United States Department of Justice Statement on the PCAST Report: *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*

In September 2016, the President's Council of Advisors on Science and Technology ("PCAST") released its report, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods.*[1] The stated purpose of the Report was to determine what additional scientific steps could be taken after publication of the 2009 National Research Council Report[2] to ensure the validity of forensic evidence used in the legal system.[3] PCAST identified what it saw as two important gaps: 1) the need for clarity about scientific standards for the validity and reliability of forensic methods; and 2) the need to evaluate specific methods to determine whether they had been scientifically established as valid and reliable.[4] The Report "aimed to close these gaps" for a number of what it described as "feature comparison methods."[5] These are methods for comparing DNA samples, latent fingerprints, firearm marks, footwear patterns, hair, and bitemarks.[6]

Unfortunately, the PCAST Report contained several fundamentally incorrect claims. Among these are: 1) that traditional forensic pattern comparison disciplines, as currently practiced, are part of the scientific field of metrology; 2) that the validation of pattern comparison methods can *only* be accomplished by strict adherence to a non-severable set of experimental design criteria; and 3) that error rates for forensic pattern comparison methods can *only* be established through "appropriately designed" black box studies.

The purpose of this statement is to address these claims and to explain why each is incorrect. After the PCAST Report was released, the Department of Justice ("Department") announced that it would not follow PCAST's recommendations.[7] The Report was criticized by a number of commentators and organizations outside of the Department for its analysis, conclusions, factual inaccuracies, and other mistakes.[8] Formally addressing PCAST's incorrect claims has become

---

[1] PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., EXEC. OFFICE OF THE PRESIDENT, FORENSIC SCI. IN CRIM. COURTS: ENSURING SCI. VALIDITY OF FEATURE COMPARISON METHODS (2016), https://obamawhitehouse.archives.gov/sitesdefault/files/microsites/ostp/PCAST/pcast_forensic_science_report_final [https://perma.cc/VJB4-5JVQ] [hereinafter PCAST REPORT].

[2] NAT'L RES. COUNCIL, NAT'L ACAD'S., STRENGTHENING FORENSIC SCI. IN THE UNITED STATES: A PATH FORWARD 122 (Nat'l Acad. Press 2009).

[3] PCAST REPORT, *supra* note 1, at 1.

[4] *Id.*

[5] *Id.* In this statement, we use the term "pattern comparison," rather than PCAST's chosen term, "feature comparison" to describe the general nature of the methods discussed.

[6] *Id.* Department of Justice laboratories do not practice what PCAST described as "bitemark analysis."

[7] 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/whitehouse-advisory-council-releases-report-critical-of-forensics-used-in-criminal-trials-a1474394743 [https://perma.cc/N9KM-NHJL].

[8] *See*, *i.e.*, 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); Letter from Michael A. Ramos, President, Nat'l Dist. Attorneys Ass'n, to President Barack Obama (Nov. 16, 2016), http://tinyurl.com/hczkt3k.; Ass'n of Firearms and Toolmark Examiners (AFTE) Response to PCAST Report on Forensic Sci. (October 31, 2016), https://afte.org/uploads/documents/AFTE-PCAST-Response.pdf; Org. of Sci. Area Committees (OSAC) Firearms and Toolmarks Subcommittee Response to the President's Council of Advisors on Sci. and Tech. (PCAST) Call for Additional References Regarding its Rep. "Forensic Sci. in Crim. Courts: Ensuring Sci. Validity of Feature-

increasingly important as a number of recent federal and state court opinions have cited the Report as support for limiting the admissibility of firearms/toolmarks evidence in criminal cases.[9] Accordingly, the Department offers its view on these claims.

I. PCAST's Claim that "Feature Comparison Methods" are Metrology

Several times throughout its Report, PCAST claimed that forensic "feature comparison methods belong to the scientific discipline of metrology."[10] (Metrology is the science of measurement and its application.) The accuracy of this assertion is critically important because if forensic pattern comparison methods are *not* metrology, then the fundamental premise PCAST used to justify its "guidance concerning the scientific standards for [the] scientific validity"[11] of forensic pattern comparison methods is erroneous. And if that premise is flawed, then key elements of the Report have limited relevance to the methods that PCAST addressed.

PCAST cited a single source in support of its linchpin claim that pattern comparison methods are metrology. That authority, the *International Vocabulary of Metrology*[12] ("VIM"), refutes the claim.

On this point, PCAST states:

> Within the broad span of forensic disciplines, we chose to narrow our focus to techniques that we refer to here as forensic "feature-comparison" methods . . . because . . . they all belong to the same broad scientific discipline, *metrology*, which is "the science of measurement and its application," in this case to measuring and comparing features.[13]

Later in its Report, PCAST claimed:

Comparison Methods (December 14, 2016), https://theiai.org/docs/20161214_FATM_Response_to_PCAST.pdf.; Org. of Sci. Area Committees (OSAC) Friction Ridge Subcommittee Response to Call for Additional References Regarding: President's Council of Advisors on Sci. and Tech. Rep. to the President (December 14, 2016), https://www.nist.gov/system/files/documents/2016/12/16/osac_friction_ridge_subcommittees_response_to_the_presidents_council_of_advisors_on_science_and_technologys_pcast_request_for_additional_references_-_submitted_december_14_2016.pdf; International Ass'n for Identification (IAI) Comments on the PCAST Report from the IAI FW/TT Sci. and Prac. Subcommittee (undated), https://theiai.org/docs/8.IAI_PCAST_Response.pdf; American Soc'y of Crime Laboratory Directors (ASCLD) Statement on September 20, 2016 PCAST Report on Forensic Sci. (September 30, 2016), https://pceinc.org/wp-content/uploads/2016/10/20160930-Statement-on-PCAST-Report-ASCLD.pdf.

[9] *U.S. v. Odell Tony Adams*, 2020 U.S. Dist. LEXIS 45125 (D. Oregon); *U.S. v. Shipp*, 2019 U.S. Dist. LEXIS 205397 (E.D.N.Y.); *U.S. v. Davis*, 2019 U.S. Dist. LEXIS 155037 (W.D. Va.); *U.S. v. Tibbs*, 2019 D.C. Super LEXIS 9 (D.C. 2019); *Williams v. U.S.*, 210 A.3d 734 (D.C. Ct. App. 2019); *U.S. v. Jovon Medley*, PWG 17-242 (D. Md., April 24, 2018); *People v. Azcona*, 2020 Cal. App. LEXIS 1173 (Cal. Ct. App.); *State v. Barquet*, DA No. 2392544-1D (Multnomah County, Oregon November 12, 2020); *People v. A.M.*, 2020 N.Y. Misc. LEXIS 2961 (Sup. Ct. Bronx, N.Y. 2020); *State v. Goodwin-Bey*, Case No. 1531-CR00555-01 (Greene County, Mo., Dec. 16, 2016).

[10] PCAST REPORT, *supra* note 1, at 23, 44 n.93, 143.

[11] *Id.* at x, 2, 4, 7, 21, 43.

[12] INT'L VOCABULARY OF METROLOGY – BASIC AND GENERAL CONCEPTS AND ASSOCIATED TERMS (VIM 3rd edition) JCGM 200 (2012), https://www.ceinorme.it/en/normazione-en/vim-en/vim-content-en.html.

[13] PCAST REPORT, *supra* note 1, at 23 (citing the VIM) (emphasis original).

> [F]eature-comparison methods belong squarely to the discipline of metrology—the science of measurement and its application.[14]

Again, the Report provided only a general citation to the VIM in support.[15]

The VIM makes no reference to forensic science or what PCAST described as "feature comparison methods." Further, the document provides no examples of the types of scientific disciplines, technologies, or applied knowledge that constitute metrology. Most fundamentally, however, the VIM's terms and definitions affirmatively *refute* PCAST's claim that "feature comparison methods" are metrology. The VIM defines "measurement" as follows:

> **Measurement**
> process of experimentally obtaining *one or more quantity values* that can reasonably be attributed to a quantity
>
> NOTE 1 Measurement *does not apply to nominal properties*.
> NOTE 2 Measurement implies *comparison of quantities or counting of entities*
> NOTE 3 Measurement presupposes a description of the quantity commensurate with the intended use of a measurement result, a measurement procedure, and a calibrated measuring system operating according to the specified measurement procedure, including the measurement conditions.[16]

The term "quantity" is defined in the VIM as follows:

> **Quantity**
> property of a phenomenon, body, or substance, where *the property has a magnitude* that can be *expressed as a number* and a reference.[17]

Finally, a "nominal property" is defined as:

> **Nominal Property**
> *property of a phenomenon, body, or substance, where the property has no magnitude*
> EXAMPLE 1 Sex of a human being
> EXAMPLE 2 Colour of a paint sample
> EXAMPLE 3 Colour of a spot test in chemistry
> EXAMPLE 4 ISO two-letter country code
> EXAMPLE 5 Sequence of amino acids in a polypeptide
> NOTE 1 *A nominal property has a value, which can be expressed in words, by alphanumerical codes, or by other means.*[18]

---

[14] *Id.* at 44.
[15] *Id.*
[16] VIM, *supra* note 12, at (2.1) (emphasis added).
[17] *Id.* at (1.1) (emphasis added).
[18] *Id.* at (1.30) (emphasis added).

