide-ranging opinions on the origin, circumstance, and characteristics of the wound given for all six images.” Mark Page et al., *Expert Interpretation of Bitemark Injuries—A Contemporary Qualitative Study*, 58 J. FORENSIC SCI. 664, 664 (2013).

53. NAS FORENSIC REPORT, *supra* note 26, at 175.

54. 9 N.E.3d 1072 (Ohio Ct. App. 2014).

55. *Id.* at 1097.


57. *Id.* at 926.
mentioned above, and both prosecutors and their experts attacked researchers without offering any foundational research.

Unfortunately, the American Board of Forensic Odontology ("ABFO") has fiercely defended bite mark analysis. To bolster its position, the ABFO conducted a study that was presented at a forensic conference in 2015. As it turned out, the study undercut the ABFO's own position. Thirty-nine ABFO-certified bite mark experts—with an average of twenty years' experience—examined one hundred bite mark photographs. Each was asked three questions:

1. Is there sufficient evidence in the presented materials to render an opinion on whether the patterned injury is a human bite mark?

58. See, e.g., Prade, 9 N.E.3d at 1098 (“As to Dr. Bush’s cadaver studies, Dr. Wright testified that cadaver skin simply cannot compare with living skin. Dr. Wright explained that cadaver skin only distorts after a bite for two to three minutes at most because, unlike live skin, no bruising, contusions, or lacerations occur. Dr. Wright also testified that using a mechanical jaw to bite is problematic because the jaw operates on a fixed hinge that cannot mimic the wider range of movement that an actual jaw is capable of.”). But see Iain A. Pretty & David Sweet, A Paradigm Shift in the Analysis of Bitemarks, 201 FORENSIC SCI. INT’L 38, 40 (2010) (noting that, while cadaver models have limitations, “there is little alternative for researchers to produce bitemarks of known origin” and the use of anesthetized pigs to create peri-mortem injuries raises a different issue—i.e., differences between pigskin and human skin).


60. The study is known as Construct Validity Bitemark Assessments Using the ABFO Bitemark Decision Tree ("Freeman/Pretty Study"). See AM. ACAD. OF FORENSIC SCI., ADVANCE PROGRAM 67TH ANNUAL SCIENTIFIC MEETING: CELEBRATING THE FORENSIC SCIENCE FAMILY 175 (2015).

2. Is it a human bite mark, not a human bite mark, or suggestive of a human bite mark?

3. Does the bite mark have distinct, identifiable arches and individual tooth marks?

The results to the first question were not reassuring. The thirty-nine experts agreed unanimously in only four out of the one hundred cases. In only twenty cases was there 90 percent or more agreement. At the end of question two—whether the mark is a human bite mark—there were only sixteen cases with 90 percent or more agreement. At the end of the third question, there were only eight cases in which at least 90 percent of the analysts agreed. Equally disturbing was the ABFO’s decision to postpone publishing the results “until the organization can tweak the design of the study and conduct it again, a process that’s expected to take at least a year.” In effect, ABFO wanted a do-over. Meanwhile, an Associated Press analysis reported that at least twenty-four men convicted or charged with murder or rape based on bite marks have been exonerated since 2000.

2. Texas Forensic Science Commission (2016)

Steven Chaney spent twenty-eight years in prison for murder based largely on bite mark evidence. When his conviction was overturned, the Innocence Project filed a complaint on his behalf with the Texas Forensic Science Commission (“TFSC”). In 2016, after a six-month

62. Id.
63. Id.
64. Id.
65. Id.
66. Id.
67. Id.
investigation, the TFSC recommended a moratorium on the admission of bite mark testimony. It found that there is no scientific basis for claiming that a particular mark can be associated to a person’s dentition: “Any testimony describing human dentition as 'like a fingerprint' or incorporating similar analogies lacks scientific support.”71 Similarly, “there is no scientific basis for assigning probability or statistical weight to an association, regardless of whether such probability or weight is expressed numerically (e.g., 1 in a million) or using some form of verbal scale (e.g., highly likely/unlikely).”72

TFSC was also alarmed that the ABFO study was not published due to “political and organizational pressures.”73 In the Commission’s view, “such a resistance to publish scientific data contradicts the ethical and professional obligations of the profession as a whole, and is especially disconcerting when one considers the life and liberty interests at stake in criminal cases.”74


In September 2016, the White House released its report on forensic science.75 Regarding bite mark analysis, it concluded that (1) appropriately designed validation studies are lacking, (2) the few available studies had “very high” false-positive rates, (3) “inappropriate closed-set designs . . . are likely to underestimate the true false positive rate,” and (4) the studies show that experts “cannot even consistently agree

---

71. Id. at 12.
72. Id. at 13.
74. See generally WHITE HOUSE PCAST REPORT, supra note 38.
on whether an injury is a human bitemark.”76 Numerous cases support the last observation.77

In sum, the courts have yet to reject bite mark evidence—a subjective method that is not supported by foundational research and lacks agreed-upon standards.78 “Perhaps no discredited forensic assay has benefitted more from criminal courts’ abdication of gatekeeper responsibilities than bite mark analysis.”79 Instead, it was the Innocence Project that spearheaded the challenges in this area, and in 2016 the Texas Forensic Science Commission became the first governmental body to seriously scrutinize the technique. Notwithstanding the NAS, PCAST, and TFSC reports, courts continue to admit bite mark evidence.80

76. Id. at 9. “PCAST finds that bitemark analysis is far from meeting the scientific standards for foundational validity.” Id.

77. See, e.g., Ege v. Yukins, 380 F. Supp. 2d 852, 878 (E.D. Mich. 2005) (“[T]he defense attempted to rebut Dr. Warnick’s testimony with the testimony of other experts who opined that the mark on the victim’s cheek was the result of livor mortis and was not a bite mark at all.”); Czapleski v. Woodward, No. C-90-0847 MHP, 1991 WL 639360, at *1–2 (N.D. Cal. Aug. 30, 1991) (noting that, while a dentist’s initial report concluded that “bite” marks found on child were consistent with dental impressions of mother, several experts later established that the marks on child’s body were postmortem abrasion marks and not bite marks); Kinney v. State, 868 S.W.2d 463, 464–65 (Ark. 1994) (noting disagreement between expert witnesses about whether injuries were from human bite marks); People v. Noguera, 842 P.2d 1160, 1165 n.1 (Cal. 1992) (“At trial, extensive testimony by forensic odontologists [sic] was presented by both sides, pro and con, as to whether the wounds were human bite marks and, if so, when they were inflicted.”); State v. Duncan, 802 So. 2d 533, 553 (La. 2001) (“Both defense experts testified that these marks on the victim’s body were not bite marks.”); Stubbs v. State, 845 So. 2d 656, 668 (Miss. 2003) (“Dr. Galvez denied the impressions found on Williams were the results of bite marks.”).

78. See Michael J. Saks et al., Forensic Bitemark Identification: Weak Foundations, Exaggerated Claims, 3 J.L. & BIOSCIENCES 1, 29 (2016) (“[R]ecent reviews of the field’s claims, as well as recent empirical findings, have underscored the lack of reliability and validity of the most fundamental claims about the ability of forensic dentists to identify the source of bite marks on human skin.”).


At the April 10, 2017 meeting of the National Commission on Forensic Science, Keith Harward described how bite mark evidence resulted in his thirty-three years of imprisonment before he was exonerated by DNA evidence. Incredibly, the next day the chairman of the National District Attorneys Association stated that his organization believes that bite mark evidence is a “reliable science.”

B. Microscopic Hair Analysis

In this examination, samples are first analyzed to identify features visible to the naked eye such as color and form, i.e., whether it is straight, wavy, or curved. Next, the sample is viewed microscopically to determine characteristics such as shaft form, hair diameter, and pigment size.

Experts have long acknowledged that a positive identification is not possible with microscopic hair analysis. Instead, examiners testify that a crime scene exemplar was “consistent with” a hair sample from the defendant. The probative value of this conclusion would, of course, vary if only a hundred people had microscopically indistinguishable hair as opposed to several million. Due to a lack of research, no one knows whether the crime scene hair could have come from 10 other persons or 100, 10,000, and so forth.


83. See generally 2 Giannelli ET AL., supra note 20, § 24.02] (discussing the techniques used to identify the human source of a hair sample, including conventional microscopy).

84. As one hair examiner wrote, “[]if a pubic hair from the scene of a crime is found to be similar to those from a known source, [the courts] do not know whether the chances that it could have originated from another source are one in two or one in a billion.” B.D. Gaudette, Probabilities and Human Pubic Hair Comparisons, 21 J. Forensic Sci. 514, 514 (1976).
often omitted from the experts’ testimony, thus making marginal evidence appear misleadingly convincing.\footnote{85}

Experts frequently went way beyond the “consistent with” language in their testimony, however, often suggesting a rare association. For example, in the Edward Honaker case, the expert testified that the crime scene hair sample “was unlikely to match anyone” other than the defendant.\footnote{86} Honaker spent ten years in prison before DNA proved him innocent.\footnote{87} In another case, an expert testified that matching hair samples were “consistent microscopically” but then elaborated: “In other words, hairs are not an absolute identification, but they either came from this individual or there is—could be another individual somewhere in the world that would have the same characteristics to their hair.”\footnote{88} This is an implicit—and extreme—probability statement that lacks any empirical support.

Although microscopic hair analysis had long been judicially accepted,\footnote{89} its validity was suspect.\footnote{90} In 1995, a federal district court in

85. Professor Berger explained the problem:

We allow eyewitnesses to testify that the person fleeing the scene wore a yellow jacket and permit proof that a defendant owned a yellow jacket without establishing the background rate of yellow jackets in the community. Jurors understand, however, that others than the accused own yellow jackets. When experts testify about samples matching in every respect, the jurors may be oblivious to the probability concerns if no background rate is offered, or may be unduly prejudiced or confused if the probability of a match is confused with the probability of guilt, or if a background rate is offered that does not have an adequate scientific foundation.

Berger, supra note 22, at 1357.


87. Id. at 59.


89. See Edward J. Imwinkelried, Forensic Hair Analysis: The Case Against the Underemployment of Scientific Evidence, 39 Wash. & Lee L. Rev. 41, 62 (1982) (stating that “[t]he massive body of case law, liberally admitting even hair evidence of low probative value, dwarfs the handful of cases excluding hair evidence”).

90. See Clive A. Stafford Smith & Patrick D. Goodman, Forensic Hair Comparison Analysis: Nineteenth Century Science or Twentieth Century Snake Oil?, 27 Colum. Hum. Rts. L. Rev. 227, 231 (1996) (“If the purveyors of this dubious science cannot do a better job of validating hair analysis than they have done so far, forensic hair comparison analysis should be excluded altogether from criminal trials.”).
Williamson v. Reynolds\textsuperscript{91} observed: “Although the hair expert may have followed procedures accepted in the community of hair experts, the human hair comparison results in this case were, nonetheless, scientifically unreliable.”\textsuperscript{92} The court also noted that the “expert did not explain which of the ‘approximately’ 25 characteristics were consistent, any standards for determining whether the samples were consistent, how many persons could be expected to share this same combination of characteristics, or how he arrived at his conclusions.”\textsuperscript{93} Williamson, who was five days from execution when the district court issued a stay, was subsequently exonerated by DNA testing.\textsuperscript{94}

The Williamson opinion—perhaps the only thorough judicial analysis of microscopic hair comparisons—was all but ignored by other courts. In 1999 in Johnson v. Commonwealth,\textsuperscript{95} the Kentucky Supreme Court upheld the admissibility of hair evidence, taking “judicial notice” of its reliability\textsuperscript{96} and thus implicitly finding its validity indisputable.\textsuperscript{97} Other courts echoed Johnson, not Williamson.\textsuperscript{98} Indeed, ten years after Williamson was decided, a 2005 decision by the Connecticut Supreme Court observed—correctly—that “[t]he overwhelming majority of courts have deemed such evidence admissible.”\textsuperscript{99}

Once again, the courts abdicated their responsibility. Indeed, hair evidence only began to be carefully scrutinized after a startling number of DNA exonerations were reported.\textsuperscript{100} A 2008 study of 200 DNA

\footnotesize
\begin{itemize}
  \item 91. 904 F. Supp. 1529 (E.D. Okla. 1995), rev’d sub nom. Williamson v. Ward, 110 F.3d 1508, 1523 (10th Cir. 1997).
  \item 92. Id. at 1558.
  \item 93. Id. at 1554.
  \item 94. See Jim Dwyer et al., Actual Innocence: Five Days to Execution and Other Dispatches from the Wrongly Convicted 146 (2000) (noting that the hair evidence was shown to be “patently unreliable”). See also John Grisham, The Innocent Man: Murder and Injustice in a Small Town 166–87 (2006) (examining Williamson’s trial, including the role played by hair analysis).
  \item 95. 12 S.W.3d 258 (Ky. 1999).
  \item 96. Id. at 263.
  \item 97. See Fed. R. Evid. 201(b) (limiting judicial notice to a “fact that is not subject to reasonable dispute”).
  \item 98. See 2 Giannelli et al., supra note 20, § 24.03, at 825 (noting the “limited impact of Daubert”).
  \item 99. State v. West, 877 A.2d 787, 808 (Conn. 2005).
  \item 100. In 1998, a Canadian judicial inquiry into the wrongful conviction of Guy Paul Morin was released. Morin’s original conviction was based, in part, on hair evidence. The judge conducting the inquiry recommended that “[t]rial judges should undertake a more critical analysis of the admissibility of hair comparison evidence as circumstantial evidence of guilt.” Fred Kaufman, Ministry of the Attorney General, Report of the Kaufman Commission on Proceedings Involving Guy Paul Morin 312 (1998);}

885
Forensic Science: Daubert’s Failure

Exonerations found that forensic evidence was the second leading type of evidence, at 57 percent—after eyewitness identifications at 79 percent—used in wrongful conviction cases.\(^\text{101}\) A subsequent investigation of trial transcripts underscored the role of hair analysis in the exoneration cases: “Of the 65 cases involving microscopic hair comparison in which transcripts were located, 25 cases, or 38%, had invalid forensic science testimony.”\(^\text{102}\) The 2009 NAS Report observed that “testimony linking microscopic hair analysis with particular defendants is highly unreliable.”\(^\text{103}\)

1. FBI Hair Review

In April 2013, the Mississippi Supreme Court, in a five-to-four decision, rejected Willie Jerome Manning’s request for a stay of execution to permit DNA testing—“potentially setting up what experts said would be a rare case in recent years in which a person is put to death with such requests unmet.”\(^\text{104}\) A week later, the court unexpectedly stayed Manning’s execution after the Department of Justice (“DOJ”) notified state officials that FBI experts had presented misleading testimony at his trial, including hair and firearms evidence.\(^\text{105}\)

Soon after, the DOJ announced that Manning was but one of 120 cases—including twenty-seven death penalty prosecutions—in which improper microscopic hair analysis had been introduced in evidence.\(^\text{106}\)

---


\(^\text{103}\) NAS FORENSIC REPORT, supra note 26, at 161.


For example, examiners claimed to connect a hair sample to a single person “to the exclusion of all others” or stated or suggested a probability for such a match from past casework.\(^\text{107}\) The FBI review came after three District of Columbia men, who had been convicted of rape or murder in the early 1980s, were exonerated through DNA testing.\(^\text{108}\) In one of these cases, the prosecutor claimed that, based on an FBI expert’s testimony, the chances that the sample came from someone else were “one in 10 million.”\(^\text{109}\)

After further investigation, DOJ reported in 2015 that “FBI examiners had provided scientifically invalid testimony in more than 95 percent of cases where that testimony was used to inculpate a defendant at trial.”\(^\text{110}\) *Commonwealth v. Perrot*\(^\text{111}\) was one of the first cases to reach the courtroom as a consequence of the DOJ review. A superior court granted Perrot a new trial in 2016, criticizing the misleading use of hair evidence. The court noted: “In discussing the ‘microscopic characteristics’ of hair, [the expert] stated that these characteristics ‘make that hair somewhat unique.’ He likened the ‘subtle’ characteristics of
hair that ‘make it somewhat unique’ to the subtle differences in a human face.”


