Texas Forensic Science Commission

Justice Through Science

FINAL REPORT ON COMPLAINT NO. 21.27; NATIONAL INNOCENCE PROJECT, UNIVERSITY OF COLORADO LAW SCHOOL (HOUSTON POLICE DEPT CRIME LAB; FIREARMS/TOOL MARKS)

April 26, 2024

I. Table of Contents

TABLE OF EXHIBITS

EXECUTIVE SUMMARY

In 1995, a Harris County jury convicted Nanon Williams of the 1992 capital murder of Adonius Collier. Testimony established that Williams and another man (Guevara) met with Collier and another man (Rasul) for an apparent drug transaction in a Houston park. During the meeting, Collier was killed by gunfire and Rasul was also shot.

At trial, a HPD firearms examiner testified that a deformed projectile recovered during Collier's autopsy was a .25 caliber bullet (consistent Williams' .25 caliber firearm). He also testified there was "no way in the world" the bullet could have been fired by Guevara's .22 caliber Derringer. Post-conviction testing established the deformed projectile recovered at autopsy was in fact a .22 caliber bullet fired from Guevara's .22 caliber Derringer.

The National Innocence Project and the Colorado Law School Criminal Defense Clinic filed a complaint with the Commission requesting an investigation into the error in Williams' case and raising questions about the validity and reliability of FATM evidence and testimony in general.

The details related to the forensic error in the Williams case were previously discussed in a 2007 report by Michael Bromwich concerning multiple issues at the Houston Police Department Crime Laboratory and Property Room (Bromwich Report). The Bromwich Report attributed the error in the William's case to multiple factors, including the fact that the bullet from autopsy was distorted, Guevara's Derringer was not originally submitted for comparison, the original examiner did not conduct a microscopic examination before characterizing the caliber of the projectile, and the laboratory lacked appropriate quality control protocols.

Landmark forensic science reports (Ballistic Imaging (2008), NAS (2009), PCAST (2016)) have called for empirical proof of the reliability and validity of FATM examination and have specifically stressed the need for appropriately designed black-box studies to provide estimates of reliability. Recently, several black box studies on the accuracy, reproducibility, and repeatability of FATM examination have been published related to bullet and cartridge case comparisons. NIST is in the process of reviewing data from 23 published studies and expects to release a Scientific Foundation Review based on this body of literature in the near future.

One area of prolonged disagreement among academics, statisticians, lawyers, and practitioners concerns the best way to calculate and discuss the significance of error rates in the FATM literature, particularly with respect to the treatment of inconclusive opinions. Many options have been proposed including treating inconclusive decisions as correct, treating them as incorrect, ignoring them altogether, etc.

The Commission agrees with authors of a recent paper (Swofford et al. 2024) that error rates calculated from black-box studies in FATM do not provide a suitable metric for representing FATM method performance because FATM examiners do not report results using a binary scale. Instead, the AFTE Range of Conclusions provide examiners with five possible options: Identification, Inconclusive (sub-classified as Type A, B, or C), or Elimination. Therefore, it is unsatisfactory

(and potentially misleading) to use false positive and false negative error rates from black box studies as a metric of performance in a feature comparison discipline that does not limit the examiner to only two choices.

The Commission proposes an approach for Texas that focuses on two concepts—method conformance and method performance—as explained by Swofford et al:

Method Performance relates to measures that reflect whether the outcome of the method can effectively distinguish between proposition of interest (*e.g.,* between same-source and different-source comparisons).

Method Conformance relates to the analyst's adherence to procedures that define the method. In FATM analysis, procedures may vary from one laboratory to the next, especially when making elimination conclusions based on accidental (individual) characteristics. Some laboratories do not allow an elimination decision based on accidental characteristics while others allow it, and still others require a firearm be available to rule out subclass characteristics. These variations in procedure impact performance.

In the proposed approach, Texas firearms examiners would present data for consideration by the trier of fact using *data obtained based on their own comparison method.* The data would be presented in two tables referred to as a "Discriminability Table" (detailing the extent to which the method can distinguish between same-source and different-source comparisons) and a "Reproducibility Table" (detailing the extent to which outcomes of the method are consistently produced between different examiners). These data provide greater transparency about the method's overall performance.

The Commission recommends Texas crime laboratories include the following in their reporting: a description of how each category of conclusions is defined under the laboratory's protocol; a statement that a same-source opinion does not imply uniqueness; a statement of whether the laboratory's protocol incorporates sub-categories of inconclusive, and, once developed, a discriminability table and reproducibility table.

FATM examiners should not report black box study error rates as a measure of accuracy in a specific case or assert that two toolmarks originate from the same source with absolute certainty. Photographic and descriptive note-based documentation of comparison conclusions should be created concurrently, in a linear sequential method (to the extent possible), sufficient to allow another properly trained analyst to understand and evaluate the work performed and to independently analyze and interpret the data and draw conclusions.

The Commission recommends the establishment of a FATM task group to assess the benefits and barriers to the new approach to FATM reporting discussed in the Report, as well as to provide feedback to the Commission on related issues of interest. Finally, the Commission adds two items to the Texas accreditation checklist: development of a blind verification policy and a policy regarding defining inter-examiner "consultation" for purposes of documentation.

I. COMMISSION BACKGROUND

A. History and Mission of the Texas Forensic Science Commission

The Texas Forensic Science Commission ("Commission") was created during the 79th Legislative Session in 2005 with the passage of HB-1068. The Act amended the Code of Criminal Procedure to add Article 38.01, which describes the composition and authority of the Commission.[1] During subsequent legislative sessions, the Texas Legislature further amended the Code of Criminal Procedure to clarify and expand the Commission's jurisdictional responsibilities and authority.[2]

The Commission has nine members appointed by the Governor of Texas.[3] Seven of the nine commissioners are scientists or medical doctors and two are attorneys (one prosecutor nominated by the Texas District and County Attorney's Association and one criminal defense attorney nominated by the Texas Criminal Defense Lawyer's Association).[4] The Commission's Presiding Officer is Jeffrey Barnard, MD. Dr. Barnard is the Chief Medical Examiner of Dallas County and Director of the Southwestern Institute of Forensic Sciences in Dallas.

B. Commission Jurisdiction

Texas law requires the Commission to "investigate in a timely manner, any allegation of professional negligence or professional misconduct that would substantially affect the integrity of:

(A) the results of a forensic analysis conducted by a crime laboratory;

(B) an examination or test that is conducted by a crime laboratory and that is a forensic examination or test not subject to accreditation; or

[1] TEX. CODE CRIM. PROC. art. 38.01.
[2] *See e.g.*, Acts 2013, 83rd Leg. ch. 782 (S.B. 1238) §§ 1-4 (2013); Acts 2015, 84th Leg. ch. 1276 (S.B. 1287) §§ 1-7 (2015); TEX. CODE CRIM. PROC. art 38.01 § 4-a(b).
[3] TEX. CODE CRIM. PROC. art. 38.01 § 3.
[4] *Id.*

(C) testimony related to an analysis, examination, or test described by paragraph (A) or (B)."[5]

The term "forensic analysis" is defined as a medical, chemical, toxicological, ballistic, or other examination or test performed on physical evidence, including DNA evidence, for the purpose of determining the connection of the evidence to a criminal action.[6]

1. Accreditation Jurisdiction

The Commission is charged with accrediting crime laboratories that conduct forensic analyses of physical evidence.[7] The term "crime laboratory" includes a public or private laboratory or other entity that conducts a forensic analysis subject to article 38.35 of the Code of Criminal Procedure.[8] As part of its accreditation jurisdiction, the Commission may "validate or approve specific forensic methods or methodologies," and "establish procedures, policies, standards, and practices to improve the quality of forensic analyses conducted in this state."[9] Firearm and toolmark analysis (FATM) is a discipline for which accreditation has been required since 2003.[10]

2. Jurisdiction Applicable to the Complaint

The forensic analysis involved in this complaint was performed by the Houston Police Department Crime Laboratory before it was accredited.[11] Accordingly, this report does not include any assessment of professional negligence or misconduct by any person or entity. When the Commission investigates an unaccredited entity or unaccredited discipline, the investigation is limited to: (1) observations of the Commission regarding the integrity and reliability of the

[5] TEX. CODE CRIM. PROC. art. 38.01 § 4(a)(3).
[6] TEX. CODE CRIM. PROC. art. 38.35(a)(4).
[7] TEX. CODE CRIM. PROC. art. 38.01 § 4-d(b).
[8] *Id.* at art. 38.35(a)(1).
[9] *Id.* at art. 38.01(b-1).
[10] TEX. CODE CRIM. PROC. art. 38.35(d)(1).
[11] In 2012, the City of Houston transitioned from having a forensic laboratory housed within HPD to an independent local government corporation. The Houston Forensic Science Center (HFSC) now provides forensic services for the City's law enforcement agencies and is governed by an appointed board.

analysis, examination, or test conducted; (2) best practices identified by the Commission during the course of the investigation, and (3) other recommendations that are relevant, as determined by the Commission.[12]

3. Limitations of this Report

The Commission's authority contains important limitations. For example, no finding by the Commission constitutes a comment upon the guilt or innocence of any individual.[13] The Commission's written reports are not admissible in civil or criminal actions.[14] The Commission does not have the authority to subpoena documents or testimony; information received during any investigation is dependent on the willingness of affected parties to submit relevant documents and respond to questions posed. Information gathered in this report was not subject to standards for the admission of evidence in a courtroom. For example, no individual testified under oath, was limited by either the Texas or Federal Rules of Evidence (*e.g.*, against the admission of hearsay) or was subject to cross-examination under a judge's supervision.

II. BACKGROUND AND SUMMARY OF THE COMPLAINT

A. Complaint and Investigative Decision by the Commission

This report contains observations and recommendations regarding a joint complaint filed by the University of Colorado Law School Clinical Programs and the Innocence Project, Inc. on behalf of Nanon Williams. The Commission accepted the complaint for investigation at its October 22, 2021, quarterly meeting.[15]

---

[12] TEX. CODE CRIM. PROC. art 38.01 § 4 (b-1).
[13] TEX. CODE CRIM. PROC. art. 38.01 § 4(g).
[14] *Id*. at § 11.
[15] While the Commission typically aims to release its reports within 2-3 quarterly meetings of receipt depending on the complexity of the issues, we delayed the drafting and release of this report in anticipation of NIST's publication of a foundational review of scientific literature in firearms analysis. As of this writing, the report has not yet been issued for public comment. The Commission directs the reader's attention to NIST's website where scientific foundation reports are published: https://www.nist.gov/forensic-science/interdisciplinary-topics/scientific-foundation-reviews\.

B. Summary of the Complaint

The complaint concerns the 1992 capital murder of Adonius Collier in Houston, Texas. Nanon Williams was arrested for the crime, along with a co-defendant, Vaal Guevara. The complaint alleges the Houston Police Department Crime Laboratory failed to adequately examine the firearm evidence and mischaracterized the caliber of the recovered projectile at the defendant's trial. The complaint alleges the examiner testified erroneously by characterizing a projectile recovered at autopsy as a .25 caliber bullet (consistent with a firearm carried by the defendant) when in fact the projectile was a .22 caliber bullet (consistent with a firearm carried by the testifying co-defendant turned accomplice witness). The examiner erroneously testified there was "no way in the world" the projectile in question could have been fired from Guevara's .22 caliber weapon.

The complaint also details longstanding questions concerning the validity and reliability of FATM evidence as expressed in landmark reports issued by three separate committees of scientists.[16] Complainant also points to a series of recent court decisions limiting the scope of testimony that may be offered by FATM examiners.

The complaint asserts: 1) the methodology employed by FATM examiners consists of a subjective process applied to an unsubstantiated assumption (that each firearm leaves a set of unique markings on ammunition fired from it); 2) the NAS and PCAST Reports each raise significant concerns about FATM; 3) FATM has a strong potential for high error rates; 4) FATM

[16] National Research Council Committee on Identifying Needs in the Forensic Science Community, *Strengthening Forensic Science in the United States: A Path Forward*, The National Academies Press, Washington, D.C., USA, 2009, https://doi.org/10.17226/12589; National Research Council Committee to Assess the Feasibility, Accuracy, and Technical Capability of a National Ballistics Database, *Ballistic Imaging*, The National Academies Press, Washington, D.C., USA, 2008, https://nap.nationalacademies.org/catalog/12162/ballistic-imaging; President's Council of Advisors on Science and Technology, Report to the President, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Method*s, U.S. Executive Office of the President, Washington, D.C., USA, 2016.

has not been validated through appropriate testing; and 5) FATM encompasses many different assays, each requiring separate validation (*e.g.,* deformed projectiles, ejector marks, handheld toolmarks).

The complaint asks the Commission to review the evidence and related testimony presented in the defendant's case. The complaint further requests the Commission suggest appropriate limits on the conclusions of FATM examiners in traditional bullet-to-firearm matching testimony and determine what conclusions – if any – can be proffered by other toolmark assays. (*See*, **Exhibit A Complaint)**.

C. Facts of Criminal Case

1. Trial Testimony

The following facts are contained in the findings of fact and conclusions of law issued by the trial court pursuant to the post-conviction writ of habeas corpus filed by Nanon Williams. The Commission summarizes aspects of the record that are relevant to the firearms analysis only; for a full understanding of the case, the reader should refer to the complete record.

On July 26, 1995, Williams was convicted of the 1992 capital murder of Adonius Collier (murder committed during the course of a robbery by shooting him with a firearm). Based on the jury's answers to the special issues, Williams was sentenced to death.[17]

The evidence at trial established that Vaal Guevara lived in a Houston apartment located next door to a woman named Winn and that Winn sometimes dated Guevara. On May 13, 1992, an individual named "Xavier" (later identified as Patrick Smith) drove Williams, Guevara, and Winn to Adonius Collier's apartment for an apparent drug transaction. According to Guevara and Winn,

[17] Following the United States Supreme Court's decision in *Roper v. Simmons*, 543 U.S. 551, 125 S.Ct. 1183, 161 L.Ed.2d 1 (2005), the district court commuted Williams's sentence to life in prison because he was 17 at the time of the offense.

who testified for the prosecution, Guevara carried a .22 Magnum Davis Derringer, and Williams had a .25 semiautomatic pistol tucked in his shorts and a pistol grip shotgun in his Raiders jacket.[18] Collier was present in the apartment with a woman named Stephanie Anderson and a man named Emmade Rasul. The two groups proceeded to a park in Houston in two separate vehicles to conduct the drug transaction.[19]

Once at the park, Collier and Rasul exited one vehicle and Williams and Guevara exited the other. The four men talked for a moment and then went into the woods. The remaining witnesses stayed in their cars until they heard a series of gunshots (between two and four gunshots). Winn testified she saw Rasul running from the woods followed by Guevara who tried unsuccessfully to catch Rasul. Guevara got into the vehicle with Winn, carrying his .22 caliber Magnum Davis Derringer, and claimed his gun had jammed. Two or three minutes later, Williams got into the car carrying the Raiders jacket and a shotgun.

Co-defendant Guevara testified at the defendant's trial in exchange for a plea of guilty to illegal investment in drugs and an agreed punishment recommendation of ten years.[20] He testified that at the time of the offense he was with Rasul while the defendant was with Collier approximately 8-10 feet away, but he could not hear what they were saying. Guevara testified he had the money and Rasul showed him the drugs. According to Guevara, there was gunfire from his right side where the defendant and Collier were standing. Guevara testified he was holding his Derringer when the shooting started but he did not fire the first shot. Guevara testified he saw Collier staggering towards him and that he shot his Derringer in the direction of Collier but does not know whether he hit him. Guevara further testified that his gun misfired when he attempted

[18] *Ex parte Nanon Williams*, Findings of Fact, Conclusions of Law and Order (May 3, 2001) at Finding 14.
[19] *Id.* at Finding 16.
[20] Guevara was convicted of illegal investment and sentenced to ten years imprisonment five days after the defendant was convicted and sentenced to death. *Id.* at Finding 128.

to shoot a second time and that he ran in the same direction as Rasul. Guevara testified that while he was running, he heard a loud noise and turned and saw Collier lying on the ground in front of Williams who had a shotgun in his hands. Guevara testified that upon returning to the apartment, the defendant took a semiautomatic weapon out of his pocket.[21]

Rasul testified that during the drug transaction, Williams shot Rasul in the face. Rasul claimed he was two feet away from Collier and Williams was the first person to shoot. As he ran away, Rasul was also shot in the heel of his foot and a bullet was later removed from his foot at the hospital. Rasul testified he never saw Guevara with a weapon, and he did not know what happened to Collier.[22]

The .22 Magnum Davis Derringer was later recovered by police from Guevara's apartment.[23] However, at the time of the trial, it had not been submitted to the laboratory for testing or comparison to the fired evidence.[24]

The assistant medical examiner testified Collier died as a result of a shotgun wound to the left temple and cheekbone fired from not more than five or six feet away and that Collier was alive when he was shot. The assistant medical examiner further testified he could not account for the presence of a bullet (EB1) which was recovered during the autopsy.[25]

An HPD firearms examiner testified that the bullet recovered from the medical examiner during autopsy (EB1) was a .25 automatic. He testified there was "no way in the world" the bullet was shot from Guevara's .22 Magnum Davis Derringer. (During post-conviction testing, this

[21] *Id.* at Finding 21.
[22] *Id.* at Finding 22.
[23] *Id.* at Finding 23.
[24] *Id.* at Finding 28.
[25] *Id.* at Finding 24. The medical examiner testified he had no explanation for the presence of the bullet fragment (EB1) found in the container containing shotgun pellets; that he x-rayed the deceased's head before autopsy and did not see a bullet; and that he saw no signs the deceased was shot with a projectile other than the shotgun blast. He also testified Collier's brain was severely shredded by the shotgun blast resulting in an inability to trace out wound tracks. (Finding 29). The trial court found that EB1 came from Collier's head (Finding 33).

opinion was shown to be incorrect). He also testified the bullet recovered at autopsy (EB1) was extremely mutilated and a large part of the bullet's mass was missing. He testified the bullet recovered from Rasul at the hospital (EB2) was a .25 automatic full metal jacketed bullet which also could not have been fired from a Derringer. Finally, he testified he had no knowledge regarding the functional condition of the .22 Derringer because he never checked it.[26]

At the time of the trial, "Xavier" (*aka,* Patrick Smith), the shotgun brought to the scene by Williams, and the .25 caliber semiautomatic Williams carried on the night of the offense had not yet been located.[27]

On July 26, 1995, Williams was convicted of capital murder. Williams' direct appeal was denied in an unpublished opinion in May of 1997. *Williams v. State,* No. 72, 179 (Tex. Cr. App. 1997), *cert. denied, Williams v. Texas,* 522 U.S. 1030 (1997).

2. Post-Conviction Writ Proceedings

On January 7, 1998, the trial court granted habeas counsel's motion for the independent testing of certain ballistic evidence (bullets labelled EB1 (extracted from autopsy) and EB2 (extracted from Rasul's foot) and a .22 Magnum Davis Derringer). Before the items were released for independent testing, the HPD firearms examiner test-fired the .22 Derringer and summarized his findings in a January 15, 1998, letter to the Harris County District Attorney's Office, viz:

> After completing a microscopic comparison of the test firings to the fired bullets, it is my opinion that the fired jacketed bullet (EB1) …was fired in the bottom barrel of the .22 Magnum Davis Derringer [serial number]. The fired jacketed lead bullet (EB2) was not fired in the .22 Magnum Davis Derringer and is consistent with a .25 auto.[28]

[26] *Id.* at Finding 25.
[27] *Id.* at Finding 26.
[28] *Id.* at Finding 7.