According to the VIM, "measurement" is a process for obtaining a "quantity value." A "quantity" is the property of a phenomenon, body, or substance that has a magnitude expressed as a number. Measurement, however, does not apply to "nominal" properties—features of a phenomenon, body, or substance that have no magnitude. "Nominal" properties have a value expressed in words, codes, or by other means.

As their reflexive description makes clear, forensic pattern comparison methods *compare* the features/characteristics and overall patterns of a questioned sample to a known source; they do not *measure* them.[19] Any measurements made during the comparison process involve general class characteristics. However, the distinctive features or characteristics that examiners observe in a pattern form the primary basis for a source identification conclusion. These features or characteristics are not "measured."

During the examination process, forensic examiners initially focus on the general patterns observed in a trace sample. Next, they look for successively more detailed and distinctive features or characteristics. Once those properties are observed and documented, a visual comparison is made between one or more trace samples and/or one or more known sources. The method of comparison is observational, not based on measurement. Correspondence or discordance between class and sub-class features or characteristics of a trace sample and a known source are documented in "nominal" terms—not by numeric values. Finally, examination conclusions are provided in reports and testimony in words (nominal terms), not as measurements (magnitudes).

The conclusion categories described in the Department's Uniform Language for Testimony and Reports (ULTRs) illustrate this point.[20] Pattern comparison ULTR conclusions are reported and expressed in nominal terms such as "source identification," "source exclusion," "inclusion," "exclusion," and "inconclusive." Conclusions offered by examiners in the traditional forensic pattern disciplines are not expressed or reported as a measurement or a magnitude. To the contrary, the ULTRs specifically describe the nominal nature of the conclusions offered, along with restrictions on the use of certain terms that might otherwise imply reliance on measurement or statistics. For example, the following language is taken from the Department's Latent Print Discipline ULTR:

> A conclusion provided during testimony or in a report is ultimately an examiner's decision and is *not based on a statistically-derived or verified measurement or comparison* to all other friction ridge skin impression features. Therefore, an examiner shall not:
>
> - assert that a 'source identification' or a 'source exclusion' conclusion is based on the 'uniqueness' of an item of evidence.

[19] *See, e.g.,* BRADFORD T. ULERY, ET AL., ACCURACY AND RELIABILITY OF FORENSIC LATENT PRINT DECISIONS, 108 PROC. OF THE NAT'L ACAD. OF SCI. 7733, 7733 (May 10, 2011) ("Latent print examiners compare latents to exemplars, using their expertise rather than a quantitative standard to determine if the information content is sufficient to make a decision.").

[20] *See* U.S. DEP'T OF JUST., UNIFORM LANGUAGE FOR TESTIMONY AND REPORTS (ULTRs), www.justice.gov/forensics.

- use the terms 'individualize' or 'individualization' when describing a source conclusion.
- assert that two friction ridge skin impressions originated from the same source to the exclusion of all other sources.[21]

A separate limitation in all Department pattern ULTRs directs that "[a]n examiner shall not provide a conclusion that includes a statistic or numerical degree of probability except when based on relevant and appropriate data."[22]

Aside from PCAST's reference to the VIM, it offers a *single argument*—confined to a footnote—that pattern comparison methods are metrology:

> That forensic feature-comparison methods belong to the field of metrology is clear from the fact that NIST—whose mission is to assist the Nation by "advancing measurement science, standards and technology," and which is the world's leading metrological laboratory—is the home within the Federal government for research efforts on forensic science. NIST's programs include internal research, extramural research funding, conferences, and preparation of reference materials and standards . . . Forensic feature-comparison methods involve determining whether two sets of features agree within a given measurement tolerance.[23]

This statement is both a non-sequitur and factually inaccurate. PCAST's claim that NIST is the "world's leading metrological laboratory" and "is the home within the Federal government for research efforts on forensic science" has no logical nexus to its further claim that forensic pattern comparison methods—as currently practiced—are metrology. Obviously, a laboratory's status as a leader in the field of metrology and the fact that it conducts forensic research does not somehow transform the subject matter studied into metrology. In addition, PCAST's claim that "feature-comparison methods involve determining whether two sets of features agree within a given measurement tolerance" is simply not accurate.

As noted, the features or characteristics in a pattern are not "measured" and determined to be "within a given measurement tolerance." Rather, the combination of class characteristics and distinctive sub-class features within patterns are visually analyzed, compared, and evaluated for correspondence or discordance with a known source. An examiner *does* form an opinion whether "two sets of features agree";[24] however, that opinion is *not* based on whether those features agree "within a given *measurement* tolerance." Instead, examiners analyze, compare, evaluate, and

---

[21] *See Id*. (Emphasis added).
[22] *Id.*
[23] PCAST REPORT, *supra* note 1, at 44 n.93.
[24] To "agree," the features observed in the compared samples need not be identical. For example, in latent print examination, due to the pliability of skin, two prints from the same source will not appear to be identical. Surface type, transfer medium, and development method—among other factors—will affect the appearance of the friction ridge features. Because of these factors, examiners must determine whether the observed differences are within the range of variation that may be seen in different recorded impressions from the same source. This also applies to facial comparison—the same face will appear different when the subject's expression changes.

express their conclusions in nominal terms—not magnitudes. Therefore, contrary to PCAST's claim, forensic pattern comparison disciplines—as currently practiced—are *not* metrology.

From a legal perspective, however, that fact has no bearing on their admissibility. The Supreme Court made clear in *Daubert v. Merrell Dow Pharms., Inc.* and *Kumho Tire Co. v. Carmichael*[25] that judges "cannot administer evidentiary rules under which a gatekeeping obligation depend[s] upon a distinction between 'scientific' knowledge and 'technical' or 'other specialized' knowledge'[26] . . . ." The Court emphasized that trial judges, as part of their gatekeeping function, should not attempt to compartmentalize and shoehorn expert testimony into separate and mutually exclusive bins or boxes of knowledge that is then rigidly analyzed as "scientific," "technical," or "specialized."[27] As the Court noted, such efforts would range from difficult to impossible and would inevitably produce no clear lines of distinction capable of case-specific application.

To emphasize this point, the *Kumho Tire* Court cautioned, "We do not believe that Rule 702 creates a schematism that segregates expertise *by type* while mapping certain kinds of questions to certain kinds of experts. Life and the legal cases that it generates are too complex to warrant so definitive a match."[28] Rather than promoting impractical efforts at binning separate categories of knowledge, the Court stressed that the touchstone for the admissibility of expert knowledge under FRE 702—whatever its epistemic underpinning—is relevance and reliability.[29]

Reliable evidence must be grounded in *knowledge*, whether scientific, technical, or specialized in nature.[30] The term knowledge " 'applies to any body of known facts or to any body of ideas inferred from such facts or accepted as truths on good grounds.' "[31] The Court hastened to add that no body of knowledge—including scientific knowledge—can or must be "known" to a certainty.[32] In addition, the *Kumho Tire* Court stressed that the assessment of reliability may appropriately focus on the personal knowledge, skill, or experience of the expert witness.[33]

---

[25] 509 U.S. 579 (1993); 526 U.S. 137 (1999).
[26] *See Kumho Tire*, 526 U.S. 137 at 148.
[27] *See* Thomas S. Kuhn, *Reflections on my Critics*, *in* CRITICISM AND THE GROWTH OF KNOWLEDGE 231, 263 (Imre Lakatos & Alan Musgrave eds., Cambridge Univ. Press, 1965) ("Most of the puzzles of normal science are directly presented by nature, and all involve nature indirectly. Though different solutions have been received as valid at different times, *nature cannot be forced into an arbitrary set of conceptual boxes*.") (Emphasis added).
[28] *Kumho Tire,* at 151 (emphasis added).
[29] *Daubert*, 509 U.S. at 589; *see also U.S. v. Mitchell*, 365 F.3d 215, 244 (3rd Cir. 2004) ("That a particular discipline is or is not 'scientific' tells a court little about whether conclusions from that discipline are admissible under Rule 702 . . . Reliability remains the polestar."); *U.S. v. Herrera*, 704 F.3d 480, 486 (7th Cir. 2013) ("[E]xpert evidence is not limited to 'scientific' evidence, however such evidence might be defined. It includes any evidence created or validated by expert methods and presented by an expert witness that is shown to be reliable.").
[30] *Id*. at 590 (emphasis added). *See also Restivo v. Hessemann*, 846 F.3d 547, 576 (2d Cir. 2017) ("Rule 702 'makes no relevant distinction between 'scientific' knowledge and 'technical' or 'other specialized' knowledge, and 'makes clear that any such knowledge might become the subject of expert testimony.' *Kumho Tire Co.,* 526 U.S. at 147.").
[31] *Daubert, supra* note 25, at 590 (citing WEBSTER'S THIRD NEW INT'L DICTIONARY 1252 (Merriam-Webster Inc.1986).
[32] *Id.*
[33] *Kumho Tire*, 526 U.S. at 150 ("[T]he relevant reliability concerns may focus upon personal knowledge or experience.").