In June 2016, the DOJ released proposed guidelines concerning hair testimony. Documentation purporting to support the validity and reliability of hair evidence accompanied the guidelines.113 Listing several studies, the FBI concluded:

Based on these and other published studies, microscopic hair comparison has been demonstrated to be a valid and reliable scientific methodology. These studies have also shown that microscopic hair comparisons alone cannot lead to personal identification and it is crucial that this limitation be conveyed both in the written report and in testimony.114

The White House PCAST Report, however, challenged the supporting documentation, which discussed only a handful of studies from the 1970s and 1980s but did not comment on subsequent studies that found “substantial flaws in the methodology and results of the key papers.”115 Moreover, “PCAST’s own review of the cited papers [found] that these studies do not establish the foundational validity and reliability of hair analysis.”116

112. Id. at *32. The court also noted:

[The expert] asserted that the hairs “matched” and showed a “strong association.” In discussing the chance that the hair found on the victim’s bed came from someone other than Perrot, [the expert] conceded the possibility, adding that during his ten years of experience “it’s extremely rare that I will have known hair samples from two different people that I can’t tell apart.” [The expert] made these statements of confidence, despite being unable to recall at trial the length or diameter of the one hair found on the bed.

Id.


115. WHITE HOUSE PCAST REPORT, supra note 38, at 13.

116. Id. DOJ cited Max M. Houck & Bruce Budowle, Correlation of Microscopic and Mitochondrial DNA Hair Comparisons, 47 J. Forensic Scis. 964 (2002). DOJ SUPPORTING DOCUMENTATION, supra note 114, at 8. This FBI
The bottom line, again, is the judiciary’s dereliction in failing to curb the misuse of hair microscopy testimony. The Innocence Project’s track record of DNA exonerations brought this issue to the fore. Indeed, the three exonerations in the District of Columbia triggered the FBI review. Yet, DOJ’s proposed guidelines were based on “foundational research” that PCAST questioned.

C. Arson Investigations

For decades, arson investigators came from the “old school” of investigators—those who used intuition and a number of rules of thumb to determine whether a fire was incendiary. Critics complained that instead of being rooted in science, the approach was based on folklore that had been passed down from generation to generation—without any empirical testing. A government report noted, as early as 1977, that common arson indicators had “received little or no scientific testing” and that “[t]here appears to be no published material in the scientific literature to substantiate their validity.” Proponents of a science-based approach to arson investigations waged an uphill battle through the 1980s, finally winning a major victory when the National Fire Protection Association (“NFPA”) published its Guide for Fire and Explosion Investigations (“NFPA 921”) in 1992.

1. Willingham Case

Although NFPA 921 would later become the bible for arson investigations, Cameron Todd Willingham was convicted for the
arson-murders of his young children weeks before the guide was published. Willingham, who was executed twelve years later, is the poster boy for junk science in arson investigations.

At trial, Deputy Fire Marshall Vasquez testified that “[t]he fire tells a story. I am just the interpreter . . . . And the fire does not lie. It tells me the truth.” He also testified that he had found numerous so-called “indicators” for arson during his post-fire investigation of Willingham’s house. One such indicator was a low burning fire. He told the jury that “[a]ll fire goes up,” and thus, burn patterns on the floor and lower walls suggested that an accelerant was used. This reasonable

has “become the de facto national standard for fire scene examination and analysis.”

121. See Frontline: Death by Fire (PBS television broadcast Oct. 19, 2010) (detailing the case of Cameron Todd Willingham); David Grann, Trial by Fire: Did Texas Execute an Innocent Man?, NEW YORKER (Sept. 7, 2009), https://www.newyorker.com/magazine/2009/09/07/trial-by-fire [https://perma.cc/Z5PU-P7BS] (noting that after Willingham’s execution, the Innocence Project commissioned a panel of fire experts that, after reviewing the evidence supporting the conviction, “concluded that ‘each and every one’ of the indicators of arson had been ‘scientifically proven to be invalid’”); Hall, supra note 70 (“The 893-page report [on the Willingham case], released in April 2011, was anticlimactic for people looking for proof that Texas had executed an innocent man.”); Steve Mills & Maurice Possley, Texas Man Executed on Disproved Forensics: Fire that Killed His 3 Children Could Have Been Accidental, CHI. TRIB. (Dec. 9, 2004), http://articles.chicagotribune.com/2004-12-09/news/0412090169_1_cameron-todd-willingham-arson-fire-fire-scene [https://perma.cc/2CLW-25SK] (“[The fire investigators] used rules of thumb that have since been shown to be false. There was no evidence to support a conclusion that the fire was intentionally set. Just an unsupported opinion.”).


123. See id. at 224-68 (recording the direct examination of Manuel Vasquez). A second expert’s testimony essentially tracked Vasquez’s. See id. at 156-85 (recording the testimony of Douglas Fogg).

124. Id. at 248. Vasquez testified that there was “char burning, like, for example, this is the bottom here. It’s burned down here at the bottom. That is an indicator in my investigation of an origin of fire because it’s the lowest part of the fire.” Id. at 239; see also Willingham, 897 S.W.2d at 354 (“An expert witness for the State testified that the floors, front threshold, and front concrete porch were burned, which only occurs when an accelerant has been used to purposely burn these areas. This witness further testified that this igniting of the floors and thresholds is typically employed to impede firemen in their rescue attempts.”).


126. Id. at 256 (“So when I found that the floor is hotter than the ceiling, that’s backwards, upside down. It shouldn’t be like that. The only reason that the
notion, however, has its limitations—especially when a fire occurs in a contained area, such as a house with its windows shut:

Due to buoyancy, a thermal plume initially rises once a fire is ignited. As the fire continues, the plume reaches the ceiling, which causes it to spread outward towards the walls. When it reaches the walls, the combustion products press down from the ceiling creating an upper level, which continues to increase in depth and temperature. Eventually thermal radiation replaces convection as the principal method of heat transfer.\(^\text{127}\)

At this point, every combustible surface in the room will spontaneously burst into flames. This transition phenomenon, known as the onset of “flashover,” can occur within minutes.\(^\text{128}\) After flashover, the entire room is burning, including the lower walls and floor. Flashover has been described as the point at which the fire transitions from a “fire in a room” to a “room on fire.”\(^\text{129}\) At trial, prosecution witnesses acknowledged that there was an explosion, which could be explained by flashover.\(^\text{130}\) Consequently, a low burning fire is not necessarily indicative of the use of an accelerant.\(^\text{131}\)

Moreover, many of Vasquez’s other “indicators”—including what he called “pour patterns” and “puddle configurations,”\(^\text{132}\) which appear as splotchy areas on the floor—can appear after flashover in an accidental fire.\(^\text{133}\) Additional indicators, such as alligatoring—large shiny

\(^\text{127}\) Giannelli, \textit{supra} note 117, at 227.

\(^\text{128}\) \textit{Id.}\n
\(^\text{129}\) \textit{Lentini, supra} note 117, at 77.

\(^\text{130}\) \textit{See Willingham Transcript, supra} note 122, at vol. XI, 75 (testimony of Mary Diane Barbe) (“The windows, the electricity started crackling and popping, and the top of the—well, I was facing the side of the house, and it just blew out. The flames just blew out . . . . All the windows and the front room was engulfed.”); \textit{id.} at 96 (testimony of Brandy Barbe) (“We was running towards the house, me and my mother, we was fixing to go and try to get in, and that’s when it was an explosion . . . .”). Vasquez mentioned flashover in his testimony, but he did not appear to understand its implications. \textit{See id.} at vol XII, 47–48.

\(^\text{131}\) Paul C. Giannelli & Kimberly Gawel, \textit{Arson Evidence}, 47 C\textsc{rim.} L. \textsc{ Bull.} 1241, 1250–51 (2011) (identifying “flashover” as an alternative reason for “low burning” fires).

\(^\text{132}\) \textit{Willingham Transcript, supra} note 122, at vol. XI, 244–45 (“You can see that on the burnt patterns on this puddle configuration on Exhibit No. 36. This is a strong indicator of a liquid. . . . [The sunlight] just lights up the puddle configurations, the burnt trailers, the pour patterns on that floor.”).

\(^\text{133}\) Giannelli & Gawel, \textit{supra} note 131, at 1242–44.
charred blisters on burned wood—can also be explained by flashover.\textsuperscript{134} The flashover phenomenon also accounts for another fact that Vasquez thought incriminatory. Willingham told investigators that he had attempted to save his daughters, but he was forced to run from his home without shoes because the heat was too great. According to Vasquez, the burn debris on the floor made it impossible that Willingham would not have had burns on his feet.\textsuperscript{135} However, Willingham’s feet would not have been burnt if he left his home before flashover.\textsuperscript{136}

Charring under an aluminum threshold of an interior door provided another clue according to Vasquez.\textsuperscript{137} But again, this “indicator” may occur in a flashover. Other perceived “indicators”—melted bed springs,\textsuperscript{138} multiple points of origins,\textsuperscript{139} and brown stains on a concrete floor—\textsuperscript{140} were also consistent with an accidental fire.\textsuperscript{141} Finally, Vasquez relied on the presence of “crazed glass”—spider-web patterns on the

\begin{itemize}
\item \textsuperscript{134} Id. at 1246.
\item \textsuperscript{135} “There was fire on the floor . . . . He had no injuries on his feet.” Willingham Transcript, \textit{supra} note 122, at 267.
\item \textsuperscript{136} Giannelli, \textit{supra} note 127, at 228.
\item \textsuperscript{137} Willingham Transcript, \textit{supra} note 122, at 251.
\item \textsuperscript{138} “[T]he springs were burned from underneath. This indicates there was a fire under this bed because of the burn underneath the bed.” \textit{Id.} at 241.
\item \textsuperscript{139} \textit{Id.} at 255 (“Multiple areas of origin indicate—especially if there is no connecting path, that they were intentionally set by human hands.”). There are two problems here. First, there could have been one origin, according to independent experts. \textit{DOUGLAS J. CARPENTER ET AL., REPORT ON THE PEER REVIEW OF THE EXPERT TESTIMONY IN THE CASES OF STATE OF TEXAS V. CAMERON TODD WILLINGHAM AND STATE OF TEXAS V. ERNEST RAY WILLIS} 12 (2006). Second, even if the fire scene had shown multiple points of origin, this would not necessarily indicate an intentional fire. \textit{LENTINI, supra} note 117, at 513–14.
\item \textsuperscript{140} Willingham Transcript, \textit{supra} note 122, at vol. XI, 248–49. Fire experts reviewing the evidence from Willingham’s trial pointed out that “[t]he behavior of concrete in fires, including the development of various colors, has been extensively studied.” \textit{CARPENTER ET AL., supra} note 139, at 18. These experts concluded that there is simply “no scientific basis for Mr. Vasquez's statement about the brown discoloration being an indication of the presence of accelerants.” \textit{Id.}
\item \textsuperscript{141} Vasquez’s testimony also demonstrated other misconceptions. A common one is that arson fires burn hotter and faster than “normal” fires: “You know, [an accelerant] makes the fire hotter. It’s not a normal fire.” Willingham Transcript, \textit{supra} note 122, at vol. XI, 249. However, the temperature of burning wood and burning gasoline are nearly identical, so to claim that a fire using liquid accelerants burns “hotter” than a wood fire is wrong. \textit{LENTINI, supra} note 117, at 501–02.
\end{itemize}
windows—as an indication of arson.\textsuperscript{142} Arson investigators long believed that crazed glass resulted from a fire that burned fast and hot and that the presence of crazed glass indicated that the fire was fueled by a liquid accelerant. Yet, subsequent research demonstrated that crazing occurs from rapid cooling—rather than rapid heating—when water from fire hoses is sprayed on heated windows.\textsuperscript{143}

In retrospect, one of the numerous debris samples submitted for laboratory analysis contributed the most damning piece of evidence.\textsuperscript{144} The debris sample—collected from an area near the front door—was the \textit{only} sample that tested positive for a chemical commonly used in charcoal lighter fluids.\textsuperscript{145} This finding can be explained by the fact that a charcoal grill and lighter fluid were on the front porch at the time of the fire.\textsuperscript{146} In fact, that the other samples yielded negative results supported Willingham’s case.\textsuperscript{147}

Numerous nationally recognized experts reviewed the arson testimony presented at Willingham’s trial and found it seriously flawed. The first examination of the record by an independent expert was part of Willingham’s petition for habeas corpus and was also submitted to the governor and the Board of Pardons and Parole days before Willingham’s execution. It concluded: “On first reading, a contemporary fire origin and cause analyst might well wonder how anyone

\begin{itemize}
  \item \textsuperscript{142} “The pieces of broken window glass on the ledge of the north windows to the northeast bedroom disclosed a crazed ‘spider webbing’ condition. This condition is an indication that the fire burned fast and hot.” \textsc{Carpenter et al., supra note 139}, at 18 (citing Vasquez’s written report on the Willingham fire at 4).
  
  \item \textsuperscript{143} \textsc{Lentini, supra note 117}, at 478 (“It is unclear why anyone ever thought that crazing of glass indicated rapid heating.”).
  
  \item \textsuperscript{144} In closing argument, the defense counsel referred to a “dozen samples.” \textsc{Willingham Transcript supra note 122}, at vol. XIII, 20.
  
  \item \textsuperscript{145} \textit{Id.} at vol. XI, 220–21 (documenting testimony by expert stating there was “no distinguishing characteristic” in the charcoal lighter fluid from plastic container on porch “that was not present” in the front-door threshold sample tested); \textit{id.} at vol. XIII, 20–21 (“They sent [the samples] to the lab and what did they find? Nothing, not a trace of anything at all except on the very front of the front porch where the charcoal lighter fluid was.”); \textit{id.} at vol. XIII, 45 (“The [accelerant] is gone, except [on] the threshold; it burned away . . . ”).
  
  \item \textsuperscript{146} \textit{Id.} at vol. XII, 14–15 (noting that, although photographs show a grill, Vasquez apparently did not know of the grill’s presence); \textit{id.} at vol. XII, 16 (acknowledging that a fire-damaged charcoal lighter fluid container was found on the front porch).
  
  \item \textsuperscript{147} The prosecutor would later say that he “‘never did understand why they weren’t able to recover’ positive tests in these parts.” \textsc{Grann, supra note 121}. At trial, he argued that, except in the threshold, the “liquid . . . burned away in that destructive madness created by Cameron Todd Willingham.” \textsc{Willingham Transcript, supra note 122}, at vol. XIII, 45.
\end{itemize}
could make so many critical errors in interpreting the evidence.”

Nevertheless, a stay was denied, and Willingham was put to death. Subsequent evaluators agreed that the trial evidence was junk science. For example, five independent experts prepared a forty-three-page report, finding that “each and every one of the indicators relied upon have since been scientifically proven to be invalid.”

In May 2006, the Innocence Project petitioned the Texas Forensic Science Commission to review the arson testimony in Willingham’s and Ernest Ray Willis’ cases. The TFSC is not authorized to determine guilt or innocence. Instead, the Innocence Project argued that the State Fire Marshall Office should have reinvestigated arson cases in which its experts testified after NFPA 921 was published in 1992—a full twelve years before Willingham’s execution. TFSC retained its own independent consultant, Dr. Craig Beyler, another nationally-recognized expert, to review the arson evidence. His fifty-one-page report dissected the expert testimony, concluding:

The investigations of the Willis and Willingham fires did not comport with either the modern standard of care expressed by NFPA 921, or the standard of care expressed by fire investigation texts and papers in the period 1980–1992. The investigators had poor understandings of fire science and failed to acknowledge or apply the contemporaneous understanding of the limitations of fire indicators. Their methodologies did not comport with the scientific method or the process of elimination. A finding of arson could not be sustained based upon the standard of care expressed by NFPA 921, or the standard of care expressed by fire investigation texts and papers in the period 1980–1992.


