SCIENTIFIC VALIDITY AND ERROR RATES: A SHORT RESPONSE TO THE PCAST REPORT

*Ted Robert Hunt**

INTRODUCTION

In *Daubert v. Merrell Dow Pharmaceuticals*,[1] the U.S. Supreme Court set the standard for admitting scientific evidence in federal court. The Court ruled that testimony concerning scientific evidence must be founded, in part, on scientific knowledge supported by "appropriate validation—i.e. 'good grounds,' based on what is known."[2] It also instructed that "in a case involving scientific evidence, *evidentiary reliability* will be based upon *scientific validity*."[3] The task of determining scientific validity, and therefore legal reliability, fell to trial judges.[4]

The Court offered some "general observations" about the types of things trial judges might take into account when assessing whether a theory or technique amounts to scientific knowledge.[5] These five observations are now widely known as the "*Daubert* factors."[6] Importantly, however, the Court declined to adopt a strict legal or scientific litmus test for establishing scientific validity. Instead, its general observations were framed by bookend admonitions that sought to dissuade the rigid application of those factors.[7] To that end, the Court stated: "we do not presume to set out a

* Senior Advisor on Forensic Science, U.S. Department of Justice. Prior to joining the U.S. Department of Justice, Mr. Hunt was Chief Trial Attorney at the Jackson County Prosecutor's Office in Kansas City, Missouri, where he served as a prosecuting attorney for over twenty-five years.

This Article was prepared as a companion to the *Fordham Law Review* Reed Symposium on Forensic Expert Testimony, *Daubert*, and Rule 702, held on October 27, 2017, at Boston College School of Law. The Symposium took place under the sponsorship of the Judicial Conference Advisory Committee on Evidence Rules. For an overview of the Symposium, see Daniel J. Capra, *Foreword: Symposium on Forensic Testimony,* Daubert*, and Rule 702*, 86 FORDHAM L. REV. 1459 (2018).

1. Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579 (1993).
2. *Id.* at 590.
3. *Id.* at 591 n.9.
4. *Id.* at 589.
5. *Id.* at 593–94.
6. *See generally* Harvey Brown, *Eight Gates for Expert Witnesses*, 36 HOUSTON L. REV. 743 (1999); Harvey Brown & Melissa Davis, *Eight Gates for Expert Witnesses: Fifteen Years Later*, 52 HOUSTON L. REV. 1 (2014); John B. Meixner & Shari Seidman Diamond, *The Hidden* Daubert *Factor: How Judges Use Error Rates in Assessing Scientific Evidence*, 2014 WIS. L. REV. 1063 (2014).
7. *Daubert*, 509 U.S. at 593.

definitive checklist or test,"[8] as "[t]he inquiry envisioned by [Federal Rule of Evidence] 702 is, we emphasize, a flexible one."[9]

In September 2016, the President's Council of Advisors on Science and Technology (PCAST) released a report titled *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* ("Report").[10] At that time, the Department of Justice (DOJ) issued a short statement in support of PCAST's efforts to advance the reliability of forensic science.[11] However, the DOJ also noted that PCAST had overstepped its role as a science and technology advisory council by making recommendations about the courtroom use of forensic science.[12] The DOJ stated, "[w]hile we appreciate [PCAST's] contribution to the field of scientific inquiry, the [DOJ] will not be adopting the recommendations related to the admissibility of forensic science evidence."[13]

Much of the Report describes PCAST's view of how it believes the scientific validity of forensic feature-comparison methods should be established. To develop its novel position on this issue, PCAST co-opted the term "scientific validity" from the *Daubert* decision and divided it into two parts: "foundational validity"[14] and "validity as applied."[15] PCAST then equated its new term, foundational validity, with *Daubert*'s term, scientific validity.[16] Next, PCAST described foundational validity as the scientific benchmark that corresponds to the legal requirement, in Rule 702,[17] that evidence must be based on "reliable principles and methods."[18]

After the Report's release, some advocates have urged that it be used to exclude or limit the use of forensic feature-comparison evidence in criminal cases.[19] Defense attorneys who have enlisted this strategy generally cite PCAST's conclusion that some forensic methods are not reliable or have

8. *Id.*
9. *Id.* at 594.
10. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., EXEC. OFFICE OF THE PRESIDENT, FORENSIC SCIENCE IN CRIMINAL COURTS: ENSURING SCIENTIFIC VALIDITY OF FEATURE-COMPARISON METHODS (2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final [https://perma.cc/VJB4-5JVQ] [hereinafter PCAST REPORT].
11. Gary Fields, *White House Advisory Council Is Critical of Forensics Used in Criminal Trials*, WALL ST. J. (Sept. 20, 2016, 4:25 PM), https://www.wsj.com/articles/white-house-advisory-council-releases-report-critical-of-forensics-used-in-criminal-trials-1474394743 [https://perma.cc/N9KM-NHJL].
12. *Id.*
13. *Id.*
14. PCAST REPORT, *supra* note 10, at 43.
15. *Id.*
16. *Id.* at 4–5, 43; *see* Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 590–91 n.9 (1993).
17. FED. R. EVID. 702.
18. PCAST REPORT, *supra* note 10, at 4–5, 43.
19. *See, e.g.*, Eric Alexander Vos, *Using the PCAST Report to Exclude, Limit, or Minimize Experts*, CRIM. JUST. (Am. Bar Ass'n, New York, N.Y.), Summer 2017, at 15, https://www.americanbar.org/content/dam/aba/publications/criminal_justice_magazine/v32/VOS.authcheckdam.pdf [https://perma.cc/6LSC-5R6C].

not been sufficiently validated.[20] The DOJ strongly disagrees with this position. It also disagrees with PCAST's novel and purportedly exclusive "litmus test" for determining scientific validity. The DOJ is firmly committed to only using valid and reliable forensic methods.