As the *Daubert* and *Kumho Tire* decisions made clear, an expert's opinion may—but need not—be derived from or verified by measurement or statistics. Experience, either alone or in conjunction with knowledge, skill, training, or education, provides an equally legitimate legal foundation for expert testimony. This fact is reflected in the Comment to FRE 702, which states:

> Nothing in this amendment is intended to suggest that experience alone—or experience in conjunction with other knowledge, skill, training, or education—may not provide a sufficient foundation for expert testimony. To the contrary, the text of Rule 702 expressly contemplates that an expert may be qualified on the basis of experience. In certain fields, experience is the predominant, if not sole, basis for a great deal of reliable expert testimony.[34]

Finally, a forensic expert's reasoning process is typically inductive,[35] (and thereby potentially fallible) and her opinion may be offered in categorical form.[36] In the domain of forensic science, a "source identification"[37] conclusion is the result of an inductive reasoning process[38] that makes

---

[34] FED. RULE OF EVIDENCE 702 advisory committee's note to 2000 amendment.

[35] *See* NEWTON C.A. DA COSTA & STEVEN FRENCH, SCI. AND PARTIAL TRUTH: A UNITARY APPROACH TO MODELS AND SCI. REASONING 130-159 (Oxford Univ. Press 2003) for a formal treatment of pragmatic inductive inference.

[36] *See* FED. RULE OF EVIDENCE 704 (the "Ultimate Issue Rule"); *see also U.S. v. Sherwood*, 98 F.3d 402, 408 (9th Cir. 1996) (fingerprint source identification); *U.S. v. Williams*, 2013 U.S. Dist. LEXIS 120884 (D. Hawaii); *U.S. v. McClusky*, 954 F. Supp. 2d 1224 (D. N. M. 2013); *U.S. v. Davis*, 602 F. Supp.2d 658 (D. Md. 2009) (forensic DNA source attribution); *Revis v. State*, 101 So.3d 247 (Ala. Ct. App. 2011) (firearms/toolmarks source identification).

[37] Eoghan Casey & David-Olivier Jaquet-Chiffelle, *Do Identities Matter?* 13 POLICING: A JOURNAL OF POL'Y & PRAC. 21, 21 (March 2019) ("Identification is the decision process of establishing with sufficient confidence (not absolute certainty), that some identity-related information describes a specific entity in a given context, at a certain time.").

[38] *See* COLIN AITKEN ET AL., COMMUNICATING AND INTERPRETING STAT. EVIDENCE IN THE ADMIN. OF CRIM. JUST., I. FUNDAMENTALS OF PROBABILITY AND STAT. EVIDENCE IN CRIM. PROC., GUIDANCE FOR JUDGES, LAWYERS, FORENSIC SCIENTISTS AND EXPERT WITNESSES, ROYAL STAT. SOC'Y 14 (November 2010), http://www.rss.org.uk/Images/PDF/influencing-change/rss-fundamentals-probability-statistical-evidence.Pdf.

> Most inferential reasoning in forensic contexts is inductive. It relies on evidential propositions in the form of empirical generalisations . . . and it gives rise to inferential conclusions that are ampliative, probabilistic and inherently defeasible. This is, roughly, what legal tests referring to "logic and common sense" presuppose to be the lay fact-finder's characteristic mode of reasoning. Defeasible, ampliative induction typifies the eternal human epistemic predicament, of reasoning under uncertainty to conclusions that are never entirely free from rational doubt.

PAUL ROBERTS & COLIN AITKEN, COMMUNICATING AND INTERPRETING STAT. EVIDENCE IN THE ADMIN. OF CRIM. JUST., 3. THE LOGIC OF FORENSIC PROOF — INFERENTIAL REASONING IN CRIM. EVIDENCE AND FORENSIC SCI., GUIDANCE FOR JUDGES, LAWYERS, FORENSIC SCIENTISTS AND EXPERT WITNESSES, ROYAL STAT. SOC'Y 43 (March 2014), https://www.maths.ed.ac.uk/~cgga/Guide-3-WEB.pdf.

> Events or parameters of interest, in a wide range of academic fields (such as history, theology, law, forensic science), are usually not the result of repetitive or replicable processes. These events are singular, unique, or one of a kind. It is not possible to repeat the events under identical conditions and tabulate the number of occasions on which some past event actually occurred. The use of subjective probabilities allows us to consider probability for events in situations such as these.

COLIN AITKEN & FRANCO TARONI, STAT. AND THE EVALUATION OF EVIDENCE FOR FORENSIC SCIENTISTS 22-23 (Wiley 2nd Ed. 2004); *See also* DA COSTA, *supra* note 35, at 8-20 for a formal treatment of pragmatic probability;

no claim of certainty.[39] During an examination, two items are examined for a sufficient combination of corresponding features. If correspondence is observed,[40] an examiner must determine whether that correspondence provides extremely strong support for the proposition that the items came from the same source and extremely weak or no support for the proposition that the items came from a different source.[41]

If an examiner determines that there *is* sufficient correspondence such that she (based on her knowledge, training, experience, and skill) would not expect to find the same combination of features repeated in another source and there is insufficient disagreement to conclude that the combination of features came from a different source, then the examiner inductively infers (from the observed data) that the items originated from the same source.[42]

Importantly, however, an examiner makes no claim that the observed combination of corresponding features (class and individual characteristics) are "unique"[43] in the natural world, or that the examiner can universally "individualize"[44] the item or person from which the pattern originated. In addition, given the limitations of inductive reasoning, an examiner cannot logically "exclude all other" potential sources of the item.[45] Accordingly, ULTR documents that authorize

---

"Probability can be 'objective' (a logical measure of chance, where everyone would be expected to agree to the value of the relevant probability) or 'subjective,' in the sense that it measures the strength of a person's belief in a particular proposition."

[39] *See* N. Malcolm, *Certainty and Empirical Statements*, 51 MIND, 18-46, 41 (1942) ("If any statement is capable of demonstrative proof, then it is not an empirical statement, but an a priori statement.")

[40] Christophe Champod & Ian Evett, *A Probabilistic Approach to Fingerprint Evidence*, J. OF FORENSIC IDENTIFICATION, 101-22, 103 (2001) ("The question for the scientist is not 'are this mark and print identical' but, 'given the detail that has been revealed and the comparison that has been made, what inference might be drawn in relation to the propositions that I have set out to consider.'").

[41] *See* WILLIAM THOMPSON ET AL., FORENSIC SCI. ASSESSMENTS: A QUALITY AND GAP ANALYSIS (2017), at 66 (2017), https://mcmprodaaas.s3.amazonaws.com/s3fs.public/reports/Latent%20Fingerprint%20Report%20FINAL%209_14.pdf?i9xGS_EyMHnIPLG6INIUyZb66L5cLdlb.

> Because ridge features have been demonstrated to be highly variable, an examiner may well be justified in asserting that a particular feature set is rare, even though there is no basis for determining exactly how rare. And an examiner may well be justified in saying that a comparison provides "strong evidence" that the prints have a common source, even though there is no basis for determining exactly how strong.

[42] *See* David Kaye, *Probability, Individualization, and Uniqueness in Forensic Sci. Evidence: Listening to the Academies*, 75 BROOK. L. REV. 1163, 1176 (2010) ("In appropriate cases . . . it is ethical and scientifically sound for an expert witness to offer an opinion as to the source of the trace evidence. Of course, it would be more precise to present the random-match probability instead of the qualitative statement, but scientists speak of many propositions that are merely highly likely as if they have been proved. They are practicing rather than evading science when they round off in this fashion.").

[43] Champod, *supra* note 40, at 103 ("Every entity is unique; no two entities can be 'Identical' to each other because an entity may only be identical to itself. Thus, to say 'this mark and this print are identical to each other' invokes a profound misconception: they might be indistinguishable but they cannot be identical.").

[44] Kaye, *supra* note 42, at 1166 ("[I]ndividualization—the conclusion that 'this trace came from this individual or this object'—is not the same as, and need not depend on, the belief in universal uniqueness. Consequently, there are circumstances in which an analyst reasonably can testify to having determined the source of an object, whether or not uniqueness is demonstrable." The Department uses the term "identification" rather than "individualization.").