Another firearms examiner, Ronald Singer, former Chief Criminalist of the Tarrant County Medical Examiner's Office, testified during a post-conviction writ hearing that EB1 was a .22 Magnum caliber bullet, and it was fired from the bottom barrel of the Derringer.[29]

The trial court found, based on the post-conviction testing by HPD and subsequent review by Singer, that EB1 was incorrectly characterized as a .25 caliber bullet fragment and instead should have been characterized as a .22 caliber bullet. The trial court also found EB1 was fired from Guevara's .22 Magnum Davis Derringer and that Guevara shot Collier in the head with his .22 Magnum Davis Derringer.[30]

Because Collier was shot with both a shotgun and a pistol, a significant focus on appeal was disagreement among forensic pathologists regarding whether his death was caused by injuries from the impact of the EB1 bullet or the shotgun blast (including whether an answer to this question was possible to discern), given that different weapons were in the possession of each co-defendant.

After two evidentiary hearings in 1998 and 2000, the trial court recommended relief in May 2001 on Mr. Williams's claim that his trial counsel was ineffective for failing to retain a firearms expert. The trial court found that proper ballistics testing would have caused counsel to seek out independent pathologist testimony on the possibility that the EB1 bullet, not the shot gun, caused the decedent's death, which in turn, would have changed the type and strength of the cross-examination of Guevara, the jury's assessment of Guevara, and the prosecution's closing argument. The Court of Criminal Appeals denied relief in an unpublished decision the following year. *Ex parte Williams*, No. 46, 736-02 (Tex. Cr. App 2002).

[29] *Id.* at Finding 34.
[30] *Id.* at Findings 35 and 36.

Williams then filed a federal writ of habeas corpus after his state proceedings were exhausted. After initial litigation about the limitations of federal habeas review, the federal district court held an evidentiary hearing and granted relief, concluding in relevant part as follows:

> Had the State performed a competent forensic inquiry, the parties would have been able to debate before the jury Guevara's role as a potential killer, rather than a mere witness. Whether by inadvertence or by incompetence, the State's untrustworthy investigation hampered trial counsel's performance. [The HPD firearms examiner] misidentified the type of pistol shot and the source of EB-1, and thus the shooter. . . . The administration of criminal justice cannot countenance such fundamental errors.

Williams v. Thaler, 756 F. Supp. 2d 809, 828 (S.D. Tex. Nov. 24, 2010), *rev'd*, 684 F.3d 597 (5th Cir. 2012). The Fifth Circuit reversed, holding that it "[could] not conclude that there was no reasonable basis for the state court's denial of Williams's habeas petition." *Williams v. Thaler*, 684 F.3d 597, 599 (5th Cir. 2012).

3. The Bromwich Investigation of Houston Police Department Crime Lab

In 2005, the City of Houston commissioned an investigation of the Houston Police Department (HPD) Crime Laboratory and Property Room in response to serious questions about the quality of the forensic work performed in the crime laboratory. Michael Bromwich was appointed as the independent investigator. Bromwich and his investigative team reported their conclusions in a final report published in June of 2007 ("Bromwich Report"). The Bromwich Report included the historical review of thousands of cases in many disciplines (DNA, Controlled Substances, Trace Evidence, Toxicology, Questioned Documents, and Firearms). It also provided a detailed review of the role forensic science played in the cases of four specific defendants. One of those detailed case reviews included the fired bullet evidence in Williams' case.

The Bromwich Report detailed the crime laboratory's pretrial analysis of the firearms evidence, the trial testimony of the HPD examiner regarding the crime laboratory's evaluation of

the firearm evidence, the post-conviction discovery and reporting of the mischaracterization of EB1, and the post-conviction proceedings in state and federal court. Portions of the Report are excerpted and attached as **Exhibit B**. In sum, it concluded as follows:

- EB1 was [mischaracterized] by HPD as a .25 caliber bullet because distortions caused the bullet fragment to exhibit apparent similarities in class characteristics to those of a .25 caliber bullet.

- The error was exacerbated by the subsequent submission of EB2 (the .25 caliber bullet extracted from Rasul's foot) which had similar class characteristics.

- The failure of investigators to submit Guevara's .22 Derringer with a request to compare the bullet evidence to the firearm contributed to the [mischaracterization].

- One of the HPD examiners may have reached the wrong conclusion partly because he had only EB1 available at the time of his examination. Additionally, he performed a visual examination, not a microscopical examination.[31]

- The [mischaracterization] was perpetuated by the laboratory's lack of appropriate quality assurance protocols at the time and during the years that preceded accreditation.

- HPD firearms examiners were allowed to co-sign reports of other examiners without personally reviewing the evidence that was the subject of the report. That practice was contrary to generally accepted forensic science principles and obviated the purpose of secondary signatures on each report.

4. Commission Observations re: Errors in the Williams Case

The fired bullet analysis in this case included a significant error (the mischaracterization of EB1 as a .25 caliber bullet) and related testimony. The error, though ultimately acknowledged by HPD, went undetected by the laboratory for an extended period. As with all Commission investigations, the potential impact of this error on the legal remedies available to Williams (if any) is for state and federal courts with jurisdiction to decide.

[31] At the time, it was not uncommon for firearm examiners to initially determine class characteristics with a cursory visual examination. However, deformation can change the apparent class characteristics of a bullet or fragment. As a result, firearms examiners in current practice would conduct careful measurements and/or careful microscopic examination and *not* rely on a cursory visual examination.

In the period since Williams' case was tried and the Bromwich Report was issued, Texas crime laboratories and the firearms discipline in general have experienced significant evolution. Included among the developments of the last 20 years are the following:

1. The Texas Legislature required accreditation of firearm and toolmark analysis for admission of the evidence and related testimony in criminal actions.[32]

2. The Houston Forensic Science Center (HFSC), a local non-government corporation, was established to provide forensic services to the City of Houston. The Houston area now has four accredited laboratories with analysts licensed by the Commission to perform firearms analysis, including: HFSC; Harris County Institute of Forensic Sciences (HCIFS); Harris County Sheriff's Office (HCSO); and Texas Department of Public Safety (DPS) Regional Crime Laboratory in Houston.

3. The Organization of Scientific Area Committees (OSAC) of the National Institute for Standards and Technology (NIST) was established to develop and promote implementation of standard methods across all laboratories engaged in forensic analysis, including FATM.

4. NIST undertook a multi-year study to document the scientific foundations of firearms analysis. By evaluating the literature on error rates, the study will make observations regarding the core assertion in firearms analysis—that examiners can reliably determine whether a specific gun was used in a crime by examining bullets and cartridge cases under a comparison microscope. (As of this writing, the draft report is still in NIST internal review and pending release.)

5. HFSC published the results of its blind quality control program in firearms comparison, which showed no false positives/ false negatives but highlighted the need for increased transparency regarding the significance of almost half (40%) of the results, which were deemed inconclusive by examiners.

6. At the 2024 American Academy of Forensic Sciences annual meeting, a group of authors with expertise in statistics and a cross-section of forensic disciplines offered a potential solution for understanding and reporting inconclusive decisions and error rates in forensic comparison disciplines including but not limited to FATM.[33]

[32] TEX. CODE CRIM. PROC. art. 38.35(d)(1).

[33] The AAFS presentation was made by Swofford, H. as lead author. *See*, **Exhibit E:** Swofford, H., Lund, S.; Iyer, H.; Butler, J.; Soons, J.; Thompson, R.; Desiderio, V.; Jones, J.P.; Ramotowski, R. *Inconclusive Decisions and Error Rates in Forensic Science*, Forensic Science International: Synergy 8 100472 (2024)*:* doi.org/10.1016/j.fsisyn.2024.100472. A CSAFE presentation of the paper by Swofford may be accessed free of charge at https://learn.forensicstats.org/product?catalog=WB240227

III. PRIOR COMMISSION REPORTS AND LABORATORY REQUESTS FOR GUIDANCE ON REPORTING OF INCONCLUSIVES IN FATM COMPARISON

A. Prior Commission Reports on FATM

This complaint is not the first time the Commission has been asked to investigate opinions rendered in FATM cases. In one 2015 investigation, a laboratory disclosed that an examiner mistakenly excluded a group of cartridge cases as having been fired from the submitted reference firearm.[34] In another 2016 investigation, a defense attorney filed a complaint alleging an examiner erroneously identified a submitted weapon as having fired ammunition components recovered at an autopsy.[35] Though the prosecutor in the second case did not join in the defense attorney's complaint, he expressed similar concerns as the defense attorney from the perspective of a legal end-user of the FATM analysis. The erroneous identification significantly complicated the investigation in an already complex capital murder case involving juvenile defendants. Fortunately, the errors identified in both cases were not the result of professional misconduct and did not have an impact on the ultimate disposition of the criminal matters because the cases turned on other evidence. Both investigations provided opportunities for reflection within the laboratories that were the subject of the investigations and among the larger community of FATM examiners.

B. Requests From Labs Regarding Non-consensus PT and Inconclusive Results

In 2022, the Commission investigated a series of non-consensus proficiency testing results at the Fort Worth Police Department Crime Laboratory (FWPDCL). At the time, FWPDCL had a policy of treating *all* inconclusive firearm comparison opinions reached by proficiency test-takers as "correct," even where consensus opinion and expected results of the test provider were

[34] *See*, *Laboratory Self-Disclosure 14.01* (2015): https://www.txcourts.gov/media/1441009/14-01-final-report-ifl-firearms-section-lab-self-disclosure-20151105.pdf
[35] *See*, *Complaint 14.08* (2016): https://www.txcourts.gov/media/1453994/blazek-swifs-final-investigative-report-041916.pdf

"identification" or "elimination." The Commission published a report describing the problems inherent in a policy that accepts *all* inconclusive results as "correct" regardless of the test provider's expected result, including the need for review and root cause analysis of all non-consensus results in accordance with accreditation requirements.[36] The FWPDCL approach (subsequently changed) provided an incentive for examiners to report an "inconclusive" response in proficiency testing because such results would never be viewed as "incorrect," thus undermining the entire purpose of proficiency monitoring.

After the FWPDCL report, the Commission received other non-consensus PT results reported in the firearms discipline from additional Texas laboratories. The laboratories acknowledged the difficulty in communicating the significance of inconclusive results including but not limited to performance monitoring settings and asked for the Commission's guidance.

As with many other issues that have arisen in forensic science since the Commission was created by the Legislature, Texas crime laboratories have demonstrated a willingness to address and collaborate on the most challenging questions facing the profession. In particular, FATM examiners in Texas have led the way in acknowledging the need for increased transparency in documentation, seeking funding for enhanced technologies (*e.g.*, VCM), and providing essential input on the drafting of this report and its implementation.[37] Texas firearms examiners care deeply about helping end-users understand the significance and limitations of their work. They understand the trier of fact is only able to perform its essential role with an appropriate understanding of the FATM testing and the inferences that may or may not be drawn from the testing depending on the

[36] *See, Laboratory Self-Disclosure 22.17 Final Report on Fort Worth Police Department Crime Laboratory No. 22.17 (Proficiency Testing)* (2023). https://www.txcourts.gov/media/1456474/2217-fwpd-draft-report-033123-1.pdf

[37] There are 23 accredited FSSP's with FATM sections in Texas at the current time. Eight of the 23 are Texas DPS regional laboratories. For a complete list, search the Commission's database here: https://fsc.txcourts.gov/.

case circumstances. The Commission is confident in the potential for meaningful evolution in the discipline *because of* the dedication and professionalism consistently exhibited by Texas FATM examiners and legal stakeholders committed to these issues.

IV. BASICS OF FATM ANALYSIS

The purpose of forensic firearm examination is to assess the value of observed similarities and differences between questioned and known items of ammunition to help determine whether or not compared items may have a common source, *i.e.*, whether they may have been fired from the same recovered weapon. FATM examiners are also sometimes asked to compare fired ammunition from one crime scene to other fired ammunition from a different crime scene. In either case, the question analysts are attempting to answer is how much support the results of their comparisons provide for the proposition that the recovered bullets or cartridge casings were fired from either a recovered firearm or another unknown firearm. Due to the solemnity of this task and its potential repercussions on justice, life and liberty, the legal system in Texas has a strong interest in understanding how well firearms analysts are able to distinguish non-mated comparisons from mated comparisons using the test methods currently available.

The main tool used by FATM analysts during the comparison process historically has been comparison microscopy (CM), where the specimens are placed on two separate stages and viewed simultaneously, side-by-side in the optical path under variable lighting conditions and orientations. As discussed later in this report, new advances in the form of virtual comparison microscopy (VCM) and related imaging technologies have the potential to capture more information and

reduce some aspects of the subjectivity inherent in traditional comparison using CM.[38] But for purposes of this report, we refer to firearms comparison utilizing CM.

Some researchers have likened the firearms comparison analytical process to diagnostic testing.[39] FATM examiners first look for general features, such as the caliber of the bullet recovered from autopsy. Then, using CM, they look for tool mark impressions that the action of a firearm has produced on the surface of a bullet or cartridge. In particular, the firing pin, chamber, and breech face of a firearm may leave marks on a cartridge case, while the rifling, and arrangement of spiral grooves in a firearm barrel, may leave impressions and engraving on a bullet.[40] These accidental characteristics are then used to address the question of whether recovered ammunition parts may have been fired from a particular firearm.

A. Classification of Characteristics

Examiners follow an approach that goes from the general to the particular. They generally start by looking for *class characteristics*. "Class characteristics refer to measurable features that characterize a group of sources, such as firearms of a certain caliber and/or with a certain rifling (*e.g.*, with right-or left-hand twisted grooves inside the barrel). Such features pertain to well-defined, intentional design aspects of the manufacturing process of the firearm."[41] When class

[38] *See, Forensic Optical Topography: A Landscape Study,* p. 5 (NIJ 2016), https://forensiccoe.org/private/6548091ac44c2 ; Chapnick C, Weller TJ, Duez P, Meschke E, Marshall J, Lilien R. *Results of the 3D Virtual Comparison Microscopy Error Rate (VCMER) Study for Firearm Forensics*. J Forensic Sci. 2021 Mar; 66(2):557-570 (2020). doi: 10.1111/1556-4029.14602.

[39] Cuellar, M., Vanderplas, S., Luby, A., Rosenblum, M., comparing FATM to mammography in *Methodological Problems in Every Black-Box Study of Forensic Firearm Comparisons*, arXiv 2403.17248. https://doi.org/10.48550/arXiv.2403.17.248.

[40] For a thorough summary of the basic parts of firearms (*e.g.*, barrels, chamber, breech face, firing pin, extractor, ejector) and ammunition (bullet, propellant, primer, cartridge cases), a description of the physical process that takes place when a trigger is pulled and a gun is fired, a detailed description of the types of toolmarks that may be left on ballistic evidence by firing (cartridge case markings, firing pin impressions, ejector marks, bullet markings), and a brief description of concepts in the manufacture of both firearms and ammunition, *see* Chapter 2 of a 2008 Ballistic Imaging Report by the National Academy of Sciences. Chapter 2 is titled "Firearms and Ammunition: Physics, Manufacturing, and Sources of Variability." (**Exhibit C**).

[41] *See*: Affidavit of Biedermann, A., Budowle, B., Champod, C. Section 3.1 (2022) submitted in a federal case entitled *US v. Kaevon Sutton* (2018 CF1 009709*).*

characteristics of questioned and known specimens are found to correspond, an examination may proceed further by focusing on *accidental characteristics*, historically known as "individual" characteristics.[42] "This term refers to marks attributed to random imperfections or irregularities of various surfaces of the firearm, such as the inner surface of the barrel. The surface features arise accidentally during manufacturing but also in the subsequent use of the firearm, cleaning and maintenance operations, corrosion, and damage. Generally, configurations of accidental marks tend to vary considerably (*e.g.*, National Research Council, 2008) between sources (firearms) and therefore, are widely used by the examiner to help discriminate between candidate sources."[43]

The Association of Firearm and Toolmark Examiners (AFTE) defines three categories of characteristics as follows: [44]

> **"*Class" Characteristics***: Measurable features of a specimen which indicate a restricted group source. They result from design factors and are determined prior to manufacture.
>
> "***Subclass" characteristics[45]:*** Features that may be produced during manufacture that are consistent among items fabricated by the same tool in the same approximate state of wear. These features are not determined prior to manufacture and are more restrictive than class characteristics.
>
> "***Individual" [accidental] characteristics:*** Marks produced by the random imperfections or irregularities of tool surfaces. These random imperfections or irregularities are produced incidental to manufacture and/or caused by use, corrosion, or damage.

---

[42] The firearms community has used the term "individual" historically to indicate that the marks arise randomly in the firearm and are not shared between guns based on the gun's class. While the historical use of this term makes some intuitive sense, as stated throughout this report, the Commission urges Texas FSSP's to shift to using the terms "accidental" over "individual" as different firearms may actually share some subset of common or "accidental" characteristics.

[43] *Supra* n.30.

[44] *See*, AFTE Glossary (6th Ed. 2013: Version 6.091922).

[45] As noted in OSAC FATM Subcommittee Draft Document entitled, *Standard Scale of Source Conclusions and Criteria for Toolmark Examinations*, subclass characteristics are manufactured toolmarks that sometimes repeat virtually unchanged from one manufactured item to another over a limited run of manufactured items. When these characteristics are present on or near the working surfaces of tools, *it is possible for these toolmarks to be mistakenly interpreted as individual characteristics (thus resulting in the identification of a toolmark to a tool other than the one that produced the mark)*.

B. AFTE Examination Process, Theory of Identification and Range of Conclusions

Most accredited forensic laboratories that perform firearms examinations in the United States employ the AFTE Examination Methodology, Theory of Identification, and Range of Conclusions. We summarize their component parts as follows:

1. Examination Process

The examination process begins with the evaluation of the class characteristics discussed above and ends with verification.[46] The process provides an order to the examination and decision-making engaged in by FATM examiners but does not impose rules or proscribe minimum (qualitative or quantitative) criteria to guide decision-making or mitigate the potential impact of human factors.[47] The four-step process is outlined below:

Evaluation:

> The initial examination phase evaluates if the observed class characteristics are the same between two compared elements of ammunition (i.e., two unknown specimens, or an unknown and known specimen). If the specimens are suitable for examination and the class characteristics are the same, then it is possible that the toolmarks were produced utilizing the same tool (such as a firearm). If they are different, then the two specimens can be eliminated as being produced by the same tool.

Comparison:

> If the same class characteristics are observed between two specimens, then a comparative examination is performed utilizing a comparison microscope. The methodology utilized in the examination process is "pattern matching" of the subsequent characteristics. This comparison is conducted to determine: 1) if any marks present are potential subclass characteristics from a particular manufacturing process and/or individual characteristics, and 2) the level of correspondence of any individual characteristics.

---

[46]. *See*, afte.org/resources/swggun-ark/summary-of-the-examination-method (last accessed April 23, 2024).

[47] The Commission does not focus on the potential impact of human factors in the current FATM examination process, except by acknowledging the importance of transparent documentation in each step of decision-making. A recommendation for clear documentation (photographic and note-based) utilizing a linear sequential unmasking approach is provided in Section XI. The Commission made the same recommendation in its report in the Webster (friction ridge) case, which is found at https://www.txcourts.gov/media/1457687/fir-complaint-2216-rsa-latent-prints.pdf

Conclusion:

> If sufficient agreement of individual characteristics is observed between two specimens, an identification conclusion is rendered. If all of the discernible class characteristics are the same but sufficient agreement or disagreement of the individual characteristics is not observed, then an inconclusive (no-conclusion) determination is rendered. In exceptional situations, an elimination conclusion may be rendered on observed differences in individual characteristics.

Verification:

> A verification process is employed to ensure proper conclusions are rendered. As outlined in a laboratory's quality assurance policy, a mechanism should be in place to determine which cases will require verification.[48]

2. AFTE Theory of Identification

As discussed in the conclusion step above, AFTE guidance provides that an examiner may offer an opinion that a specific tool or firearm was the source of a specific set of toolmarks or a particular bullet striation pattern when "sufficient agreement" exists in the pattern of two sets of marks:[49]

- This "sufficient agreement" is related to the significant duplication of random toolmarks as evidenced by the correspondence of a pattern or combination of patterns of surface contours.