To date, the DOJ has largely responded to the Report through the filing of legal briefs in criminal cases. While overwhelmingly successful in litigation, these responses are not widely circulated.[21] To clarify the DOJ's position, this Article is a short response to the Report's discussion of scientific validity. The focus is on PCAST's use of the term foundational validity, its views on error rates, and the proposed application of these concepts to forensic feature-comparison methods. First, Part I explains the standards for scientific validation of forensic methods, including those set forth by PCAST, international organizations, and other countries. Then, Part II describes the problems with PCAST's view and demonstrates how it is inconsistent with mainstream and international scientific thought. Finally, this Article concludes that the Supreme Court's standard for the admission of scientific evidence, as outlined in *Daubert*, is appropriate in the context of forensic evidence.

I. STANDARDS FOR SCIENTIFIC VALIDATION

According to PCAST, foundational validity for a forensic feature-comparison method "requires that [the method] be shown, based on empirical studies, to be *repeatable*, *reproducible*, and *accurate*, at levels that have been measured and are appropriate to the intended application."[22] This statement is correct and consistent with mainstream scientific thought.[23] However, PCAST's discussion of validation does not end there. Instead, it takes the extraordinary step of purporting to impose a novel, non-severable, nine-part test that prescribes the *exclusive* experimental design and mandatory criteria for validating "subjective feature-comparison methods."[24] This claim puts PCAST at odds with mainstream scientific thought.

20. *See, e.g.*, United States v. Pitts, No. 16-CR-550, 2018 U.S. Dist. LEXIS 30589, at *9–11 (E.D.N.Y. Feb. 26, 2018) (fingerprints); United States v. Casaus, No. 14-cr-00136-CMA-09, 2017 U.S. Dist. LEXIS 212945, at *1–3 (D. Colo. Dec. 29, 2017) (fingerprints); United States v. North, No. 1:16-cr-309-WSD, 2017 U.S. Dist. LEXIS 190935, at *8 (N.D. Ga. Nov. 17, 2017) (gunshot residue); United States v. Bonds, No. 15 CR 573-2, 2017 U.S. Dist. LEXIS 166975, at *4–6 (N.D. Ill. Oct. 10, 2017) (fingerprints).

21. *See, e.g.*, *Casaus*, 2017 U.S. Dist. LEXIS 212945, at *3–5; *North*, 2017 U.S. Dist. LEXIS 190935, at *7–9; *Bonds*, 2017 U.S. Dist. LEXIS 166975, at *5–13.

22. PCAST REPORT, *supra* note 10, at 4–5.

23. *See infra* Part I.B–C.

24. PCAST REPORT, *supra* note 10, at 46.

A. PCAST's Novel Validation Litmus Test

PCAST describes experiments that meet *each* of the nine requirements as "appropriately designed studies."[25] Its litmus test for establishing foundational validity is as follows:

> Scientific validation studies—intended to assess the validity and reliability of a metrological method for a particular forensic feature-comparison application—must satisfy a number of criteria.
>
> (1) The studies must involve a sufficiently large number of examiners and must be based on sufficiently *large* collections of *known* and *representative* samples from *relevant* populations to reflect the range of features or combinations of features that will occur in the application. In particular, the sample collections should be:
>
> (a) representative of the quality of evidentiary samples seen in real cases. (For example, if a method is to be used on distorted, partial, latent fingerprints, one must determine the *random match probability*—that is, the probability that the match occurred by chance—for distorted, partial, latent fingerprints; the random match probability for full scanned fingerprints, or even very high quality latent prints would not be relevant.)
>
> (b) chosen from populations relevant to real cases. For example, for features in biological samples, the false positive rate should be determined for the overall US population and for major ethnic groups, as is done with DNA analysis.
>
> (c) large enough to provide appropriate estimates of the error rates.
>
> (2) The empirical studies should be conducted so that neither the examiner nor those with whom the examiner interacts have any information about the correct answer.
>
> (3) The study design and analysis framework should be specified in advance. In validation studies, it is inappropriate to modify the protocol afterwards based on the results.
>
> (4) The empirical studies should be conducted or overseen by individuals or organizations that have no stake in the outcome of the studies.
>
> (5) Data, software and results from validation studies should be available to allow other scientists to review the conclusions.
>
> (6) To ensure that conclusions are reproducible and robust, there should be multiple studies by separate groups reaching similar conclusions.[26]

---

25. *Id.* at 9 ("As noted above, the foundational validity of a subjective method can only be established through multiple, *appropriately designed* black-box studies." (emphasis added)). PCAST claimed to have reviewed 2100 scientific studies and found only three studies to be "appropriately designed"—two latent print studies and one firearms and toolmarks study—according to its newly-minted criteria for establishing what it described as "foundational validity." *Id.* at 96, 111. For a list of PCAST references, see Office of Sci. & Tech. Pol'y, *PCAST Documents & Reports*, WHITE HOUSE: PRESIDENT BARACK OBAMA, https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports [https://perma.cc/BJF4-EZXY] (last visited Feb. 26, 2018).