[45] Champod, *supra* note 40, at 104-105.

a "source identification"[46] conclusion also prohibit claims that two patterns originated from the same source "to the exclusion of all other sources." They also preclude assertions of absolute/100% certainty, infallibility, or an error rate of zero.[47] Federal courts have found these limitations to be reasonable and appropriate constraints on expert testimony.[48]

The empirically-informed inductive process through which a qualified forensic pattern examiner forms and offers an opinion is the product of technical and specialized knowledge under Rule 702,[49] grounded in science, but ultimately based on an examiner's training, skill, and experience—*not* statistical methods or measurements. Moreover, the classification of a "source identification," "source exclusion," "inconclusive," or other conclusion is ultimately an examiner's *decision*. Thus, PCAST's claim that the traditional forensic pattern comparison disciplines—as currently practiced—are metrology is plainly incorrect.

II. PCAST's Claim that Forensic "Feature Comparison" Methods Can *Only* be Validated Using Multiple "Appropriately Designed" Independent Black Box Studies

In its Report, PCAST claimed that it compiled and reviewed more than 2,000 forensic research papers.[50] From that number—based on its newly-minted criteria—PCAST determined that only

---

> We cannot consider the entire population of suspects - the best we can do is to take a *sample*… We use our observations on the sample, whether formal or in formal, to draw inferences about the *population.* No matter how large our sample, it is not possible for us to say that we have eliminated every person in the population with certainty. . . . This is the classic scientific problem of *induction* that has been considered in the greatest depth by philosophers.

[46] *See also* Kaye, *supra* note 42, at 1185 ("Radical skepticism of all possible assertions of uniqueness is not justified. Absolute certainty (in the sense of zero probability of a future contradicting observation) is unattainable in any science. But this fact does not make otherwise well-founded opinions unscientific or inadmissible. Furthermore, whether or not global uniqueness is demonstrable, there are circumstances in which an analyst can testify to scientific knowledge of the likely source of an object or impression.").

[47] https://www.justice.gov/olp/uniform-language-testimony-and-reports.

[48] *U.S. v. Hunt*, 2020 U.S. Dist. LEXIS 95471 (W.D. Okla. 2020) ("The Court finds that the limitations . . . prescribed by the Department of Justice are reasonable, and that the government's experts should abide by those limitations."); *U.S. v. Harris*, 2020 U.S. Dist. LEXIS 205810 (D.C. 2020) ("This Court believes . . . that the testimony limitations as codified in the DOJ ULTR are reasonable and should govern the testimony at issue here. Accordingly, the Court instructs [the witness] to abide by the expert testimony limitations detailed in the DOJ ULTR.").

[49] *See e.g.*, *U.S. v. Harris*, 2020 U.S. Dist. LEXIS 205810 (D.C. 2020) (firearms/toolmarks); *U.S. v. Hunt*, 2020 U.S. Dist. LEXIS 95471 (W.D. Okla. 2020); latent prints); *U.S. v. Johnson*, 2019 U.S. Dist. LEXIS 39590 (S.D.N.Y. 2019) (firearms/toolmarks); *U.S. v. Simmons*, 2018 U.S. Dist. LEXIS 18606 (E.D. Va. 2018) (firearms/toolmarks); *U.S. v. Otero*, 849 F. Supp. 2d 425 (D. N.J. 2012) (firearms/toolmarks); *U.S. v. Mouzone*, 696 F. Supp. 2d 536 (D. Md. 2009) (firearms/toolmarks); *U.S. v. Glynn*, 578 F. Supp. 2d 567 (S.D.N.Y. 2008) (firearms/toolmarks); U.S. v. *Monteiro*, 407 F. Supp. 2d 351 (D. Mass. 2006) (firearms/toolmarks); *U.S. v. Herrera*, 704 F.3d 480 (7th Cir. 2013) (latent prints); *U.S. v. Baines*, 573 F.3d 979 (10th Cir. 2009) (latent prints); *U.S. v. Mosley*, 339 Fed. Appx. 568 (6th Cir. 2009) (latent prints); *U.S. v. Mitchell*, 365 F.3d 215 (3rd Cir. 2004) (latent prints); *U.S. v. Jones*, 2003 U.S. App. LEXIS 3396 (4th Cir. 2003) (latent prints); *U.S. v. Navarro-Fletes*, 49 Fed. Appx. 732 (9th Cir. 2002) (latent prints); *U.S. v. Mercado-Gracia*, 2018 U.S. Dist. LEXIS 192973 (D. N.M. 2018) (latent prints); *U.S. v. Bonds*, 2017 U.S. Dist. LEXIS 166975 (N.D. Ill. 2017) (latent prints); *U.S. v. Kreider*, 2006 U.S. Dist. LEXIS 63442 (W.D.N.Y. 2006) (latent prints); *U.S. v. Plaza*, 188 F. Supp. 2d 549 (E.D. Pa. 2002) (latent prints).

[50] PCAST REPORT, *supra* note 1, at 2.

three of those 2,000+ studies were "appropriately designed"—two for latent prints and one for firearms/toolmarks.[51] According to PCAST, "the foundational validity of a subjective method can *only* be established through multiple, appropriately designed black box studies."[52] To be "appropriately designed," a study must adhere to a strict set of six, non-severable criteria.[53] PCAST claimed that absent conformity with each of these requirements a "feature-comparison" method cannot be considered scientifically valid.[54]

PCAST's six criteria for an "appropriately designed" black box study are 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.[55]

[51] *Id.* at 91 (latent prints) (firearms/toolmarks) at 111.
[52] *Id.* at 9 (emphasis original).
[53] *Id.* at 52-53.
[54] *Id.* at 68.
[55] *Id.* at 52-53 (emphasis original).

To be clear, none of these criteria standing alone are novel or controversial. However, PCAST failed to cite a single authority that supports its sweeping claim that the collective and *non-severable* application of *all* of these experimental design requirements in multiple black box studies is the *sine qua non* for establishing the scientific validity of forensic "feature comparison" methods. Indeed, the sources that PCAST did cite only serve to undermine its position. In footnote 118 of its Report, PCAST claimed: "The analogous situation in medicine is a clinical trial to test the safety and efficacy of a drug for a particular application."[56] This is a reference to *post hoc* changes in the analysis of a study that may compromise its validity. PCAST offered a handful of Food and Drug Administration (FDA) validation guidance documents to support its analogy.[57] However, the first two cited sources refute PCAST's claim that method validation studies must adhere to a strict set of mandatory criteria. On that point, the documents offer the following disclaimer in bold and prominent display: "**Contains Non-Binding Recommendations**." Additionally, the first two cited sources include a call-out box that states, in relevant part:

> This guidance represents the Food and Drug Administration's (FDA's) current thinking on this topic. It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. *You can use an alternative approach if the approach satisfies the requirements of the applicable statutes and regulations*. If you want to discuss an *alternative approach*, contact the FDA staff responsible for implementing this guidance.[58]

Similarly, the first page of *Statistical Principles for Clinical Trials*, contains nearly identical language:

> This guidance represents the Food and Drug Administration's (FDA's) current thinking on this topic. It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. *An alternative approach may be used if such approach satisfies the requirements of the applicable statutes and regulations*.[59]

Moreover, the *Adaptive Designs* document states, "The use of the word *should* in Agency guidance *means that something is suggested or recommended, but not required.*"[60] In addition, the *Design Considerations for Pivotal Clinical Investigations for Medical Devices* document states:

---

[56] *Id.* at 52.

[57] U.S. DEP'T OF HEALTH AND HUM. SERV. FOOD AND DRUG ADMIN. CTR. FOR DEVICES AND RADIOLOGICAL HEALTH, and THE CTR. FOR BIOLOGIC EVALUATION AND RES., DESIGN CONSIDERATIONS FOR PIVOTAL CLINICAL INVESTIGATIONS FOR MED. DEVICES: GUIDANCE FOR INDUSTRY, CLINICAL INVESTIGATORS, INST. REV. BOARDS AND FOOD AND DRUG ADMIN. STAFF (November 7, 2013); U.S. DEP'T OF HEALTH AND HUM. SERV., CTR. FOR DEVICES AND RADIOLOGICAL HEALTH, and CTR. FOR BIOLOGICS EVALUATION AND RES.: ADAPTIVE DESIGNS FOR MED. DEVICE CLINICAL STUD. (July 27, 2016); and U.S. DEP'T OF HEALTH AND HUM. SERV., CTR. FOR DRUG EVALUATION AND RES., and CTR. FOR BIOLOGICS EVALUATION AND RES.: GUIDANCE FOR INDUSTRY E9 STAT. PRINCIPLES FOR CLINICAL TRIALS (September 1998).

[58] DESIGN CONSIDERATIONS, *supra* note 57, at 4; ADAPTIVE DESIGNS, *supra*, note 57, at 2 (emphasis added).

[59] GUIDANCE FOR INDUSTRY, *supra* note 57, at 1 (emphasis added).

[60] ADAPTIVE DESIGNS, *supra* note 57, at 2 (emphasis added).