- Significance is determined by the comparative examination of two or more sets of surface contour patterns comprised of individual peaks, ridges, and furrows. Specifically, the relative height or depth, width, curvature and spatial relationship of the individual peaks, ridges, and furrows within one set of surface contours are defined and compared to the corresponding features in the second set of surface contours.

- Agreement is significant when the agreement in individual characteristics exceeds the best agreement demonstrated between toolmarks known to have been produced by different tools and is consistent with agreement

[48] Notably, this provision does not require any degree of blind verification. The Commission first recommended laboratories consider incorporating blind verification in its 2016 FATM report. The last 8 years should have given FSSP's sufficient time to implement, and thus the recommendation is repeated in this report and added as an accreditation checklist item.

[49] *See*, AFTE Theory of Identification at https://afte.org/about-us/what-is-afte/afte-theory-of-identification (last accessed April 23, 2024).

demonstrated by toolmarks known to have been produced by the same tool.

- The statement that "sufficient agreement" exists between two toolmarks means that the agreement of individual characteristics is of a quantity and quality that the likelihood another tool could have made the mark is so remote as to be considered a *practical impossibility*.[50] [emphasis added]

- Currently the interpretation of individualization/identification is subjective in nature, founded on scientific principles and based on the examiner's training and experience.

3. AFTE Range of Conclusions

Based on the AFTE Theory of Identification, there are four categories of examination outcomes (AFTE Range of Conclusions Possible When Comparing Toolmarks) typically used by firearm examiners in the microscopic comparison of fired bullets. Laboratory protocols determine how final conclusions are reported.[51] The conclusions are:

> **Identification**: Agreement of a combination of individual characteristics and all discernible class characteristics where the extent of agreement exceeds that which can occur in the comparison of toolmarks made by different tools and is consistent with the agreement demonstrated by toolmarks known to have been produced by the same tool.
>
> **Inconclusive**:
>
> A. Some agreement of individual characteristics and all discernible class characteristic, but insufficient for an identification.
>
> B. Agreement of all discernible class characteristic without agreement or disagreement of individual characteristics due to an absence, insufficiency, or lack of reproducibility.

[50] Per AFTE, "practical impossibility" is defined as "[a] phrase, which currently cannot be expressed in mathematical terms, that describes an event that has an extremely small probability of occurring in theory, but which empirical testing and experience has shown will not occur. In the context of firearm and toolmark identification, "practical impossibility" means that based on 1) extensive empirical research and validation studies, and 2) the cumulative results of training and casework examinations that have either been performed, peer reviewed, or published in peer-reviewed forensic journals, no firearms or tools other than those identified in any particular case will be found that produce marks exhibiting sufficient agreement for identification." *See*, AFTE Glossary (6th ed.) at https://afte.org/resources/afte-glossary

[51] See, afte.org/about-us/what-is-afte/afte-range-of-conclusions (last accessed April 23, 2024). *See also*, AFTE Journal Volume 24, Number 3 (1992).

C. Agreement of all discernible class characteristics and disagreement of individual characteristics, but insufficient for an elimination.

Elimination: Significant disagreement of discernible class characteristics and/or individual characteristics.

Unsuitable: Unsuitable for examination.

V. THE CALL FOR RESEARCH TO EVALUATE THE VALIDITY OF FATM COMPARISON METHODS

AFTE acknowledges that decision-making in FATM comparison involves subjective judgments (*i.e.,* interpretation) by examiners.[52] There is no defined number of quality and quantity of corresponding and non-corresponding features required to make an identification, and the quality and quantity may vary from one comparison to another because some firearms are known to mark better than others. In addition, elements of ammunition recovered from crime scenes may be damaged, fragmented, distorted, and /or corroded to varying degrees. Studies show that different examiners assign different evidential values to observed features, and at times disagree about what exactly constitutes similarities and differences (in accidental characteristics) for a given pair of compared items.[53]

As noted by Scurich et al., it is possible that "trained and experienced examiners develop a highly tuned intuitive ability to distinguish marks made by the same tool versus marks made by different tools." The only way to demonstrate this, however, is through strong empirical proof that could "demonstrate a technique's validity even without an explanation for why or how the

[52] The Commission notes that all forensic disciplines involve subjective judgment to varying degrees.

[53] *See, generally*, Affidavit of Biedermann et al. *supra* n. 29 Section 3.1. *See also*, Monson, K., Smith, E., Peters, E., *Repeatability and reproducibility of comparison decisions by firearms examiners*, Journal of Forensic Sciences, 68: 1721-1740 (2023). https://doi.org/10.1111/1556-4029.15318.

technique works."[54] The need for empirical proof for all forensic disciplines was the focus of two landmark forensic science reports in 2009 and 2016.

A. 2009 NAS Report

In 2009, the National Research Council, National Academy of Sciences released a report titled *Strengthening Forensic Science in the United States, A Path Forward* ("NAS Report") after a multi-year congressionally mandated study. Regarding the AFTE Theory of Identification described above, the NAS Report pointed out that understanding the extent of agreement in marks made by different tools and the extent of variation in marks made by the same tool is a challenging task. The NAS Report concluded that "[t]he validity of the fundamental assumptions of uniqueness and reproducibility of firearms-related toolmarks has not yet been fully demonstrated." It further urged that significant research would be needed to scientifically determine the degree to which firearms-related toolmarks are unique or to quantitatively characterize the probability of uniqueness. The report recommended additional studies be performed to understand "variability" among tools and guns, "the reliability and repeatability of the methods", and how many points of similarity are necessary for a given confidence in the result."[55]

The NAS Report also noted that many forensic science guidance documents "lack the level of specificity" to ensure consistency and rigor in practice:

> Often there are no standard protocols governing forensic practice in a given discipline. And, even when protocols are in place (*e.g.*, [Scientific Working Group] standards), they often are vague and not enforced in any meaningful way.[56]

[54] Scurich, N. Faigman, D., Albright, T., *Scientific Guidelines for Evaluating the Validity of Forensic Feature Comparison Methods*, pnas.org/doi/10.1073/pnas.2301843120 (2023).
[55] NAS Report at 154.
[56] *Id*. at p.6.

B. 2016 PCAST Report

In 2016, the President's Council of Advisors on Science and Technology issued a report titled "*Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature Comparison Methods*." ("PCAST Report"). The report focused on the scientific validity of several feature comparison methods by examining their foundational validity and validity as applied.[57] Regarding "foundational validity" the PCAST Report determined that as of the report's publication in 2016, FATM analysis fell short of the criteria for foundational validity because at the time there had only been a single appropriately designed study to measure validity and estimate reliability.[58] The Report stressed the need for additional, appropriately designed black-box studies to provide estimates of reliability.[59]

VI. ANSWERING THE CALL FOR RESEARCH AND EMPIRICAL DATA: PROGRESS & AREAS FOR IMPROVEMENT SINCE PCAST

Significant time and effort have been dedicated by researchers and practitioners alike to responding to both NAS and PCAST's call for more and better constructed empirical studies and related publications in the FATM discipline. In a recent PNAS[60] article, Scurich et al., while generally critical of FATM black box study design, note some improvements over the last decade:

> Only in the last decade have FATM studies utilizing a fundamentally appropriate design been conducted. This design—known as sample-to-sample design—gives the participant one "known" item and one "unknown" item and asks the participant

[57] PCAST defined "foundational validity" as the *scientific* standard corresponding to the legal standard of evidence being based on "reliable principles and methods." "Validity as applied" means the *scientific* standard corresponding to the legal standard of an expert having "reliably applied the principles and methods." President's Council of Advisors on Science and Technology, Report to the President, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*, U.S. Executive Office of the President, Washington, D.C., USA, 2016. p. 43. (Emphasis original).

[58] Other "set-based analyses" studies (*e.g.*, closed set/partly open set) were shown to have a drastically lower rate of inconclusive examinations and false positives due to the design of those studies.

[59] The OSAC Firearms and Toolmarks Subcommittee (among other stakeholder groups) disagreed with the PCAST conclusion that firearms analysis fell short of the criteria for foundational validity and issued an extensive formal response. *See*, *e.g.*, OSAC Firearms and Toolmarks Subcommittee's Response to the President's Council of Advisors on Science and Technology's (PCAST) Request for Additional Information – Submitted December 14, 2016.

[60] PNAS is an acronym for Proceedings of the National Academy of Sciences of the United States of America.

> to determine whether the unknown item came from the same source as the known item. The participant makes a judgment and then puts those items away. She is then presented with additional items to compare in the same fashion. In this way, each comparison is independent, which makes calculating performance metrics relatively straightforward. To date, only five studies have utilized the sample-to-sample design….[61]

It is beyond the scope of this report (or the current resources of the Commission) to assess whether the quality of any particular post-PCAST study (or body of studies as a whole) should be graded as sufficient or insufficient.[62] Even if the Commission were inclined to make that assessment, the effort would be duplicative of five years of study on this subject pending publication at NIST.

A. NIST Scientific Foundational Review

The NIST Scientific Foundational Review related to FATM examination began in October of 2019 and gathered literature with a focus on error rate studies. The goal of the study was to

---

[61] *See*, Scurich et al., *supra* n. 43, citing D. P. Baldwin, S. J. Bajic, M. Morris, D. Zamzow, "*A study of false-positive and false-negative error rates in cartridge case comparisons*" (US Department of Energy, Ames Laboratory, Iowa, 2014), https://apps.dtic.mil/ sti/pdfs/ADA611807.pdf.; M. A. Keisler, S. Hartman, A. Kilmon, M. Oberg, M. Templeton, *Isolated pairs research study*. *AFTE J.* 50, 56–58 (2018): K. L. Monson, E. D. Smith, E. M. Peters, *Accuracy of comparison decisions by forensic firearms examiners*. *J. Forensic Sci.* 68, 86–100 (2023); B. A. Best, E. A. Gardner, *An assessment of the foundational validity of firearms identification using ten consecutively button-rifled barrels. AFTE J.* 54, 28–37 (2022); M. Guyll, S. Madon, Y. Yang, K. A. Burd, G. Wells, *Validity of forensic cartridge-case comparisons*. Proc. Natl. Acad. Sci. U.S.A. 120, e2210428120 (2023).

[62] Indeed, there has been a robust debate about the design of the post PCAST studies or the calculation of the FPR. *See e.g.*, Cuellar et al., *Methodological Problems in Every Black Box Study of Firearms Comparisons*, March 2024 available at https://arxiv.org/abs/2403.17248; Rosenblum et al., *Misuse of statistical method results in highly biased interpretation of forensic evidence in Guyll et al.* (2023) Law, Probability and Risk, 2024, 23, 1–6 https://www.researchgate.net/publication/377331008;*Misuse_of_statistical_method_results_in_highly_biased_interpretation_of_forensic_evidence*; Scurich et al., *Scientific guidelines for evaluating the validity of forensic feature-comparison methods*, PNAS, October 10, 2023, VOL. 120, NO. 41, https://www.pnas.org/doi/epdf/10.1073/pnas.2301843120; Khan & Carriquiry (2023) *Shining a Light on Forensic Black-Box Studies, Statistics and Public Policy*, 10:1, 2216748, DOI: 10.1080/2330443X.2023.2216748; Dorfman & Valiant, *A Re-analysis of Repeatability and Reproducibility in the Ames-USDOE-FBI Study, Statistics and Public Policy* (2022), https://www.tandfonline.com/doi/full/10.1080/2330443X.2022.2120137; Faigman et al., *The Field of Firearms Forensics is Flawed*; *The matching of bullets to guns is subjective, and courts are starting to question it because of testimony from scientific experts*, Scientific American, May 2022, https://www.scientificamerican.com/article/the-field-of-firearms-forensics-is-flawed; Khan & Carriquiry (2023) *Hierarchical Bayesian non-response models for error rates in forensic black-box studies*. Phil. Trans. R. Soc. A 381:20220157 https://doi.org/10.1098/rsta.2022.0157; and Dror & Scurich, (*Mis)use of Scientific Measurements in Forensic Science*, Forensic Science International: Synergy, Volume 2, 2020, Pages 333-338 https://www.sciencedirect.com/science/article/pii/S2589871X20300553.

evaluate the scientific foundations of firearm examinations and reliability of conclusions drawn. The objective was to answer the question: *What empirical data exist to support or refute the claims and methods that firearm practitioners use to analyze evidence?*[63]

NIST formulated the supporting claims in firearm examination into an overall claim and six supporting subclaims.

> **Overall Claim**: A conclusion of common origin between two compared toolmarks can be made when there is sufficient correspondence of distinctive (or unique) features called individual characteristics, and these conclusions are extremely accurate when rendered by a competent or qualified examiner.
>
> **Division into Supporting Subclaims:**
>
> - The surfaces of firearm parts produced by manufacturing tools are unique. The parts of interest are those that interact as tools on ammunition components.
> - Upon loading and firing, these parts of interest produce marks on the surfaces of ammunition components that are unique to the firearm.
> - Upon loading and firing, the marks on the surfaces of ammunition are also reproducible from one shot to the next and for different ammunition.
> - With normal use, the unique surfaces on the firearm parts are stable over time and over many firings resulting in reproducible, unique marks on ammunition over time and over many firings.
> - Although there are limits to accuracy due to human performance, to the resolution of optical microscopes, and to variations in the dynamic process of firing, error rates are low when conclusions are drawn about common origin.
> - The AFTE Theory of Identification and its use of the term "practical impossibility" are consistent with measured error rates.

The NIST Foundational Review narrowed its focus to 23 published studies on FATM. The report is currently undergoing internal review at NIST and should be released for public comment

[63] *See*, NIST Scientific Foundation Reviews, NISTIR 8225 Draft, available at https://doi.org/10.6028/NIST.IR.8225-draft.

sometime in 2024. While the Commission (and many others) will view this report with great interest once it is released, the Commission's work in this report is intended as a path forward for firearms examiners in Texas pending the outcome of the NIST Foundational Review. *The Commission believes the recommendations offered in this report are sound and should be implemented regardless of the outcome of the NIST Foundational Review, though additional measures may be recommended after it is published.*

B. Relentless Disagreement About How to Calculate Error Rates and How Their Significance Should Be Expressed to the Trier of Fact

Improvements in certain aspects of FATM study design post-PCAST have not resolved extensive disagreement about the best way to calculate and discuss the significance of error rates, particularly with respect to inconclusives.[64] As noted by Swofford et al., "when deliberating on this issue, nearly every possible option has been proposed including inconclusive decisions be ignored altogether, inconclusive decisions always be considered correct, inconclusive decisions always be considered incorrect, inconclusive decisions be considered correct in some situations and incorrect in other situations, and inconclusive decisions be considered neither correct nor incorrect."[65]

While error rates are one consideration for admissibility of scientific evidence under Texas law, the Court of Criminal Appeals (CCA) has held that they are not required for a forensic technique to be admissible.[66] For example, in a 2017 capital murder case, when a lack of laboratory

[64] Other methodological concerns raised include inadequate sample size; non-representative samples; invalid confidence intervals for error rates; and missing data. *See supra* n. 51. *See also*, Cuellar et al. *supra* n.28.

[65] For an extensive discussion of the various options offered, *see* Swofford et al, *supra* n. 33, particularly Table 2. One option suggested by some researchers is to approach FATM data in a similar way as the FDA does when evaluating diagnostic tools in medicine, which would require setting the inconclusive responses to positive (and then to negative) and reporting the range. While borrowing from established approaches in other industries makes some sense, the Commission believes the approach suggested by Swofford et al. would be more useful in Texas cases.

[66] *Williams v. State*, 2017 Tex. Cr. App. Unpub. LEXIS 906 (Tex. Cr. App. 2017).

test method or analyst-specific error rates was presented as an argument against FATM admissibility on appeal, the CCA rejected the argument in favor of admissibility. During testimony, a DPS FATM examiner "acknowledged that a precise casework error rate could not be measured but pointed out that consecutive-manufacture and proficiency studies provided error rates in the context of controlled studies, and noted the known error rates could then be used to estimate casework error rates."[67]

The DPS examiner's testimony represents common testimony by firearms examiners when they are questioned about the existence and significance of error rate data derived from published black box studies.[68] Examiners endeavor to provide accurate information to the trier of fact about the nature and limitations of published research by acknowledging that error rates from those studies provide some useful information but are not an exact fit for the question of interest to the court. Texas firearms examiners are often expected to answer questions about the implications of the results of black box studies in which they may (or may not) have participated and for which data generated may (or may not) reflect the performance of the methodology set forth in their own comparison procedures, which is the issue of greatest interest to the trier of fact.[69] Finally, though many firearms examiners have developed an understanding of certain aspects of statistics, they are generally not statisticians and may not have the requisite skill set or knowledge base to respond accurately to questions about choices made by the authors regarding experimental design or the statistical models used in evaluating the data collected.

[67] *Id.*

[68] As described by NIST, black box studies are used by researchers to measure the reliability of methods that rely mainly on human judgment (such as in many forensic science feature comparison/pattern matching disciplines like FATM).

[69] Because the guidance published by AFTE sets forth an overall process for identification rather than a specific protocol, it is not possible to know which specific protocol was followed by each black box study participant in such a way that one could generalize the performance data from the study to a specific case.

C. Why Existing Error Rates Calculated from Black Box Studies in FATM Do Not Provide a Suitable Metric for Representing FATM Method Performance

As articulated by Swofford et al., (2024) "[t]he focus on error rates as a primary measure of method performance is generally satisfactory when experts report results using a binary scale, such as identification or exclusion [or elimination]" Koehler et al. describe the situation in a similar way: "If examiners always reached either an identification conclusion (*i.e.,* that two patterns originated from the same source) or an elimination, it would be a simple matter to compute, say, a false-positive error rate. It would be the number of times the examiner reached a 'same source' conclusion divided by the number of sample pairs that were known to have been produced by a different source."[70] However, the AFTE Range of Conclusions provides FATM examiners with five possible options, including:

- Identification;
- Inconclusive (sub-classified as Type A, B, or C); or
- Elimination.[71]

In making an argument for a new approach, Swofford et al. asks the reader to consider the following "hyperbolic example," which is intended to illustrate the point that it is unsatisfactory (and potentially misleading) to use false positive and false negative error rates from black box studies as the metric of performance in any feature comparison discipline (such as FATM) that does not limit examiners to only two choices. In Texas, the reader does not need to imagine a "hyperbolic example," because the Commission discussed this exact example with a previous

[70] Koehler, J., Mnookin, J., Saks, M., *The scientific reinvention of forensic science* (2023), pnas.org/doi/full/10.1073/pnas.2301840120.

[71] The Commission recognizes that unsuitable for comparison is also an option but for purposes of this report limits the discussion to the five main interpretation categories.

FWPDCL proficiency testing policy (since changed) in a report approved at its April 2023 quarterly meeting.[72]

In the following (2x3) discriminability table created by Swofford et al., Method 1 represents a method that results in the examiner issuing an "inconclusive" opinion for all comparisons, while Method 2 yields no inconclusive opinions.

Table 1a. A (2x3) discriminability table representing performance metrics where only inconclusives were reported.

Method 1	Identification	Inconclusive	Exclusion
Mated Comparisons	0%	100%	0%
Non-Mated Comparisons	0%	100%	0%

Table 1b. A (2x3) discriminability table representing performance metrics relating to hypothetical Method 2 where all reported outcomes for mated comparisons are "Identification" and all non-mated comparisons are "Exclusion."

Method 2	Identification	Inconclusive	Exclusion
Mated Comparisons	100%	0%	0%
Non-Mated Comparisons	0%	0%	100%

Swofford et al. explain that both of these examples achieved false positive and false negative rates of 0%, which ordinarily would be indicators of reliability.[73] However, Method 1 achieved this by allowing the examiner to call everything "inconclusive" regardless of the expected result from the PT provider, while Method 2 reached conclusion opinions consistent with ground truth in all instances. As Swofford et al. observed, the two methods are distinct in terms of their usefulness. The Commission agrees and made a similar observation in the FWPDCL report.[74] The usefulness of a method that essentially encourages examiners to default to inconclusive is nil if the purpose of proficiency testing is to periodically assess an examiner's analytical abilities. Because

[72] See, *supra* n. 25.