149. Carpenter et al., supra note 139, at 3.

150. The expert evidence in both cases was comparable, but Willis was lucky. His death penalty conviction was overturned on procedural grounds, and the prosecutor subsequently refused to reindict him after Dr. Hurst wrote the same type of critical report in Willis’s case that he had written in Willingham’s. Willis, who had spent seventeen years on death row, was subsequently exonerated on actual innocence grounds. See Mary Alice Robbins, New-York Based Innocence Project Attacks Texas Arson Convictions, Tex. Law., May 8, 2006, Factiva, Doc. No. TEXASL0020060508e2580000l.


Once Beyler’s report became public, a political firestorm erupted, and the governor, who was in the midst of a re-election battle, abruptly replaced commission members two days before a meeting was scheduled to consider the Beyler report.153 The newly appointed chair, a prosecutor, promptly cancelled the meeting,154 raising the specter of a cover-up.155 Next, the Attorney General issued an opinion finding that the TFSC was prohibited “from considering or evaluating specific items of evidence that were tested or offered into evidence prior to [its creation in 2005].”156

The TFSC eventually produced a report—one that did not directly deal with the Willingham and Willis cases. Nevertheless, the Report’s recommendations and statements indicated that the Willingham arson


154. Hoppe, supra note 153 (noting that the new chair was “known as one of the toughest law-and-order prosecutors in the state”).

155. See Jennifer Emily, Texas Forensic Science Commission Refuses to End Inquiry into Willingham Arson Case, DALL. MORNING NEWS (Sept. 18, 2010), https://www.dallasnews.com/news/crime/2010/09/18/Texas-Forensic-Science-Commission-refuses-to-5315 [https://perma.cc/3RP5-GSH6] (“Perry’s replacements were seen by some as a political maneuver intended to change the outcome of the commission’s decision.”); Christy Hoppe, Perry Ousts Officials Before Arson Hearing: He’s Assailed as New Chair Delays Session on Flawed Case that Led to Execution, DALL. MORNING NEWS, Oct. 1, 2009, at 1A; Dave Mann, Fire and Innocence, TEX. OBSERVER (Dec. 3, 2009), https://www.texasobserver.org/fire-and-innocence/ [https://perma.cc/SB7T-AGYB] (“Then in late September, Perry booted three members off of the Texas Forensic Science Commission, which was investigating the Willingham and Willis cases, just three days before a crucial hearing on scientists’ findings. Perry’s new appointees promptly canceled the hearing and have yet to reschedule it. Even conservative commentators cried cover-up, suggesting that Perry, in a tough battle for re-election, was trying to subvert an investigation that might prove he oversaw the execution of an innocent man.”).

investigation was seriously flawed. Its first recommendation was “that fire investigators adhere to the standards of NFPA 921.” In addition, the report reviewed a number of arson indicators that were used in the Willingham and Willis cases. Citing Vasquez’s testimony, the report undermined his opinions concerning (1) V-patterns as an indicator of origin, (2) pour patterns, (3) low/deep burning, (4) multiple separate points of origin, (5) spalling, (6) burn intensity, and (7) crazed glass. It also observed that “testimony, such as Vasquez’s response to a question regarding Willingham’s state of mind, is an example of the type of testimony that experts should avoid as falling outside of their field of expertise.” The report even encouraged lawyers to “aggressively pursue admissibility hearings in arson cases.”

Despite the opinions of all the independent experts, the State Fire Marshal vigorously defended its investigation. In a breathtaking letter, the office asserted that “[i]n reviewing documents and standards in place then and now, we stand by the original investigator’s report and conclusions.” This left the TFSC incredulous.

2. Han Tak Lee Case

Unfortunately, Willingham’s case was not an outlier. In the 1989 trial of Han Tak Lee, the expert also relied on the old “myths” to declare the fire incendiary: (1) greater intensity and heat, (2) burn patterns, (3) alligatoring, (4) melted metal in bed frames, and (5) crazed glass.

158. Id. at 22–28.
159. Id. at 36.
160. Id. at 48.
162. TFSC Report, supra note 157, at 16 (“This appears to be an untenable position in light of advances in fire science. The fires in these cases occurred two decades ago; there are few circumstances in which an investigation could not be improved with the benefit of twenty years of controlled scientific experiment and practical experience.”).
163. Lee petitioned for a writ of habeas corpus in 2010, based in part on “inaccurate and unreliable evidence.” Lee v. Tennis, No. 4:CV-08-1972, 2010 WL 3812160, at *2 (M.D. Pa. Sept. 22, 2010). Although the district court denied Lee’s petition, the Third Circuit reversed and remanded the case to the district court. Lee v. Glunt, 667 F.3d 397, 407–08 (3d Cir. 2012) (“If Lee’s expert’s independent analysis of the fire scene evidence—applying principles from new developments in fire science—shows that the fire expert testimony at Lee’s trial was fundamentally unreliable, then Lee will be entitled to federal habeas relief on his due process claim.”).
In addition, the investigation was “hobbled by an incomplete and inaccurate understanding” of flashover. After serving twenty-five years, Lee was released from prison in 2014.


After the publication of NFPA 921 in 1992, the kind of testimony presented in the Willingham and Lee cases should have vanished from the courtroom. But arson investigators balked. According to one expert, “[t]he initial response to NFPA 921 in the fire investigation community was overwhelmingly negative.” Babick v. Berghuis is illustrative. In that case, Andrew Babick was convicted of arson-murder for a 1995 house fire and was sentenced to two terms of life imprisonment without the possibility of parole. He later sought habeas relief, claiming ineffective assistance of counsel and prosecutorial misconduct. In 2010, the Sixth Circuit rejected these claims.

In dissent, however, Judge Merritt chastised the defense attorney for not contesting the arson evidence in “this strange junk science case.” One prosecution expert testified that: (1) char marks on the porch were evidence of an accelerant, (2) a “line of demarcation” in a burn pattern on a carpet was “suspicious” because “it should not have burned the carpeting on these jagged edges,” and (3) the burns were “not normal” and were “unnatural.” Another prosecution expert

165. Id.
167. Id. at 40.
168. 620 F.3d 571 (6th Cir. 2010). See generally Marc Price Wolf, Habeas Relief from Bad Science: Does Federal Habeas Corpus Provide Relief for Prisoners Possibly Convicted on Misunderstood Fire Science?, 10 MINN. J.L. SCI. & TECH. 213 (2009) (detailing the shift of the principles in the field of fire investigation and how past convictions may use habeas relief).
169. Id. at 574.
170. Id. at 580.
171. 620 F.3d at 582-83 (Merritt, J., dissenting).
172. Id. at 581 (quoting transcript).
stated that “low burning” and other “unnatural” patterns indicated the presence of an accelerant. 173 Both experts “testified—in direct contrast to the NFPA guide—that they were so confident in their reading of burn patterns that the absence of any laboratory confirmation of accelerant had no effect on their testimony.” 174

4. Dog-Sniff Evidence

More alarming, in Judge Merritt’s view, was dog-sniff evidence. 175 The NFPA guide provides: “Research has shown that canines have been alerted to pyrolysis products that are not produced by an ignitable liquid” and a positive canine alert without laboratory confirmation “should not be considered validated.” 176 The lab tests had not detected accelerants in the house debris. Yet, a dog handler testified that “his dog, Samantha, was ‘1000 times’ more effective at detecting fire starters or liquid accelerants than a laboratory test on burnt material.” 177 In short, the “jury was misled into trusting Samantha over the arson forensic lab.” 178

A more recent arson-dog case involved James Hebshie, who was convicted of arson and mail fraud in 2006. A federal district court granted his habeas petition based on ineffective assistance of counsel grounds. 179 In the court’s view, if a Daubert hearing had been requested on the canine evidence, there was a “‘reasonable probability’ that the Court would have excluded the canine testimony or severely limited it.” 180 Without a challenge from the defense, the dog handler testified

---

173. Id.
174. Id.
175. Id. at 580.
176. NFPA 921, supra note 119, § 17.5.4.7 (describing the role of canine investigations as “assisting with the location and collection of samples” for laboratory testing).
177. Babick, 620 F.3d at 580.
178. Id. at 581. See also United States v. Myers, No. 3:10-00039, 2010 WL 2723196, at *3–4 (S.D.W. Va. July 8, 2010) (granting motion in limine to prohibit expert testimony of a canine handler because the alert had not been confirmed by lab testing, conflicted with the Fire Guide, and did not meet the Daubert standards).
180. Id. at 124 (quoting Strickland v. Washington, 466 U.S. 668, 694 (1984)).
that his dog, Billy, “was 97% accurate.”\footnote{181} Indeed, the handler testified to “an almost mystical account of Billy’s powers and her unique olfactory capabilities.”\footnote{182} The court explained:

[The handler] went on and on about what he understood about Billy, as if his relationship with Billy somehow enhanced the reliability and probative value of the results—that she was unique, that he could “read her face,” that he was with her 365 days a year, that he knew her personality, “the way her eyes shifted,” the ways her ear shifted, etc.\footnote{183}

Moreover, the handler focused on one area as the origin of the fire and testified that the dog had not alerted anywhere else on the premises. However, the handler had limited the dog’s access to that one area.\footnote{184} In addition, a dog’s failure to alert has no evidential value: “[T]he scientific literature cast doubt on the significance of the dog’s failure to alert (false negatives) and even raised concerns about canine ‘proficiency’ testing, concerns counsel never raised.”\footnote{185} Indeed, the term “accelerant-detection” dog was misleading because the dog is trained to alert to many common materials that are not accelerants; the site of the fire was a convenience store which sold lighter fluid and lighters.\footnote{186}

5. Post-Daubert Cases

The courts’ response to bogus arson evidence is mixed.\footnote{187} It is not hard to find cases citing discredited arson indicators after Daubert, such
as pour patterns or puddle configurations,\textsuperscript{188} melted bedsprings,\textsuperscript{189} concrete spalling,\textsuperscript{190} fire load,\textsuperscript{191} and “fast and hot” burn.\textsuperscript{192} Decided in 1998, \textit{Michigan Millers Mutual Insurance Corp. v. Benfield}\textsuperscript{193} is considered the “first serious challenge to the ‘old school’ of fire investigators.”\textsuperscript{194} In that case, the Eleventh Circuit ruled that arson testimony “is subject

\begin{itemize}
\item[188.] See, e.g., State v. Allen, No. 22835, 2009 WL 2096295, ¶ 114 (Ohio Ct. App. July 17, 2009) (noting that investigator testified to “an irregular burn pattern on the floor which through all my experience and training it appears to be an irregular pour pattern [sic], an ignitable liquid pour pattern”); State v. Wolf, 891 N.E.2d 358, 360 (Ohio Ct. App. 2008) (noting that a firefighter testified “he observed ‘pour patterns’ located on the floor throughout the mobile home; that pour patterns are burn marks that look like puddles that result from ignitable liquids . . . being poured out of containers . . .”); Colburn v. State, 990 So. 2d 206, 209–10 (Miss. Ct. App. 2008) (“This pour pattern, [the fire investigator] explained, was indicative of flammable liquid being poured in the area . . . . On cross-examination [the fire investigator] did admit that the State Crime Laboratory was unable to identify ignitable liquids in the three debris samples taken from the pour pattern area.”); State v. Henderson, 125 P.3d 1132, 1137 (Mont. 2005) (finding that the trial court “did not err in allowing [a firefighter] to identify in the photographs and diagrams the pour patterns he had observed at the scene”).


\item[190.] See, e.g., State v. Amodio, 915 A.2d 569, 576 (N.J. Super. Ct. App. Div. 2007) (“They washed the floor and observed areas of spalling in the concrete underneath the door. This was an indication that a flammable liquid had been employed in that area.”); McCord v. Gulf Guar. Life Ins. Co., 698 So. 2d 89, 95 (Miss. 1997) (“The arson investigator . . . testified that he found five different areas of spalling and concluded arson to be the cause of the fire.”).

\item[191.] See, e.g., Wise v. State, 719 N.E.2d 1192, 1200 (Ind. 1999) (noting that a fire investigator testified that a fire was intentionally set based on several factors, including that “the fire burned too fast for its fuel load”); Carter v. State, 516 S.E.2d 556, 560 (Ga. Ct. App. 1999) (“[T]he arson investigator deduced there must have been an accelerant or some kind of extra fuel load.”).

\item[192.] See, e.g., People v. Klaft, No. 289522, 2010 WL 2076956, at *5 (Mich. Ct. App. May 25, 2010) (“[B]oth [investigators] testified that they believed, based on the fast and hot nature of the fire, that it was set intentionally.”); State v. Walters, 813 P.2d 857, 858 (Idaho 1990) (noting that a fire investigator testified that “it was a hot, fast fire as opposed to a small or as opposed to a slow, smoldering fire, yes, the evidence suggests to me that it was deliberately set”); State v. Cutlip, No. 99-L-149, 2001 WL 687493, at *2 (Ohio Ct. App. 2001) (noting that a fire department lieutenant testified to a list of factors including that “the fire was fast and hot” and “that such observations are typical of a fire started by someone pouring an accelerant and fighting it”).

\item[193.] 140 F.3d 915 (11th Cir. 1998).


\end{itemize}
to Daubert’s inquiry regarding the reliability of such testimony.” 195 Some federal courts after Benfield cited NFPA 921. 196 Yet, a 2011 article on the subject began with the passage: “Fire researchers have shattered dozens of arson myths in recent years. So why do American courts still lag behind?” 197 And a 2013 survey of 586 public sector fire investigators found that some myths endure: “Nearly 40 percent did not know that crazed glass is caused by rapid cooling, not rapid heating. Twenty-three percent think puddle-shaped burns indicate the use of an accelerant. Eight percent still believe that alligator blistering implies that a fire burned fast and hot.” 198

The TFSC Report did more than the courts to curb flawed arson testimony. And it took the execution of an innocent man to trigger that report. 199 In addition, the resistance to change is all-too-familiar: Rules based on science “were slow to take hold, as veteran investigators clung to what now are considered disproven theories. In some police and fire departments, investigators were openly hostile to the updated science.” 200

D. Comparative Bullet Lead Analysis

For more than three decades, FBI experts testified about Comparative Bullet Lead Analysis (“CBLA”), a technique first used in the investigation into President Kennedy’s assassination. 201 CBLA compares trace chemicals found in bullets at crime scenes with ammu-

195. Benfield, 140 F.3d at 920.

196. See e.g., Fireman’s Fund Ins. Co. v. Canon U.S.A., Inc., 394 F.3d 1054, 1058 (8th Cir. 2005) (finding the district court did not abuse its discretion where the court concluded that evidentiary support of arson theory advanced by experts was inadequate because they did not examine their theory “against empirical data obtained from fire scene analysis and appropriate testing,” in violation of NFPA 921); Ind. Ins. Co. v. Gen. Elec. Co., 326 F. Supp. 2d 844, 850–51 (N.D. Ohio 2004) (holding that cause-and-origin expert’s failure to properly collect evidence violated NFPA 921).