26. PCAST REPORT, *supra* note 10, at 52–53.

To be clear, none of the listed criteria is novel or controversial. All are well-known aspects of good experimental design and sound scientific practice. Each can play an important role in the validation process. However, what *is* novel and controversial is PCAST's claim that a single experimental design and the non-severable use of these criteria is the *only* way to establish the scientific validity of "subjective" forensic feature-comparison methods.[27]

To that end, PCAST states: "the *sole* way to establish foundational validity is through multiple independent 'black-box' studies that measure how often examiners reach accurate conclusions across many feature-comparison problems involving samples representative of the intended use. In the absence of such studies, a feature-comparison method cannot be considered scientifically valid."[28] This position is out of step with mainstream scientific thought.

B. The International Scientific Standard for Scientific Validation

The International Organization for Standardization (ISO) is the preeminent body for developing and publishing consensus international standards. ISO is composed of 161 national standards bodies from countries across the world.[29] Subject matter experts from various fields develop these standards. An ISO International Standard represents "a global consensus on the state of the art in the subject of that standard."[30]

ISO/IEC 17025 is the standard that governs the general requirements for the competence of testing and calibration laboratories.[31] This standard guides the core scientific activities and management operations of labs engaged in a diverse range of activities.[32] These activities include clinical testing, research, and forensic science, among others.[33] Identical accreditation requirements apply to all labs, regardless of whether they test clinical samples, groundwater, or forensic evidence.[34]

ISO does not recognize or use PCAST's term, foundational validity, in any of its standards or definitions. Instead, the non-bifurcated term, validation, is used to describe the process of determining whether a method

27. *Id.* at 68.
28. *Id.* (emphasis added).
29. *See All About ISO*, ISO, https://www.iso.org/about-us.html [https://perma.cc/CP5P-BE7M] (last visited Feb. 26, 2018).
30. INT'L ORG. FOR STANDARDIZATION, GUIDANCE FOR ISO NATIONAL STANDARDS BODIES: ENGAGING STAKEHOLDERS AND BUILDING CONSENSUS 2 (2010), https://www.iso.org/files/live/sites/isoorg/files/archive/pdf/en/guidance_nsb.pdf [https://perma.cc/8NWL-JLAT].
31. *See ISO/IEC 17025:2017*, ISO, https://www.iso.org/obp/ui/#iso:std:iso-iec: 17025: ed-3:v1:en [https://perma.cc/C4V5-2RU4] (last visited Feb. 26, 2018).
32. *Id.* § 1.
33. *Id.*
34. *Id.*

is fit for its intended purpose.[35] For example, ISO generally defines validation as "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled."[36] Likewise, in the context of ISO/IEC 17025, validation is defined as when "the specified requirements are adequate for an intended use."[37] Section 7.2.2 governs the validation of test methods.[38] It states that "validation shall be as extensive as is necessary to meet the needs of the given application or field of application."[39]

A separate note to section 7.2.2 provides a non-exclusive, non-prescriptive list of techniques that can be used—either alone or in combination with others—to validate a method.[40] These techniques include: the evaluation of bias and precision using reference standards or reference materials, systematic assessment of the factors influencing the result, evaluation of method robustness through variation of controlled parameters, comparison of results achieved with other validated methods, inter-laboratory comparisons, evaluation of measurement uncertainty based on theoretical principles of the method, and practical experience in performing the sampling or test method.[41]

In direct contrast to PCAST's validation litmus test, the ISO does not prescribe *how* labs must validate their methods, *which* criteria must be included, or *what* experimental design must be used. Instead, "[t]he performance characteristics of validated methods, as assessed for the intended use, shall be relevant to the customer's needs and consistent with specified requirements."[42] The selection of those specified requirements and experimental designs are the responsibility of each laboratory.[43]

C. The Holistic, Flexible Approach to Scientific Validation

The American Association for the Advancement of Science (AAAS) recently published a study on latent fingerprint examination.[44] The authors

35. *See id.* § 3.9; *ISO/IEC 9000:2015* § 3.8.13, ISO, https://www.iso.org/obp/ui/#iso:std:iso:9000:ed-4:v1:en [https://perma.cc/7E5R-MMDH] (last visited Feb. 26, 2018).

36. *ISO/IEC 9000:2015*, *supra* note 35, § 3.8.13.

37. *ISO/IEC 17025:2017*, *supra* note 31, § 3.9.

38. *Id.* § 7.2.2.

39. *Id.* § 7.2.2.1.

40. *Id.* § 7.2.2.1 n.2.

41. *Id.*

42. *Id.* § 7.2.2.3.

43. JOHN W. CRESWELL, RESEARCH DESIGN: QUALITATIVE, QUANTITATIVE, AND MIXED METHODS APPROACHES 21 (4th ed. 2014) ("In planning a research project, researchers need to identify whether they will employ a qualitative, quantitative, or mixed methods approach. This approach is based on bringing together a worldview or assumptions about research, a specific design, and research methods. Decisions about choice of an approach are further influenced by the research problem or issue being studied, the personal experiences of the researcher, and the audience for whom the researcher writes.").