[73] Other authors have suggested a similar approach, *see* Letter to the Editor, by Weller and Morris. doi: 10.1016/j.fsisyn.2020.10.004. *See also*, Monson et al. (2022) discussed later in this report, which presents data in a similar way.

[74] *Id.*

of the various options available to FATM examiners per the AFTE Range of Conclusions (including three different sub-categories of inconclusive), false positive and false negative rates alone "do not accurately convey how successfully one could use the method output to distinguish non-mated comparisons from mated comparisons and therefore do not adequately characterize method performance."[75]

The Commission observes that there are many circumstances in which a FATM examiner's decision to report an inconclusive result in controlled studies, proficiency testing, or case work *is* the most appropriate and professionally responsible decision. The Commission agrees with Swofford et al. that inconclusive decisions can be appropriate or inappropriate depending on case circumstances. Indeed, in a circumstance where evidentiary quality is poor and limited features are observable, if an examiner were to push ahead and make a binary call (either identification or elimination) he or she could risk violating the Texas Code of Professional Responsibility for Forensic Analysts and Crime Laboratory Managers, which requires reported conclusions to be based on sufficient data.[76] Texas examiners are keenly aware of their obligations under the Code; indeed they are tested on the Code (and its application) when they apply to be licensed by the Commission.[77]

---

[75] *See*, *supra* n. 22 at p. 5.

[76] *See*, 37 Tex. Admin. Code Section 651.219 (Code of Professional Responsibility (effective May 2018). https://texreg.sos.state.tx.us/public/readtac$ext.TacPage?sl=R&app=9&p_dir=&p_rloc=&p_tloc=&p_ploc=&pg=1&p_tac=&ti=37&pt=15&ch=651&rl=219.

[77] The Texas Forensic Science Commission's General Forensic Analyst Licensing Exam covers seven different topics that include professional responsibility, root cause analysis, human factors, statistics for forensic practitioners, evidence handling, *Brady*/Michael Morton Act, and expert testimony.

VII. SOLUTION FOR TEXAS FATM EXAMINERS: REPORTING BOTH METHOD PERFORMANCE AND METHOD CONFORMANCE TO THE TRIER OF FACT[78]

A. Proposed Approach

In forensic science, end-users of forensic results are provided a report of a forensic test conducted by a particular analyst. They are then tasked with making inferences and decisions about the truth of various propositions in question (*e.g.,* whether or not two patterns could have originated from the same source). As stated by Swofford et al., forensic science end-users have an interest in understanding the answers to the following three questions:

(1) What method did the analyst apply when conducting the forensic examination?

(2) How effective is that method at discriminating between the propositions of interest?[79]

(3) How relevant is the data describing the discriminability (*i.e.,* diagnostic capacity) of that method (generally) to the examination in the case at hand (specifically)?[80]

To answer these questions, information about whether the analyst conformed to a particular method as well as measures relating to the performance of the method are needed. In this context, Swofford et al. offer the following defined terms:

- *Method conformance* relates to assessments of whether the outcome of a particular method is the result of the analyst's adherence to the procedures that define that method.

- *Method performance* relates to measures that reflect the extent to which the outcome of a particular method can effectively distinguish between different propositions of interest (*e.g.,* between same-source and different-source comparisons).

---

[78] The Commission is grateful to Henry Swofford and his co-authors at NIST who developed the approach in **Exhibit E.** Their work is incorporated extensively in this and other sections of the Commission's report.

[79] The Commission notes that in order for one to assess the efficacy of a method in discriminating between propositions of interest, one first must clearly define the propositions of interest. The propositions may change depending on the case. H_1 will often be "the item was fired in the submitted firearm." However, H_2 may vary. For example, a common one is, "the item was fired in an unknown firearm." However, if considering the general population of "unknown" firearms, then even agreement of class characteristics (in the absence of disagreement regarding accidental characteristics) provides some weight in favor of H_1, where an examiner using the AFTE Range of Conclusions would opine inconclusive. Thus, implications of the approach must be carefully considered.

[80] *See*, Swofford *et al*, *supra* n. 33 p. 3.

In the approach suggested by Swofford et al., instead of referencing binary error rates generated from black box studies, Texas firearms examiners would present two tables of data for consideration by the trier of fact using data obtained based on their own comparison methods. The first table (referred to as a "(2x3) discriminability table") provides information about the discriminability of the method. The second table (referred to as a "(3x3) reproducibility table") provides information about the reproducibility of the method. The term "discriminability" refers to the extent to which the outcomes of a method can accurately distinguish between non-mated (*i.e.*, different source) and mated (*i.e.*, same source) comparisons. The term "reproducibility" refers to the extent to which outcomes of a method are consistently produced between different examiners. Both measures of discriminability and reproducibility provide important information about the performance of a method. Measures of reproducibility provide the "gauge by which measures of discriminability (based on outcomes from multiple analysts generally) are relevant to an outcome by a particular analyst (specifically) as well as the adequacy of the procedures that define the method."[81]

B. First Table Type (2x3): Method Discriminability Where Ground Truth Is Known.

The Commission agrees with Swofford et al. that this method provides "greater transparency about the method's performance and enables users of the information to more effectively discriminate between propositions of interest (*i.e.*, mated versus non-mated)."[82] In this

81 Per Swofford, et al., *supra* n. 33,p.4 n.7: "this is particularly important when analyst vary in their performance and measures of discriminability and reproducibility are based on aggregate outcomes from multiple analysts."

82 Swofford et al. *supra* n. 33, p. 5, , citing, T.J. Weller, M.D. Morris, Commentary on: I. Dror, N Scurich *"(Mis)use of Scientific Measurements in Forensic Science"* Forensic Science International: Synergy (2020) https://doi. org/10.1016/j.fsisyn. 2020.08.006; A. Biedermann, K.N. Kotsoglou, *Forensic Science and the Principle of Excluded Middle: "Inconclusive" Decisions and the Structure of Error Rate Studies*, Forensic Sci. Int.: Synergy 3 (2021) 1–11. https://doi.org/10.1016/j.fsisyn.2021.100147: M. Guyll, S. Madon, Y. Yang, K.A. Burd, G. Wells, *Validity of Forensic Cartridge-Case Comparisons*, Proc Natl Acad Sci USA 120 (2023). https://doi.org/10.1073/pnas.2210428120.

section, the Commission provides three examples of what this might look like. The first uses data from HFSC's blind quality control program, and the second two use data from the most recent CTS Firearms Proficiency Test #23-5262.

1. HFSC Blind QC Firearms Data in a (2x3) Discriminability Table

In this example, HFSC's quality division, in collaboration with HPD and CSAFE, created mock casework samples and ran them through HFSC's system as they would with real casework.[83] Analysts reported comparison results consistent with HFSC firearms protocols which incorporate the AFTE Range of Conclusions: *Identification, Elimination, Inconclusive,* and *Unsuitable*. Conclusions of identification were based on individual characteristics, while conclusions of elimination were based on class or accidental characteristics. An inconclusive decision indicates an inadequate correspondence of accidental and/or class characteristics needed to make an identification or elimination decision.

HFSC included 558 total items, 529 of which were deemed suitable for comparison where analysts reported identification, inconclusive or elimination conclusions.[84] Ground truth was reached by examiners in 333, or 59.7% of the comparisons.[85] The (2x3) discriminability table

[83] The HFSC study was published as an exploratory study, and thus has certain limitations. The following are key parameters of the study: 11 examiners participated; 365 ground truth ID comparisons/143 ground truth elimination comparisons were incorporated; fired evidence was created using firearms slated for destruction, staff's personally owned firearms, or firearms from the laboratory's reference collection; the laboratory chose challenging cases to better define the limitations of analysis; 23% of all comparisons were created with two different firearms of the same class. HFSC's decision parameters and key definitions are set forth in Neuman et al., at page 966.

[84] All information related to the HFSC study is derived from the following article: Neuman et al., *Blind Testing in Firearms: Preliminary Results from a Blind Quality Control Program,* Journal Forensic Science, 67: 964-974 (2022). doi: 10.1111/1556-4029.15102. A total of 558 comparisons were made by HFSC examiners in the blind QC program, but those deemed insufficient for comparison by examiners are not included in this table.

[85] Table 2 provides a helpful starting point for reporting, and with some additional effort, even more information could be provided regarding results for particular types of comparisons. For example, in the HFSC data, bullets were the main contributors to inconclusive decisions. Per Neuman et al., "these results indicate firearms examiners routinely reach a correct determination of ground truth identification for cartridge cases and bullets (more sensitivity) but may have more difficulty discriminating elimination in bullets compared with cartridge cases (less specificity)." *Id.* at 971 FSSP's could include an appendix to providing further detail regarding the differences in outcomes for cartridge casings vs. bullets which may be helpful to the trier of fact in assessing the weight of evidence in a given case.

shows the breakdown of reporting for mated (same source) and non-mated (different source) pairs, respectively:[86]

Table 2a. A (2x3) discriminability table with performance metrics for HFSC results from blind testing data.

HFSC Firearms Comparisons	Reported as: **Identification**	Reported as: **Inconclusive**	Reported as: **Elimination**
Ground Truth: Mated (Same Source)	69%	31%	0
Ground Truth: (Non-Mated) Different Source	0	74%	26%

The table shows a few helpful pieces of information for the trier of fact. First, no identifications were declared for true nonmatching pairs, and no eliminations were declared for true matching pairs.[87] However, it also shows that when ground truth was same source, examiners reported inconclusive findings at a rate of 31%, and when the ground truth was different source, examiners reported inconclusive findings at a rate of 74%. When asked about this imbalance, FATM examiners explain that identifications are easier to make based on accidental characteristics alone. When an examiner makes an elimination conclusion, the examiner is essentially stating the examined items *could never have been fired* by the firearm in question, and there is reluctance within the community to make that kind of statement if there is no divergence in class characteristics because variability in accidental characteristics occurs for any number of reasons (environment, passage of time, etc.). Thus, legal end-users (prosecutors and defense attorneys alike) should *take great care* to understand the potential significance (or lack thereof) of

[86] From Swofford *et al, supra n. 33:* For feature comparison disciplines, this can be accomplished using a (2x3) discriminability table or equivalent rate parameters reflecting the occurrence of identification, exclusion, and inconclusive decisions as they relate to ground-truth of the compared items. A (2x3) discriminability table is used in this discussion; however, this recommendation would create a 2x5, 2x7, or 2x9 scale, if the FSSP's test method allowed for an expanded range of inconclusive results, such as the sub-categories of inconclusive described in AFTE's Range of Conclusions.

[87] While the false positive/false negative error rates from the HFSC blind study (and the CTS data described below) are encouraging, they should not be read to imply that there is such a thing as a 0% error rate. An oft-cited critique of forensic science practice historically is that examiners in various disciplines would testify to a 0% error rate. This assumed infallibility is difficult to overcome, and examiners should be cautious not to imply that FATM (or any forensic assay) has a 0% error rate when discussing the data from (2x3) discriminability tables, even where no false positive/false negatives were reported in the particular data set.

inconclusive opinions because published data (HFSC data and data in other published studies) show that inconclusive opinions occur more frequently among comparisons for which ground truth is different source than among those for which ground truth is same source.[88]

2. CTS Proficiency Test Data in a (2x3) Discriminability Table

The second and third (2x3) discriminability tables below derive from the results of CTS Proficiency 23-5262. In this scenario, participants were told that police recovered four bullets from a crime scene and seized a CZ 75 P-07 firearm from the suspect. According to the scenario, the suspect was apprehended later that day. Three rounds of PMC .40 S&W 180 grain FMJ-FP ammunition (consistent with the bullets found at the scene) were test-fired with the suspect's firearm and the bullets were collected. The analyst was asked to compare the recovered bullets from the scene with those test-fired from the suspect's firearm and report their findings.

Following is the (2x3) discriminability table from three comparisons where the participants were given known test-fired bullets from the suspect's firearm, but ground truth for all three items was that they were fired from a *different firearm of the same brand*:[89] The first table reflects data from all participants nationwide:

Table 2b. A (2x3) discriminability table with performance metrics for *all participants* in CTS 23-5262: Items 2,3,5.

CTS Comparisons (Different Firearm, Same Brand)	**Reported as: Identification**	**Reported as: Inconclusive**	**Reported as: Elimination**
Ground Truth: Mated (Same Source)	None	None	None
Ground Truth: Non-Mated (Different Source)	19%	49%	32%

[88] HFSC is currently engaged with researchers in an effort to ascertain whether the decision criteria adopted for inconclusive calls differed for "discovered" vs. "undiscovered" blind FATM tests. In the HFSC FATM blind study, 20% of bullet comparison test items and 18% of cartridge case comparisons were "discovered" by examiners as part of the testing program. The extent to which examiner knowledge of the type of case they are working (real case work vs. proficiency testing cases) impacts decision-making is an issue frequently raised by academics in critiquing existing external proficiency testing programs. This issue will be the subject of subsequent publication by HFSC and may help inform design of future blind quality control programs.

[89] Proficiency data here are from a very small sample (4 items compared). Lab(s) would need to gather much more data to represent a broader range of casework to report the (2x3) discriminability table across a range of casework.

The same data for Texas-only participants is provided in the following (2x3) discriminability table:

Table 2c. A (2x3) discriminability table with performance metrics for *Texas participants* in CTS 23-5262: Items 2,3,5

CTS Comparisons (Different Firearm, Same Brand)	**Reported as: Identification**	**Reported as: Inconclusive**	**Reported as: Elimination**
Ground Truth: Mated (Same Source)	N/A	N/A	N/A
Ground Truth: Non-Mated (Different Source)	0	67%	33%

Notably (and unlike the data for the rest of the participants) *no Texas examiners* reported ground truth (non-mated) comparisons as identification.[90] The CTS data capture important information for the gatekeeper and jury because they show the possible challenges in testing when there are different firearms of the same brand.[91] This information should assist gatekeepers in making decisions regarding reliability given the circumstances of the case before them. It is also important to note (and not reflected in the above table) that when examiners were asked to compare bullets fired from a Desert Eagle 40 S&W (a third firearm of a different brand and with different class characteristics), the results were closer to ground truth, and much more aligned with the false positive and false negative data reported in the HFSC blind study and other black box studies with low false positive and false negative rates:

Table 2d. A (2x3) discriminability table with performance metrics for *all participants* in CTS 23-5262: Item 4 only

CTS Comparisons (Different Firearm, Different Brand)	**Reported as: Identification**	**Reported as: Inconclusive**	**Reported as: Elimination**
Ground Truth: Mated (Same Source)	N/A	N/A	N/A
Ground Truth: Non-Mated (Different Source)	0.4%	1.1%	98.2%

[90] CTS provided data for 18 Texas participants. The Commission acknowledges this is a small sample size. To confirm that Texas FSSPs outperform the rest of the participant group in terms of avoiding false positive errors, one would need to collect more data from all Texas FATM examiners for similarly challenging CTS tests over an extended period.

[91] Texas firearms examiners point out that the extent of difficulty in comparing recovered bullets from different firearms of the same brand depends in part on the brand; other challenging scenarios may include between firearms utilizing certain manufacturing processes.

Following is data for Texas CTS participants where the comparison was to bullets fired from a different brand of firearm (the Desert Eagle 40 S&W). *All Texas examiners* reported an elimination conclusion where ground truth was non-mated (different source).

Table 2e. A (2x3) discriminability table with performance metrics for *Texas participants* in CTS 23-5262: Item 4 only

CTS Comparisons (Different Firearm, Different Brand)	**Reported as: Identification**	**Reported as: Inconclusive**	**Reported as: Elimination**
Ground Truth: Mated (Same Source)	None	None	None
Ground Truth: Non-Mated (Different Source)	0%	0%	100%

In sum, the Commission observes that including a (2x3) discriminability table, such as that represented with the data above, provides greater transparency about the method's performance.[92] Information regarding the discriminability of the method should also help the trier of fact assess what weight to give to the method's result in a given criminal case. Reporting the information in the (2x3) discriminability table above would be a significant step toward helping end-users better understand the strengths and limitations of FATM comparison. However, additional information regarding reproducibility of decisions among examiners/laboratories when the method is applied is needed to complete the picture.

[92] While the Commission believes the table is a step forward for firearms examination, it is important to note that no table is perfect. For example, the (2x3) discriminability table presented here only presents the perspective of pairwise comparisons leading to a matrix of conclusions. *In casework, an examiner may be presented with no known firearm and multiple bullet fragments, each of which has similar class characteristics but insufficient information to justify that the fragments all came from the same bullet.* This common scenario (and the process an examiner uses to analyze the fragments) is not necessarily reflected in the (2x3) discriminability table, and thus the working group proposed in the Recommendation section might consider approaches (or qualifying language) to reflect the limitations inherent in pairwise comparison data in the FATM discipline. Additionally, there may be some limitations regarding the extent to which the data in a single table are reflective of different types of cases, including, for example, those based on hits from a database search, which may involve close nonmatch samples, and those involving guns produced using newer manufacturing processes (*e.g.*, 3D guns, ghost guns).

C. Second Table Type (3x3): (Reproducibility of Decisions Among Examiners/Laboratories)

In addition to understanding how a comparison method performs when ground truth is known, all parties (FSSPs, lawyers, and judges alike) need to understand how reproducible the method is. A well-defined method will lead to a high proportion of consistent outcomes between examiners when viewing the same evidence.[93] If a judge has the data presented in the (2x3) method discriminability table above about how the laboratory's method performs, and also has data showing examiner consistency when utilizing the method, then he can begin to make a more informed assessment of reliability (and if appropriate, issue cautionary instructions or limiting language) in the case before him.[94] If the 3x3 reproducibility table shows a lack of consistency between examiners, the lawyers and court should know this information in assessing the value of the forensic evidence. A lack of consistency among examiners may signal that the comparison method itself is too loosely defined. In such a case, the trier of fact would need to be cautious in relying on method performance data from the (2x3) discriminability table alone as applicable to the criminal case before him because the method itself may or may not have been followed by the examiner in the case.[95] In other words, an FSSP's inability to establish reproducibility among examiners means that a (2x3) discriminability table may not be as informative as it could be regarding the performance of any single examiner.

A laboratory can create a (3x3) reproducibility table without necessarily having ground truth about whether the comparisons are mated (same source) or non-mated (different source), because reproducibility measures how often examiners (or laboratories) are consistent in their

---

[93] Swofford *et al*. *supra* n. 33 p. 6 (Table 4)..

[94] *See* "Statement of the Texas Forensic Science Commission Regarding 'Alternate Firearms Opinion Terminology'" at: https://www.txcourts.gov/media/1453352/tfsc-statement-re-firearms-terminology-document.pdf

[95] For example, per Swofford *et al*., *supra* n. 33, p.5: "the approaches for assessing conformance might not be sufficient (*i.e.*, outcomes have been improperly assessed as conforming)."

conclusions after applying the test method, not how often their conclusions are aligned with ground truth. Inconsistent outcomes reflect the extent of variability between analysts (or laboratories) and the degree to which interpretations might vary due to subjectivity relating to analyses of quality, quantity, similarity or rarity of features, or different decision thresholds. Just as method reproducibility is important for lawyers and judges to understand, it is similarly important for quality managers and those in the laboratory responsible for training. If the data show it is common for different analysts to reach different decisions for a given input, a deeper dive into the reason for this would be called for. One example of a root cause might be a section of a procedure that is unclear (too "loose") or lacking in specificity, and which could be cured by revisions to a standard operating procedure. Alternatively, if the data show outliers in the form of one or two individuals, this might point to an isolated training or competency issue previously undetected.

The Commission agrees with Swofford et al. that "forensic service providers that do not have well documented and detailed step-by-step procedures that define their method, including conditions for method application and decision criteria for results for which performance data can be associated are unlikely to be able to meaningfully support a claim that the outcome of their examination is the product of a reliable method"[96]

Table 3a. A basic (3x3) reproducibility table adopted from Swofford et al. The table reflects consistency (or lack thereof) between multiple applications of the same method. Once a laboratory has established internal reproducibility, the same information could be collected across FSSPs.