198. Hansen, supra note 166, at 42–43.

199. For a fuller discussion of the Cameron Todd Willingham case, see Giannelli, supra note 117.


nition found in a suspect’s possession. This technique was used when firearms or “ballistics” identification could not be employed. FBI experts used various analytical techniques, first neutron activation analysis (“NAA”), and then inductively coupled plasma atomic emission spectrometry (“ICP-AES”), to determine the concentration levels of seven elements—arsenic, antimony, tin, copper, bismuth, silver, and cadmium—in the bullet lead alloy of the suspect’s bullets and those recovered from the crime scene. Statistical tests were then used to compare the elements in each bullet and determine whether the fragments and suspect’s bullets were “analytically indistinguishable” for each of the elemental concentration averages. Exactly what the phrase “analytically indistinguishable” meant was the main issue—i.e., did such a finding mean that the bullet fragments came from a small or large universe? Obviously, the probative value of the test results would differ if only a hundred bullets had the same chemical composition as opposed to several million.

The published cases revealed disparate and often inconsistent interpretive conclusions provided by FBI experts. In some, experts testified that two exhibits were “analytically indistinguishable.” In other cases, examiners concluded that samples could have come from the same “source” or “batch.” In still others, they stated that the samples came from the same source. The testimony in numerous cases went much further and referred to a “box” of ammunition—typically fifty loaded cartridges, sometimes twenty. For example, two specimens:

1. Could have come from the same box;
2. Could have come from the same box or a box manufactured on the same day;
3. Were consistent with their having come from the same box of ammunition;
4. Probably came from the same box or

204. See United States v. Davis, 103 F.3d 660, 673-74 (8th Cir. 1996); People v. Lane, 628 N.E.2d 682, 689-90 (Ill. App. Ct. 1993).
(5) Must have come from the same box or from another box that would have been made by the same company on the same day.209

Several other statements that differ appear in the published opinions. An early case reported that the specimens “had come from the same batch of ammunition: they had been made by the same manufacturer on the same day and at the same hour.”210 One case reports the expert’s conclusion with a statistic.211 In another case, the expert used the expressions “such a finding is rare”212 and “a very rare finding.”213 In still another case, the expert “opined that the same company produced the bullets at the same time, using the same lead source. Based upon DOJ records, she opined that an overseas company called PMC produced the bullets around 1982.”214


The technique was not seriously challenged until a retired FBI examiner, William Tobin, began questioning the procedure in scientific

209. See Davis, 103 F.3d at 666 (“An expert testified that such a finding is rare and that the bullets must have come from the same box or from another box that would have been made by the same company on the same day.”); Commonwealth v. Daye, 587 N.E.2d 194, 207 (Mass. 1992); State v. King, 546 S.E.2d 575, 584 (N.C. 2001) (“[The expert] opined that, based on her lead analysis, the bullets she examined either came from the same box of cartridges or came from different boxes of the same caliber, manufactured at the same time.”).


212. Davis, 103 F.3d at 666.

213. Id. at 667.

214. People v. Villarta, No. H021354, 2002 WL 66887, at *6 (Cal. Ct. App. Jan. 17, 2002). In later years, the testimony became more limited. A 2002 FBI publication states the conclusion as follows: “Therefore, they likely originated from the same manufacturer’s source (melt) of lead.” Charles A. Peters, The Basis for Compositional Bullet Lead Comparisons, FORENSIC SCI. COMM'NS (July 2002) (emphasis added), https://archives.fbi.gov/archives/about-us/lab/forensic-science-communications/lsc/july2002/peters.htm [https://perma.cc/889Y-4TGL]. Testimony to the same effect has also been proffered. Transcript of Record at 6, Commonwealth v. Wilcox, No. 00CR2727 (Ky. Cir. Ct. Jefferson County Feb. 28, 2002) (trial testimony of Charles Peters, FBI examiner) (“Well, bullets that are analytically indistinguishable likely come from the same molten lead sources of lead, uh, as opposed to bullets that have different composition come from different, uh, melts of lead.”).
and legal journals as well as in court testimony. As a result, the FBI asked the National Academy of Sciences (“NAS”) to review the technique. The 2004 NAS Report undercut the FBI testimony, stating: “The available data do not support any statement that a crime bullet came from a particular box of ammunition. In particular, references to ‘boxes’ of ammunition in any form should be avoided as misleading under Federal Rule of Evidence 403.” Perhaps the most disturbing case is State v. Earhart, a capital murder case in which the CBLA evidence apparently played a significant role. The transcript contains the following expert testimony: “We can—from my 21 years experience of doing bullet lead analysis and doing research on boxes of ammunition down through the years I can determine if bullets came from the same box of ammunition . . . .” However, the NAS Report found that the


218. 823 S.W.2d 607, 614 (Tex. Crim. App. 1991) (en banc) (“[The expert] concluded that the likelihood that two .22 caliber bullets came from the same batch, based on all the .22 bullets made in one year, is approximately .000025 percent, ‘give or take a zero.’ He subsequently acknowledged, however, that the numbers which he used to reach the .000025 percent statistic failed to take into account that there are different types of .22 caliber bullets made each year—.22, .22 long, and .22 long rifle. [The expert] ultimately testified that there could be several hundred thousand bullets per batch, but with some variation in the elemental composition within the batch.”).

219. See Earhart v. Johnson, 132 F.3d 1062, 1067 (5th Cir. 1998) (“Given the significant role the bullet evidence played in the prosecution’s case, we shall therefore assume Earhart could have made a sufficient threshold showing that he was entitled to a defense expert under Texas law.”).

220. Transcript of Record at 5248–49, State v. Earhart, No. 4064, Dist. Ct. Lee County, 21st Judicial Dist., Texas (testimony of John Riley); see also id. at 5258 (“Well, bullets that are—that have analytically indistinguishable compositions or compositions that are generally similar typically are found within the same box of ammunition and that is the case that we have here. Now, bullets that are the same composition can also be found in other boxes of ammunition, but it’s most likely those boxes would have been
amount of bullets that can be produced from a melt “can range from the equivalent of as few as 12,000 to as many as 35 million 40-grain, .22 caliber longrifle bullets . . . .” Earhart was executed before the report was published.222

2. Post-Report Developments

Much of the FBI testimony rested on a database, which the Bureau had built up over the span of several years. Although the NAS committee frequently asked for this data during its year-long investigation, the FBI did not turn it over until it was too late to analyze for its report.223 The two statisticians who served on the NAS committee later wrote that their subsequent inspection of the data “identified several peculiarities.”224 First, the database was incomplete. The FBI claimed to have a “complete data file” of some 71,000+ measurements, but it only turned over 64,869. Moreover, only ICP-AES’s measurements were included; a different analytical method—NAA—had been used before 1997. Both techniques measured the same elements, and thus the results from either technique would have been appropriate for comparison. Additionally, the numbering system for the bullets was “highly inconsistent and rather unexpected,” suggesting that some bullet measurements were deleted.225 Additionally, “a rough investigation of the manufactured at the same place on or about the same date.”). But see testimony of Charles Peters, FBI examiner, Commonwealth v. Wilcox, Kentucky, Feb. 28, 2002 (testifying during a Daubert hearing: “We have never testified, to my knowledge, that that bullet came from that box. We’d never say that. All we are testifying is that bullet, or that victim fragment or something, the bullet, either came from that box or the many boxes that were produced at the same time.” Transcript at 1-2 (emphasis added)).

221. NAS REPORT, supra note 217, at 6.


223. See Clifford H. Spiegelman & Karen Kafadar, Data Integrity and the Scientific Method: The Case of Bullet Lead Data as Forensic Evidence, 19 CHANCE, no. 2, 2006, at 17, 22 (“During the open sessions of the committee meetings, the FBI claimed to have a ‘complete data file’ of some 71,000+ measurements. Following repeated requests from the Committee, the FBI submitted at its last meeting a CD-ROM that contained two data files with a combined total of 64,869 bullet (not 71,000+) measurement records. This data set could not be analyzed in time for the release of the report . . . .”); Giannelli, supra note 215.

224. Spiegelman & Kafadar, supra note 223, at 17, 22.

225. Id. (“[T]he numbering system of the bullets was highly inconsistent and rather unexpected (e.g., the bullets from a suspect in a particular case might be numbered Q13A, Q13B, Q13C, Q14A, Q14B, Q14C, . . . leading one to wonder what happened to bullets Q01, Q02, . . . Q12).” (omissions in original)). Other illustrations of incomplete data were noted: “[W]hile most of the bullets indicated three measurements, about 30 bullets had six or more
measurement error indicated many measurement errors that exceeded the FBI’s claimed analytical precision of 2–5%.”

Finally, “only 15% of the 1,079 cases listed in these two files had measurements from [National Institute of Standards and Technology] . . . making it impossible to determine the frequency of matches . . . in a case.”

Accordingly, the “missing data and the inconsistent precisions” undermined the Bureau’s public claims. These authors were puzzled by the FBI’s failure to disclose data: “The scientific method is important for science generally; forensic science is no exception. . . . [T]here is evidence in this paper suggests that, at least for [CBLA], forensic science failed in the requirement to share the material, methods, and data to reach conclusions with the scientific community.”

The FBI’s response to the NAS Report was also troubling. The Bureau quickly put out a press release, obscuring the report’s findings. The release highlighted the committee’s conclusion that the FBI was using appropriate instrumentation and suitable elements for comparison. Yet, these aspects of CBLA were never seriously questioned. Rather, the interpretation of the data was disputed. Only one sentence in the press release addressed this critical issue: “Recommendations by the [NAS] included suggestions to improve the statistical analysis, quality control procedures, as well as expert testimony.”

The news media read the report quite differently—e.g., “Study Shoots Holes in Bullet Analyses by FBI,” “Report Finds Flaws in FBI Bullet Analysis,”

measurements. . . . [O]nly about 50% of the bullets in this data set were identified as having come from one of the four major bullet manufacturers in the United States [i.e., Cascade Cartridge, Inc.; Federal; Remington; Winchester]; the ‘complete data file’ of 71,000 bullets may yield a higher proportion of bullets from these four manufacturers.”

226. Id.
227. Id.
228. Id.
229. Id. at 24; see also Giannelli, supra note 215.
231. Id.
“FBI Lab Under Scrutiny Again,”\textsuperscript{234} and “Report Questions the Reliability of an F.B.I. Ballistics Test.”\textsuperscript{235}

The Bureau also included the following passage in the press release: “The basis of bullet lead compositional analysis is supported by approximately 50 peer-reviewed articles found in scientific publications beginning in the early 1970’s. Published research and validation studies have continued to demonstrate the usefulness of the measurements of trace elements within bullet lead.”\textsuperscript{236} In contrast, the NAS Report pointed out that there were “very few peer-reviewed articles on homogeneity and the rate of false positive matches” and “[o]utside reviews have only recently been published.”\textsuperscript{237}

Over a year later, the FBI discontinued CBLA testing\textsuperscript{238} and issued another, similar press release. Once again, the release minimized the problems, citing the following reason for its decision: “While the FBI Laboratory still firmly supports the scientific foundation of bullet lead analysis, given the costs of maintaining the equipment, the resources necessary to do the examination, and its relative probative value, the FBI Laboratory has decided that it will no longer conduct this exam.”\textsuperscript{239} Nevertheless, Dwight Adams, the laboratory director, had written a private memorandum to the FBI Director a month earlier specifying different reasons for abandoning the technique, including the following comments: (1) “We cannot afford to be misleading to a jury” and (2) “We plan to discourage prosecutors from using our previous results in future cases.”\textsuperscript{240} The press release did not reflect either concern.


\textsuperscript{236} FBI Nat’l Press Office, \textit{supra} note 230.

\textsuperscript{237} NAS REPORT, \textit{supra} note 217, at 100.


907
In the wake of the NAS Report, several state courts excluded CBLA evidence. Surprisingly, the FBI supplied affidavits in several cases supporting prosecutors’ efforts to sustain convictions based on the technique. In one affidavit, the FBI cited the NAS report but ignored that the report had faulted the Bureau’s statistical methods. The chair of the NAS committee criticized the affidavit because it did “not discuss the statistical bullet-matching technique, which is key and probably the most significant scientific flaw found by the committee.” The affidavit was also misleading because it estimated that the maximum number of .22-caliber bullets in a batch of lead was 1.3 million, when the NAS committee found that the number could be as high as 35 million.

On November 18, 2007, 60 Minutes aired a segment on CBLA. In an interview, the now-retired FBI lab director acknowledged that testimony about boxes was “misleading and inappropriate.” That broadcast, along with a Washington Post investigation, questioned the FBI’s response to the NAS Report. The main problem was that only the FBI had records of all the cases in which its experts had testified, and the Bureau had declined to disclose the names of those cases.

241. See Ragland v. Commonwealth, 191 S.W.3d 569, 580 (Ky. 2006) (noting that “[i]f the FBI Laboratory that produced the CBLA evidence now considers such evidence to be of insufficient reliability to justify continuing to produce it, a finding by the trial court that the evidence is both scientifically reliable and relevant would be clearly erroneous”); Clemons v. State, 896 A.2d 1059, 1070, 1078 (Md. 2006) (“CBLA is not admissible under the Frye-Reed standard because it is not generally accepted within the scientific community as valid and reliable.”); “Based on the criticism of the processes and assumptions underlying CBLA, we determine that the trial court erred in admitting expert testimony based on CBLA because of the lack of general acceptance of the process in the scientific community.”); State v. Behn, 868 A.2d 329, 331 (N.J. Super. Ct. 2005) (finding the technique was “based on erroneous scientific foundations”). But see Commonwealth v. Fisher, 870 A.2d 864, 871 (Pa. 2005) (“The CBLA evidence, at best, established a possible connection between Appellant and the bullets recovered from the victim’s body.”); see also United States v. Davis, 406 F.3d 505, 509 (8th Cir. 2005) (“Davis’s trial counsel cannot be said to be ineffective for failing to challenge the FBI’s methodology on a basis that was not advanced by the scientific community at the time of trial.”).


243. Id.

244. 60 Minutes: Evidence of Injustice (CBS television broadcast, Nov. 16, 2007), https://www.youtube.com/watch?v=H4g62cpRz7M [https://perma.cc/XCE55HP].

245. Id. at 5:26.

246. Solomon, supra note 242 (“Hundreds of defendants sitting in prisons nationwide have been convicted with the help of an FBI forensic tool that
Instead, the Bureau relied on the NAS Report, its own press releases, and pro forma letters sent to prosecution and defense organizations to notify defendants. This method of communication was grossly inadequate because the letters neither highlighted the problem, nor its significance. A few days after the 60 Minutes expose, Senator Patrick Leahy, the Chairman of the Senate Judiciary Committee, sent a letter to the FBI Director noting that the Bureau’s letters gave “the false impression that these discredited tests had continuing reliability.”

Here, the flaws are many: Lack of foundational research, failure to make a database available to outside scientists, and ignoring the FBI’s own protocols by presenting inconsistent and misleading testimony. Moreover, the reluctance to confess error and take timely corrective action violated basic scientific norms. After decades of use, a federal district court in 2003 excluded CBLA evidence under the Daubert standard for the first time.

II. Misleadingly Presented Techniques

A. Firearms & Toolmark Identifications

Firearms identifications, popularly known as “ballistics,” is another long-established forensic discipline. It developed in the early part of the last century, and by the 1930s courts were admitting evidence based on this technique. “Subsequent cases have followed these precedents, admitting evidence of bullet, cartridge case, and shot shell iden-


248. John Solomon, Leahy Pursues Forensic-Test Answers: Attorney General Is Told to Prepare For Senate Inquiry, WASH. POST (Nov. 22, 2007), https://www.pressreader.com/usa/the-washington-post/20071122/281539401609068 [https://perma.cc/BKY5-SFFH] (quoting Leahy). Leahy also wrote: “The new revelations about bullet-lead analysis are just the latest examples of the Department’s inadequate efforts to ensure that sound forensic testing is utilized to the maximum extent to find the guilty rather than merely obtain a conviction. Punishing the innocent is wrong and allows the guilty party to remain free.” Id.; see also Giannelli, supra note 215.