44. *See* WILLIAM THOMPSON ET AL., FORENSIC SCIENCE ASSESSMENTS: A QUALITY AND GAP ANALYSIS (2017), https://mcmprodaaas.s3.amazonaws.com/s3fs-public/reports/Latent%20Fingerprint%20Report%20FINAL%209_14.pdf?i9xGS_EyMHnIPLG6INIUyZb66L5cLdlb [https://perma.cc/C9K2-T6QG].

disagreed with PCAST's premise that *only* those research papers "intentionally and appropriately designed" should be considered when assessing evidential support for method validation.[45] Instead, the AAAS used the concept of "convergent validity" to draw conclusions regarding scientific validity from the body of relevant literature as a whole.[46] This conceptual approach acknowledges that various studies will have different strengths and weaknesses.[47] It also recognizes that some studies can reinforce others and collectively support conclusions not warranted on the basis of a single study.[48]

Others share this same general point of view. For example, one group of experts has observed: "There is *no one best way* to study a phenomenon of interest. Each methodological choice involves trade-offs."[49] Trade-offs, in turn, require flexibility, and flexibility is required by the pull of competing interests, existing resources, and countless other operational considerations.[50] The international scientific community, through ISO/IEC 17025, acknowledges these realities by observing that "[v]alidation is always a balance between costs, risks and technical possibilities."[51] This balancing requires a realistic assessment of the object of inquiry, the nature of the analysis, and the specifications for a given application.

Many feature-comparison methods rely on human interpretation and judgment. In the United Kingdom, the Forensic Science Regulator publishes the Forensic Code of Practice and Conduct ("Code"), which states:

> The functional and performance requirements for interpretive methods are less prescriptive than for measurement-based methods. They concentrate on the competence requirements for the staff involved and how the staff shall demonstrate that they can provide consistent, reproducible, valid and

45. *Id.* at 44.
46. *Id.*
47. *Id.*
48. *Id.* at 94.
49. 1 DAVID L. FAIGMAN ET AL., MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY, STATISTICS & RESEARCH METHODS § 1:22 (2010) (emphasis added); *see also ISO/IEC 17025:2005* § 5.4.5.3. n.3, ISO, https://www.iso.org/obp/ui/#iso:std:iso-iec:17025:ed-2:v1:en [https://perma.cc/ER8Y-CRNY] (last visited Feb. 26, 2018) ("Validation is always a balance between costs, risks and technical possibilities. There are many cases in which the range and uncertainty of the values (e.g. accuracy, detection limit, selectivity, linearity, repeatability, reproducibility, robustness and cross-sensitivity) can only be given in a simplified way due to lack of information.").
50. GEOFFREY MARCZYK ET AL., ESSENTIALS OF RESEARCH DESIGN AND METHODOLOGY 137 (2005) ("The most obvious limitation of studies that employ a randomized experimental design is their logistical difficulty. Randomly assigning participants in certain settings (e.g., criminal justice, education) may often be unrealistic, either for logistical reasons or simply because it may be considered inappropriate in a particular setting. Although efforts have been made to extend randomized designs to more real-world settings, it is often not feasible. In such cases, the researcher often turns to quasi-experimental designs.").
51. *ISO/IEC 17025:2005*, *supra* note 49, § 5.4.5.3 n.3.

> reliable results that are compatible with the results of other competent staff.[52]

Similar to ISO, the Code provides a non-prescriptive, non-exclusive combination of measures that may be used to validate interpretive methods.[53] These include blind confirmation by a second examiner, inter-laboratory comparisons and proficiency tests, and the in-house use of competency tests.[54] The Code also states that an interpretive method "shall require only the relevant subset of . . . parameters and characteristics for measurement-based methods."[55]

Finally, an equally-flexible view of validating interpretive methods is shared by Australia's National Association of Testing Authorities (NATA). NATA recognizes that the validation of interpretive methods "is more challenging and less proscriptive than it is for analytical methods."[56] However, validity can be established "if the analyst or examiner repeatedly obtains correct results for positive and negative known tests."[57] In addition, NATA correctly concedes that certain validation parameters "are not relevant in subjective tests."[58]

Unlike PCAST, these scientific bodies do not require that multiple, independent black-box studies be performed to establish the scientific validity of forensic feature-comparison methods. Instead, they all promote a holistic, flexible, and pragmatic approach to validation.[59] This approach considers the body of *all* relevant evidence that bears upon a method's accuracy and precision. It is also consistent with the view that interpretive methods require flexible, non-prescriptive validation criteria.[60] Finally, it understands that validation is always a balance of competing interests and that various experimental techniques may be used when assessing a method's fitness for a particular use.

II. CONCERNS WITH THE PCAST APPROACH

The DOJ fully agrees with PCAST that the feature-comparison methods used by forensic experts must be scientifically valid and reliable. The

52. FORENSIC SCI. REGULATOR, CODES OF PRACTICE AND CONDUCT FOR FORENSIC SCIENCE PROVIDERS AND PRACTITIONERS IN THE CRIMINAL JUSTICE SYSTEM § 20.9.1 (2016).

53. *Id.*

54. *Id.*

55. *Id.* § 20.9.2.