REPRODUCIBILITY	**Identification**	**Inconclusive**	**Exclusion**
Identification	*Consistent*	Inconsistent	Inconsistent
Inconclusive	Inconsistent	*Consistent*	Inconsistent
Elimination	Inconsistent	Inconsistent	*Consistent*

[96] Swofford *et al*. *supra* n. 33 p. 8..

Table 3b. A reproducibility table using combined data from mated and non-mated bullet comparisons sets in Monson et al. (2023)[97]. The table reflects the extent to which different examiners are consistent with each other in performing bullet comparisons employing the AFTE Theory of Identification. Once a laboratory has established internal reproducibility, the same information could be collected across FSSPs in Texas.

REPRODUCIBILITY (BULLETS)	**Second Evaluation: Identification**	**Second Evaluation: Inconclusive**	**Second Evaluation: Elimination**
First Evaluation: Identification	601 (21.6%)	109 (3.9%)	13 (0.5%)
First Evaluation: Inconclusive	93 (3.3%)	841 (30.3%)	379 (13.6%)
First Evaluation: Elimination	14 (0.5%)	451 (16.2%)	277 (10%)

The green highlighted blocks indicate consistency in opinions between the first and second evaluations of the same evidence by two different examiners. For example, in 601 of the comparisons, both the first and second evaluations yielded an identification opinion; in 841, inconclusive; and in 277, elimination. In all other table squares, the first and second evaluations yielded different opinions (*e.g.*, in 93 comparisons, the first evaluation yielded an inconclusive opinion while the second evaluation yielded an identification opinion, and so on.) The blocks in the far opposite corners of the table show the number of times the first evaluation swung all the way from an elimination opinion to identification in the second evaluation (14 times), or vice versa from identification to elimination (13 times).

We include the Monson et al. reproducibility data as one example of how an FSSP would populate a (3x3) table. It is notable that in this particular study (Monson et al.) examiners applied the AFTE Theory of Identification framework, but it is unclear the extent to which specific laboratory protocols differed among participants and to what extent. To be clear, the purpose of recommending the 3x3 reproducibility table for Texas FSSPs is to focus on reproducibility data for the laboratory (or laboratory system) comparison method actually used in Texas criminal

[97] Monson, K., Smith, E., Peters, E., *Repeatability and Reproducibility of Comparison Decisions by Firearms Examiners*, Journal of Forensic Sciences, 68: 1721-1740 (2023). https://doi.org/10.1111/1556-4029.15318.

casework. The Commission would hope (and even expect) to see better reproducibility within a single Texas laboratory or even a large laboratory system than what was observed in the Monson study.[98]

It could also be valuable (especially with a lens toward risk management from a quality system perspective) to separate reproducibility data for mated and non-mated comparisons because it may show the extent to which the process verification in scenarios with disagreement between a first and second examiner could yield a more accurate result. Using the Monson data as an example, the first table shows reproducibility (from first to second examiner) with mated bullet comparisons while the second table shows the same data for non-mated comparisons:

Table 3c. A reproducibility table using data from mated bullet comparisons sets in Monson et al. (2023)[99]. The table reflects the extent to which different examiners are consistent with each other in performing bullet comparisons for ground truth mated comparisons employing the AFTE Theory of Identification.

REPRODUCIBILITY (BULLETS)	**Second Evaluation: Identification**	**Second Evaluation: Inconclusive**	**Second Evaluation: Elimination**
First Evaluation: Identification	85.4%	12.9%	1.7%
First Evaluation: Inconclusive	45.0%	45.5%	9.5%
First Evaluation: Elimination	31.7%	61.0%	7.3%

[98] As previously stated with respect to the (2x3) discriminability table, laboratories should include Inconclusive A-C sub-categories in their reproducibility tables to the extent they utilize those categories in casework. Thus, the table would expand to a (5x5) reproducibility table.

[99] Monson, K., Smith, E., Peters, E., *Repeatability and reproducibility of comparison decisions by firearms examiners*, Journal of Forensic Sciences, 68: 1721-1740 (2023). https://doi.org/10.1111/1556-4029.15318.

Table 3d. A reproducibility table using data from non-mated bullet comparisons sets in Monson et al. (2023)[100]. The table reflects the extent to which different examiners are consistent with each other in performing bullet comparisons for ground truth non-mated comparisons employing the AFTE Theory of Identification.

REPRODUCIBILITY (BULLETS)	Second Evaluation: Identification	Second Evaluation: Inconclusive	Second Evaluation: Elimination
First Evaluation: Identification	0.0%	94.7%	5.3%
First Evaluation: Inconclusive	0.7%	67.2%	32.1%
First Evaluation: Elimination	0.1%	60.8%	39.1%

Separating the data out shows a number of interesting points. For example, no false identifications were reproduced in the second examination, which could provide helpful information to inform considerations such as the importance of verification (preferably blind, as discussed in the recommendations section) as a quality control measure.

D. The Importance of High-Quality Standards Development in Reducing Variability and Strengthening Evidence-Base of FATM Methodology

The OSAC Registry is a repository of selected published and proposed standards for forensic science. These documents contain minimum requirements, best practices, standard protocols, terminology, or other information to promote valid, reliable, and reproducible forensic results. The standards on the Registry have undergone a technical and quality review process that actively encourages feedback from forensic science practitioners, research scientists, human factors experts, statisticians, legal experts, and the public.

The Firearms and Toolmarks Subcommittee of OSAC has written several standards and best practice recommendations. These documents are at various stages of development and may be found at the subcommittee's website here: https://www.nist.gov/organization-scientific-area-committees-forensic-science/firearms-toolmarks-subcommittee.

[100] *Id.*

The development and use of standard methods both within a laboratory and across laboratories is an important step toward reducing variability and allowing for aggregate measures of performance to be represented as generalized measurements of performance, which is something that the results of current black box studies are currently unable to do for the reasons previously discussed. A review of documents across OSAC subcommittees demonstrates that some subcommittee standards contain far greater detail and specificity than others. The Commission urges the Forensic Science Standards Board (FSSB), which is responsible for approving standards for the OSAC Registry, to insist on clearly defined standards that will in turn *be useful to* laboratories in developing clearly defined protocols. The OSAC "Mandatory Requirements for Standards Development" document should be strictly adhered to if high-quality standards are a priority of OSAC. When standards are so vague as to capture any and all comparison approaches, they do not actually help practitioners establish method conformance, which means the FATM community also loses the opportunity to establish evidence-base (defined as empirical data reflecting the performance of the method under varying conditions.)[101]

In sum, method standardization "strengthens the evidence-base supporting the validation of those methods and reduces the resource burdens that would otherwise be placed on individual laboratories to accomplish the studies independently."[102] This is an especially powerful concept in Texas which has 23 accredited firearm and toolmark laboratories represented by city, county, private and state laboratories (8 of the 23 are DPS regional laboratories). The ability to share method validation data would be especially helpful but can only be done if methods are sufficiently specific and well-defined to be implemented successfully across laboratories.

[101] Swofford *et al*. *supra* n. 33, p. 8, n. 16.
[102] Swofford *et al*. *supra* n. 33, p. 8.

E. Gathering Data for the Discriminability and Reproducibility Tables

1. Blind Quality Control

One question that may arise as laboratories consider how they would produce method discriminability and reproducibility tables in their reporting is whether the data collected for the (2x3) discriminability table must be the result of a blind quality control program such as the one conducted by HFSC. Before answering this question, the Commission briefly addresses the benefits and obstacles to the integration of blind quality control programs in crime laboratories in Texas and nationwide. "Blind proficiency tests involve samples that are submitted through the normal analysis pipeline as if it were real casework. In blind testing, the examiners conduct the analysis under the assumption that they are working on real samples. Only after the work is complete do they learn that a case was a proficiency test."[103]

The 2009 NAS report recommended that forensic proficiency testing programs include blind tests where appropriate. In 2016, the PCAST report issued an even stronger recommendation:

> PCAST believes that test-blind proficiency testing of forensic examiners should be vigorously pursued, with the expectation that it should be in wide use, at least in large laboratories, within the next five years.[104]

PCAST's expectation of widespread test-blind quality control programs within five years has not come to pass, in part because the federal government has put few resources toward the initiative. In Texas, both HFSC and HCIFS have blind quality control programs, with additional efforts underway at Texas DPS and the Jefferson County Regional Crime Laboratory. HFSC began its program in 2015 and currently supports the most robust program in a non-federal forensic laboratory, with blind testing operational in the following divisions: biology, digital forensics,

---

[103] Mejia, R., Cuellar, M., Salyards, J., *Implementing blind proficiency testing in forensic laboratories: Motivation, obstacles, and recommendations,* Forensic Science International: Synergy 2 293-298 (2020). https://doi.org/10.1016/j.fsisyn.2020.09.002.

[104] PCAST report *supra* n. 46 at p.59.

forensic multimedia, latent print comparison, latent print processing, firearms, toxicology, and seized drugs. The main challenge with widespread adoption of blind quality control programs in state, county and city laboratories is not a lack of desire; it is a lack of resources. The logistical requirements for a successful blind quality control program are extensive. Many laboratories lack enough quality division staff to sustain a successful program. The Commission encourages increased funding to expand opportunities for blind quality control programs across Texas and nationwide.

Returning to the question of data collection for method performance tables in FATM, though it is by far the best approach, the Commission does not take the position that laboratories *must* have a FATM blind testing program to construct and report the (2x3) discriminability and reproducibility tables, because existing resources would make this impossible for all but HFSC. However, any data reported in the discriminability and reproducibility tables must be representative of the methodology used in casework.

2. Limitations on Using PT Monitoring Data for Discriminability and Reproducibility Tables

Historical criticism of using external proficiency testing results as a measure of performance is that: (1) they do not sufficiently test the limits of the methodology utilized in the laboratory (*i.e.,* they are too easy compared to actual casework); (2) they do not necessarily reflect the laboratory's test method (analysts are forced to adjust their approach to meet the particulars of the provider's testing); and (3) analysts behave differently when administered external proficiency tests because they know they are being tested, and they know a non-consensus PT result may be viewed as a non-conformity by the laboratory's management system.

Notwithstanding common historical perceptions that proficiency testing is "easy," CTS reported the test presented in the (2x3) discriminability table above (CTS 23-5262) as "more

challenging than originally intended." Regardless of what was intended, reporting these data (as limited as they are) in a (2x3) discriminability table provides legal stakeholders with some insight into possibly challenging scenarios, such as comparisons between recovered bullets from different firearms of the same brand (depending on the brand), or between firearms utilizing certain manufacturing processes, using currently available test methods. Information regarding method performance is critically important to the ability of attorneys and gatekeepers to do their jobs properly within the adversarial system.

The Commission reiterates that external proficiency testing data would not be useful for a (2x3) discriminability or (3x3) reproducibility table if it does not reflect the same methodology (decision thresholds or other factors) that examiners use in real casework. A laboratory should also have more data for its tables than what is generated by CTS. The data should come from the laboratory's method validation, which should represent a mix of case difficulty and sample types seen in casework.[105]

VIII. TESTIMONY AND REPORTING

Complainant asks the Commission to "set appropriate limitations" on FATM testimony. The Commission understands this request stems from observations that, historically, some FATM testimony has overstated the strength of the evidence whenever it (wrongly) implied uniqueness, or that the examiner's expert opinion derived from a comparison of the toolmark(s) to all other toolmarks in the world, as no such database currently exists. Forensic science best serves the legal system in cases involving firearm/toolmark evidence when the examiners supply the best available

[105] Ideally, the data reported in (2x3) discriminability and (3x3) reproducibility tables for a particular case should be limited to circumstances that are most like the case at hand (*e.g.,* if the case is "complex" or "difficult" then the data in the 2x3 and 3x3 tables should be based on similar quality evidence rather than just "easy" or more straightforward comparisons). If this level of detail is not practical, the laboratory should at least notate the difficulty/complexity mix of cases represented in the tables.

scientifically justified information in a manner that successfully conveys the expert's understanding of the evidence.[106] The Commission observes the following regarding the current state of FATM testimony guidance, and notes that this same observation applies to many forensic disciplines.[107]

A. Conclusion-Based vs. Strength of Evidence-Based Reporting

There are two main ways to report forensic results for comparison disciplines: conclusion-based and evidence-based. In a conclusion-based scale such as the AFTE Range of Conclusions, examiners report one of three categorical decisions when comparing an unknown to a known: identification, inconclusive, or elimination. Sometimes the inconclusive category is further broken down into subcategories, such as permitted in the AFTE Range of Conclusions. To the Commission's knowledge, all FATM sections in Texas accredited laboratories currently use the categorical approach.

Many experts have urged forensic examiners to move away from conclusion-centric reporting toward an evidence-based reporting approach, which has been adopted by DNA laboratories that have validated probabilistic genotyping software. The software generates a likelihood ratio, which describes the probability of the observed scientific findings given two mutually exclusive propositions. As noted by Kaye et al., this approach "has won widespread

[106] Kaye et al., *Toolmark-Comparison Testimony: A Report to the Texas Forensic Science Commission*, (May 2022).
[107] The Commission is grateful to David Kaye and the Forensic Science Standards Practicum at Yale University Law School, students of which authored the attached *Toolmark-Comparison Testimony: A Report to the Texas Forensic Science Commission* (May 2022). (**Exhibit D**). It contains a comprehensive discussion on the following subjects: testimony limitations imposed by courts; voluntary standards governing toolmark comparisons and testimony; inconclusives and a more graduated reporting scale; and possible modes of testimony.

endorsement from statistical[108] and scientific or laboratory associations[109] and agencies[110] as well as from scholars of law and statistics."[111] The Commission notes that the OSAC is in the early stages of providing guidance to subcommittees (including FATM) regarding interpretation scales. This initiative was recently generated by the FSSB upon the realization that subcommittees of OSAC were proposing interpretation standards that had lacked consistency in approach—some are conclusion-based, others are evidence-based and still others combine the two approaches (indeed, this is the case for the FATM subcommittee's ASB *Standard Scale of Source Conclusions and*

[108] *Am. Stat. Ass'n Position on Statistical Statements for Forensic Evidence*, Am. Stat. Ass'n 1, 2-4 (Jan. 2, 2019), https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf : To evaluate the weight of any set of observations made on questioned and control samples, it is necessary to relate the probability of making these observations if the samples came from the same source to the probability of making these observations if the questioned sample came from another source in a relevant population of potential sources. . . . We . . . strongly advise forensic science practitioners to confine their evaluative statements to expressions of support for stated hypotheses: *e.g.*, the support for the hypothesis that the samples originate from a common source and support for the hypothesis that they originate from different sources.

[109] Colin Aitken et al., *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses* (2010), http://www.rss.org.uk/Images/PDF/influencing-change/rss-fundamentals-probability-statistical-evidence.pdf (committee of the Royal Statistical Society); *Ass'n of Forensic Sci. Providers, Standards for the Formulation of Evaluative Forensic Science Expert Opinion,* 49 Sci. & Just. 161 (2009); *Eur. Network of Forensic Sci. Insts., ENFSI Guideline for Evaluative Reporting in Forensic Science* 10 (2015), http://enfsi.eu/wp-content/uploads/2016/09/m1_guide line.pdf ("Evaluative reports should address the probability of the findings given the propositions and relevant background information and not the probability of the propositions given the findings and background information."); cf. Royal Society, *Forensic DNA Analysis: A Primer for Courts* 36 (2017) ("Likelihood ratios are generally accepted as being the most appropriate method for evaluating the evidential strength of DNA profiles.").

[110] Subcomm. on Reporting and Testifying of the National Commission on Forensic Science. Nat'l Comm'n on Forensic Sci., *Views of the Commission: Statistical Statements in Forensic Testimony*, U.S. Dep't Justice (Feb. 9, 2017), https://www.justice.gov/archives/ncfs/page/file/965931/download Forensic science practitioners should confine their evaluative statements to the support that the findings provide for the claim linked to the forensic evidence."); Nat'l Inst. of Forensic Sci. Austl. N.Z., *An Introductory Guide to Evaluative Reporting* 6 (2017), available at https://www.anzpaa.org.au/forensic-science/ourwork/products/publications: The fundamental principles of evaluative reporting or interpretation are . . . (iii) that the role of the expert is to comment on the probability of their findings, given these propositions and not on the propositions themselves. It is this last principle that allows the factfinders to combine aspects of evidence they hear during the course of the trial with their judgement in their deliberations. This framework of evidence evaluation is commonly referred to as evaluative reporting but may also be referred to as the likelihood ratio approach, logical thinking, or Bayesian inference.

[111] *E.g.*, Edward K. Cheng, *The Burden of Proof and the Presentation of Forensic Results*, 130 Harv. L. Rev. F. 154, 161-62 (2017) ("Scholars have long argued in favor of presenting forensic results using likelihood ratios, and indeed some forensic communities in Europe have embraced them . . .. The key is that likelihood ratios present a clear path to improving the use of forensics testimony in court.") (footnotes omitted); Colin G.G Aitken & 30 coauthors, *Expressing Evaluative Opinions: A Position Statement*, 51 Sci. & Just. 1 (2011), http://dx.doi.org/10.1016/j.scijus.2011.01.002.

Criteria for Toolmark Examiners).[112] The creation of FSSB guidance to subcommittees on interpretation scales will take time and will benefit from the active engagement of practitioners, statisticians, human factors experts and other stakeholders.

One challenge is that when no probabilistic software exists to calculate a likelihood ratio (such as in DNA), an evidence-based reporting approach requires one of two things from forensic practitioners: the manual generation of a quantitative (numerical) likelihood ratio or the generation of a qualitative probability statement such as, "[I]t is far more probable that this degree of similarity in features would occur when comparing [the questioned impressions] with the defendant's [tool] than with [some other tool]."[113] As noted by Kaye et al.:

> Strength-of-evidence testimony does not require experts to draw a sharp line between the overall similarity of paired samples that establishes (in the mind of the examiner) that the pair originated from the same tool. Both qualitative and quantitative likelihood ratios range from marking evidence as highly supportive for one source hypothesis to depicting evidence as highly supportive of the alternative source hypothesis.

However, as also noted by Kaye et al., "[l]imited psychological research on such scales has been done to investigate how forensic science practitioners understand terms such as 'moderate support' and 'strong support,'[114] and how lay individuals use them."[115]

---

[112] *See e.g.*, Patteet, J., Champod, C., *Striated toolmarks comparison and reporting methods: Review and perspectives, Forensic Science International*, 357 111997 (2024), https://doi.org/10.1016/j.forsciint.2024.111997 (encouraging the adoption of a LR-based probabilistic framework for reporting forensic findings).

[113] Kaye et al., citing: NIST Expert Working Group on Human Factors in Latent Print Analysis, *Latent Print Examination and Human Factors: Improving the Practice Through a Systems Approach* 134 (David H. Kaye ed. 2012); cf. David H. Kaye, *Likelihoodism, Bayesianism, and a Pair of Shoes*, 53 Jurimetrics J. 1 (2012) (discussing footwear-impression testimony).

[114] Kaye et al., citing: Thomas Busey et al., *Validating Strength-of-support Conclusion Scales for Fingerprint, Footwear, and Toolmark Impressions*, J. Forensic Sci. (2022), https://doi.org/10.1111/1556-4029.15019; Elmarije K.van Straalenab et al., *The Interpretation of Forensic Conclusions by Criminal Justice Professionals: The Same Evidence Interpreted Differently*, 313 Forensic Sci. Int'l (2020).