> Although the Agency has articulated policies related to design of studies intended to support specific device types, and a general policy of tailoring the evidentiary burden to the regulatory requirement, *the Agency has not attempted to describe the different clinical study designs that may be appropriate to support a device pre-market submission, or to define how a sponsor should decide which pivotal clinical study design should be used to support a submission for a particular device*. This guidance document describes different study design principles relevant to the development of medical device clinical studies that can be used to fulfill pre-market clinical data requirements. *This guidance is not intended to provide a comprehensive tutorial on the best clinical and statistical practices for investigational medical device studies*.[61]

Finally, PCAST's purportedly mandatory criteria for pattern comparison method validation is inconsonant with the regulatory definition of "Valid Scientific Evidence" in the FDA's *Design Considerations* document:

> Valid scientific evidence is evidence from well-controlled investigations, partially controlled studies, studies and objective trials without matched controls, well-documented case histories conducted by qualified experts, and reports of significant human experience with a marketed device, from which it can fairly and responsibly be concluded by qualified experts that there is reasonable assurance of the safety and effectiveness of a device under its conditions of use. *The evidence required may vary according to the characteristics of the device, its conditions of use, the existence and adequacy of warnings and other restrictions, and the extent of experience with its use*. Isolated case reports, random experience, reports lacking sufficient details to permit scientific evaluation, and unsubstantiated opinions are not regarded as valid scientific evidence to show safety or effectiveness. Such information may be considered, however, in identifying a device the safety and effectiveness of which is questionable.[62]

The FDA's validation guidance clearly states that no single experimental design is either essential or required. To the contrary, the documents take pains to stress that it may be appropriate to utilize various study designs when validating medical devices or clinical drugs. The FDA also emphasized the non-binding nature of its guidance, which contains no prescriptive requirements or mandatory criteria. Finally, the applicable federal regulation instructs that "valid scientific evidence" may be generated by a variety of study designs and that the evidence required for validation may vary by the nature of the device, the conditions of use, and experience.

a. Forensic Laboratory Standards

Laboratory accreditation standards in the field of forensic science address the issue of method validation. The international standard applicable to all testing and calibration laboratories—

[61] DESIGN CONSIDERATIONS, *supra* note 57, at 4 (emphasis added).
[62] *Id.* at 9 (quoting 21 CFR 860.7(c)(1)) (emphasis added).

including crime labs—is ISO 17025.[63] This document guides the core activities and management operations of laboratories engaged in a diverse range of scientific inquiry. This includes clinical testing and diagnostics, research and development, and forensic science, among many other fields. Identical requirements apply to all testing and calibration laboratories, regardless of whether they analyze clinical samples, groundwater, or forensic evidence.

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."[64] A method has been validated per ISO/IEC 17025 when "the specified requirements are adequate for an intended use."[65] Section 7.2.2 of ISO 17025 is the applicable requirement for validating test methods. It provides that "validation shall be as extensive as is necessary to meet the needs of the given application or field of application."[66]

In contrast to PCAST's prescriptive stance, ISO does not dictate *how* labs must validate their methods, *which* criteria must be employed, or *what* experimental design must be followed. Instead, ISO simply requires that "[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." The selection of those requirements, the chosen experimental design, and the extent of the validation performed, is the responsibility of each laboratory. The pragmatic and flexible nature of method validation is also emphasized by other international scientific organizations.[67]

---

[63] 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].
[64] *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].
[65] ISO/IEC 17025:2017, *supra* note 63, § 3.9.
[66] *Id.* § 7.2.2.1.
[67] For example, 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 reliable results that are compatible with the results of other competent staff.

FORENSIC SCI. REGULATOR, CODES OF PRAC. AND CONDUCT FOR FORENSIC SCI. PROVIDERS AND PRAC. IN THE CRIM. JUST. SYS. § 20.9.1 (2016).

Like ISO, the Code sets forth a non-prescriptive, non-exclusive combination of measures that may be used to validate interpretive methods. These include blind confirmation by a second examiner, inter-laboratory comparisons and proficiency tests, and the in-house use of competency tests. The Code also states that an interpretive method "shall require only the relevant subset of . . . parameters and characteristics for measurement-based methods." § 20.9.1 & .2. 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." However, validity can be established "if the analyst or examiner repeatedly obtains correct results for positive and negative known tests." In addition, NATA correctly concedes that certain validation parameters "are not relevant in subjective tests." NAT'L ASS'N OF TESTING AUTHS., TECHNICAL NOTE 17: GUIDELINES FOR THE VALIDATION AND VERIFICATION OF QUANTITATIVE AND QUALITATIVE TEST METHODS § 5 (2013) at § 5-5.1.

b. Authorities in Experimental Design

The need for pragmatic flexibility in validating test methods is also stressed by authoritative sources in the field of experimental design. These experts advise that there are no rigid rules and that the most suitable approach depends on a variety of factors and circumstances. For example, Westgard, in his classic text, *Basic Method Validation*, states, "Method validation should be a standard laboratory process, but the process *need not be exactly the same for every laboratory or for every method validated* by a laboratory."[68] He also emphasizes the individual nature of validation: "Develop a validation plan on the basis of the characteristics of the test and method that will be critical for its successful application *in your laboratory*."[69] Finally, Westgard notes that "[e]ach laboratory situation may be different, therefore different adaptations are possible in different laboratories. The approach we advocate is to maintain the principles of the method validation process, while making the experimental work as efficient and practical as possible."[70]

Creswell, another leading expert on research design, emphasizes the contingent nature of various approaches and decisions:

> 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.[71]

A group of legal academics has also observed, "There is no one best way to study a phenomenon of interest. Each methodological choice involves trade-offs."[72] Trade-offs, in turn, require flexibility, which is necessitated by the pull of competing interests, existing resources, and countless operational considerations.[73]

Perhaps most notably, a leading treatise in the field of metrology states, "The situation regarding the frequency of validation is comparable for the situation for the appropriate amount of validation; there are *no firm and generally applicable rules*, *and only recommendations can be offered* that help the person responsible for validation with a competent assessment of the particular situation."[74]

---

[68] WESTGARD, BASIC METHOD VALIDATION198 (Westgard QC Inc., 3rd ed. 2008) (emphasis added).
[69] *Id*. at 203 (emphasis added).
[70] *Id.* at 205.
[71] JOHN W. CRESWELL, RESEARCH DESIGN: QUALITATIVE, QUANTITATIVE, AND MIXED METHOD APPROACHES 21 (4th ed. 2014).
[72] 1 DAVID L. FAIGMAN ET AL., MODERN SCI. EVIDENCE: THE LAW AND SCI. OF EXPERT TESTIMONY, STAT. & RES. METHODS § 1:22 (2010).
[73] GEOFFREY MARCZYK ET AL., ESSENTIALS OF RES. 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.").
[74] CZICHOS ET AL., SPRINGER HANDBOOK OF METROLOGY AND TESTING 86 (Springer 2011) (emphasis added).

On this point, the American Association for the Advancement of Science (AAAS) recently published a study on latent fingerprint examination.[75] The authors disagreed with PCAST's premise that only those research projects "intentionally and appropriately designed" should be considered when assessing evidential support for method validation.[76] Instead, the AAAS discussed the concept of "convergent validity," an approach that draws conclusions about method validity from the body of relevant literature as a whole. This approach acknowledges that various study designs have different strengths and weaknesses.[77] It also recognizes that some studies can reinforce others and collectively support conclusions not otherwise warranted.[78]

In sum, the sources cited by PCAST, the relevant international standard, and noted authorities in the fields of experimental design all refute its claim that only multiple black studies that strictly adhere to its six non-severable criteria may be used to validate forensic pattern comparison methods. Instead, they emphasize the absence of strict rules, the need for pragmatic flexibility, and an adaptive, context-based approach for testing a method's fitness for purpose.

III. PCAST's Claim that Error Rates for Forensic Pattern Comparison Methods Must be Established Using *Only* Black Box Studies

The Department fully agrees with PCAST's statement that "all laboratory test and feature comparison analyses have non-zero error rates."[79] That said, the more difficult questions are: *Can* such rates be accurately determined? *How* can that be accomplished? And to *whom, where,* and to *what* activities may such rates be validly applied?

PCAST claimed that error rates for subjective forensic pattern comparison methods must be *solely* determined through black box studies.[80] It also asserted that forensic examiners who took no part in those studies should testify that those study-derived rates apply to their work in the case at hand.[81] There are significant practical and scientific problems with these specious claims. Most fundamentally, no single error rate is generally applicable to all laboratories, all examiners, and all cases in which a particular method is used. Error rates derived from any given study are the output of numerous different inputs. Rates will vary depending on a multitude of factors immanent in a study's design, participants, rules, execution, and the model chosen for data generation and statistical summation.