56. NAT'L ASS'N OF TESTING AUTHS., TECHNICAL NOTE 17: GUIDELINES FOR THE VALIDATION AND VERIFICATION OF QUANTITATIVE AND QUALITATIVE TEST METHODS § 5 (2013).

57. *Id.* § 5.1.

58. *Id.* § 5.

59. CRESWELL, *supra* note 43, at 10 (noting that in a pragmatic approach, "[t]here is a concern with applications—what works—and solutions to problems. Instead of focusing on methods, researchers emphasize the research problems and use all approaches available to understand the problem").

60. *Id.* at 11 (noting that when using the pragmatic philosophical approach to research, "[i]ndividual researchers have a freedom of choice. In this way, researchers are free to choose the methods, techniques, and procedures of research that best meet their needs and purposes").

empirical demonstration of accuracy and precision is a critical part of scientific validation. However, the DOJ rejects PCAST's novel premise that the scientific validity of subjective forensic feature-comparison methods can only be established by strict adherence to its non-severable nine-part litmus test. The DOJ also disagrees with PCAST's assertion that rate of error for these methods can only be established through the use of black-box studies. PCAST's nine-part test and approach to error rates puts PCAST at odds with mainstream international scientific thought.

A. Erroneous Exclusivity of the PCAST's Litmus Test

Before the release of the Report in September 2016, the DOJ was unaware of any discipline-specific multipart litmus test claimed by any group—scientific or regulatory—to be the *only* way to establish scientific validity. PCAST not only failed to cite the origin of its test, but it also failed to identify when, where, or how its test had been previously described or if its test was ever fully used prior to publication. Thus, PCAST's targeted application of this test to forensic feature-comparison methods appears to be unprecedented.

It is important to note that PCAST's position on method validation, through the use of black-box studies, is not a scientific imperative. It merely represents one view—an extremely narrow view—of the appropriate means by which empirical data can be generated and used to assess scientific validity. Mainstream scientific thought, however, is not so narrow and prescriptive. Instead, it is consistent with the view that *all* available information, evidence, and data derived from a multitude of studies—diverse and varied in experimental design—can be appropriately considered when assessing method accuracy, precision, and fitness for an intended use.[61] "Only through replications, using *various designs and methods*, do scientists gain confidence that a hypothesis has been sufficiently corroborated."[62] PCAST's insistence on the use of a single inflexible experimental design and mandatory set of criteria is thus inconsistent with mainstream scientific thought.

B. Error Rates

One of the five "observations" made by the *Daubert* Court about whether a theory or technique has attained the status of "scientific knowledge" was its "known or potential rate of error."[63] Unfortunately, some commentators discuss error rates with a specious and superficial simplicity. They treat the concept as if it were self-defining and had a uniform meaning and

61. *See supra* Part I.B–C.
62. FAIGMAN ET AL., *supra* note 49, § 1:22 (emphasis added).
63. Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 594 (1993).

application to multiple different methods.[64] Many of these critiques are misleading and unhelpful.

An error rate is not a static concept. It has neither a uniform definition nor a fixed existence. Instead, error rates are highly dynamic and dependent upon a wide range of human choices, assumptions, and values that relate to the particular application, object, and variables chosen (or ignored) for measurement. Other factors that affect a given rate include the definition of "error," and the time, place, and manner in which measurements are made.[65]

These choices have a direct impact on both the data collected and the resulting rate. Different choices, assumptions, and values will lead to different rates. Moreover, established rates will constantly change based on new facts, applications, and human intervention after error detection and remediation. A rate derived from one application—given a series of choices—will not replicate in a separate but related context.[66] As a result, a calculated error rate is, at best, a highly generalized proxy for the true value at any given moment in time. Determining a reasonably accurate error rate is like to trying to hit a moving target.

C. PCAST's Views on Error Rates

In its discussion of foundational validity, PCAST emphasizes the importance of determining error rates for forensic feature-comparison methods. The Report correctly states that "all laboratory tests and feature comparison analyses have non-zero error rates."[67] However, it also purports to describe exactly *how* these rates must be calculated.

---

64. *See, e.g.*, NAT'L RESEARCH COUNCIL, NAT'L ACADS., STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES: A PATH FORWARD 122 (2009) (stating that the estimation of error rates requires "rigorously developed and conducted scientific studies" without explaining the appropriate experimental design, scope, or execution of such studies); Erin Murphy, *The New Forensics: Criminal Justice, False Certainty, and the Second Generation of Scientific Evidence*, 95 CALIF. L. REV. 721, 795–97 (2007) (calling for an unspecified error rate threshold for the admissibility of forensic evidence without explaining how error should be defined, determined, or how evidence-excluding rates should be applied to different forensic disciplines); Munia Jabbar, Note, *Overcoming* Daubert*'s Shortcomings in Criminal Trials: Making the Error Rate the Primary Factor in* Daubert*'s Validity Inquiry*, 85 N.Y.U. L. REV. 2034, 2037 (2010) (calling for an error rate to be the "primary factor in the validity inquiry under *Daubert*" for forensic evidence without explaining how error rates should be defined, determined, or applied to different forensic disciplines).

65. *See* MARCZYK ET AL, *supra* note 50, at 178–92 (discussing threats to the generalizability of research findings, including sample characteristics, experimental conditions and circumstances, as well as the timing of the assessment and measurement).