[115] Kaye et al., citing Eleanor Arscott et al., *Understanding Forensic Expert Evaluative Evidence: A Study of the Perception of Verbal Expressions of the Strength of Evidence*, 57 Sci. & Just. 221 (2017); Kristy A. Martire & Gary Edmund, *How Well Do Lay People Comprehend Statistical Statements from Forensic Scientists*, in Handbook of Forensic Statistics 201 (David Banks et al. 2021); Kristy A. Martire & Ian Watkins, *Perception Problems of the Verbal Scale: A Reanalysis and Application of a Membership Function Approach*, 44 Sci. & Just. 264 (2015); Kristy

The Commission encourages the FSSB to pursue its current initiative to provide guidance to all forensic disciplines of the OSAC regarding interpretation scales. When that initiative is complete, the results should be implemented by subcommittees in their interpretation documents, such as the Firearms and Toolmarks Subcommittee ASB Document entitled, "*Standard Scale of Source Conclusions and Criteria for Toolmark Examinations*."[116]

While the Commission generally supports the move from conclusion-centric to evidence-centric reporting, we reserve a specific recommendation on this subject until such time as the OSAC has completed the work described above. In the meantime, the Commission recommends all Texas crime laboratories currently using some form of the AFTE Range of Conclusions (which is everyone) include the following in their reporting:

1. A clear description of how each reporting category (Same-source opinion; Inconclusive opinion (especially sub-categories of Inconclusive where used); Elimination opinion) is defined under the laboratory's protocol.

2. A clear statement that when providing a same-source opinion, the examiner has observed agreement of all discernible class characteristics and a sufficient correspondence of accidental characteristics such that both unknown and reference items lead to an examiner's decision of having originated from the same source. The reporting and testimony should be clear that "sufficient correspondence" is not strictly defined but rather comprises a combination of characteristics that in the opinion of the examiner, lead to a decision of same source.

3. A clear statement that a same-source opinion does not imply uniqueness, *i.e.*, it is not a statement that other sources could not have similar accidental features.

4. A clear statement of whether the laboratory's protocol incorporates subcategories of Inconclusive. (Note: an examiner's opinion about subcategories of inconclusive should

A Martire et al., *On the Interpretation of Likelihood Ratios in Forensic Science Evidence: Presentation Formats and the Weak Evidence Effect*, 240 Forensic Sci. Int'l 61 (2014); Kristy A. Martire et al., *The Expression and Interpretation of Uncertain Forensic Science Evidence: Verbal Equivalence, Evidence Strength, and the Weak Evidence Effect*, 37 Law & Hum. Behav. 187 (2013); W.C. Thompson et al., *Perceived Strength of Forensic Scientists' Reporting Statements About Source Conclusions*, 17 Law, Probability & Risk 133 (2018), http://doi.org/10.1093/lpr/mgy012; William C. Thompson & Eryn J. Newman, *Lay Understanding of Forensic Statistics: Evaluation of Random Match Probabilities, Likelihood Ratios, and Verbal Equivalents*, 39 Law & Hum. Behav. 332 (2015).

[116] This document is not yet on the OSAC Registry of Standards but is at the Academy Standards Board still under development.

be transparent in the record. An Inconclusive C opinion may be closer to Elimination than it is to Inconclusive A and thus may have a different (and potentially exculpatory) meaning for attorneys and the trier of fact than an Inconclusive A opinion.)[117]

5. Once developed, a (2x3) discriminability table and (3x3) reproducibility table with laboratory method performance data as described in this report (reflecting the discriminability of the method and reproducibility of decisions among analysts when the method is applied). (*See* Recommendations section for further guidance.)

Laboratory reporting and testimony in FATM should *refrain from*: (1) reporting black box study error rates as a measure of accuracy in the case (rather method performance will be addressed in the (2x3) discriminability table and (3x3) reproducibility tables when developed); (2) citing the number of examinations conducted by the examiner in his or her career as a direct measure for the accuracy of a conclusion provided; or (3) asserting that two toolmarks originated from the same source with absolute or 100% certainty, or use the expressions 'reasonable degree of scientific certainty,'[118] or "practical impossibility."[119]

IX. EMERGING TECHNOLOGY: VIRTUAL COMPARISON MICROSCOPY

The field of firearms identification is undergoing a change in technology and capability with the introduction of optical topography. This technology provides a three-dimensional view of the surface of a bullet or cartridge case at resolutions that capture the full range of subclass and individual characteristics. The technology offers an additional method to the comparison microscope for one-to-one firearm comparisons.[120]

[117] We emphasize that the table may be 2x5, 2x7 or 2x9 depending on the sub-categories of inconclusive used in the laboratory's protocol. A laboratory that incorporates Inconclusive A-C may have a 2x5 table.
[118] See, *United States Department of Justice Uniform Language for Testimony and Reporting for the Forensic Firearms/Toolmarks Discipline Pattern Examination* (effective 5.18.23)
[119] This is one of the subjects currently being evaluated by NIST. While the NIST Scientific Foundation Review is not yet published, the agency has made clear in public comments that it will recommend the community move away from this terminology as it is unsupported by existing published data.
[120] *Forensic Optical Topography: A Landscape Study* (NIJ 2016).

A. Potential Advantages of VCM Technology

A key advantage of VCM imaging is that it allows high-definition scans of the actual surface topography of a sample with high repeatability.[121] Although VCM images are now routinely used to rank samples in a database against crime scene evidence, VCM imaging has not yet replaced the use of traditional comparison microscopy for assessing whether two samples may have been fired from the same firearm. Compared to traditional images, these VCM data are more repeatable and not sensitive to lighting conditions and can be compared by both algorithms and examiners.[122]

Optical topography may also address some of the limitations inherent in traditional approaches, such as depth-of-focus, specular reflection, and lack of three-dimensional data. The instrument can be used in several ways: to build and search a reference database, to serve as a compliment to the comparison microscope, to supplement image data taken from the comparison microscope, to make comparison decisions, and to clarify the basis on which an examiner has made a particular comparison decision.[123]

VCM also allows for the creation of digital models and documentation, which can be stored, shared, and reviewed by experts and stakeholders. Working with virtual scans allows a second examiner to independently verify the conclusions of the original examiner (at least with some currently available vendor software). It is also possible to hide the work of the first examiner and more efficiently complete the verification process blind.

[121] *Id.*
[122] *A Century of Ballistics Comparison Giving Way to Virtual 3D Methods* (NIJ 2022).
[123] *See*, *supra* n. 109.

B. Barriers to Implementation

There are barriers to the implementation of VCM technology. The equipment is expensive, and training is mostly limited to that provided by the manufacturers. The laboratory must put significant resources into developing a plan for deployment, validation, and quality control. Protocols must also be created to delineate when optical topography is to be used and the procedures for each application.[124] An appropriate validation study must also be conducted prior to use in casework to determine baseline capacity, laboratory accuracy, and examiner proficiency.[125]

X. LIMITATIONS OF REPORTING IN HANDHELD TOOL ANALYSIS

Common tools like screwdrivers, chisels, crowbars, etc. may be used during the commission of a crime, and striated toolmarks are frequently found at crime scenes. If a suspect tool is available, the question arises whether that particular tool was used to create the toolmarks. To answer this question, the forensic examiner generates experimental toolmarks with the suspect tool in the laboratory and compares them with the marks found at the crime scene, using a comparison microscope.[126]

Manual toolmark comparison necessarily includes subjective judgments. Various factors can have a major impact on toolmark impressions (the angle of attack of a tool, the substrate material, the depth of a toolmark in the substrate material, and the axial rotation of the tool). The parameter that has by far the most prominent influence on the variability of a toolmark, is the angle of attack.[127] For these reasons, inconclusive decisions in handheld tool comparisons are common.

---

[124] *See*, *supra* n. 109.

[125] *Id*.

[126] Baiker, M., Pieterman, R., Zoon, P., *Toolmark variability and quality depending on fundamental parameters: Angle of Attack, toolmark depth, and substrate material*, Forensic Science International 251; 40-49 (2015).

[127] *Id.*

Though the general concept of toolmark comparison applies to both firearm analysis and handheld toolmark analysis, the factors described above highlight differences in application. Data on method performance that applies to firearms (such as the information from black box studies, or other firearms performance studies) cannot be extrapolated as metrics for handheld toolmark performance. Indeed, perhaps because it is far less common than firearm toolmark comparison,[128] little data exists from black box studies or proficiency testing to establish performance metrics for this assay. Researchers are developing methods to objectively compare toolmarks and thus improve the reliability of toolmark comparisons. One such example is reflected in an article published by Cuellar et al.[129] The authors posit that they have developed an objective method to perform forensic toolmark comparisons that present results using likelihood ratios and address common problems with factors such as the angle of attack and direction of the mark.

The Commission encourages FATM examiners in Texas to explore the use of methods such as the one described by Cuellar et al. To the extent examiners continue to use methods of comparison and reporting using a traditional (conclusion-based) scale, they should—at a minimum—conform to the guidance on testimony for FATM described in this report. Because of additional complicating factors such as angle of attack and direction of mark, it may be challenging to create (2x3) discriminability and (3x3) reproducibility tables for the analysis. Examiners who issue opinions regarding handheld tools should be particularly cautious to not imply uniqueness in their testimony, or that different tools are incapable of having similar accidental characteristics.

[128] One Texas DPS regional laboratory estimated 12 total requests per year, which pales in comparison to their volume of firearms casework.

[129] Cuellar, M., Gao, S., Hofmann, H., *Revolutionizing Forensic Toolmark Analysis: An Objective and Transparent Comparison Algorithm,* (2023). https://doi.org/10.48550/arXiv.2312.00032.

XI. COMMISSION RECOMMENDATIONS

The Commission makes the following recommendations:

1. A statewide FATM task group shall be established (led by Commission staff, HFSC and DPS but to include FATM practitioners from a range of Texas laboratories, academic members with subject matter expertise (*e.g.*, Commissioner Patrick Buzzini), lawyers (representing defense and prosecution) and author(s) of the Swofford et al. paper, *Inconclusive Decisions and Error Rates in Forensic Science).*

   The purpose of the group will be to provide feedback to the Commission on benefits and barriers to the method performance/method conformance reporting suggestions set forth by Swofford et al., among other issues. The work of the task group may include the following:

   - Developing a plan (with suggested timeline) for implementing (2x3) discriminability and (3x3) reproducibility table reporting in Texas FATM sections including identifying areas requiring clarification and possible barriers to implementation;

   - Creating a path for expansion of HFSC blind quality control program to other Texas FATM sections;

   - Suggesting model reporting and testimony guidance for Texas;

   - Suggesting ways to facilitate OSAC Registry standards implementation with an eye toward improving method standardization;

   - Assessing the "objective" toolmark method proposed by Cuellar et al. and a path (including timeline) for implementation if possible;

   - Other related recommendations as the group sees fit.[130]

2. FATM laboratories in Texas should implement the recommendations set forth in Section VIII of this report with respect to "Reporting and Testimony" pending the results of the work by the task group referenced above. They include:

   - Providing a clear statement of how each reporting opinion category (Same source; Inconclusive (A-C, if applicable); Elimination) is defined under the laboratory's protocol.

[130] The Commission does not discuss "trigger pull" analysis in this report but notes that HFSC recently paused trigger pull analysis to assess concerns regarding potential misunderstanding of significance by legal stakeholders. This is the type of issue the working group may consider (and make recommendations to the Commission on).

- Explaining that when issuing a same-source opinion, the examiner has observed agreement of all discernible class characteristics and a sufficient correspondence of accidental characteristics such that both unknown and reference items lead to an examiner's decision of having originated from the same source. The reporting and testimony should be clear that "sufficient correspondence" is not strictly defined but rather comprises a combination of characteristics that in the opinion of the examiner, lead to a decision of same source.

- Providing a clear statement that a same source opinion does not imply uniqueness, *i.e.*, it is not a statement that other sources could not have similar accidental features.

- Providing a clear statement when the laboratory's protocol incorporates subcategories of Inconclusive (A-C, or some other variation). Legal end-users need to understand the difference between Inconclusive A and Inconclusive C given their definitions in AFTE documents. An Inconclusive C opinion may be closer to Elimination than it is to Inconclusive A and thus may have a different (and potentially exculpatory) meaning for attorneys and the trier of fact than an Inconclusive A opinion.

- Providing (2x3) discriminability[131] and (3x3) reproducibility tables with method performance data (reflecting both discriminability of the method and reproducibility of decisions among analysts when the method is applied) as soon as it is available per recommendation #1 above.

3. Laboratory reporting and testimony in FATM should ***not***:

    - Report black box study error rates as a measure of accuracy in the specific case (rather method performance is addressed in the (2x3) discriminability and (3x3) reproducibility tables, reflecting the discriminability and reproducibility data for the method applied);

    - Cite the number of examinations conducted by the examiner in his or her career as a direct measure for the accuracy of a conclusion provided;[132] or

    - Assert that two toolmarks originated from the same source with absolute or 100% certainty, or use the expression "reasonable degree of scientific certainty";

[131] We emphasize that the table may be 2x5, 2x7 or 2x9 depending on the sub-categories of inconclusives used in the laboratory's protocol. A laboratory that incorporates Inconclusive A-C may have a 2x5 table.

[132] It is also important to refrain from statements predicating conclusions on examiner experience. Research (including but not limited to FBI/Ames II and the HFSC Blind Study) shows that experience-related variables (such as certification status and years of experience) do not corollate with performance.

- Notwithstanding AFTE guidance on this subject, use the term "practical impossibility" due to its potential to confuse criminal justice end-users;

4. Records should be created concurrently during the examination of evidence and during technical review that would allow another analyst with proper training and experience to understand and evaluate the work performed and to independently analyze and interpret the data and draw conclusions. Additionally, documentation should be sufficient to clearly demonstrate the analyst followed the protocols of the method utilized. To aid in this practice, FSSPs should develop a policy that includes a combination of photographic[133] and descriptive note-based documentation of comparison conclusions, documenting features observed in the questioned evidence first, *before* viewing and documenting features in the known sample, followed by comparing the samples and documenting features relied on to reach the opinion.[134]

5. Laboratories should seek funding (and the federal government should make available significant funding) to incorporate VCM as soon as practicable.

6. Laboratories should seek funding (and the federal government should make available significant funding) to create meaningful blind quality control programs with the support of NIST (or other federal resources) as soon as practicable.

The following items will be added to the ANAB checklist of TFSC accreditation requirements:

7. All FATM laboratories shall have a policy providing for some level of blind verification with a clear minimum number of verifications set forth in policy. While blind verification may not be necessary (or an efficient use of resources) in each and every case, the blind verification policy should cater to potentially problematic comparisons. To be effective, the policy should not be reserved for identification conclusions only, but rather include other conclusion categories (inconclusive or exclusion).

   The statewide FATM task group will review and recommend approval (or suggest revisions) to blind verification plans developed by Texas laboratories, metrics for which may vary depending on size and resources of the laboratory. Approved plans would then be used as evidence of conformance for ANAB and A2LA assessors. The group should provide a suggested timeline for this work by the Commission's October

[133] The Commission understands that two-dimensional photographs alone may not capture all data observed under the CM and thus examiners must use their professional judgment to determine how to best represent linear decision-making in the case record. Specialized software may assist examiners in this process.

[134] The Commission understands the FATM community feels their ability to apply a GYRO-type method of documenting features (such as that utilized by the friction ridge community) is limited by available technology and certain inherent differences between the disciplines. We recognize there may be some limitations that apply; however, linear sequential unmasking is an important goal to work toward in all forensic comparative disciplines and the FATM community should endeavor to implement LSU as soon as practicable.

2024 quarterly meeting. Based on the group's recommendation, the Commission will inform ANAB and A2LA when to begin assessing to this requirement.

8. All FATM laboratories shall have a policy on the documentation of consultations occurring during casework. The statewide FATM task group should consider and adopt a recommended definition of "consultation" by the Commission's October 2024 quarterly meeting. For purposes of generating discussion, a starting point for a possible definition is provided here: "A consultation is an evaluation by a second qualified examiner of specific items or data to assist examiners in developing a mutually agreeable opinion or interpretation[135] of that item or data."

[135] The rationale for not including disagreements here is because accreditation standards already require disagreements to be documented per ANSI National Accreditation Board Accreditation Requirements for Forensic Testing and Calibration 7.5.1.5 (Document Number AR3125, February 1, 2023), (If an observation, data, or calculation is rejected, the reason, the identity of the individual(s) taking the action and the date shall be recorded in the technical record*)*. Because this definition of consultation was offered to the Commission by ANAB as a starting point for discussion, Commission staff will include ANAB (and A2LA) in discussions about how to modify it to suit FSSP needs.

EXHIBIT A

INNOCENCE PROJECT

Executive Director
Christina Swarns

Co-Founders & Special Counsel
Barry C. Scheck, Esq.
Peter J. Neufeld, Esq.

October 6, 2021

By electronic mail

Texas Forensic Science Commission
1700 North Congress Avenue, Suite 445
Austin, Texas 78701

Dear Commissioners:

Please accept this complaint, filed on behalf of our client, Nanon McKewn Williams, and on behalf of the Innocence Project, Inc. We ask that the Texas Forensic Science Commission ("the Commission") exercise its statutory mandate specifically to investigate and report on the firearms evidence used to convict Mr. Williams 25 years ago, and, more broadly, to investigate and report "the integrity and reliability" of toolmark and firearms analysis ("Firearm Toolmark Evidence" or "FTE") as used in criminal proceedings. Tex. Crim. Proc. Code Ann. § art. 38.01(4)(b-1)(1).[1]

The Innocence Project is a national litigation and public policy organization dedicated to exonerating wrongfully convicted persons through DNA testing and improving the criminal justice system to prevent future miscarriages of justice. To date, 375 people in the United States, including 20 who served time on death row, have been exonerated by DNA testing. One lesson to be drawn from these exonerations is that the misapplication of forensic sciences is one of the leading causes of wrongful conviction, contributing to the original wrongful conviction in nearly half of the DNA exoneration cases. As this complaint outlines, no published peer reviewed data, and no valid proficiency testing, supports FTE evidence. This, along with the fact that FTE is entirely subjective, greatly increases the risk of wrongful conviction.

Given the lack of published data supporting FTE and the growing concerns from both courts and the scientific community about the unsubstantiated claims and exaggerated testimony made by FTE practitioners in criminal trials, FTE represents an ideal and critical opportunity for this Commission to bring to bear its statutory mandate to "advance the integrity and reliability of forensic science" in Texas. *See* Tex. Crim. Proc. Code Ann. § art. 38.01(4)(a-1). We thus ask that, in addition to examining the evidence used to convict Mr. Williams, this Commission undertake a thorough investigation of all toolmark assays, including both FTE and handheld toolmark evidence. Our request is that this Commission set appropriate limits on the conclusions of FTE examiners in traditional bullet-to-firearm matching testimony, and to determine what conclusions, if any, can be proffered in other toolmark

[1] Firearms/toolmarks is enumerated as an accredited field of forensic science. *See* 37 Tex. Admin. Code § 28.145, which may thus be conducted out of an accredited laboratory, giving the Commission additional jurisdiction. *See* Tex. Crim. Proc. Code Ann. § art. 38.01(4)(a)(3).

assays. Doing so will not only advance this body's statutory mission, but also help ensure that no innocent Texans are incarcerated as a result of unreliable evidence and overstated testimonial conclusions by toolmark examiners. In addition, by providing guidance to the forensic community the Commission will assist Texas forensic analysts, forensic technicians, and crime laboratory management "guard[] against the use of non-valid methods in casework, the misapplication of validated methods or improper testimony regarding a particular analytical method or result." Tex. Admin. Code Ann Title 37.15(c)(2).