---

[75] *See* THOMPSON ET AL., *supra* note 41.
[76] *Id.* at 44. ("[W]e consider all studies that examine the accuracy of latent print examiners, rather than focusing just on those that are 'intentionally and appropriately designed' for a particular purpose. Our goal is to draw conclusions from the literature as a whole, recognizing (consistent with the concept of convergent validity) that studies will have different strengths and limitations, and that the literature as a whole will have strengths and limitations.").
[77] *Id.* ("Our goal is to draw conclusions from the literature as a whole, recognizing (consistent with the concept of convergent validity) that studies will have different strengths and limitations, and that the literature as a whole will have strengths and limitations.").
[78] *Id.* at 94. ("[We] determined that the evaluation of individual publications, one at a time, was not an effective approach to reviewing this literature. This atomistic approach ignores the concept of convergent validity- i.e., the possibility that various publications, each with distinct limitations when considered by itself, can reinforce each other and collectively support conclusions that would not be warranted on the basis of a single article.").
[79] PCAST REPORT, *supra* note 1, at 3, 29.
[80] *Id*. at 46, 51, 111, 112, 116, 143, 147, 150.
[81] *Id.* at 56, 66, 112, 147, 150. *But see* ULERY ET AL., *supra* note 19, at 7734 ("Combining [experimental study] results among multiple agencies with heterogeneous procedures and types of casework would be problematic.").

In the experimental context, inputs are the assumptions and choices that researchers make and the actions they take to answer the questions of interest. These include: the study's internal design—its structure and scope; its experimental conditions; its participants—including their number, experience, and skill; how they are selected; their risk tolerance or aversion; whether they know they are being tested; the requirements of the laboratory quality systems in which they work; how closely test conditions mimic those requirements/systems; instructions researchers provide to participants; the number and type of comparisons conducted; the nature of the test samples used; how representative those samples are to evidence encountered during actual casework; how different answers are classified; and the statistical model(s) selected and employed to describe the results—to name a few.

Similar points were recently made by a well-known academic psychologist and commentator. Although noting the desirability of valid error rates, he also conceded that practical and scientific problems with generating such rates abound:

> Providing "an error rate" for a forensic domain may be misleading because it is a function of numerous parameters and depends on a variety of factors. An error rate varies by difficulty of the decision. . . . Error rates are going to be higher for difficult cases, but lower for easier cases . . . An error rate will also vary across individuals. Some experts have higher error rates, and others, lower error rates. This can be a function of training background . . . as well as cognitive aptitude, motivation, ideology, experience, etc. Therefore, error rates may give insights into forensic domains in general, but may say very little about a specific examiner's decision in a particular case. Hence, an average error rate for an average expert, in an average case, may not be informative (may even be misleading) for evaluating a specific expert examiner, doing a specific case.[82]

The American Association for the Advancement of Science's (AAAS) recent report, *Forensic Science Assessments: A Quality and Gap Analysis – Latent Fingerprint Examination*,[83] also cautioned against generalizing study-derived error rates to unrelated case scenarios. The report stated, "[I]t is unreasonable to think that the 'error rate' of latent fingerprint examination can meaningfully be reduced to a single number or even a single set of numbers."[84] The AAAS found that "[t]he probability of error in a particular case may vary considerably depending on the difficulty of the comparison. Factors such as the quality of the prints, the amount of detail present, and whether the known print was selected based on its similarity to the latent will all be important."[85]

The AAAS also noted that black box studies "can in principle determine the relative strength of different analysts and the relative difficulty of different comparisons, however the relationship of

---

[82] Itiel Dror, *The Error in "Error Rates": Why Error Rates Are So Needed Yet So Elusive*, 65 JOURNAL OF FORENSIC SCIENCES 5, 15-16 (2020).
[83] THOMPSON ET AL., *supra* note 41, at 46. With relevance to the points raised in Section I, the AAAS report stated, "Because the characteristics of fingerprints are unlikely to be statistically independent, it will be difficult to determine the frequency of any particular combinations of features. While research of this type is important, it is unlikely to yield quick answers." At 22.
[84] *Id.* at 45.
[85] *Id.* at 58.

such findings to the error rate in a specific case is problematic."[86] One concern was that study participants know they are being tested, which could affect their performance.[87] Another was that decision thresholds used by participants in controlled studies may differ from those used during actual casework. In sum, the report concluded that "the existing studies generally do not fully replicate the conditions that examiners face when performing casework."[88] Consequently, "the error rates observed in these studies *do not necessarily reflect the rate of error in actual practice*."[89]

PCAST also claimed that forensic examiners should testify that error rates from black box studies apply to their individual casework. This raises additional concerns about the relevance of rates generated by a discrete reference class of study participants to *all* forensic examiners who practice that method. This, in turn, raises larger questions about the overall external validity of black box studies. PCAST failed to squarely address these fundamental concerns about scientific relevance and general applicability.

As alluded to in the AAAS Report, the reference class of examiner-participants in a given black box study cannot be used as a valid proxy for the class of *all* such examiners.[90] Allen and Pardo have separately noted, "The reference class problem demonstrates that objective probabilities based on a particular class of which an item . . . [in our context, an examiner] is a member cannot typically (and maybe never) capture the probative value of that evidence for establishing facts relating to a specific event."[91] They continue, adding, "There is only one empirically objective reference class—the event itself. Among the various other reference classes, there is no other unique class that will capture the probative value of the evidence."[92] In short, error rates will vary based on the chosen reference class of examiners. As such, rates generated by examiners who participate in a given study cannot be generalized to and adopted by different examiners as *their* local error rate for unrelated casework scenarios.[93]

---

[86] *Id*.
[87] *Id.* S*ee also* ULERY ET AL., *supra* note 19, at 7734 ("Ideally, a study would be conducted in which participants were not aware that they were being tested.").
[88] THOMPSON ET AL., *supra* note 41, at 46.
[89] *Id.* (Citing Haber and Haber, 2014; Koehler, 2017; Thompson et al., 2014) (emphasis added); *see also* Ulery et al., *supra* note 19, at 7734 ("There is substantial variability in the attributes of latent prints, in the capabilities of latent print examiners, in the types of casework received by agencies, and the procedures used among agencies. *Average measures of performance across this heterogeneous population are of limited value*—but do provide insight necessary to understand the problem a scope future work.") (Emphasis added); BALDWIN ET AL., A STUDY OF FALSE POSITIVE AND FALSE NEGATIVE ERROR RATES IN CARTRIDGE CASE COMPARISON 18 (2014), https://www.ncjrs.gov/pdffiles1/nij/249874.pdf: ("This finding does not mean that 1% of the time each examiner will make a false-positive error. Nor does it mean that 1% of the time laboratories or agencies would report false positives, since this study did not include standard or existing quality assurance procedures, such as peer review or blind reanalysis.").
[90] *See generally*, Allen, Ronald; Pardo, Michael, *The Problematic Value of Mathematical Models of Evidence*, 36 J. OF LEGAL STUD. 107-140, 122 (January 2007) ("[G]enerally if not always there is a practically unbounded set of reference classes with probabilities within those reference classes ranging from zero to one, and nothing privileges any particular class.").
[91] *Id.* at 114.
[92] *Id.* at 123.
[93] *See also* ULERY ET AL., *supra* note 19, at 7738 ("The rates measured in this [latent print black box] study provide useful reference estimates that can inform decision making and guide future research: the results are *not representative of all situations, and do not account for operational context and safeguards*.") (Emphasis added).

A concern closely related to the reference class problem is the external or ecological validity of error rates generated through black box studies. External validity refers to whether an experiment accurately and adequately represents the subject matter, activities, and types of individuals studied. "If a study is externally valid, its findings can be generalized to other populations (of people, objects, organizations, times, places, etc.)."[94] Conversely, if a study lacks external validity, its findings cannot be generalized and applied to different people, places, and circumstances.

It is beyond dispute that black box studies do not reflect the numerous factors at play in actual casework. The reasons are many: They are performed outside of a laboratory's quality assurance system; there is no verification and review by a second examiner;[95] study directives may deviate from participants' work-related procedures and protocols;[96] sample quantity, quality, and analytical difficulty may differ from that typically encountered in actual casework; classification decisions may be dictated by study directives; and participants know they are being tested. In addition, black box studies may include a wide range of participants with differing levels of knowledge, skill, experience, training, and risk tolerance/aversion.[97] On this point, it is important to note that in pattern comparison black box studies performed to date, false positive errors have clustered among a small number of participants.[98] Moreover, in one latent print black box study

---

[94] FAIGMAN ET AL., *supra* note 72, at § 5:39.

[95] *See* BALDWIN ET AL., *supra* note 89, at 18:

> The study was specifically designed to allow us to measure not simply a single number from a large number of comparisons, but also to provide statistical insight into the distribution and variability in false-positive error rates. The result is that we can tell that the overall fraction is not necessarily representative of a rate for each examiner in the pool. Instead, examination of the data shows that the rate is a highly heterogeneous mixture of a few examiners with higher rates and most examiners with much lower error rates. *This finding does not mean that 1% of the time each examiner will make a false-positive error. Nor does it mean that 1% of the time laboratories or agencies would report false positives, since this study did not include standard or existing quality assurance procedures, such as peer review or blind reanalysis. What this result does suggest is that quality assurance is extremely important in firearms analysis and that an effective QA system must include the means to identify and correct issues with sufficient monitoring, proficiency testing, and checking in order to find false-positive errors that may be occurring at or below the rates observed in this study.*

(Emphasis added).