Toolmark comparison, a related discipline, was also accepted during this period. At the time Daubert was decided, the FBI’s position was clear: “Firearms identification is the Forensic Science discipline that identifies a bullet, cartridge case or other ammunition component as having been fired by a particular firearm to the exclusion of all other firearms.” Yet, the examination, by means of a comparison microscope, is subjective and without a meaningful standard.

1. Post-Daubert Cases

The courts gave short shrift to the initial post-Daubert challenges to firearms and toolmark identifications. In 2005, however, the legal landscape changed abruptly. In United States v. Green, the district judge questioned the foundational basis of firearms identifications. The court wrote that the expert “declared that this match could be made ‘to the exclusion of every other firearm in the world.’ . . . That conclusion, needless to say, is extraordinary, particularly given [his] data and methods.” Moreover, the expert could not cite any reliable error rates and admitted that he relied mainly on his subjective judgment. In addition, “[t]here were no reference materials of any specificity, no national or even local database on which he relied. And although he relied on his past experience with these weapons, he had no notes or pictures memorializing his past observations.” In the end, the court restricted the expert’s testimony; he could only explain the ways in which the casings were similar but not that they came from a specific weapon “to the exclusion of every other firearm in the world.” In the court’s view, that conclusion “stretches well beyond [the expert’s] data and methodology.”

250. GIANNELLI ET AL., supra note 20, § 14.06, at 772 (citations omitted).
251. Id. at § 14.12.
253. See, e.g., United States v. Hicks, 389 F.3d 514, 526 (5th Cir. 2004) (stating that “the matching of spent shell casings to the weapon that fired them has been a recognized method of ballistics testing in this circuit for decades”); United States v. Foster, 300 F. Supp. 2d 375, 376 n.1 (D. Md. 2004) (“Ballistics evidence has been accepted in criminal cases for many years . . . . In the years since Daubert, numerous cases have confirmed the reliability of ballistics identification.”); United States v. Santiago, 199 F. Supp. 2d 101, 111 (S.D.N.Y. 2002) (“The Court has not found a single case in this Circuit that would suggest that the entire field of ballistics identification is unreliable.”).
255. Id. at 107.
256. Id.
257. Id. at 109.
258. Id.
A few weeks later, a different district judge in United States v. Monteiro found that the technique “is largely a subjective determination based on experience and expertise.” Importantly, the court also concluded that the theory on which the expert relied was “tautological.” The Association of Firearm and Toolmark Examiners (“AFTE”), the leading organization of examiners, proposed the theory. Under this theory, the examiner may declare a positive identification if (1) there is “sufficient agreement” of marks between the crime scene and test bullets; and (2) there is “sufficient agreement” when the examiner says there is. In short, the “sufficient agreement” threshold is “in the minds eye of the examiner and is based largely on training and experience.” The court would not admit the evidence unless the expert could better document the examination.

Together, Green and Monteiro should have served as a shot across the bow. But they did not; courts continued to admit the same evidence as before.

260. Id. at 355.
261. Id. at 370.
263. See Itiel E. Dror, How Can Francis Bacon Help Forensic Science? The Four Idols of Human Biases, 50 Jurimetrics 93, 104 (2009) (“The potential problem here is the nonscientific nature of the identification criteria. If the comparison of toolmarks enables conclusions about common origin when the unique surface contours of two toolmarks are in ‘sufficient agreement,’ what is the scientific definition and measurement of what constitutes such ‘sufficient agreement’? It seems that it is more in the eye of the beholder than strict scientific measures because it is determined without specific quantification and criteria.”).
265. See, e.g., United States v. Williams, 506 F.3d 151, 161–62 (2d Cir. 2007) (upholding admissibility of firearms identification evidence); United States v. Natson, 469 F. Supp. 2d 1253, 1261 (M.D. Ga. 2007) (“According to his testimony, these toolmarks were sufficiently similar to allow him to identify Defendant’s gun as the gun that fired the cartridge found at the crime scene. He opined that he held this opinion to a 100% degree of certainty. . . . The Court also finds [the expert’s] opinions reliable and based upon a scientifically valid methodology. Evidence was presented at the hearing that the toolmark testing methodology he employed has been tested, has been subjected to peer review, has an ascertainable error rate, and is generally accepted in the scientific community.”).

In 2008, NAS published a report on computer imaging of bullets.\textsuperscript{266} Although firearms identification was not the primary focus of the investigation, a section of the report commented on the subject.\textsuperscript{267} After surveying the literature on uniqueness, reproducibility, and permanence of individual characteristics, the report noted that “most of these studies are limited in scale and have been conducted by firearms examiners (and examiners in training) in state and local law enforcement laboratories as adjuncts to their regular casework.”\textsuperscript{268} The report found that the “validity of the fundamental assumptions of uniqueness and reproducibility of firearms-related toolmarks has not yet been fully demonstrated.”\textsuperscript{269} The report went on to caution:

Conclusions drawn in firearms identification should not be made to imply the presence of a firm statistical basis when none has been demonstrated. Specifically . . . examiners tend to cast their assessments in bold absolutes, commonly asserting that a match can be made “to the exclusion of all other firearms in the world.” Such comments cloak an inherently subjective assessment of a match with an extreme probability statement that has no firm grounding and unrealistically implies an error rate of zero.\textsuperscript{270}

Citing this report, the district court in \textit{United States v. Glynn}\textsuperscript{271} ruled that the expert would only be permitted to testify that it was “more likely than not” that recovered bullets and cartridge cases came from a particular weapon.\textsuperscript{272} The court also commented: “Based on the \textit{Daubert} hearings . . . the Court very quickly concluded that whatever else ballistics identification analysis could be called, it could not fairly

\begin{itemize}
  \item \textsuperscript{266}Nat’l Research Council of the Nat’l Acads., Ballistic Imaging (Daniel L. Cork et al. eds., 2008).
  \item \textsuperscript{267}Id. at 70.
  \item \textsuperscript{268}Id. at 81. The report also stated: “Additional general research on the uniqueness and reproducibility of firearms-related toolmarks would have to be done if the basic premises of firearms identification are to be put on a more solid scientific footing.” Id. at 82.
  \item \textsuperscript{270}Id. at 82.
  \item \textsuperscript{271}578 F. Supp. 2d 567 (S.D.N.Y. 2008).
  \item \textsuperscript{272}Id. at 575.
\end{itemize}
be called ‘science.’”^{273} The court further noted that “[t]he problem is compounded by the tendency of ballistics experts . . . to make assertions that their matches are certain beyond all doubt, that the error rate of their methodology is ‘zero,’ and other such pretensions.”^{274}


As noted previously, NAS issued its forensic report the following year in 2009. That Report summarized the state of the research as follows:

Because not enough is known about the variabilities among individual tools and guns, we are not able to specify how many points of similarity are necessary for a given level of confidence in the result. Sufficient studies have not been done to understand the reliability and repeatability of the methods . . . . Individual patterns from manufacture or from wear might, in some cases, be distinctive enough to suggest one particular source, but additional studies should be performed to make the process of individualization more precise and repeatable.^{275}

In a different passage, the report—citing firearm and toolmark identifications—observed that “[m]uch forensic evidence . . . is introduced in criminal trials without any meaningful scientific validation, determination of error rates, or reliability testing to explain the limits of the discipline.”^{276}

AFTE rejected these findings out of hand, arguing that NAS “ignore[d] extensive research supporting the scientific underpinnings of the identification of firearm and toolmark evidence . . . .”^{277} The court in United States v. Otero^{278} accepted the AFTE’s position, citing studies which it was ill-equipped to evaluate.^{279} A subsequent review of the oft-cited studies by two scientists concluded:

---

273. *Id.* at 570.

274. *Id.* at 574.


276. *Id.* at 107–08.


278. 849 F. Supp. 2d 425, 437–38 (D.N.J. 2012) (“The Court’s analysis of the proposed testimony according to the Daubert factors leads it to conclude that [the] expert report and opinion are admissible under Rule 702.”).

279. *Id.* at 438; *see also* NAS FORENSIC REPORT *supra* note 26, at 153–55.
Exaggerated and unfounded implications relating to rates of error inferred from even the best of existing experiments in the field of firearms/toolmarks, generally self-described as ‘validation studies’, typically result from statistical, metallurgical and/or psychological (cognitive) deficiencies in the design and conduct of the experiments, and frequently lead to unjustified inferential extrapolation to universal assumption for the practice domain.\textsuperscript{280}

Other courts took an important, but still limited, step of restricting examiner testimony by precluding the expert from making gross over-statements such as declaring a match to the exclusion, either practical or absolute, of all other weapons.\textsuperscript{281} Similarly, some courts forbade experts from testifying that they hold their opinions to a “reasonable degree of scientific certitude.”\textsuperscript{282} That term has long been required by courts in many jurisdictions for the admission of expert testimony. Incredibly, the phrase has no scientific meaning and the claim of certainty is unsupported by empirical research. Thus, it is grossly misleading. Indeed, the National Commission on Forensic Science rejected it.\textsuperscript{283} Still other courts went off on a quixotic tangent, substituting the phrase “reasonable degree of ballistic” certitude.\textsuperscript{284} Changing “scientific

\textsuperscript{280} Clifford Spiegelman & William A. Tobin, \textit{Analysis of Experiments in Forensic Firearms/Toolmarks Practice Offered as Support for Low Rates of Practice Error and Claims of Inferential Certainty}, 12 LAW, PROB. & RISK 115, 115 (2013).

\textsuperscript{281} See, \textit{e.g.}, United States v. Ashburn, 88 F. Supp. 3d 239, 249 (E.D.N.Y. 2015) (“Nor can [the expert] testify that a match he identified is to ‘the exclusion of all other firearms in the world,’ or that there is a ‘practical impossibility’ that any other gun could have fired the recovered materials.”); United States v. Taylor, 663 F. Supp. 2d 1170, 1180 (D.N.M. 2009) (“[The expert] also will not be allowed to testify that he can conclude that there is a match to the exclusion, either practical or absolute, of all other guns.”).

\textsuperscript{282} See, \textit{e.g.}, \textit{Ashburn}, 88 F. Supp. 3d at 249 (“[T]he court joins in precluding this expert witness from testifying that he is ‘certain’ or ‘100%’ sure of his conclusions that certain items match.”); United States v. Willock, 696 F. Supp. 2d 536, 549 (D. Md. 2010) (“[The expert] shall state his opinions and conclusions without any characterization as to the degree of certainty with which he holds them.”); People v. Robinson, 2 N.E.3d 383, 402 (Ill. App. Ct. 2013) (“[T]he judicial decisions uniformly conclude toolmark and firearms identification is generally accepted and admissible at trial. Accordingly, we conclude the trial court did not err in ruling the testimony in this case was admissible . . . particularly where the trial judge barred the witnesses from testifying their opinions were ‘within a reasonable degree of scientific certainty.’”).


\textsuperscript{284} Taylor, 663 F. Supp. 2d at 1180 (“He may only testify that, in his opinion, the bullet came from the suspect rifle to within a reasonable degree of certainty in the firearms examination field.”); United States v. Cerna, No. CR 08-0730 WHA, 2010 WL 3448528, at *4 (N.D. Cal. 2010) (allowing
certainty” to “ballistic certainty” merely underscores the courts’ scientific incompetence.

However, even these modest limitations were rejected by other courts. For example, in United States v. Casey, the district court declined “to follow sister courts who have limited expert testimony based upon the 2008 and 2009 NAS reports and, instead, remains faithful to the long-standing tradition of allowing the unfettered testimony of qualified ballistics experts.”


The 2016 White House PCAST report agreed with the 2009 NAS Report’s characterization of the scientific research on firearms and toolmarks identification: “We find that many of these earlier studies were inappropriately designed to assess foundational validity and estimate reliability. Indeed, there is internal evidence among the studies themselves indicating that many previous studies underestimated the false positive rate by at least 100-fold.” In addition, PCAST found only one of the post-2009 studies sufficiently rigorous. The Defense Department’s Forensic Science Center commissioned the study, which was conducted by an independent testing lab—the Ames Laboratory, a Department of Energy national laboratory affiliated with Iowa State University. In this study, “[t]he false-positive rate was estimated at 1

experts “to testify that a particular bullet or cartridge case was fired from a particular firearm ‘to a reasonable degree of certainty in the ballistics field’” (quoting United States v. Diaz, No. 05-00167, 2007 WL 485967, at *14 (N.D. Cal. 2007)); Commonwealth v. Pytou Heang, 942 N.E.2d 927, 945 (Mass. 2011) (stating that “the expert may offer that opinion to a ‘reasonable degree of ballistic certainty’”).

285. See, e.g., Fleming v. State, 1 A.3d 572, 590 (Md. Ct. App. 2010) (“[N]otwithstanding the current debate on the issue, courts have consistently found the traditional method [of firearms identification] to be generally accepted within the scientific community, and to be reliable.”); People v. Givens, 912 N.Y.S.2d 855, 857 (Sup. Ct. 2010) (“This Court was unable to find any cases where firearms and toolmark identification was found to be unreliable or no longer scientifically acceptable.”).


287. Id. at 400; see also United States v. Sebbern, No. 10 Cr. 87(SLT), 2012 WL 5988813, at *7 (E.D.N.Y. Nov. 30, 2012); State v. Languois, 2 N.E.3d 936, 950 (Ohio Ct. App. 2014) (“Our conclusion on this issue finds support in the decisions of other appellate districts in Ohio, notwithstanding the recent criticisms in scientific reports and the limitations some federal courts have imposed on the testimony of firearms experts. These decisions hold that the methodology of comparatively analyzing and testing bullets and shell cases recovered from crime scenes is reliable.”); State v. Jones, 303 P.3d 1084, at ¶ 75 (Wash. Ct. App. 2013) (expert testimony comparing bunter marks on the base of shell casings found at the crime scene to shell casings found in Jones’s home admissible under the Frye standard).

288. WHITE HOUSE PCAST REPORT, supra note 38, at 11.
in 66, with a confidence bound indicating that the rate could be as high as 1 in 46.”289 The study had not been published in a scientific journal. According to the PCAST Report, more than one study is required and studies should be published in peer-reviewed scientific literature. Consequently, “the current evidence still falls short of the scientific criteria for foundational validity.”290

AFTE quickly retorted, expressing their “disappointment in the PCAST’s choice to ignore the research that has been conducted” and claiming that “[d]ecades of validation and proficiency studies have demonstrated that firearm and toolmark identification is scientifically valid . . . .”291 However, when PCAST later invited stakeholders to submit validation studies that it may have overlooked, no studies satisfying PCAST’s criteria were offered.292

The lessons here are familiar. For years, an entrenched forensic discipline vigorously guarded its turf by rejecting the conclusions of the outside scientific community.293 It published a journal which was “peer-reviewed” by other members of its discipline. The journal, which is advertised as “the Scientific Journal” of AFTE, was not generally available until 2016. The discipline claimed to be a “science” but did not hold itself to the normative standards of science. The AFTE “Theory of Identification” is “clearly not a scientific theory, which the National Academy of Sciences has defined as ‘a comprehensive explanation of some aspect of nature that is supported by a vast body of evidence.’ . . . More importantly, the stated method is circular.”294 Only recently, after two NAS reports, have some courts begun to limit misleading testimony. Many have not. Thus, the courts’ competence to deal with flawed research remains extant.295

289. Id.
290. Id.
292. President’s Council of Advisors on Sci. and Tech., An Addendum to the PCAST Report on Forensic Science in Criminal Courts 7 (2017) (“Several respondents wrote to PCAST concerning firearms analysis. None cited additional appropriately designed black-box studies similar to the recent Ames Laboratory study.”).
293. See William A. Tobin et al., Absence of Statistical and Scientific Ethos: The Common Denominator in Deficient Forensic Practices, 4 STAT. & PUB. POL’Y 1, 9 (2016) (“P)ractitioners remain intractable even after years of critical scholarly papers, ad hoc committees of the National Academy of Sciences (NAS), position statements from the U.S. Department of Justice . . . .”) (citation omitted).
294. White House PCAST Report, supra note 38, at 60.
295. Tobin et al., supra note 293, at 9 (“T)he purported ‘validation studies’ typically proffered to courts are seriously flawed [and] have no external validity . . . .”).
In 2005, the district court in Green cautioned: “The more courts admit this type of toolmark evidence without requiring documentation, proficiency testing, or evidence of reliability, the more sloppy practices will endure; we should require more.” Over a decade later, a concurring opinion in Williams v. United States concluded: “As matters currently stand, a certainty statement regarding toolmark pattern matching has the same probative value as the vision of a psychic: it reflects nothing more than the individual’s foundationless faith in what he believes to be true.” In short, there is a “lost decade” during which the discipline summarily dismissed criticisms when it should have lead the effort for more rigorous research.