MR. NANON MCKEWN WILLIAMS: AMENDED COMPLAINT

Errors in Mr. Williams' trial are indicative of larger problems in the Firearm Toolmark Field

The recent loss of ANSI accreditation by the Washington D.C. crime lab has parallels to the scandal that led to the shuttering of the Houston Police Department Crime Lab, the crime lab responsible for the evidence used to convict Mr. Williams.[2] Earlier in 2021, ANAB (ANSI National Accreditation Board) received "credible evidence" of the "concealment of evidence," "misrepresentations," and "fraudulent behavior" during an ANAB audit of the Firearms Examination Unit ("FEU") of the Forensic Science Laboratory Division within the Department of Forensic Sciences ("DFS") for the District of Columbia.[3] Detailed in ANAB's forthcoming Final Report, this evidence led to an April 2nd letter to Dr. Jennifer Smith, Director of the DFS, that placed the accreditation of FEU on suspension for 30 days—the first step within the removal process of the lab's ANSI accreditation.[4] Now finalized, this discreditation led to Dr. Smith's late-May resignation from DFS as D.C. began attempts to purge the "substantial issues" in methodology and misconduct that led to the lab's discreditation.[5]

Mr. Williams was convicted upon evidence proffered by similarly egregious oversight and abuse inside the now-defunct Houston Police Department Crime Lab. Detailed below is Mr. Williams' story—an accounting of the Houston PD Crime Lab's blatant failure to adequately test firearm toolmark and ballistics evidence, a trial that hinged upon over-stated firearms and ballistics testimony, and the unjustly prejudicial impact this evidence had on Mr. Williams' conviction. Following Mr. Williams' story is outline of the scientific evidence supporting our request for a broader investigation into the scientific underpinning of firearms and toolmark evidence. The evidence discussed in that section demonstrates that a case analogous to Mr. Williams's would today face the combination of myriad evidence detailing the lack of data supporting FTE with the resulting shifts in both special instructions from the Court and the admissibility, or inadmissibility, of expert testimony on the subject. Such obstacles for the State would drastically change the trial landscape of a defendant fighting accusations similar to those for which Mr. Williams remains incarcerated.

Nanon Williams was Wrongfully Convicted Based on Egregiously Faulty Firearm Toolmark and Ballistics Testimony

Mr. Williams was imprisoned almost 30 years ago, wrongfully convicted of capital murder in the death of Adonius Collier in Houston, TX, on May 13, 1992. Mr. Williams' trial was riddled with

---

[2] American National Standards Inc. (ANSI) is an independent national accreditation corporation long recognized for its role in government oversight. The corporation's website is http://www.ansi.org.

[3] Letter from Pamela L. Sale, VP of Forensics, ANSI National Accreditation Board, to Dr. Jennifer Smith, Director (Former), D.C. Dept. of Forensics (April 2, 2021).

[4] *Id.*

[5] Ryan Sprouse, Nick Boykin, et al., *Director of DC's Department of Forensic Sciences Resigns Amid District's Crime Lab Losing Accreditation*, WUSA9 (May 20, 2021), *available at* https://www.wusa9.com/article/news/local/dc/washington-dc-crime-lab-forensic-sciences-jenifer-smith/65-9f396f3f-18ee-449b-97f7-86c985608d59. *See also* Letter from Karl A. Racine, Attorney General, District of Columbia, to Daniel W. Lucas, Inspector General, District of Columbia (March 22, 2021). This letter briefly details the full timeline of the D.C. Crime Lab saga, beginning with a USAO case audit.

constitutional defects, including faulty ballistics evidence that was deeply prejudicial and in violation of his right to a fair trial.

In May of 1992, Adonius Collier was first shot by a .22-caliber pistol and then by a 12-gauge shotgun. Later arrested for the crime were 17-year-old Nanon Williams and his co-defendant, 21-year-old Vaal Guevara. Mr. Guevara quickly started cooperating with the police. His self-serving testimony at trial placed the shotgun in Mr. Williams' hands. Mr. Guevara also initially avoided confessing that he carried a .22-caliber handgun that evening. Though originally unidentified on x-Ray, a .22-caliber slug was found during Mr. Collier's autopsy. This slug was misidentified by the prosecution's ballistics expert, Mr. Robert Baldwin, as a .25-caliber bullet and labeled "EB-1."[6]

At Mr. Williams' trial, Mr. Baldwin, a Houston police department criminalist, testified that EB-1 was fired from a .25-caliber pistol linked to Mr. Williams, not from the .22-caliber pistol that Vaal Guevara eventually admitted to carrying and firing during the incident that culminated in the death of Mr. Collier. Importantly, on cross-examination, Mr. Baldwin confessed to never test-firing either pistol. This would indeed have been impossible as Mr. Williams' .25-caliber pistol was not yet entered into evidence. Yet, Mr. Baldwin stood by his conclusions unequivocally.[7]

During subsequent state habeas proceedings, the Court ordered the release of Mr. Guevara's .22-caliber pistol and all other ballistics evidence. The Harris County District Attorney's office ordered the evidence re-tested, concluding that EB-1 was fired by Mr. Guevara's .22-caliber handgun and not Mr. Williams' .25-caliber pistol—directly contradicting Mr. Baldwin's highly prejudicial trial testimony. Mr. Williams then presented Mr. Baldwin's own recantation of his trial testimony and his updated conclusion that EB-1 came from Mr. Guevara's pistol. Accompanying this recantation was an affidavit from another criminologist, Mr. Ronald Singer, summarizing the findings of his independent ballistics review. Mr. Singer opined that Mr. Collier suffered two wounds to the head—one from EB-1, fired from a .22-caliber weapon—and one from a shotgun. Mr. Singer found that EB-1 was not fired from Mr. Williams' pistol.[8]

Mr. Singer further stated that the .22-caliber slug from Mr. Collier's skull was readily distinguishable from a .25-caliber round. Mr. Singer testified that Mr. Baldwin's testimony showed "at best…extreme carelessness on his part and at worst calls into question his expertise."[9] At this stage, Mr. Williams also presented affidavits from two jurors saying they would have reached a different verdict had the correct ballistics evidence been presented at trial. [10]

Mr. Williams' trial counsel also later stipulated to have been ineffective in her failure to consult either an independent ballistics expert or medical examiner, resulting instead in the admission of faulty

---

[6] *See Williams v. Thaler*, No. 10-20876, 2011 WL 2526559, at *11 (C.A. 5. June 17, 2011).

[7] *Williams v. Texas*, Cause No. 63442, Tr. at 25:7-26:19 (July 18, 1995). Mr. Baldwin's testimony may also be found excerpted in the Bromwich Report, *infra* note 13, at 222-25.

[8] *Ex Parte Nanon McKewn Williams*, No. 634442-A (Tex. Dist. 2001).

[9] *Williams v. Texas*, Post-Conviction Writ Hearing Trans, No. 634442-A (Tex. Dist. 2000), at 22:17-25:18. During this second State Post-Conviction hearing, Ron Singer was re-called and gave testimony summarizing the findings contained in his affidavit from two years earlier—it was from this affidavit to which Singer referred that this direct quote was taken. Singer's words can be found excerpted in the Bromwich Report, *infra* note 13, at 226-27.

[10] *Ex Parte Nanon McKewn Williams*, *supra* note 8, at 10.

evidence and the bolstering of false, self-interested testimony.[11] Mr. Williams' trial counsel never challenged Mr. Baldwin's original ballistics assessment nor did she seek an independent examination of the ballistics evidence. In fact, Mr. Williams' trial counsel ultimately admitted to her ineffective assistance after espousing her belief that further questioning the ballistics results would have led her to an exculpatory defense.[12]

This assertion is well-supported—Mr. Baldwin's unrebutted testimony bolstered Mr. Guevara's account that Mr. Williams lethally shot Mr. Collier, regardless of whether death resulted from the shotgun wound or from EB-1. This faulty line of reasoning strongly contributed to Mr. Williams' conviction for capital murder and the death sentence that followed. Mr. Williams remains convicted on indisputably faulty evidence and potentially faulty forensic science.

Finally, Vaal Guevara cooperated with law enforcement against Mr. Williams in exchange for dismissal of his own capital murder charge and a ten-year plea deal. Contrary to Mr. Guevara's perjured trial testimony, it was Mr. Guevara who carried the .22-caliber pistol that fired the .22-caliber slug found in Mr. Collier's skull during the autopsy. Combined with the faulty ballistics evidence, Mr. Guevara's false statements that Mr. Collier was shot twice by Mr. Williams—and never with the .22-caliber pistol—were catastrophic to Mr. Williams' defense. This combination of Mr. Guevara's perjured testimony with faulty and misleading forensic evidence, only exacerbated by unscientific over-statements by Mr. Baldwin, resulted in a constitutionally defective trial proceeding for Mr. Williams.

Evolving Practices and New Findings Allow Current Review of Forensic Evidence

In June 2005, after a years-long independent investigation into the Houston Police Department (HPD) Crime Laboratory's operating practices, Michael Bromwich published the "Bromwich Report," detailing a "crisis" within the HPD Crime Lab. The Report referred specifically to Mr. Williams' case and eventual 1995 conviction as an example of the failings of the now-defunct HPD Crime Lab's Firearms Division.[13] Among other problems, the investigation found that HPD Crime Lab policies provided inadequate oversight that ultimately led to the demonstrably false testimony of an HPD ballistics examiner during Mr. Williams' trial, identified above and within the report as Mr. Robert Baldwin. Mr. Baldwin was one of three HPD Crime Lab examiners who failed to correctly identify the source-weapon of a bullet fragment significant to Mr. Williams' conviction.[14]

The Report concluded that policies in place at the time of Mr. Williams' conviction provided insufficient oversight to correct the errors made by three of the lab's ballistics examiners during the investigation of Mr. Collier's murder. Describing a cursory review process requiring no independent verification by reviewing examiners, the Report eviscerates the rubber-stamping of FTE reviews of the

---

[11] *Williams v. Thaler*, 756 F. Supp. 2d 809, 819 (S.D. Tex. 2010), *rev'd*, 459 F. App'x 327 (5th Cir. 2012), *opinion withdrawn and superseded on reconsideration*, 684 F.3d 597 (5th Cir. 2012), and *rev'd* 684 F.3d 597 (5th Cir. 2012).

[12] *Williams v. Texas*, Post-Conviction Writ Hearing Trans, No. 634442-A (Tex. Dist. 1998), at 117:18, 23-24, 118:1-6.

[13] Michael H. Bromwich, Independent Investigator, Fried, Frank, Harris et. al. LLP, *Final Report of the Independent Investigator for the Houston Police Department Crime Laboratory and Property Room 4* (2007) (hereinafter "Bromwich Report").

[14] *Id.* at 14.

type characterized by Mr. Williams' case.[15] Cited as one of only four case studies within the Report, the inequities marking the Crime Lab's investigations during Mr. Williams' case were among the many motivating factors in the closure of the HPD Crime Lab and the inception of HCIFS—the independent crime lab now responsible for such forensic examination.

Not only was Mr. Williams' case an exemplar of forensic mismanagement and prejudice, but the field of firearm toolmark evidence—the field upon which his conviction turned—has now widely been criticized as *unscientific*, as detailed below.

Crime Laboratory Scandals across the Country Have Demonstrated the Unreliability of FTE

The mishandling of firearms toolmark evidence appears to be a problem of growing proportion, affecting vast numbers of suspects—many of whom are innocent. Numerous large-scale metropolitan crime laboratories have been shuttered based on their faulty forensic conclusions—firearm toolmark errors appear to be pervasive nationwide:

- The Houston Police Department Crime Lab: Though mainly motivated by the scandalous past of its DNA unit, the eventual closure of the HPD Crime Lab followed publication of the damning Bromwich Report. Detailing only four case studies of procedural unfairness grounded in faulty forensic evidence, the Bromwich Report included Mr. Williams' case as an example of the faulty firearm toolmark and ballistic methods within the unit and the prejudicial testimony at his trial.[16]

- The Detroit Police Department Crime Lab: In September of 2008, the Detroit Police Department shut down its crime lab following an audit into its firearm examination unit that revealed "erroneous or false" conclusions in 10% of the cases examined.[17] A report detailing the findings of this audit reads "if this 10 percent error rate holds, the negative impact on the judicial system would be substantial, with a strong likelihood of wrongful convictions and a valid concern about numerous appeals."[18] It is of note that, as condemnatory as it was, this audit was conducted by the State Police and not an independent bureau or agency—raising the potential for further biases and undiscovered errors within even these scrutinized data.

- Discreditation of the D.C. Crime Lab: As mentioned above, the Washington D.C. crime lab has lost its ANSI accreditation after an audit conducted by ANAB, ANSI's accrediting board. Discovering egregious misconduct within the lab's firearms examination unit—including the bad faith suppression of information relevant to the review—ANAB in April sent the director of the Washington D.C. Department of Forensic Sciences (DFS) a letter suspending the accreditation of the entire lab. Recently finalized, the discreditation of the D.C. crime lab led to the resignation of

---

[15] *Id.* at 14 "…3) policies that, at the time the case was originally examined, permitted firearms examiners to co-sign reports of other examiners without personally reviewing the evidence that was the subject of the report."

[16] Bromwich Report, *supra* note 13, at 219-234.

[17] Nick Bunkley, *Detroit Police Lab is Closed After Audit Finds Serious Errors in Many Cases*, NEW YORK TIMES (Sept. 25, 2008), *available at* https://www.nytimes.com/2008/09/26/us/26detroit.html.

[18] *Id.*

the director of DFS at the end of May.

The DC Crime Lab scandal also surfaces grave concerns about the reliability of firearms analysis generally. Because the technique is entirely subjective, as discussed more below, multiple different conclusions were offered by practitioners examining the same evidence.[19] For a forensic technique to be considered reliable, however, experts viewing the same evidence should come to similar conclusions.

The most recent closure in D.C. has brought the questions surrounding FTE into critical focus. Not all of these erroneous results can be attributed merely to misconduct, but instead must also draw scrutiny to the shaky foundations of the field.

Longstanding Questions Concerning the Validity of Toolmark Evidence

In the past thirteen years, the field of firearm toolmark evaluation has become the cause of increasing concern among scientists, statisticians, and the legal community. Three reports issued by three separate committees of nationally recognized experts—two by the research arm of the National Academy of Science ("NAS"), and one by the President's Council of Advisors on Science and Technology ("PCAST")—have concluded that FTE lacks scientific validity.[20] The concerns around toolmark examination have led to calls from scientists and scholars for the outright exclusion of FTE,[21] and a series of decisions from courts around the country limiting the permissible testimony of firearm toolmark

[19] Keith L. Alexander, *Ballistics Work at D.C.'s Crime Lab Criticized by Forensic Experts*, Wash. Post, (March 26, 2021), *available at* https://www.washingtonpost.com/local/public-safety/dc-crime-lab-ballistics-mistake/2021/03/26/42e992aa-8c0e-11eb-a730-1b4ed9656258_story.html

[20] National Research Council, Committee on Identifying the Needs of the Forensic Sciences Community, *Strengthening Forensic Science in the United States: A Path Forward* (August 2009) *available at* https://www.ojp.gov/pdffiles1/nij/grants/228091.pdf (hereinafter "2009 NAS Report"); National Research Council, Committee to Assess the Feasibility, Accuracy, and Technical Capability of a National Ballistics Database, *Ballistics Imaging* (2008), *available at* https://www.nap.edu/catalog/12162/ballistic-imaging (hereinafter "Ballistics Imaging Report"); President's Council of Advisors on Science and Technology, *Forensic Science in Criminal Courts: Ensuring Validity of Feature-Comparison Methods* (September 20, 2016), *available at* https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf (hereinafter "PCAST Report").

[21] *See, e.g.*, Paul C. Giannelli, *Forensic Science: Under the Microscope*, 34 Ohio N.U.L. Rev. 315 (2008) (noting an unfulfilled "need for comprehensive regulation of crime laboratories…there is a critical need for independent scientific validation of forensic techniques."); William A. Tobin, *Affidavit in Virginia v. Macumber* (2011), *available at* https://afte.org/uploads/documents/swggun-azvmacumber-tobin.pdf.

examiners.[22] Statisticians have widely endorsed the NAS reports and called for reevaluation of experiment design and reporting of error rates.[23]

1. The Subjective AFTE Methodology

The first step in understanding why this undertaking is problematic is understanding the employed methodology as laid out by the the Association of Firearm and Toolmark Analysis ("AFTE") and followed by firearm toolmark examiners throughout the county. AFTE's firearm toolmark analysis methodology consists of a subjective process applied to an unsubstantiated assumption: that each gun leaves a set of unique markings on every piece of ammunition fired from it. The same unsubstantiated assumptions are made in related subdisciplines of toolmark analysis, including handheld tools. That is, AFTE's methodology assumes that every handheld tool, such as a screwdriver, leaves a "unique" mark identifiable to an individual tool on any substrate upon which an impression can be made. (See Pt. 5, *infra*, for a discussion of related assays within the field of toolmark analysis.)

In a traditional bullet-to-gun assay, a toolmark examiner purports to match spent ammunition to the firearm from which it was discharged by looking through a microscope to examine the markings the firearm left on the spent ammunition and compare them to the markings on test-fired samples from the same gun. To compare these "individual characteristics," the examiner must first be able to identify and eliminate class characteristics—markings left by all firearms of a particular make and model—and then subclass characteristics—markings left by firearms of a particular batch lot of a particular make and

---

[22] One Missouri state court, citing the PCAST Report, noted that the *only* community to declare toolmark testimony valid was the Association of Firearms and Toolmark Examiners themselves; independent scientists, on the other hand, "have uniformly concluded that firearm and toolmark analysis has not been scientifically validated." *Missouri v. Goodwin-Bey*, No. 1531-CR00555-01, at 4-5 (Mo. Cir. Ct. Green Cnty. Dec. 16, 2016). Similarly, the District of Columbia Superior Court precluded the government from eliciting toolmark testimony "based largely […] on the lack of acceptance of the discipline's foundational validity outside of the community of firearms and toolmark examiners." *United States v. Tibbs*, 2019 WL 4359486, at *1 (D.C. Sup. Ct. Sept. 5, 2019). *See also, e.g.*, *United States v. Adams,* 444 F. Supp. 3d 1248, 1266 (D. Or. 2020) (concluding that the methodology for firearm identification testimony was "largely disavow[ed]" because it did not "meet the parameters of science"); *United States v. Shipp*, 422 F. Supp. 3d 762, 783 (E.D.N.Y. 2019) (forbidding expert to "testify, to any degree of certainty that the recovered firearm is the source of the recovered bullet fragment or the recovered shell casing"); *United States v. Davis*, No. 4:18-cr-00011, 2019 WL 4306971, at *7 (W.D. Va. Sept. 11, 2019) (concluding that expert "witnesses may not testify as to a 'match,' that the cartridges bear the same 'signature,' that they were fired by the same gun, or words to that effect"); *United States v. White*, No. 17 CR. 611 (RWS), 2018 WL 4565140, at *3 (S.D.N.Y. Sept. 24, 2018) (prohibiting expert from testifying "to any specific degree of certainty as to his conclusion that there is a ballistics match between the firearms"); *State v. Raynor*, 337 Conn. 527, 541-42 (2020) (determining that trial court abused its discretion in failing to consider new criticism of firearm and toolmaker analysis); *People v. Ross*, 68 Misc. 3d 899, 917 (N.Y. Sup. Ct. 2020) ("At a foundational level, beyond comparing class characteristics forensic toolmark practice lacks adequate scientific underpinning and the confidence of the scientific community as whole."); *Williams v. United States*, 210 A.3d 734 (D.C. 2019) (error to admit testimony that it was beyond doubt that a specific bullet could be matched to a specific gun).

[23] *See* American Statistical Association, *ASA Board Policy Statement on Forensic Science Reform*, (Apr. 17, 2010), *available at* http://www.amstat.org/asa/files/pdfs/POL-Forensic_Science_Endorsement.pdf; Karen Kafadar, *Statistical Issues in Assessing Forensic Evidence*, 83 International Statistical Review 1 (April 2015); Alicia Carriquiry, "*Declaration in Support of Defendant Joseph Blacknell's Motion to Exclude Firearms & Toolmark Identification evidence Or, In the Alternative, for a Kelly Hearing*," (Nov. 21, 2011) ("In my opinion as a statistician with many years of experience, the studies that have been carried out and the (scant) data that have been collected in no way support the methods or the conclusions that are routinely drawn by firearms examiners"), *available at* https://afte.org/uploads/documents/swggun-cavblacknell-carriquiry.pdf.

model. The remaining markings are then deemed "individual characteristics," and it is these that the examiner relies upon to tie the spent ammunition to a specific gun. If the examiner makes the determination that there is "sufficient agreement" between the individual characteristics seen on two sets of ammunition, he or she declares a "match" and concludes that they were from the same gun.[24]

2. The NAS and PCAST Reports Raise Significant Concerns about FTE

Beginning in 2008 and continuing through 2017, the NAS and PCAST convened committees to closely examine concerns in the firearm toolmark (and other pattern-matching) arena. Importantly, the committees authoring the reports consisted of independent scientists and professors—with expertise in physics, chemistry, biology, materials science, engineering, biostatistics, statistics, and medicine—statisticians, medical examiners, judges, forensic practitioners, and lawyers, rather than firearms examiners, whose financial and professional stake in the continued embrace of their discipline is apparent.[25] Each committee heard testimony from forensic scientists, reviewed nearly every available journal article and study involving firearms examination, and read every article or study submitted by members of the forensic community.[26] As such, these bodies were uniquely qualified to determine whether the forensic discipline was based on a valid, reliable scientific principle or methodology.