See also ULERY ET AL., *supra* note 19, at 7735 ("In no case did two examiners make the same false positive error [out of six total in the study]. Five errors occurred on image pairs where a large majority of examiners correctly excluded; one occurred on a pair where the majority of examiners made inconclusive decisions. *This suggests that these erroneous individualizations would have been detected if blind verification were routinely performed*.") (Emphasis added).

[96] *See* ULERY ET AL., *supra* note 19, at 7734 ("Combining results among multiple agencies with heterogeneous procedures and types of casework would be problematic.").

[97] *Id.* at 7737 ("Examiner skill varied substantially."); BALDWIN, ET AL., *supra* note 89, at 18 ("[E]xamination of the data shows that the rate is a highly heterogeneous mixture of a few examiners with higher rates and most examiners with much lower error rates.").

[98] BALDWIN ET AL., *supra* note 89, at 3, 18 ("[E]xamination of the data shows that the [false positive] rate is a highly heterogeneous mixture of a few examiners with higher rates and more examiners with much lower rates"); ULERY ET AL., *supra* note 19, at 7735, 7738 (the 6 false positive errors were committed by 5 examiners from a total of 169 study participants). In addition, ("Most of the false positive errors involved latents on the most complex combination

discussed by PCAST, when a second examiner performed the verification of the first examiner's results under non-biased conditions, all false positive results reported by the first examiners were detected.[99]

A different study examined the repeatability and reproducibility of decisions made by latent print examiners.[100] Participants examined approximately one-hundred image pairs of latent prints.[101] Six false positive errors were committed by five (out of one-hundred sixty-nine) examiners in the initial test.[102] Seventy-two examiners participated in the retest.[103] None of the six false positive errors were reproduced by a different examiner in the initial test and none of the four false positive errors was repeated by the same examiner in the retest.[104] The study concluded that "blind verification [by a second examiner] should be highly effective at detecting such errors."[105]

PCAST's claim that forensic pattern comparison error rates can *only* be derived from black box studies and that examiners must testify that those rates apply to the case at hand is scientifically erroneous. Black box error rates cannot travel from place to place and equally apply from case to case. In sum, these rates cannot be generalized to different laboratories, examiners, and casework situations.[106]

a. Alternative Experimental Designs

The PCAST Report also criticized forensic studies that employed what it described as a "closed-set" experimental design. In closed-set studies, a small number of samples generate many comparisons in which the source of the questioned items is always present.[107] PCAST noted that this creates internal dependencies among comparisons. It expressed concern that this type of experimental design may underestimate false-positive error rates. PCAST focused its critique on

---

of processing and substrate included in the study."); Ulery et al., *Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners.* PLoS ONE 7(3): e32800. Doi: 10.1371/journal.pone.0032800 (2012), available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0032800.

[99] PACHECO ET AL., MIAMI-DADE RES. STUDY FOR THE RELIABILITY OF THE ACE-V PROCESS: ACCURACY & PRECISION IN LATENT FINGERPRINT EXAMINATIONS 2, 7, 66 (2014), https://www.ncjrs.gov/pdffiles1/nij/grants/248534.pdf. ("Of the 42 erroneous identifications reported in both Phase I and Phase 2, seventeen of these errors occurred during Phase 2 ACE trials. The seventeen erroneous identifications were sent to fourteen of the 63 participants for verification in Phase 3, and fifteen responses for the seventeen erroneous identifications were returned. None of the fourteen participants agreed with the initial erroneous identification; twelve participants disagreed a total of thirteen times and two participants reported an inconclusive decision.").

[100] Ulery, et al., *Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners.* PLOS ONE 7(3): e32800. Doi: 10.1371/journal.pone.0032800 (2012), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0032800.

[101] *Id.* at 3.

[102] *Id.* at 3, 6.

[103] *Id.* at 3.

[104] *Id.* at 6, 9.

[105] *Id.* at 9.

[106] *See* ULERY ET AL., *supra* note 19, at 7734 ("There is substantial variability in the attributes of latent prints, in the capabilities of latent print examiners, in the types of casework received by agencies, and the procedures used among agencies. Average measures of performance across this heterogeneous population are of limited value—but do provide insight necessary to understand the problem and scope future work.").

[107] PCAST REPORT, *supra* note 1, at 86.

several studies conducted in the firearms/toolmarks discipline. While PCAST criticized the closed-set design of these studies, it failed to consider their purpose, substance, and utility.

The studies PCAST reviewed used consecutively manufactured firearms to produce the test samples provided to participants. Consecutively manufactured firearms are known to bear subclass characteristics. These are machined marks that carry over from one manufactured part of a firearm (i.e. breech face) to the next with little variation. It should be noted that subclass characteristics are unlikely to appear in real casework. Nevertheless, using test samples made from consecutively manufactured parts creates a challenging "worst-case-scenario" of best non-matching patterns. This can create comparison scenarios for examiners that are more difficult than those typically encountered during actual casework. In addition, a number of these studies used more "questioned" than "known" samples. As a result, participants were unable to determine a few correct answers and simply deduce the rest. Finally, because these studies used samples produced by consecutively manufactured parts, it was equally important to know whether participants could correctly associate questioned samples with known sources as it was to know whether those samples would be falsely identified. As a result, the studies included at least one known source with each questioned sample.

An additional benefit of a closed-set design is that it simulates real casework. In black box studies, the questioned samples are examined independently of each other—not as a set. During actual casework, however, examiners are not faced with completely independent comparison scenarios. Questioned samples and known items are typically collected and examined as a group, a circumstance that is mimicked by closed-set study designs. If one goal of method validation is to partially replicate casework conditions, then it is important to supplement black box studies with closed-set or partially open experimental designs.

There is no question that black box studies generate valuable information about examiner performance and decision thresholds under specified experimental conditions. Nevertheless, forensic method validation cannot be performed in a singular and one-dimensional manner. Studies of various design, scope, and substance all add value in the quest to better understand the circumstances under which error occurs and how it can be minimized. These efforts have been enhanced by a variety of experimental designs that have posed different questions to seek different types of answers.

To date, there have been approximately twenty firearms/toolmarks studies primarily focused on sample classification decisions and resulting error rates.[108] These studies used various experimental designs (black box, closed-set, partially-open, set-to-set), but have all resulted in a false positive error rate ranging from 0% to just over 1.0%.[109] The overall consistency of these findings when considered as a whole is a good indicator of what the *Daubert* Court described as a method's "potential rate of error."[110] Importantly, this aggregate rate is very low, giving the

[108] *See* Appendices "A" and "B" to this statement.
[109] It is important to note that the composite upper range of approximately one percent false positive error in these studies does not mean that one percent of the time each examiner will make a false positive error, or that one percent of the time labs would report false positives, since these studies did not use standard quality assurance procedures, such as peer review and blind reexamination. *See* BALDWIN ET AL., *supra* note 89, at 18.
[110] *Daubert*, 509 U.S. at 594.

overall indication that examiners are very accurate and make few source identification errors. Finally, it is worth noting that PCAST opined that an acceptable error rate should be less than 5%.[111] The aggregate false positive error rate in firearms/toolmarks studies to date falls well below that figure.

b. The Rate of Error vs. the Risk of Error

Despite the focus on the general frequencies at which various errors occur, the overall *rate* of error has little relevance to the critical question posed in most criminal litigation: What is the *risk* that error occurred in the case at hand? A 1996 report by the National Research Council, *The Evaluation of Forensic DNA Evidence* ("NRC II"),[112] recognized this important distinction. The NRC II observed, "The 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."[113]

The NRC II committee specifically rejected a recommendation that laboratories use proficiency tests as the exclusive means for error rate determination—a proposal offered in a prior NRC committee report on forensic DNA evidence (NRC I, 1992), co-chaired by PCAST Co-Chair, Dr. Eric Lander. On this point, the NRC II 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.[114]

The committee also noted, "The 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."[115] Moreover, the NRC II found it unnecessary to debate differing estimates of error when concerns about a false inclusion can be easily resolved by retesting the evidence.[116] The NRC II's view on error rates is shared by many leading scientists, statisticians, and forensic practitioners.[117]

---

[111] PCAST REPORT, *supra* note 1, at 151-52.
[112] NAT'L RES. COUNCIL, NAT'L ACADS., THE EVALUATION OF FORENSIC DNA EVIDENCE 85–88 (1996).
[113] *Id.* at 85.
[114] *Id.* at 86 (emphasis added).
[115] *Id.* at 87.
[116] *Id.*
[117] *See, e.g.*, 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 case-by-case basis) ("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."); BERNARD ROBERTSON ET AL., INTERPRETING EVIDENCE: EVALUATING FORENSIC SCI. IN THE COURTROOM 138 (2d ed. 2016) ("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."); 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).