66. *Id.* at 180 ("Every study operates under a unique set of conditions and circumstances related to the experimental arrangement. The most commonly cited examples include the research setting and the researchers involved in the study. The major concern with this threat to external validity is that the findings from one study are influenced by a set of unique conditions, and thus may not necessarily generalize to another study, even if the other study uses a similar sample.").

67. PCAST REPORT, *supra* note 10, at 3, 29.

First, according to PCAST, a black-box study design is required if the method is "subjective."[68] Second, the calculation of false positive results must be based solely on the number of conclusive determinations, rather than the proportion of all examinations.[69] Third, only the percentage of false positives that occupy the upper bound of a 95 percent confidence interval should be reported. PCAST believes that even reporting accurate and empirically-derived lower bound false positive data would be an attempt at "obfuscation."[70] Fourth, forensic examiners, who took no part in these black-box studies, should testify that the box error rates are applicable to the case at hand.[71]

It is important to recognize that PCAST's views about *how* to calculate error rates are value laden and reflect PCAST's assumptions, choices, and attitudes about the relevant objects, variables, and methods of measurement.[72] As such, its opinions are not fixed, immutable, or even generally-accepted principles of science. Rather, PCAST's views represent one set of choices among a broader range of options. There is great diversity of scientific thought about both *whether* and *how* error rates should be determined for forensic methods.[73] In fact, ISO/IEC 17025 does

68. *Id.* at 46, 143.
69. *Id.* at 51–52.
70. *Id.* at 153.
71. *Id.* at 56, 66, 112, 147, 150.
72. KENNETH S. BORDENS & BRUCE B. ABBOTT, RESEARCH DESIGN & METHODS: A PROCESS APPROACH 93 (2005) ("Values . . . can creep into science when scientists go beyond describing and explaining relationships and begin to speculate on what ought to be. . . . On another level, this influence of values also is seen when researchers conduct research to influence the course of political and social events.").
73. *See, e.g.*, COLIN G.G. AITKEN & FRANCO TARONI, STATISTICS AND THE EVALUATION OF EVIDENCE FOR FORENSIC SCIENTISTS 424 (2004) (suggesting that proficiency tests should be used to determine error rates); JOHN S. BUCKLETON ET AL., FORENSIC DNA EVIDENCE INTERPRETATION 76–77 (2d ed. 2016) (noting that error and error rates should be examined on a per-case basis); NAT'L RESEARCH COUNCIL, NAT'L ACADS., DNA TECHNOLOGY IN FORENSIC SCIENCE 89 (1992) (suggesting that proficiency tests should be used to calculate error rates); NAT'L RESEARCH COUNCIL, NAT'L ACADS., THE EVALUATION OF FORENSIC DNA EVIDENCE 85–88 (1996) (suggesting that retesting/duplicate tests should be used to determine error rates); NAT'L RESEARCH COUNCIL, *supra* note 64, at 122 (noting that "rigorously developed and conducted scientific studies" of unspecified design and criteria are required to estimate error rates); BERNARD ROBERTSON ET AL., INTERPRETING EVIDENCE: EVALUATING FORENSIC SCIENCE IN THE COURTROOM 138 (2d ed. 2016) (noting that the possibility of lab error is a critical consideration in determining error rates for a particular study and rejecting use of past error rates in new studies); THOMPSON ET AL., *supra* note 44, at 47 (suggesting that blind test samples introduced into casework should be used); Simon A. Cole, *More Than Zero: Accounting for Error in Latent Fingerprint Identification*, 95 J. CRIM. L. & CRIMINOLOGY 985, 989 (2005) (noting that attempts to assess the error rate for latent fingerprint identification should not yield a single error rate, but many error rates showing the rate of error for different levels of latent print quantity and quality, and stating that "[o]ne key hindrance to generating this sort of information is the lack of an accepted metric for measuring either latent print quality and/or quantity or the difficulty of a comparison"); I.W. Evett et al., *Finding a Way Forward for Forensic Science in the US—A Commentary on the PCAST Report*, 278 FORENSIC SCI. INT'L 16, 22–23 (2017) (suggesting that proficiency tests should be used to determine error rates and rejecting the use of black-box studies in their calculation and courtroom presentation); Jonathan J. Koehler,

not even require that an "error rate" be calculated as part of method validation.[74] And the calculation of a single globally-applicable error rate for subjective forensic methods—determined by multiple black-box studies—is clearly not a generally-accepted scientific principle.

D. Concerns with PCAST's Views on Error Rates

PCAST's insistence on the exclusive use of black-box studies to determine error rates would severely limit opportunities to study a diverse range of questions during the validation process. It would also limit the opportunities for experimental replication. And the lack of replication is "one reason that researchers rarely place much faith in any single study, or even *any single type of study*."[75]

In addition, PCAST's exclusive reliance on black-box studies to determine error rates—and the purported need for examiners to embrace and profess those rates during testimony—raises serious questions about their external validity. External validity refers to "the representativeness of a study. If a study is externally valid, its findings can be generalized to other populations (of people, objects, organizations, times, places, etc.)."[76]

In its recent latent fingerprint report, the AAAS cautioned against extrapolating study-derived error rates to casework scenarios.[77] One concern was that study participants know that they are being tested, which may affect their performance.[78] This phenomenon is known as the "Hawthorne Effect."[79] Another concern was that the decision thresholds used by examiners in controlled studies may differ from those employed during actual casework.[80] Moreover, the AAAS noted that existing studies generally do not fully replicate the conditions that examiners face when performing casework.[81] As a result, the error rates observed in these studies do not necessarily reflect casework conditions.[82] Thus, according to

Proficiency Tests to Estimate Error Rates in the Forensic Sciences, 12 LAW, PROBABILITY & RISK 89, 90–94 (2013) (suggesting that blind proficiency tests should be used).