B. Fingerprint Examinations

Before DNA analysis, fingerprint identification was the “gold standard” in forensics. Like many other forensic disciplines, it gained judicial acceptance decades before Daubert was decided. People v. Jennings, the first reported fingerprint case, was decided in 1911. In 1984, the FBI pronounced the technique “infallible” in its official publication, which also referred to the technique as a “science.” Nevertheless, it is a subjective technique without an objective standard and typically involves partial prints with inevitable distortions.

1. Post-Daubert Cases

After Daubert, challenges to fingerprint comparison testimony were decidedly unsuccessful. One infamous case, United States v.

298. Id. at 355 (Easterly, J., concurring).
300. 96 N.E. 1077 (Ill. 1911). See generally 1 Giannelli et al., supra note 20, §16 (discussing the scientific and legal issues associated with fingerprint identification).
Havward,\textsuperscript{303} illustrates the judiciary’s lack of rigor in applying Daubert. Not only did the district court uphold the fingerprint testimony’s admissibility, it styled the technique as “the very archetype of reliable expert testimony under [the Daubert/Kumho] standards.”\textsuperscript{304} According to the court, latent print identification had been “tested” for nearly one hundred years in adversarial proceedings with the highest possible stakes—liberty and sometimes life. Yet, Daubert required scientific, not “adversarial,” testing.\textsuperscript{305} Next, in citing “peer review,” the court noted that a second fingerprint examiner also compared the prints: “In fact, peer review is the standard operating procedure among latent print examiners.”\textsuperscript{306} This statement reveals a fundamental misunderstanding of “peer review” as used in Daubert. In that case, peer review meant refereed scientific journals in which validation research is published. An amici brief submitted in Daubert by the New England Journal of Medicine and other scientific publications explained that peer review’s “role is to promote the publication of well-conceived articles so that the most important review, the consideration of the reported results by the scientific community, may occur after publication.”\textsuperscript{307}

Moreover, the court accepted the prosecution expert’s astounding claim that the “error rate for the method is zero.”\textsuperscript{308} Experts argued that, while individual examiners may make mistakes, the method itself is perfect. The dichotomy between “methodological” and “human” error rates in this context, however, is “practically meaningless”\textsuperscript{309} because

\textsuperscript{303} 117 F. Supp. 2d 848 (S.D. Ind. 2000), aff’d, 260 F.3d 597 (7th Cir. 2001).
\textsuperscript{304} Id. at 855; see also Paul C. Giannelli, Daubert Challenges to Fingerprints, 42 CRIM. L. BULL. 624, 628 (2006).
\textsuperscript{305} See Sandy L. Zabell, Fingerprint Evidence, 13 J.L. & POL’Y 143, 170 (2005) (“The argument that no latent print has ever been found to match the rolled print of a different person is . . . misleading because no systematic search for such pairs on the entire databank of millions of fingerprints has ever been performed.”).
\textsuperscript{306} Havward, 117 F. Supp. 2d at 854.
\textsuperscript{308} Havward, 117 F. Supp. 2d at 854.
\textsuperscript{309} Jennifer L. Mnookin, Fingerprint Evidence in an Age of DNA Profiling, 67 BROOK. L. REV. 13, 60 (2001). Professor Mnookin goes on to provide this analogy: “The same argument could be made of eyewitness testimony, a notoriously unreliable form of evidence. People are all distinct from one another in observable ways; therefore the theoretical error rate of eyewitness identification is zero, though in practice observers may frequently make errors.” Id.; see also Simon A. Cole, More Than Zero: Accounting for Error in Latent Fingerprint Identification, 95 J. CRIM. L. & CRIMINOLOGY 985, 1040 (2005) (stating that while a “distinction can be drawn between
the examiner is the method. Finally, the court turned *Daubert* on its head by requiring the defendant to prove the evidence was unreliable, a distortion that would be employed in later cases.

Then, *United States v. Llera Plaza* "sent shock waves through the community of fingerprint analysts." In that 2002 case, Judge Pollak ruled that fingerprint experts would not be permitted to testify that two sets of prints “matched”—that is, a positive identification to the exclusion of all other persons. This was apparently the first time in over ninety years that such a decision had been rendered. On rehearing, however, Judge Pollak reversed himself, and later cases continued to uphold the admissibility of fingerprint evidence. Nevertheless, the case captured the attention of the media with news reports, main-

310. *See* Zabell, *supra* note 305, at 172 (“But, given its unavoidable subjective component, in latent print examination people are the process.”).

311. *See* Michael J. Saks, *The Legal and Scientific Evaluation of Forensic Science (Especially Fingerprint Expert Testimony)*, 33 Seton Hall L. Rev. 1167, 1173–76 (2003) (discussing the reversal of the burden of persuasion as one of several judicial responses employed to avoid confronting the lack of empirical testing); *see also* Giannelli, *supra* note 304, at 630.


314. As Professor Mnookin has noted, however, “fingerprints were accepted as an evidentiary tool without a great deal of scrutiny or skepticism.” Mnookin, *supra* note 309, at 17. She elaborated: “Even if no two people had identical sets of fingerprints, this did not establish that no two people could have a single identical print, much less an identical part of a print. These are necessarily matters of probability, but neither the court in *Jennings* nor subsequent judges ever required that fingerprinting identification be placed on a secure statistical foundation.” Id. at 19.


316. *See*, e.g., *United States v. Abreu*, 406 F.3d 1304, 1307 (11th Cir. 2005) (“We agree with the decisions of our sister circuits and hold that the fingerprint evidence admitted in this case satisfied *Daubert.*”); *United States v. Janis*, 387 F.3d 682, 690 (8th Cir. 2004) (finding fingerprint evidence reliable); *United States v. Mitchell*, 365 F.3d 215, 244 (3d Cir. 2004) (finding that when assessed using the *Daubert* factors, the fingerprinting evidence “pass[ed] muster”); *United States v. Crisp*, 324 F.3d 261, 269 (4th Cir. 2003) (finding that the district court did not abuse its discretion by admitting fingerprint analysis as evidence); *United States v. Sullivan*, 246 F. Supp. 2d 700, 704 (E.D. Ky. 2003) (finding that the possibility that two fingers may have a portion of their print in common speaks to weight and not the admissibility of fingerprint evidence).

stream publications, and television shows giving it substantial coverage. A spate of legal articles followed, with some commentators believing that Llera Plaza I was more faithful to Daubert than Llera Plaza II. In response, the FBI adopted a “circle the wagons” attitude, fiercely defending the technique. The head of the FBI fingerprint section told 60 Minutes that the error rate was “zero,” examiners only testify to “hundred percent certainty,” and the FBI had won “forty-one out of forty-one” legal challenges to fingerprint evidence.

The appellate opinion most faithful to Daubert appeared in United States v. Crisp—unfortunately in dissent. The majority opinion upheld the admissibility of fingerprint evidence by shifting the burden of proof to the defendant and by grandfathering the technique. In dissent, Judge Michael conscientiously applied the Daubert factors. First, he noted that the “government did not offer any record of testing


See Jennifer L. Mnookin, Fingerprints: Not a Gold Standard, ISSUES IN SCI. & TECH., Fall 2003, at 47 (“Judge Pollak’s first opinion [restricting latent fingerprint individualization testimony] was the better one.”); Recent Case, United States v. Havvard, 260 F.3d 597 (7th Cir. 2001), 115 HARV. L. REV. 2349, 2352 (2002) (“Fingerprint expert testimony does not survive application of the Daubert factors . . . .”).

60 Minutes, supra note 320.

324. 324 F.3d 261 (4th Cir. 2003).

Id. at 269 (“Put simply, Crisp has provided us no reason today to believe that this general acceptance of the principles underlying fingerprint identification has, for decades, been misplaced. Accordingly, the district court was well within its discretion in accepting at face value the consensus of the expert and judicial communities that the fingerprint identification technique is reliable.”); see also Giannelli, supra note 304, at 632.
on the reliability of fingerprint identification. . . . [T]here have not been any studies to establish how likely it is that partial prints taken from a crime scene will be a match for only one set of fingerprints in the world.”

Second, as for peer review:

[a]gain, the government offered no evidence on this factor at trial. Fingerprint examiners . . . have their own professional publications. . . . But unlike typical scientific journals, the fingerprint publications do not run articles that include or prompt critique or reanalysis by other scientists. Indeed, few of the articles address the principles of fingerprint analysis and identification at all . . . .

Third, “an error rate must be demonstrated by reliable scientific studies, not by assumption.” Fourth, “the government did not establish that there are objective standards in the fingerprint examination field to guide examiners in making their comparisons.” Fifth, while acknowledging general acceptance in the fingerprint community, the judge remarked that “[n]othing in the record in this case shows that the fingerprint examination community has challenged itself sufficiently or has been challenged in any real sense by outside scientists.”

2. Madrid Train Bombing

Llera Plaza was soon eclipsed by a more sensational event—the FBI’s misidentification of Brandon Mayfield as the source of the crime scene prints in the terrorist train bombing in Madrid on March 11, 2004. More than any other event, the Mayfield affair exposed the myth of fingerprint infallibility. This debacle resulted in investigations

326. Id. at 273–74 (Michael, J., dissenting).
327. Id. at 274 (citing Epstein, supra note 321, at 644).
328. Id. The judge added: “In a 1995 test conducted by a commercial testing service, less than half of the fingerprint examiners were able to identify correctly all of the matches and eliminate the non-matches. On a similar test in 1998, less than sixty percent of the examiners were able to make all identifications and eliminations. . . . An error rate that runs remarkably close to chance can hardly be viewed as acceptable under Daubert.” Id. at 275 (citing Epstein, supra note 321, at 634–35).
329. Id. at 276.
330. Id.
by the FBI and the Inspector General of the DOJ ("IG"). One of the more troubling aspects of these reports dealt with the culture in the laboratory. The FBI internal investigation found that "[t]o disagree was not an expected response," and the IG reported that "FBI examiners did not attempt to determine the basis of the [Spanish National Police's] doubts before reiterating that they were 'absolutely confident' in the identification on April 15, a full week before the FBI Laboratory met with the SNP."

In addition to highlighting the lack of foundational research, these events raised a host of other issues, including: (1) the role of cognitive bias in subjective techniques, (2) the lack of well-defined standards, (3) the failure to administer rigorous proficiency tests, (4) the


333. See Office of the Inspector Gen., DOJ, A Review of the FBI’s Handling of the Brandon Mayfield Case: Unclassified Executive Summary 7 (2006) ("Having found as many as 10 points of unusual similarity, the FBI examiners began to ‘find’ additional features in [the print] that were not really there, but rather were suggested to the examiners by features in the Mayfield prints.").

334. Stacey, supra note 332, at 713.


337. Examiners follow a procedure known as Analysis, Comparison, Evaluation, and Verification (ACE-V). See Zabell, supra note 305, at 178 ("ACE-V is an acronym, not a methodology. It is merely the common sense description of what anyone would do if they were examining a latent and a candidate source print.").

338. See United States v. Crisp, 324 F.3d 261, 274 (4th Cir. 2003) (Michael, J., dissenting) ("Proficiency testing is typically based on a study of prints that are far superior to those usually retrieved from a crime scene."); United States v. Llera Plaza, 188 F. Supp. 2d 549, 565 (E.D. Pa. 2002) (noting that "the FBI examiners got very high proficiency grades, but the tests they took did not. . . . [O]n the present record I conclude that the proficiency tests are less demanding than they should be."); Jennifer L. Mnookin, Editorial, A Blow to the Credibility of Fingerprint Evidence, Boston Globe, Feb. 2,
manipulation of research, and (5) other instances of misidentifications. The FBI did not undertake a serious review of fingerprints until it was compelled to address the issue due to the negative publicity surrounding the Mayfield misidentification. Even then, however, the FBI still characterized the technique as “scientific.”

The scientific community continued to note the lack of research, and the courts continued to ignore this fact. Indeed, in United States v. Mitchell, 365 F.3d 215, 255 (3d Cir. 2004) (“We are deeply discomforted by Mitchell’s contention—supported by Dr. Rau’s account of events, though contradicted by other witnesses—that a conspiracy within the DOJ intentionally delayed the release of the solicitation until after Mitchell’s jury reached a verdict. Dr. Rau’s story, if true, would be a damning indictment of the ethics of those involved.”). See generally Giannelli, supra note 201 (discussing the manipulation of forensic science research, including fingerprint research, by law enforcement).

340. See Simon A. Cole, More Than Zero: Accounting for Error in Latent Fingerprint Identification, 95 J. CRIM. L. & CRIMINOLOGY 985, 999 (2005) (discussing twenty-two cases involving misattribution of latent fingerprints). The misidentification cases include some that involved (1) verification by one or more other examiners, (2) examiners certified by the International Association of Identification, (3) procedures using a sixteen-point standard, and (4) defense experts who corroborated misidentifications made by prosecution experts. Id. at 1023–25; see also CNN Presents: Reasonable Doubt (CNN television broadcast Jan. 9, 2005) (discussing the misidentification of Riky Jackson, who spent two years in prison).


342. See Donald Kennedy & Richard A. Merrill, Assessing Forensic Science, ISSUES IN SCI. & TECH., Fall 2003, at 33, 34 (“The increased use of DNA analysis, which has undergone extensive validation, has thrown into relief the less firmly credentialed status of other forensic science identification techniques (fingerprints, fiber analysis, hair analysis, ballistics, bite marks, and tool marks). These have not undergone the type of extensive testing and verification that is the hallmark of science elsewhere.”); Zabell, supra note 305, at 164 (“Although there is a substantial literature on the uniqueness of fingerprints, it is surprising how little true scientific support for the proposition exists.”).

343. See, e.g., United States v. Pena, 586 F.3d 105, 110 (1st Cir. 2009) (“The district court did not abuse its discretion. Numerous courts have found expert testimony on fingerprint identification based on the ACE-V method to be sufficiently reliable under Daubert.”); United States v. Abreu, 406 F.3d
v. Baines, decided in 2009, the head of the FBI fingerprint section testified: “As to these ‘false positives’ . . . the FBI had ‘made, on average, about one erroneous identification every 11 years.’ The total number of identifications made has been about one million per year . . . so that the known actual error rate was about one per eleven million identifications.” Problematically, he merely assumed that all the other identifications were correct, thus disqualifying his analysis. Perhaps the most troubling aspect of this testimony was the lack of self-awareness for a person who claimed to be a scientist.