IV. Conclusion

In their response to the PCAST Report, Dr. Ian Evett and colleagues wrote, "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. . . . [O]ur vision is not of the black-box/error rate but of continuous development through calibration and feedback of opinions."[118]

This sentiment reflects the current lack of scientific consensus on how—and indeed whether—error rates can or should be determined for forensic pattern comparison methods. Black box error rates, although adding to the body of knowledge, are a mere snapshot in time, place, and circumstance that capture a unique set of experimental conditions. Moreover, PCAST's notion of a single, generally applicable error rate wrongly assumes that such a figure can be generally applied to different evidence, examiners, and case circumstances.[119]

In conclusion, error rates derived from scientific studies of various size, scope, and experimental design *can and do* provide important information about the decision-making abilities and proclivities of examiner-participants. For most pattern comparison disciplines, extant studies show that examiners, on average, perform extremely well under a variety of experimental conditions. Competency and proficiency tests add to the body of knowledge by measuring how often examiners make correct decisions using known, ground truth samples. Verification by a second examiner, technical review, case controls, and other quality assurance measures used by accredited laboratories are critical components of risk management and mitigation. Lastly, as noted by the NRC, a wrongfully accused person's best insurance against false incrimination is the opportunity to have the evidence retested. In most cases, the typically non-consumptive nature of forensic pattern examination easily facilitates this final safeguard.

[118] Evett et al., *supra* note 8, at 22.

[119] MARCZYK ET AL., *supra* note 73, 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.").

APPENDIX A

Lead Author	Source	Year	Number of Participants	False Positive Rate (%)	Comparison Type Cases/Bullets
*Brundage	AFTE Journal	1998	30 (Plus 37 Informal Participants)	0	Bullets
Bunch	AFTE Journal	2003	8	0	Cartridge Cases
DeFrance	AFTE Journal	2003	9	0	Bullets
Smith	AFTE Journal	2004	8	0	Both
*Hamby	AFTE Journal	2009	507 (Includes *Brundage (1998) Participants)	0	Bullets
Lyons	AFTE Journal	2009	22	1.2[a]	Cartridge Cases
Mayland	AFTE Journal	2010	64	1.7[b]	Cartridge Cases
Cazes	AFTE Journal	2013	68 (or 69)	0	Cartridge Cases
Fadul	AFTE Journal	2013	Phase 1: 217 Phase 2: 114	Phase 1: .064[c] Phase 2: 0.18[c]	Cartridge Cases
Fadul	NIJ (NCJRS)	2013	183	0.40 [d]	Bullets
Stroman	AFTE Journal	2014	25	0	Cartridge Cases
Baldwin	NIJ (NCJRS)	2014	218	1.0	Cartridge Cases
Kerkhoff	Science & Justice	2015	11	0	Both
Smith	JFS	2016	31	0.14 Cases 0 Bullets	Cartridge Cases Bullets
Duez	JFS	2018	46 Examiners 10 trainees	0[e]	Cartridge Cases
Keisler	AFTE Journal	2018	126	0	Cartridge Cases
*Hamby	JFS	2019	619 (Includes *Brundage (1998) + Hamby (2009) Participants)	0.053%[f]	Bullets
Smith	JFS	2020	72	0.08	Bullets

*Brundage study was continued by Hamby who added additional participants and reported the combined data in Fall 2009 and 2019.

[a] The error rate reported by the author appears to be (1-True Positive Rate). There were three false positive identifications made but the number of true negative comparisons is not reported. 259 correct positive identifications were made. The False Discovery Rate (FDR) for the study is 3/(3+259)= 1.1%.

[b] The false positive error rate in not reported by the authors. There were three false positive identifications and 178 correct positive identifications made. The False Discovery Rate (FDR) for the study is 3/(3+178)= 1.7% and is reported in the table above.

[c] The error rates reported by the authors are roughly equivalent to the False Discovery Rates (FDR) for each of the study phases (FDR = .062% and 0.18% respectively).

[d] Eleven false positives occurred. The false positive error rate in not reported by the authors. The error rate quoted is equivalent to the False Discovery Rate =11/(11+2734)= 0.40%.

[e] Two false positives were made by one trainee. None were made the qualified examiners. The false positive rate does not include the trainee errors. If trainee data is included with that submitted by examiners, the False Positive Rate is (2/112) = 1.8%.

[f] The empirically observed false positive rate is 0%. Using Bayesian estimation methods, the authors' most conservative (worst case) estimate of the average examiner false positive error rate for the study is .053% with a 95% credible interval of ($1.1x10^{-5}$%, 0.16%).

APPENDIX B

Firearms/Toolmarks – Error Rate Studies
(Bullets & Cartridge Cases)

1. Brundage, D. (Summer 1998). The Identification of Consecutively Rifled Gun Barrels, *AFTE Journal*, 30(3), 438-44 (Bullets).

2. Bunch, S.G., & Murphy, D.P. (Spring 2003). A Comprehensive Validity Study for the Forensic Examination of Cartridge Cases, *AFTE Journal*, 35(2), 201-03 (Cartridge Cases).

3. DeFrance, C.S. & Van Arsdale, M.D. (Winter 2003). Validation Study of Electrochemical Rifling, *AFTE Journal*, 35(1), 35-37 (Bullets).

4. Smith, E.D. (Fall 2004). Cartridge Case and Bullet Comparison Validation Study with Firearms Submitted in Casework, *AFTE Journal*, 36(4), 130-35 (Bullets and Cartridge Cases).

5. Hamby, J.E., Brundage, D.J., & Thorpe, J.W. (Spring 2009). The Identification of Bullets Fired from 10 Consecutively Rifled 9mm Ruger Pistol Barrels: A Research Project Involving 507 Participants from 20 Countries, *AFTE Journal*, 41(2), 99-110 (Bullets).

6. Lyons, D.J. (Summer 2009). The Identification of Consecutively Manufactured Extractors, *AFTE Journal*, 41(3), 246-56 (Cartridge Cases).

7. Mayland, B. & Tucker, C. (Spring 2012). Validation of Obturation Marks in Consecutively Reamed Chambers, *AFTE Journal*, 44(2), 167-69 (Cartridge Cases).

8. Cazes, M. & Goudeau, J. (Spring 2013). Validation Study Results from Hi-Point Consecutively Manufactured Slides, *AFTE Journal*, 45(2), 175-77 (Cartridge Cases).

9. Fadul Jr., T.G., Hernandez, G.A., Wilson, E., Stoiloff, S., & Gulati, S. (Fall 2013). An Empirical Study to Improve the Scientific Foundation of Forensic Firearm and Tool Mark Identification Utilizing 10 Consecutively Manufactured Slides, *AFTE Journal*, 45(4), 376-93 (Cartridge Cases).

10. Fadul Jr., T.G., Hernandez, G.A., Wilson, E., Stoiloff, S., & Gulati, S. (December 2013). An Empirical Study to Improve the Foundation of Firearm and Tool Mark Identification Utilizing Consecutively Manufactured Glock EBIS Barrels with the Same EBIS Pattern. https://www.ncjrs.gov/pdffiles1/nij/grants/244232.pdf (Bullets)

11. Stroman, A. (Spring 2014), Empirically Determined Frequency of Error in Cartridge Case Examinations Using a Declared Double Blind Format, *AFTE Journal*, 46(2), 157-75 (Cartridge Cases).

12. Baldwin, D.P., Bajic, S.J., Morris, M., & Zamzow, D. (April 7, 2014). A Study of False-Positive and False-Negative Error Rates in Cartridge Case Comparisons. https://apps.dtic.mil/dtic/tr/fulltext/u2/a611807.pdf (Cartridge Cases).

13. Kerkhoff, W. et al. (2015). Design and Results of an Exploratory Double Blind Testing Program in Firearms Examination, *Science & Justice*, 55, 514-19 (Bullets and Cartridge Cases).

14. Smith, T.P., Smith, A.G., & Snipes, J.B. (July 2016). A Validation Study of Bullet and Cartridge Case Comparisons Using Samples Representative of Actual Casework, *Journal of Forensic Sciences*, 61(4), 939-45 (Cartridge Cases).

15. Duez, P. et al. (July 2018). Development and Validation of a Virtual Examination Tool for Firearm Forensics, *Journal of Forensic Sciences*, Vol. 63(4), 1069-1084 (Cartridge Cases).

16. Keisler, M. et al. (Winter 2018). Isolated Pairs Research Study, *AFTE Journal*, 50(1), 56-58 (Cartridge Cases).

17. Hamby, J. et al. (March 2019). A Worldwide Study of Bullets Fired From 10 Consecutively Rifled 9MM Ruger Pistol Barrels—Analysis of Examiner Error Rates, *Journal of Forensic Sciences*, 64(2), 551-57 (Bullets).

18. Smith, J. (October 2020). Beretta Barrel Fired Bullet Validation Study, *Journal of Forensic Sciences*, 2020;00:1-10 https://onlinelibrary.wiley.com/doi/full/10.1111/1556-4029.14604 (Bullets).