74. *See ISO/IEC 17025:2017*, *supra* note 31, § 7.2.2–7.2.2.4.

75. FAIGMAN ET AL., *supra* note 49, § 5:39 (emphasis added).

76. *Id.*

77. THOMPSON ET AL., *supra* note 44, at 46.

78. *Id.*

79. The Hawthorne Effect is defined as a "tendency for subjects of research to change their behavior simply because they are being studied." W. PAUL VOGT, DICTIONARY OF STATISTICS AND METHODOLOGY 104 (1993).

80. THOMPSON ET AL., *supra* note 44, at 46.

81. *Id.*

82. *Id.*; *see also* BORDENS & ABBOTT, *supra* note 72, at 113 ("[I]t is a fallacy to assume 'that the purpose of collecting data in the laboratory is to *predict real-life behavior in the real world*.'" (quoting Douglas G. Mook, *In Defense of External Validity*, 38 AM. PSYCHOLOGIST 379, 381 (1983))). Bordens and Abbott also note that:

> [M]uch of the research conducted in the laboratory is designed to determine:
>
> 1. whether something *can* happen, rather than whether it typically *does* happen,

AAAS, "[t]his consideration provides further support for the conclusion that the error rates in black-box studies may not reflect the error rates in casework."[83]

Another concern is PCAST's belief that *only* black-box studies can validate a feature-comparison method.[84] However, black-box studies are merely "input-output research designs where *what happens in between is impossible to study or is ignored*."[85] As such, the inputs "to" and outputs "from" these studies—e.g., true positives, false positives, true negatives, false negatives, and inconclusive results—are what is examined, *not* the method by which those outputs were generated. Therefore, black-box studies—by definition—*cannot* be used to calculate the error rate for a method.

That said, black-box studies may provide some indication of how often a unique collection of examiners—assembled at a given time and place and under defined conditions and constraints—get it right, get it wrong, or simply cannot tell. However, black-box studies do not and cannot reflect the many factors at play in actual casework. This limitation directly and adversely affects the ability to extrapolate study-derived error rates to different times, places, and circumstances. In short, black-box error rates cannot travel. These error rates cannot be generalized to and adopted as the correct error rate across different circumstances. As such, black-box error rates have little relevance to the critical question posed in most litigation: What is the risk that an error occurred in the case at hand?

E. Other Approaches to Error Rates

The National Research Council's (NRC) report, *The Evaluation of Forensic DNA Evidence*,[86] recognized the critical importance of focusing on the risk of case-specific error. On this point, the NRC observed, "[t]he question to be decided is not the general error rate for a laboratory or laboratories over time but rather whether the laboratory doing DNA testing in this particular case made a critical error."[87]

> 2. whether something we specify *ought* to happen (according to some hypothesis) under specific conditions in the lab *does* happen there under those conditions, or
> 3. what happens under conditions not encountered in the real world.
>
> In each of these cases, the objective is to gain insight into the underlying mechanisms of behavior rather than to discover relationships that apply under normal conditions in the real world. It is this understanding that generalizes to everyday life, not the specific findings themselves.

BORDENS & ABBOTT, *supra* note 72, at 113.

83. THOMPSON ET AL., *supra* note 44, at 46; s*ee also* BORDENS & ABBOTT, *supra* note 72, at 114 ("Data obtained from a tightly controlled laboratory may not generalize to more naturalistic situations in which behavior occurs."). Bordens and Abbot define "laboratory" as "any research setting that is artificial relative to the setting in which the behavior naturally occurs." *Id.*

84. PCAST REPORT, *supra* note 10, at 49.

85. VOGT, *supra* note 79, at 24 (emphasis added).

86. *See* THE EVALUATION OF FORENSIC DNA EVIDENCE, *supra* note 73.

87. *Id.* at 85.

The NRC specifically rejected the proposal that laboratories use proficiency tests as the exclusive means for error rate determination, a proposal offered by a previous NRC committee on DNA co-chaired by PCAST Co-Chair, Dr. Eric Lander. The NRC committee stated:

> Estimating rates at which nonmatching samples are declared to match from historical performance on proficiency tests is almost certain to yield wrong values. When errors are discovered, they are investigated thoroughly so that corrections can be made. A laboratory is not likely to make the same error again, so the error probability is correspondingly reduced.[88]

The NRC also observed, "[t]he risk of error is properly considered case by case, taking into account the record of the laboratory performing the tests, the extent of redundancy, and the overall quality of the results."[89] Moreover, the NRC found it unnecessary to debate differing estimates of false positive error when concerns about a false match can be easily resolved by retesting the evidence.[90]

The NRC's view that the focus should be on the *risk* of error, rather than the *rate* of error, is shared by many eminent scientists, statisticians, and forensic practitioners.[91] In their recent response to the Report, Dr. Ian Evett and colleagues wrote, "[t]he notion of an error rate to be presented to courts is misconceived because it fails to recognise that the science moves on as a result of proficiency tests. . . . [O]ur vision is not of the black-box/error rate but of continuous development through calibration and feedback of opinions."[92]

The "known or potential rate of error" [93] is one of many factors that may bear upon the scientific validity of a theory or technique. However, for forensic feature-comparison methods, there is no current scientific consensus on how or whether these rates can—or should—be determined.[94] Each approach involves trade-offs with resulting strengths and weaknesses. Instead of choosing a single approach, a multifaceted parallel focus on rate estimation, error detection, and risk mitigation may be the best path forward.