Fingerprint examiners follow a procedure known as Analysis, Comparison, Evaluation, and Verification (ACE-V). The 2009 NAS report observed that since “the ACE-V method does not specify particular measurements or a standard test protocol . . . examiners must make subjective assessments throughout.” Thus, the ACE-V method is too “broadly stated” to “qualify as a validated method for this type of analysis.” The Report added that “[t]he latent print community in the United States has eschewed numerical scores and corresponding thresholds” and consequently relies “on primarily subjective criteria” in making the ultimate attribution decision. In making the decision, the examiner must draw on his or her personal experience to evaluate such factors as “inevitable variations in pressure,” but to date those factors have not been “characterized, quantified, or compared.” In addition, the Report gave short shrift to the zero-error-rate argument, finding that “claims that these analyses have zero error rates are not scien-

1304, 1307 (11th Cir. 2005) (“[T]he fingerprint evidence admitted in this case satisfied Daubert.”); United States v. Janis, 387 F.3d 682, 690 (8th Cir. 2004) (“We conclude the district court did not err in admitting the fingerprint expert’s testimony.”); United States v. Mitchell, 365 F.3d 215, 241 (3d Cir. 2004) (“We therefore accept that the error rate has been sufficiently identified to count this factor as strongly favoring admission of the [fingerprint] evidence.”).

344. 573 F.3d 979 (10th Cir. 2009).
345. Id. at 984.
346. See White House PCAST Report, supra note 38, at 53 (“The fallacy is obvious: the expert simply assumed without evidence that every error in casework had come to light.”).
347. NAS Forensic Report, supra note 26, at 139.
348. Id. at 142.
349. Id. at 141.
350. Id. at 144. Moreover, examiners lack population frequency data to quantify the rarity or commonality of a particular type of fingerprint characteristic. Id.
tifially plausible.”351 In conclusion, the Report outlined an agenda for the research it considered necessary “[t]o properly underpin the process of friction ridge identification . . . .”352

Several studies were published after the NAS Report.353 The most important was a FBI study published in 2011,354 which is discussed below.


According to the White House PCAST report, “latent fingerprint analysis is a foundationally valid subjective methodology” and the FBI “has lead the way” by conducting the black-box study.355 Nevertheless,

[the] false positive rate . . . is substantial and is likely to be higher than expected by many jurors based on longstanding claims about the infallibility of fingerprint analysis. The false-positive rate could be as high as 1 error in 306 cases based on the FBI study and 1 error in 18 cases based on a study by another crime laboratory. In reporting results of [a] latent-fingerprint examination, it is important to state the false-positive rates based on properly designed validation studies . . . .356

Moreover, “testimony asserting any specific level of increased accuracy (beyond that measured in the studies) due to blind independent verification would be scientifically inappropriate, as speculation unsupported by empirical evidence.”357

---

351. Id. at 142; see also id. at 143 (“Some in the latent print community argue that the method itself, if followed correctly . . . has a zero error rate. Clearly, this assertion is unrealistic . . . . The method, and the performance of those who use it, are inextricably linked, and both involve multiple sources of error (e.g., errors in executing the process steps, as well as errors in human judgment).”).

352. Id. at 144.

353. See White House PCAST Report, supra note 38, at 91–95.

354. Bradford T. Ulery et al., Accuracy and Reliability of Forensic Latent Fingerprint Decisions, 108 Proc. Nat’l Acad. Sci. 7733 (2011). The White House PCAST Report described the methodology of the report: “To attempt to ensure that the non-mated pairs were representative of the type of matches that might arise when police identify a suspect by searching fingerprint databases, the known prints were selected by searching the latent prints against the 58 million fingerprints in the [Automated Fingerprint Identification System] database and selecting one of the closest matching hits.” White House PCAST Report, supra note 38, at 94.


356. Id. at 9–10.

357. Id. at 96.
5. AAAS Fingerprint Report (2017)

In September 2017, the AAAS published an extensive report on fingerprint analysis. An accompanying news release, summarized the report’s findings: “Courtroom testimony and reports stating or even those implying that fingerprints collected from a crime scene belong to a single person are indefensible and lack scientific foundation . . .”

The report reached a number of conclusions. First, claims that experts can identify the source of a latent print with 100 percent accuracy, are “clearly overstated and are now widely recognized as indefensible.” Second, use of the term “identification” in reports and testimony even with qualifications “fail to deal forthrightly with the level of uncertainty that exists in latent print examination” and “cannot be justified scientifically.” Third, because of public misconceptions, experts:

should acknowledge: (1) that the conclusions being reported are opinions rather than facts (as in all pattern-matching disciplines), (2) that it is not possible for a latent print examiner to determine that two friction ridge impressions originated from the same source to the exclusion of all others; and (3) that errors have occurred in studies of the accuracy of latent print examination.

The report went on to make several recommendations. Experts should “avoid statements that claim or imply that the pool of possible sources is limited to a single person. Terms like ‘match,’ ‘identification,’ ‘individualization’ and their synonyms, imply more that the science can sustain.” In addition, experts should “be prepared to discuss forth-
rightly the results of research studies that tested the accuracy of latent print examiners on realistic known-source samples.”

Despite the ruckus created by Llera Plaza and the Mayfield fiasco, examiner testimony remained unchanged. Testimony such as “zero error rates,” “matches to the exclusion of all other fingerprints,” and “100 percent certainty”—which had been used for decades—continued, while the fingerprint community remained oblivious that such statements were scientifically implausible. As with firearms identification, there is a “lost decade” during which more research could have been conducted. As one judge noted in a 2003 dissent: “The government has had ten years to comply with Daubert. It should not be given a pass in this case.” Those words were written fifteen years ago.

On a positive note, the Mayfield incident did trigger the FBI’s black box study, which was a significant achievement. Still, this study was released 100 years after the courts first admitted fingerprint evidence. The White House PCAST report found it “distressing” that properly constructed validation studies had only been conducted recently and only one study had been published in a peer-reviewed journal. Daubert has had little effect.

III. Forensic Science Research

By now it is almost a truism that too many forensic disciplines are not grounded in science—and yet their adherents continue to claim the mantle of science. The 2009 NAS Report emphasized the “notable dearth of peer-reviewed, published studies establishing the scientific bases and validity of many forensic methods.” Indeed, the co-chair of the NAS committee, Judge Harry Edwards, later stated: “I think that the most important part of our committee’s report is its call for real

Id. at 11.
364. Id.
366. See supra note 300 and accompanying text.
367. White House PCAST Report, supra note 38, at 95.
368. Some courts did, however, place limitations on the testimony. See, e.g., United States v. Mitchell, 365 F.3d 215, 246 (3d Cir. 2004) (“Testimony at the Daubert hearing indicated that some latent fingerprint examiners insist that there is no error rate associated with their activities or that the examination process is irreducibly subjective. This would be out-of-place under Rule 702.”); Commonwealth v. Gambora, 933 N.E.2d 50, 61 n.22 (Mass. 2010) (“[O]pinions expressing absolute certainty about, or the infallibility of, an ‘individualization’ of a print should be avoided.”).
369. NAS Forensic Report, supra note 26, at 8.
science to support the forensic disciplines. 370 Not surprisingly, the report triggered extensive commentary. 371 One article cataloged the numerous ways in which forensic science has failed to develop a research culture 372 and argued that the “core values” of a scientific culture “are empiricism, transparency, and an ongoing critical perspective.” 373 Another article documented the serious problems that have arisen when law enforcement controls forensic research. 374

A. National Commission on Forensic Science (2013-17)

To its credit, the DOJ, in partnership with the National Institute of Standards and Technology (“NIST”), established the National Commission on Forensic Science in 2013. The commission’s task was to enhance the practice and improve the reliability of forensic science. 375 Early on, the commission created a subcommittee on scientific inquiry and research, which undertook the task of reviewing bibliographies of foundational literature that had been compiled by various forensic disciplines. 376 The subcommittee quickly concluded that even a “cursory review” of the bibliographies raised serious concerns. One basic problem involved the definition of foundational literature. According to the subcommittee, “[i]n some cases, it was unclear which literature citations are crucial to support the foundation of a particular forensic science


373. Id. at 742.

374. See Giannelli, supra note 201.


376. As a result of the 2009 NAS report, an Interagency Working Group—the Research Development Technology and Evaluation (RDT&E) of the National Science and Technology Council’s Subcommittee on Forensic Science—was tasked with identifying foundational research in the forensic sciences. The RDT&E committee requested Scientific Working Groups (SWGs) to address a series of discipline-specific questions. In response, literature compendiums were submitted to the RDT&E committee by several forensic working groups. Nat’l Scl. & Tech. Council, Comm. on Scl., Subcomm. on Forensic Scl., Strengthening the Forensic Sciences (2014).
discipline.”

This finding led the subcommittee to define the term. Foundational, scientific literature should consist of “original research, substantive reviews of the original research, clinical trial reports, or reports of consensus development conferences.” Tellingly, the subcommittee felt compelled to add: “While other forms of dissemination of research and practice (e.g., oral and poster presentations at meetings, workshops, personal communications, editorials, dissertations, theses, and letters to editors) play an important role in science, the open, peer-reviewed literature is what endures and forms a foundation for further advancements.”

The subcommittee’s second concern was that “[s]ome of the cited literature had not undergone a rigorous peer-review process.” Peer review by other members of a forensic discipline is not sufficient. Many of the reviewers are not scientists, and there is the problem with role bias. According to the subcommittee, foundational research should be subjected to “rigorous peer review with independent external reviewers to validate the accuracy . . . [and] overall consistency with scientific norms of practice” and “[p]ublished in a journal that is searchable using free, publicly available search engines . . . .” With few

378. Id.
379. Id.
380. Id.
381. See United States v. Crisp, 324 F.3d 261, 274 (4th Cir. 2003) (Michael, J., dissenting) (“Fingerprint examiners . . . have their own professional publications. . . . But unlike typical scientific journals, the fingerprint publications do not run articles that include or prompt critique or reanalysis by other scientists. Indeed, few of the articles address the principles of fingerprint analysis and identification at all . . . .”) See also Zabell, supra note 305, at 164 (“Although there is a substantial literature on the uniqueness of fingerprints, it is surprising how little true scientific support for the proposition exists.”).
382. Nat’l Comm’n on Forensic Sci., supra note 377, at 3 (“Published in a journal that maintains a clear and publicly available statement of purpose that encourages ethical conduct such as disclosure of potential conflicts of interest integral to the peer review process.”).
383. Id. at 2. Other publication requirements include being: (1) “[p]ublished in a journal or book that has an International Standard Number (ISSN for journals; ISBN for books) and recognized expert(s) as authors (for books) or on its Editorial Board (for journals);” and (2) “[p]ublished in a journal that is indexed in databases that are available through academic libraries and other services (e.g. JSTOR, Web of Science, Academic Search Complete, and SciFinder Scholar).” Id. at 2–3.
exceptions, the disciplines considered above have not satisfied these requirements.\textsuperscript{384}

Another recommendation—one on technical merit—provides: “All forensic science methodologies should be evaluated by an independent scientific body to characterize their capabilities and limitations in order to accurately and reliably answer a specific and clearly defined forensic question.”\textsuperscript{385} Significantly, the commission recommended that NIST be the independent scientific evaluator within the justice system.

384. Another commission document provided guidance for evaluating scientific literature. \textit{Nat’l Comm’n. on Forensic Sci., DOJ, Views of the Commission Regarding Identifying and Evaluating Literature that Supports the Basic Principles of a Forensic Science Method or Forensic Science Discipline} (2016). This guidance includes:

- Is the problem or hypothesis clearly stated?
- Is the scope of the article clearly stated as appropriate (article, case study, review, technical note, etc.)?
- Is the literature review current, thorough, and relevant to the problem being studied?
- Does this work fill a clear gap in the literature or is it confirmatory and/or incremental?
- Are the experimental procedures clear and complete such that the work could be easily reproduced?
- Are the experimental methods appropriate to the problem?
- Are the methods fully validated to the necessary level of rigor (fit for purpose)?
- Are the data analysis and statistical methodology appropriate for the problem, and explained clearly so it can be reproduced?
- Are the experimental results clearly and completely presented and discussed?
- Are omissions and limitations to the study discussed and explained?
- Are the results and conclusions reasonable and defensible based on the work and the supporting literature?
- Are the citations and references complete and accurate?
- Are the references original (primary) and not secondary?
- Are funding sources and other potential sources of conflict of interest clearly stated?

\textit{Id. at 3.}

B. White House PCAST Report (2016)

Unlike the commission, which had a broad mandate, the White House PCAST Report focused only on the validation issue. It took pains to explain the concept of validation, noting that forensic methods must be based on empirical studies and be “repeatable, reproducible, and accurate, at levels that have been measured and are appropriate to the intended application.” The report recognized that forensic methods may be either objective or subjective. Foundational validity for objective methods “can be established by studying [and] measuring the accuracy, reproducibility, and consistency of each of its individual steps.” By definition, this approach is not possible with subjective techniques because they involve significant human judgment. Consequently, validity and reliability for these methods must be based on “black-box studies”—as if a “black box” is in the examiner’s head—in which numerous examiners make decisions on many independent tests in order to determine error rates.

Importantly, the report also specified what does not qualify as validation: “[N]either experience, nor judgment, nor good professional practices (such as certification programs and accreditation programs, standardized protocols, proficiency testing, and codes of ethics) can substitute for actual evidence of foundational validity and reliability.”

Moreover, expressions of confidence by individual examiners or a consensus among practitioners about accuracy cannot substitute for “error rates estimated from relevant studies.” In sum, empirical evidence is the “sine qua non” for establishing foundational validity.

PCAST also recommended that NIST conduct scientific evaluations of the validity of current and new forensic technologies: “[t]o ensure the scientific judgments are unbiased and independent, such evaluations

---

386. White House PCAST Report, supra note 38, at 4. Here, “repeatable” means an examiner reaches the same result when analyzing the same sample. “Reproducible” means that different examiners reach the same result when analyzing the same sample. The term “accurate” means that “an examiner obtains correct results both (1) for samples from the same source (true positives) and (2) for samples from different sources (true negatives).” Finally, “reliability” means “repeatability, reproducibility, and accuracy.” Id. at 47.

387. Id. at 5.

388. Id. at 5–6.

389. Id. at 6.

390. Id.

391. Id.
should be conducted by an agency which has no stake in the outcome.”

In response, DOJ released a statement criticizing the report on the day of its release. According to DOJ, the PCAST Report “does not mention numerous published research studies which seem to meet PCAST’s criteria for appropriately designed studies providing support for foundational validity. That omission discredits the PCAST report as a thorough evaluation of scientific validity.” PCAST, in turn, invited all stakeholders to identify validity studies that it might have overlooked. “DOJ ultimately concluded that it had no additional studies for PCAST to consider.” Nor did the more than 400 papers submitted by twenty-six respondents cause PCAST to change its positions. The bottom line remained: “In science, empirical testing is the only way to establish the validity and degree of reliability of such an empirical method. Fortunately, empirical testing of empirical methods is feasible. There is no justification for accepting that a method is valid and reliable in the absence of appropriate empirical evidence.” However, most prior studies use “closed-set design.” In these studies, “the correct source of each questioned sample is always present; studies using the closed-set design have underestimated the false-positive and inconclusive rates by more than 100-fold.”

IV. INDEPENDENT SCIENTIFIC REVIEW

As discussed above, the courts have too often failed to fulfill their “gatekeeper” function under Daubert. However, the Daubert Court...
also suggested that the adversary system would serve as a complementary safeguard, noting that “[v]igorous cross-examination, presentation of contrary evidence, and careful instruction on the burden of proof are the traditional and appropriate means of attacking shaky but admissible evidence.”

Yet, these “traditional” means have also proved inadequate. After the release of the NAS Report, some commentators focused on defense counsel’s incompetence. A 2009 study of the cases of 137 convicts exonerated by DNA profiling revealed that “[d]efense counsel rarely made any objections to the invalid forensic science testimony in these trials and rarely effectively cross-examined forensic analysts who provided invalid science testimony.” One commentator summed it up this way:

Unlike the extremely well-litigated civil challenges, the criminal defendant’s challenge is usually perfunctory. Even when the most vulnerable forensic sciences—hair microscopy, bite marks, and handwriting—are attacked, the courts routinely affirm admissibility citing earlier decisions rather than facts established at a hearing. Defense lawyers generally fail to build a challenge with appropriate witnesses and new data. Thus, even if inclined to mount a Daubert challenge, they lack the requisite knowledge and skills, as well as the funds, to succeed.

Although the defense bar bears some responsibility for Daubert’s failure, there are limits to what can be expected of overburdened and chronically underfunded public defenders when dealing with expert

399. Id. at 596 (citing Rock v. Arkansas, 483 U.S. 44, 61 (1987)).

400. See Gertner, supra note 36, at 790 (“[T]he NAS Report’s concerns will not be fully met until advocacy changes.”); D. Michael Risinger, The NAS/NRC Report on Forensic Science: A Path Forward Fraught with Pitfalls, 2010 Utah L. Rev. 225, 242 (2010) (“Criminal defense lawyers . . . are supposed to be the people who recognize bogus expert claims, challenge them, move to get them excluded, and undermine those that survive exclusion by knowledgeable, thorough, and telling cross-examination. On the whole, they don’t do any of these things very well.”).

401. Garrett & Neufeld, supra note 102, at 89.

402. Neufeld, supra note 25, at S110.
testimony. Better training for defense counsel—which is sorely needed—is not sufficient. Similarly, access to defense experts—which is sorely needed—may not be adequate. Defense experts can challenge prosecution experts’ methods and opinions but do not have the funds to conduct foundational research, nor can they act as independent evaluators of foundational research on an ongoing basis.

An independent scientific review is required. NAS has published the most authoritative and independent reviews of forensic science. In addition to the forensic report, NAS issued reports on sound spectrometry (“voiceprints”), DNA profiling, polygraph testing, and bullet lead analysis. But NAS is not a governmental entity, and its work depends on outside funding. The justice system needs scientific expertise on a continuing basis—and thus institutionalized.

The National Commission’s proposal, endorsed by PCAST, tasked NIST with the responsibility of evaluating forensic disciplines on an ongoing basis. It should be adopted. NIST has the expertise and independence for this task and has been increasingly involved in forensic research. There would be a cost, but litigating validity issues across the country at Daubert and Frye hearings also has a cost. Moreover, there

403. See generally Paul C. Giannelli, Ake v. Oklahoma: The Right to Expert Assistance in a Post-Daubert, Post-DNA World, 89 CORNELL L. REV. 1305 (2004) (discussing the legal disputes over the scope of Ake—e.g., whether it applied to non-capital cases and to non-psychiatric experts).

404. Although prosecutors are ethically obligated to avoid the use of flawed forensic testimony, the National District Attorneys Association recently asserted that bite mark evidence is a “reliable science”—an untenable position. See supra note 82 and accompanying text. See generally Paul C. Giannelli & Kevin C. McMunigal, Prosecutors, Ethics, and Expert Witnesses, 76 FORDHAM L. REV. 1493 (2007).


is a significant expense associated with rectifying the past mistakes that occurred with hair,\textsuperscript{410} bullet lead,\textsuperscript{411} DNA,\textsuperscript{412} and arson cases.\textsuperscript{413}

Unfortunately, the current Attorney General did not renew the commission’s charter in April 2017.\textsuperscript{414} The independent scientists on the commission objected to this action, writing:

The Justice Department now proposes to improve forensic science by moving its oversight and development to an office within the department. This is precisely the opposite of what was recommended by the National Academy of Sciences report and the NCFS. It is a step backwards, because it reinforces the conditions that contributed to the current problems, namely, placing this discipline within the control of law enforcement and prosecutors. The Justice Department is home to many dedicated public servants including scientists whose passion for justice is unquestioned. However, DOJ is not a scientific body, and it is difficult to see how forensic science can become a true science in

\begin{enumerate}
\item[410.] See David R. Cameron, \textit{Forum: Review of FBI Lab Suggests Huge Number of Wrongful Convictions}, \textit{New Haven Register} (April 26, 2015, 5:24 PM), http://www.nhregister.com/opinion/article/Forum-Review-of-FBI-lab-suggests-huge-number-of-11353007.php [https://perma.cc/3NVM-2UN8] (“The FBI review has identified roughly 2,500 cases that fit those criteria. The review is still in its early stages; thus far, it has considered 268 trials involving 284 defendants. It has found that lab examiners gave flawed testimony regarding the comparison of hairs in 257 of the 268 trials—more than 95 percent. Almost all of the examiners over that period—26 of 28—presented flawed testimony.”).
\item[411.] See supra note 248 and accompanying text.
\item[413.] See supra Section 2.C.
\end{enumerate}
that environment. Science flourishes when free and independent; only then can the tools and technology that it creates be truly reliable.415

The AAAS concurred, also stressing that independence “cannot be overstated” and expressing concern about the “inherent conflict of interest in having law enforcement overseeing the work of forensic labs on which police and prosecutors rely to win and defend convictions.”416 The American Academy of Forensic Science also opposed the formation of an Office of Forensic Science within DOJ.417 Instead of heeding this advice, the Attorney General appointed a prosecutor instead of a scientist to head the working group within the DOJ.418

These recent events should be put in context. The 2009 NAS Report recommended the creation of an independent federal entity—the National Institute of Forensic Sciences—to oversee the field, including the establishment of a research agenda.419 If adopted, this proposal would have wrest control of forensic science from law enforcement. The report provided the following justification: Some federal entities were “too wedded” to the status quo and “have failed to pursue a rigorous research agenda to confirm the evidentiary reliability of methodologies used in a number of forensic science disciplines.”420 As a result, these


419. NAS FORENSIC REPORT, supra note 26, at 19 (Recommendation 1(c): “promoting scholarly, competitive peer-reviewed research and technical development in the forensic science disciplines”).

420. Id. at 18.
“agencies are not good candidates to oversee the overhaul of the forensic science community . . . .”421 There is little question that the NAS was referring to National Institute of Justice and the FBI Laboratory. The report noted that, although both had provided “modest leadership” in forensic science, “neither entity has recognized, let alone articulated, a need for change or a vision for achieving it.”422 Consequently, “advancing science in the forensic science enterprise is not likely to be achieved within the confines of DOJ.”423 In fact, law enforcement had manipulated science in the past by shaping the research agenda, limiting access to data, attacking experts who disagreed with its positions, and “spin-n ing” negative reports.424 When Congress did not authorize the creation of the National Institute of Forensic Sciences, DOJ, to its credit, established the NCFS. Most importantly, independent scientists were appointed to the commission.425 Placing science back under the DOJ now is a major and unjustified retreat.

CONCLUSION

This Article explained how the judiciary’s failure to fulfill its gatekeeper role can be traced back to its refusal to demand and properly evaluate foundational research—i.e., Daubert’s first factor, empirical testing. This failure has been systemic. Flawed forensic techniques such as bite mark analysis, microscopic hair comparisons, arson evidence, and comparative bullet lead analysis were routinely admitted into evidence without foundational research. In addition, firearms, toolmark, and fingerprint examiners repeatedly presented overstated and misleading conclusions. This Article also argued that the justice system may be institutionally incapable of applying Daubert in criminal cases because it does not have access to independent scientific expertise on an ongoing basis, and endorsed the NCFS and PCAST recommendation that NIST should be tasked with this responsibility.

Even if an independent scientific review is not institutionalized, PCAST, NCFS, and AAAS have provided guidance for courts dealing with admissibility challenges. First, the flawed techniques discussed in

421. Id.

422. Id. at 16. The Report also stated: “Neither has the full confidence of the larger forensic science community. And because both are part of a prosecutorial department of the government, they could be subject to subtle contextual biases that should not be allowed to undercut the power of forensic science.” Id.

423. Id. at 18.

424. See Giannelli, supra note 201.

425. Having served on the NCFS, the Author believes that there should have been more independent scientists on the Commission.
this Article should be excluded. If used at all, bite mark analysis should be limited to exclusions and perhaps to closed universe situations.426 For hair analysis, mitochondrial DNA analysis is far superior to microscopy. Arson evidence should comport with NFPA 921 and the AAAS report. As noted above, the FBI has abandoned comparative bullet lead analysis.

Second, courts should focus, as Daubert requires, on foundational research. According to PCAST, “neither experience, nor judgment, nor good professional practices (such as certification programs and accreditation programs, standardized protocols, proficiency testing, and codes of ethics) can substitute for actual evidence of foundational validity and reliability.”427 The NCFS concurred.428

Third, subjective methods can be empirically tested. Such research has been conducted. PCAST identified studies in fingerprint and firearms identification that meet stringent standards.429 These studies show an error rate, which should be presented to the jury.430 However, more than one study is needed.

Fourth, in ruling on admissibility in firearms, toolmark, and fingerprint examination cases, courts should appreciate that there has been a “lost decade”—or two—during which rigorous research was not conducted.431 Instead, the disciplines examined in this article vigorously

426. See, e.g., State v. Lambright, No. M2012-02538-CCA-R3-CD, 2014 WL 46839, at *6 (Jan. 7, 2014) (“Dr. Tabor said that, considering the number of teeth that the victim’s sister had, she would not have been capable of producing the bite mark found on the victim’s nose and upper lip. It was Dr. Tabor’s expert medical opinion that a two-year-old was not capable of producing the nature, severity, number, and orientation of bites sustained by the victim.”).


428. See supra note 379 and accompanying text (emphasizing the importance of published peer review research).

429. See supra note 367 and accompanying text (noting that closed-set studies are not sufficiently robust).

430. If examiners claim that there is no error rate, they should be required to explain why not. See White House PCAST Report, supra note 38, at 19 (“In testimony, examiners should always state clearly that errors can and do occur, due both to similarities between features and to human mistakes in the laboratory.”).

431. This depends on when the clock started ticking. Daubert was decided in 1993. In 1995, the first challenge to handwriting testimony was decided. United States v. Starzecpyzel, 880 F. Supp. 1027 (S.D.N.Y. 1995). As noted above, the Williamson case on microscopic hair analysis was decided the same year. See supra notes 92–93 and accompanying text. Llera Plaza was decided in 2002. See supra notes 312–315 and accompanying text (addressing the admissibility of fingerprint testimony). And Green was decided in 2005. See supra notes 254–258 and accompanying text (addressing firearms identification testimony). Moreover, during this period numerous courts restricted the use of handwriting identification. See, e.g., United States v.
resisted the views of independent scientists, and they were typically supported by prosecutors. For example, the wrongful execution of Cameron Todd Willingham, which triggered numerous scientific reviews, was not enough to persuade the Texas Fire Marshal Office that its evidence was flawed,\textsuperscript{432} and it took a serendipitous event—the Madrid train bombing—to provoke fingerprint research.

In short, forensic evidence is in a “Catch-22” situation: Only the federal government has the resources to fund the needed independent research, but it has no incentive to do so as long as evidence continues to be admitted without proper limitations. Until more scientifically sound studies are published and peer-reviewed by independent scientists, courts should follow the approach adopted in \textit{United States v. Glynn},\textsuperscript{433} which permitted the expert to testify only that it was “more likely than not” that recovered bullets and cartridge cases came from a particular weapon.\textsuperscript{434}

Fifth, the presentation of expert testimony needs to be controlled. Once again, PCAST made several recommendations, including:

\begin{itemize}
\item Statements suggesting or implying greater certainty are not scientifically valid and should not be permitted. In particular, courts should never permit scientifically indefensible claims such as: “zero,” “vanishingly small,” “essentially zero,” “negligible,” “minimal,” or “microscopic” error rates; “100 percent certainty” or proof “to a reasonable degree of scientific certainty;” identification “to the exclusion of all other sources;” or a chance of error so remote as to be a “practical impossibility.”\textsuperscript{435}
\item The NCFS also recommended against the use of the phrase “reasonable degree of scientific certainty”\textsuperscript{436} and the 2009 NAS report criticized the use of “zero error rates” and claims of infallibility.\textsuperscript{537} The recent AAAS fingerprint report found no scientific justification for statements of “identity” or “practical certainty” and cautioned against the use of terms such as “match,” “identification,” and “individualization.”\textsuperscript{438}
\end{itemize}

\begin{flushleft}
\textsuperscript{432} \textit{See supra} note 161 and accompanying text.
\textsuperscript{433} 578 F. Supp. 2d 567 (S.D.N.Y. 2008).
\textsuperscript{434} \textit{Id.} at 575.
\textsuperscript{435} \textit{White House PCAST Report, supra} note 38, at 19.
\textsuperscript{436} \textit{See supra} note 283 and accompanying text.
\textsuperscript{437} \textit{See supra Section III.B.3.}
\textsuperscript{438} \textit{AAAS Fingerprint Report, supra} note 39, at 11.
\end{flushleft}
Unfortunately, there is little reason to believe that examiners will give up their claims that there is a scientific foundation for their discipline. A subjective method without a meaningful protocol can hardly claim to be a science. This is not a new issue, as an editorial in the prestigious scientific journal, Science, entitled “Forensic Science: Oxymoron?” and written by the editor-in-chief, made the same point fifteen years ago. Similarly, the 2009 NAS Report commented: “The law’s greatest dilemma in its heavy reliance on forensic evidence . . . concerns the question of whether, and to what extent, there is science in any given forensic science discipline.” After Daubert hearings, one court “very quickly concluded that whatever else ballistics identification analysis could be called, it could not fairly be called ‘science.’” The same is true of fingerprint examinations.

Courts should also guard against attempts to introduce claims of “science” through the backdoor by means of circumlocutions such as statements that firearms and fingerprint identifications are subjective techniques that are “based on science.” This is misleading. Many things are “based on science”—e.g., riding a bike, throwing a curve ball, and flying a kite.

Sixth, proficiency testing issues will continue to be litigated. These tests have long been suspect. They are not conducted blind and are not challenging.

439. See Kennedy, supra note 339 (discussing the cancellation of a National Academies project designed to examine various forensic science techniques, including fingerprinting, because the Departments of Justice and Defense insisted on a right of review that the Academy had refused to other grant sponsors).

440. NAS FORENSICS SCIENCES REPORT, supra note 26, at 9.

441. United States v. Glynn, 578 F. Supp. 2d 567, 570 (S.D.N.Y. 2008). See also United States v. Starzecpyzel, 880 F. Supp. 1027, 1038 (S.D.N.Y. 1995) (“[F]orensic document examination, despite the existence of a certification program, professional journals and other trappings of science, cannot, after Daubert, be regarded as ‘scientific . . . knowledge.’’); id. at 1041 (“[W]hile scientific principles may relate to aspects of handwriting analysis, they have little or nothing to do with the day-to-day tasks performed by [Forensic Document Examiners]. . . . [T]his attenuated relationship does not transform the FDE into a scientist.”).

442. For example, a fingerprint examiner from New Scotland Yard testified in one case that the FBI proficiency tests were deficient: “It’s not testing their ability. It doesn’t test their expertise. I mean I’ve set these tests to trainees and advanced technicians. And if I gave my experts these tests, they’d fall about laughing.” United States v. Llera Plaza, 188 F. Supp. 2d 549, 558 (E.D. Pa. 2002). The district court agreed, noting that “the FBI examiners got very high proficiency grades, but the tests they took did not. . . . [O]n the present record I conclude that the proficiency tests are less demanding than they should be.” Id. at 565. Similarly, in a trial involving handwriting comparisons, the court wrote:
NCFS “during its seventh meeting on August 10, 2015 that he has been under commercial pressure to make proficiency tests easier.”

There were aspects of Mr. Cawley’s testimony that undermined his credibility. Mr. Cawley testified that he achieved a 100% passage rate on the proficiency tests that he took and that all of his peers always passed their proficiency tests. Mr. Cawley said that his peers always agreed with each others’ results and always got it right. Peer review in such a “Lake Woebegone” environment is not meaningful.

United States v. Lewis, 220 F. Supp. 2d 548, 554 (S.D.W. Va. 2002); see supra note 338 (discussing fingerprint proficiency testing).