---

88. *Id.* at 86.
89. *Id.* at 87.
90. *Id.*
91. *See, e.g.*, BUCKLETON ET AL., *supra* note 73, at 76–77 ("Our view is that the possibility of error should be examined on a per-case basis and is always a legitimate defence explanation for the DNA result. . . . The answer lies, in our mind, in a rational examination of errors and the constant search to eliminate them."); ROBERTSON ET AL., *supra* note 73, at 138 ("It is correct . . . to say that the possibility of error by a laboratory is a relevant consideration. It is wrong, however, to assume that the probability of error in a given case is measured by the past error rate. The question is what the chance of error was on this occasion."); Evett et al., *supra* note 73, at 22 ("The notion of an error rate to be presented to courts is misconceived because it fails to recognise that the science moves on as a result of proficiency tests. . . . To repeat then, our vision is not of the black-box/error rate but of continuous development through calibration and feedback of opinions.").
92. Evett et al., *supra* note 73, at 22.
93. Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 594 (1993).
94. *See supra* note 73 and accompanying text.

In sum, error rates derived from studies of various size, scope, and experimental design can provide important information about the general decision-making thresholds of forensic examiners under defined experimental conditions. Intra- and inter-laboratory studies using known samples provide additional information about the ability of local systems to generate valid and reliable results. Competency and proficiency tests add to the body of knowledge by measuring how often forensic examiners get the right answer using known—ground truth—samples. The use of technical review, case controls, and other quality assurance measures are critical components of risk assessment and mitigation. Finally, as noted by the NRC, a wrongfully-accused person's best insurance against false incrimination is the opportunity to have the evidence retested.[95] The typically non-consumptive nature of feature-comparison testing readily facilitates the reanalysis of questioned evidence in most cases.

CONCLUSION

Twenty-five years ago, the U.S. Supreme Court declared that scientific evidence must be both valid and reliable to be admissible.[96] The Court offered a number of observations about the type of considerations that it thought were important in that determination.[97] However, it was quick to emphasize that pragmatic flexibility—rather than a normative and scientific rigidity—was the analytical disposition that should guide the trial court's inquiry.[98] To that end, the Court stated, "we do not presume to set out a definitive checklist or test."[99] Six years later, the Court further advised that "the law grants a district court the same broad latitude when it decides *how* to determine reliability as it enjoys in respect to its ultimate reliability determination."[100] In each case, the trial court has broad discretion to determine whether the *Daubert* factors are a reasonable measure of reliability.[101] These statements make it clear that a single litmus test, or an inflexible set of criteria, was not what the Court had in mind when it tasked trial judges with assessing scientific validity.

The Supreme Court's desire to infuse legal gatekeeping with pragmatic flexibility is consistent with that same general disposition in mainstream scientific thought. ISO/IEC 17025 contains a non-mandatory, non-exclusive set of experimental options for validating scientific methods.[102] Scholars and commentators generally recognize that there is no one best way to study a phenomenon of interest.[103] However, it is clear that a convergent approach to evaluating scientific validity makes the best use of

95. THE EVALUATION OF FORENSIC DNA EVIDENCE, *supra* note 73, at 81.
96. *Daubert*, 509 U.S. at 590–94.
97. *Id.* at 593–94.
98. *Id.* at 593.
99. *Id.*
100. Kumho Tire Co. v. Carmichael, 526 U.S. 137, 142 (1999).
101. *Id.* at 152.
102. *ISO/IEC 17025:2017*, *supra* note 31, § 7.2.2.
103. *See supra* Part I.C.

all available evidence bearing upon the fitness of a particular method for an intended use.[104]

The same holds true for assessing the rate and risk of error.[105] There is no consensus scientific view of how—or even whether—error rates can or should be determined.[106] Thus, a convergent, holistic path forward makes the most sense.[107] This approach considers a variety of published studies with a diversity of design, laboratory-based experiments, inter-laboratory studies, competency and proficiency tests, case-specific technical reviews, quality controls, and liberal re-examination of the evidence by defense experts.[108] All of these activities contribute to a general understanding of the various types and frequency of errors encountered during casework.

In conclusion, the DOJ strongly believes that pragmatic flexibility—the hallmark of both the Federal Rules of Evidence and mainstream scientific thought—must be maintained. Checklists and inflexible litmus tests are inconsistent with both legal and scientific standards and best practice.

104. *See supra* Part I.C.
105. *See supra* Part II.
106. *See supra* Part II.
107. *See supra* Part II.
108. *See supra* Part II.