In the end, the conclusions of these committees were uniform: the "fundamental assumptions" underlying firearms examination have not been proved; the theory of toolmark identification is "not a scientific theory"; the method is subjective; and there is insufficient empirical evidence establishing validity and estimating reliability of firearms examinations.[27] In short, the committees concluded that FTE consists of applying a subjective methodology to an unvalidated assumption and lacks the studies necessary to demonstrate that it produces reliable, repeatable results.

First, the entire undertaking of FTE rests on an unsubstantiated assumption: that each firearm leaves unique markings on ammunition discharged from it. As the committees noted, "the validity of the fundamental assumptions of uniqueness and reproducibility of firearms-related toolmarks has not yet been demonstrated," and "[a] significant amount of research would be needed to scientifically determine the degree to which firearms related toolmarks are unique or even to quantitatively characterize the probability of uniqueness."[28]

---

[24] *See generally*, The Association of Firearm and Toolmark Analysis, *Summary of the Examination Method*, *available at* https://afte.org/resources/swggun-ark/summary-of-the-examination-method; PCAST Report, *supra* note 20, at 104.

[25] *See, e.g.*, PCAST Report, *supra* note 20, at v–ix; 2009 NAS Report, *supra* note 20, at v–xiii; Ballistics Imaging Report, *supra* note 20, at v–vii, xi–xvi, 312–322.

[26] *See, e.g.*, PCAST Report, *supra* note 20, at 2, 155–160; 2009 NAS Report, *supra* note 20, at xx, 2–3; Ballistics Imaging Report, *supra* note 20, at xiii–xvi; *see also* President's Council of Advisors on Science and Technology, *An Addendum to the PCAST Report on Forensic Science in the Criminal Courts* 2 (2017) (hereinafter "PCAST Addendum").

[27] Ballistics Imaging Report, *supra* note 20, at 3; PCAST Report, *supra* note 20, at 47, 60, 104, 111 and 113; 2009 NAS Report, *supra* note 20, at 154.

[28] Ballistics Imaging Report, *supra* note 20, at 3, 81; *see also* 2009 NAS Report, *supra* note 20, at 154 ("[N]ot enough is known about the variabilities between individual tools and guns" for individualization.); PCAST Report, *supra* note 20, at 59-60 (criticizing the field of toolmark examination's "theory" of individualization as based on assumptions rather than scientific data on the frequency of toolmark characteristics or an "empirical demonstration of accuracy").

Second, the methodology applied in toolmark analysis is entirely subjective and unproven. No specific protocol defines what constitutes "sufficient agreement," leaving examiners in each case to exercise their own judgment based on their own experience. Thus, the method is entirely circular: it "declares that an examiner may state that two toolmarks have a 'common origin' when their features are in 'sufficient agreement,' [but] defines 'sufficient agreement' as occurring when the examiner considers it a 'practical impossibility' that the toolmarks have different origins."[29]

Third, there do not currently exist studies sufficient to evaluate the reliability of the proffered methods.[30] To be "foundationally valid," a field must utilize a method that has been subject to "empirical testing by multiple groups, under conditions appropriate to its intended use."[31] It must show through studies that the method is "repeatable and reproducible."[32] Because the method is "subjective," foundational validity and reliability "can *only* be established through multiple independent black-box studies."[33] In the absence of appropriate research, the committees concluded that examiners' testimony about so-called matches, "cloak[s] an inherently subjective assessment of a match with an extreme probability statement that has no firm grounding and unrealistically implies an error rate of zero."[34]

Simply put, these committees found that firearms examination "falls short of the scientific criteria for foundational validity."[35] "Without appropriate estimates of [the method's] accuracy, an examiner's statement that two samples are similar—or even indistinguishable—is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact."[36] Until the toolmarks field has shown, through empirical research rather than unsupported assertions, that its underlying theory—that toolmarks are unique—is true, that an examiner can follow a proven methodology to declare a "match," and that its examiners produce accurate results when applying that methodology, such testimony cannot be admitted into evidence.

3. <u>FTE by its Nature Has a Strong Potential for High Error Rates</u>

Significant problems inherent to the FTE methodology create the strong potential for error. Significantly, the so-called "individual characteristics" of toolmarks are actually comprised of non-unique

---

[29] PCAST Report, *supra* note 20, at 60.

[30] 2009 NAS Report, *supra* note 20, at 154 ("Sufficient studies have not been done to understand the reliability and repeatability of the methods.").

[31] PCAST Report, *supra* note 20, at 5.

[32] *Id*. at 47.

[33] "Black-box" studies are studies "with many examiners making a series of independent comparison decisions between a questioned sample and one or more known samples that may or may not contain the source." *Id.* at 110. Because these studies best replicate case work, they are the "only" studies appropriate for assessing scientific validity and estimating reliability. *Id.* at 106.

[34] Ballistics Imaging Report, *supra* note 20, at 82 ("Conclusions drawn in firearms identification should not be made to imply the presence of a firm statistical basis when none has been demonstrated.").

[35] PCAST Report, *supra* note 20, at 11.

[36] *Id.* at 6.

marks, making it difficult for examiners to discern them.[37] Indeed, early studies showed that bullets fired from different guns shared almost as many similarities as bullets fired from the same gun.[38]

Moreover, "[t]he most seminal, but problematic, obstacle for toolmarks examiners . . . is discerning subclass from purported 'individual' characteristics."[39] Indeed, it is often impossible for firearm examiners to distinguish these marks from each other. "Without personal knowledge of the individual and subclass characteristics produced by a particular manufacturing run, an examiner cannot generally distinguish between [class and subclass characteristics] for most forming processes."[40] Thus, there is a significant risk that an examiner without firsthand knowledge of the subclass characteristics common to a specific production run will identify a single gun as the source of marks on a bullet or cartridge, when in reality tens, hundreds, or even thousands of guns from a batch could have produced the same patterns.[41] It is worth nothing that the problems posed by subclass characteristics will only increase as time goes on because of advancements in manufacturing processes result in larger batches of weapons being produced in the same run, thereby increasing the risk of misidentification.[42]

In addition to these problems is the effect on examiners' conclusions of cognitive bias, that is, the human tendency to interpret data so that it confirms expectations and discount data that appears to conflict with those expectations.[43] The risk is that the "observer's conclusions become contaminated with a pre-existing expectation and perception, reducing the observer's objectivity and laying the groundwork for selective attention to evidence."[44] Scientists have long acknowledged that cognitive bias "can lead to perceptual distortion, inaccurate judgment, or illogical interpretation,"[45] specifically because it causes decisionmakers to "seek information that they consider supportive of a favored hypothesis or existing beliefs and to interpret information in ways that are partial to those hypotheses or beliefs."[46] Biasing contextual information has been documented to cause serious mistakes and misidentifications across a wealth of forensic disciplines.[47] Bias in an unavoidable product of human

---

[37] *See United States v. Monteiro,* 407 F. Supp. 2d 351, 360-61 (D. Mass. 2006) (discussing literature).

[38] Alfred A. Biasotti, *A Statistical Study of Individual Characteristics of Fired Bullets*, 4 J. Forensic Sci. 34 (1959) (finding that 15-20 percent of marks on bullets fired by *different* Smith & Wesson .38 Special revolvers matched, compared with 21-38 percent of marks fired by the same revolver).

[39] William A. Tobin, *Affidavit in Virginia v. Macumber*, *supra* note 21, at 8.

[40] *Id.* at 35.

[41] Alfred Biasotti & John Murdock, *Criteria for Identification or State of the Art of Firearm & Toolmark Identification*, 16 AFTE J 16, 18 (1984) ("We can have remarkable reproduction on many hundred or even thousands of individual items."); M.S. Bonfanti & J Dekinder, *The Influence of Manufacturing Processes on the Identification of Bullets & Cartridge Cases- A Review of the Literature*, 39 Sci. & Justice 3, 5 (1999) (noting that one tool, thanks to manufacturing improvements, may now make batches of hundreds or thousands of barrels); Gene C. Rivera, *Subclass Characteristics in Smith & Wesson SW40VE Sigma Pistols*, 39 AFTE J 247, 250 (2007) ("Anywhere between a couple of hundred to one thousand slides could be machined before the broach is resharpened.").

[42] *See United States v. Willock*, 696 F.Supp.2d 536 (M.D.N.D. 2010).

[43] *See generally*, Itiel E. Dror, *Cognitive and Human Factors in Expert Decision Making: Six Fallacies and the Eight Sources of Bias*, 92 Anal. Chem. 7998-8004 (2020).

[44] Paul Bieber, *Fire Investigation and Cognitive Bias*, Wiley Encyclopedia of Forensic Science (2014).

[45] Working Group on Human Factors in Latent Print Analysis, *Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach*, National Institute of Justice, at 10 (2012).

[46] R. S. Nickerson, *Confirmation Bias: A Ubiquitous Phenomenon in Many Guises*, Review of General Psychology 2, p. 177 (1998).

[47] *See generally*, Saul Kassin et al., *The forensic confirmation bias: Problems, perspectives, and proposed solutions*, 2 J. of Applied Research in Memory and Cognition 45-52 (2013).

decision making—it cannot be "willed away"—and is particularly problematic in subject forensic technique like FTE.[48] And yet AFTE's methodology makes no effort to mitigate the influence of irrelevant data from examiners' conclusions.

4. FTE Has Not Been Validated through Appropriate Testing

A "scientific theory," PCAST explained, is a "comprehensive explanation of some aspect of nature that is supported *by a vast body of evidence*."[49] The PCAST Report was unequivocal: experience, judgment, and years of use in court cannot establish scientific validity and a degree of reliability.[50] "The *only* way to establish the scientific validity and degree of reliability of a subjective forensic feature comparison method—that is, one involving significant human judgment—is to test it empirically by seeing how often examiners actually get the right answer."[51] At present, however, there exist *no studies* establishing that AFTE's firearm toolmark examination methodology creates repeatable and reproducible results.

Indeed, the single study deemed "appropriate" by PCAST—a black box study commissioned and funded by the Defense Department's Forensic Science Center and conducted by the Ames Laboratory, a Department of Energy national laboratory affiliated with Iowa State University—estimated a false-positive error rate between 1 in 66 and 1 in 46.[52]

A second black box study conducted by the Ames Laboratory for the FBI and released several years after the PCAST Report in 2020, shows unambiguously that toolmark examiners cannot accurately carry out firearm comparisons, such that even the same examiner looking at the same evidence will often reach different results.[53] Indeed, it reports astounding error rates and alarming problems with the undertaking, showing that, with respect to bullets, examiners were unable to repeat their *own* conclusions (repeatability) 21% of the time for known matches and 35.3% of the time for known non-matches, and were unable to repeat the conclusions of *other* examiners (reproducibility) 32.2% of the time for known matches and almost 70% of the time for known non-matches.[54]

Since PCAST, AFTE has relied heavily on a series of more recent studies[55] to claim validation of the discipline, but these studies all fail in this task for a variety of reasons. First, many emerging

---

[48] 2009 NAS Report, *supra* note 20 at 122 ("cognitive biases are not the result of character flaws; instead, they are common features of decisionmaking, and they cannot be willed away")

[49] PCAST Report, *supra* note 20, at 60 (emphasis added).

[50] *Id.* at 6.

[51] PCAST Addendum, *supra* note 26, at 1.

[52] *Id.* at 11 (noting that the study is available as a report to the Federal government, but has not been peer reviewed or published in a scientific journal). These error rates were—appropriately—calculated based solely on *conclusive* examinations, i.e., without regard to inconclusives, a common problem with FTE studies, as discussed further, below.

[53] Stanley J. Bajic et al., *Report: Validation Study of the Accuracy, Repeatability, and Reproducibility of Firearms Comparison*s, Oct. 7, 2020, Ames Laboratory-US DOE, Technical Report #ISTR-5220.

[54] *Id.* Results were similarly abysmal for cartridge casings, with disagreement for the same examiner at 24.4% for matches and 37.8% for non-matches and for different examiners at 36.4% for matches and 59.7% for non-matches.

[55] *See, e.g.,* Chad Chapnick et al., *Results of the 3D Virtual Comparison Microscopy Error Rate (VCMER) Study for firearm forensics*, 66 J. Forensic Sci. 557-70 (2020) (study of cartridge cases only, counts inconclusive results, uses Virtual Comparison Microscopy rather than the AFTE methodology); Zhe Chen et al., *Pilot study on deformed bullet correlation*, 306 Forensic Sci. Int'l (2020) (uses Congruent Matching Profile Segments (CMPS)

studies rely on examination tools and/or protocols other than the AFTE theory of examination, such as 3D Virtual Comparison Microscopy and Congruent Matching Profile Segments, which is not part of the AFTE methodology and not used in casework. As such, these studies cannot establish the validity of the AFTE theory of examination.

Many studies also suffer from design flaws. One such design flaw is the use of "closed sets." A closed-set study includes a match for every sample, so that examiners can simply "match each bullet to the standard that is closest."[56] Such studies do not replicate casework. By contrast, in open set studies (as in casework) "there is no guarantee that the correct source is present—and thus no guarantee that the closest match is correct."[57]

Moreover, nearly every study involving toolmark examination mishandles the treatment of "inconclusive" results, failing to count these conclusions as incorrect, even though there exists a ground truth answer of match or non-match. The failure to treat inconclusives as mistakes results in misleadingly low error rates that do not reflect reality.[58] This treatment of inconclusives is akin to giving a student 90% on a test where he answers 10% percent of the questions incorrectly and skips the rest.[59] An inconclusive result that does not match ground truth must be considered: it is an error and must be counted as such.

5. FTE Encompasses Many Different Assays, Each Requiring Separate Validation

Another fundamental problem with the studies purporting to validate the field of toolmark examination is that the field is comprised of many different assays, each of which requires separate validation. While there is, at best, a single appropriately designed black box study providing some limited insight on examiners' rates of error in associating bullets fired from the same firearm, there are no such studies in any other subdiscipline of FTE. And even if there were studies sufficient to establish the validity of matching pristine bullets to other pristine bullets fired from the same gun—which there are not—such studies would not validate the matching of deformed bullets or ejector mark, a far more common casework example. Before such an undertaking could be considered valid, studies would need to be undertaken to quantify the effect of the damage on the ability to make a comparison, i.e., establish how much of a sample is necessary for an examination to be valid and the effects, if any, of

---

profile comparison method rather than the AFTE methodology); Jaimie A. Smith, *Beretta barrel fired bullet validation study*, 66 J. Forensic Sci. 547-556 (2020) (study was completed by only 74 of the 110 participants, has elements of a closed set study design and allows for inconclusive results); M. A. Keisler, *Isolated pairs research study*, 50.1 AFTE J. 56-58 (2018) (small study of only .40 caliber cartridge cases; otherwise well-designed); Pierre Duez et al., *Development and Validation of a Virtual Examination Tool for Firearm Forensics*, 63 J. Forensic Sci. 1069-1084 (2017) (study of cartridge cases only, counts inconclusive results, uses Virtual Comparison Microscopy); Tasha Smith et al., *A Validation Study of Bullet and Cartridge Case Comparisons Using Samples Representative of Actual Casework*, 61 J. Forensic Sci. 939-46 (2016) (study was completed by only 34 of 47 participants and allows inconclusive results).

[56] PCAST Report, *supra* note 20, at 108.

[57] *Id*.; *see also Tibbs*, 2019 WL 4359486, at *33-38.

[58] *See generally*, Itiel E. Dror & Nicholas Scurich, *(Mis)use of scientific measurements in forensic science*, 2 Forensic Sci. Int'l, 333-38 (2020); Heike Hofmann et al., *Treatment of inconclusives in the AFTE range of conclusions*, 19 Law, Prob. & Risk 317-64.

[59] *See Adams,* 444 F. Supp. 3d at 1265 ("There seems to be no real negative consequence for reaching an answer of inconclusive. Since the test takers know this, and know they are being tested, it at least incentivizes a rate of false positives that is lower than real world results.").

the deformation on the marks to be compared. Similar studies are necessary to validate the practice of matching ejector markings on bullet casings to firearms.

The inapplicability of one discipline to another is even more apparent when considering the field of handheld toolmarks, i.e., marks left on a surface by tools *other than* guns, such as screwdrivers, wire cutters, and pliers. At present, examiners rely largely on firearm toolmark studies to justify matches of handheld tools to marks. As discussed above, however, the firearms studies are insufficient to validate bullet-to-gun associations and are plainly incapable of validating an entirely separate, much more challenging, assay. There are myriad additional variables at play with handheld tools. These variables include the arm pressure on the tool used to make the mark, the arm angle, the speed at which the tool was used, and the innumerable variety of tools examiners claim—without data—the ability to associate to marks.

Moreover, there are no studies establishing toolmark experts are capable of reliably distinguishing marks made from one class of tool as opposed to another. In other words, there is no evidence an examiner can look at a mark on something like a windowsill and determine it was made by a screwdriver rather than a crowbar or a chisel. Finally, as opposed to the relatively controlled situation of a bullet travelling through the barrel of a gun, handheld tools can leave different types of marks depending on the sensitivity and characteristics of the substrate—or surface—of contact, and there are no studies examining what constitute reliable substrates for faithfully recording a supposed toolmark.[60] Nor are there standards—or research—as to how much information must be included in a supposed toolmark to reach evidentiary value.

Conclusion

On behalf of Mr. Williams, we ask the Commission to examine the evidence used to convict him at trial and to undertake a thorough investigation of all toolmark evidence, including both FTE and handheld toolmark evidence. Our request is that this Commission set appropriate limits on the conclusions of FTE examiners in traditional bullet-to-firearm matching testimony, and determine what conclusions—if any—can be proffered in other toolmark assays.

[60] *See* 2009 NAS Report, *supra* note 20, at 154-55 (specifically distinguishing between the utility of toolmarks made by firearms versus hand tools, due to the numerous confounding variables associated with use of hand tools); *see also United States v. Smallwood*, No. 5:08-CR-38, 2010 WL 4168823 (W.D. Ky. Oct. 12, 2010), *aff'd* 456 F. App'x 563 (6th Cir. 2012) ("[W]hile a firearm can generally only be used in one way, by pulling the trigger, a tool can be used in a number of ways . . . . As a result, this Court does not think that precedent in firearm identification is applicable to tool mark identification.").

Very Truly Yours,

*_/s/ Violeta Chapin*_____
Violeta Chapin, Supervising Attorney
Clinical Professor of Law (Atty. Reg. 41090)
University of Colorado Law School
Boulder, CO 80309
violeta.chapin@colorado.edu
(303) 492-5830

Joseph Craver, Student Attorney
Alex Espina, Student Attorney
Criminal & Immigration Defense Clinic
University of Colorado Law School
Boulder, CO 80309

*_/s/ M. Chris Fabricant*_____
Tania Brief
Innocence Project, Inc.
40 Worth Street, Suite 701
New York, New York 10013
(212) 364-5997

EXHIBIT B

EXHIBIT B: RELEVANT EXCERPTS FROM REPORT OF MICHAEL BROMWICH

a. Bromwich Report: "The Crime Lab's Pretrial Analysis of Firearms Evidence"

A fragment of an extensively deformed and fragmented small caliber bullet (identified as EB1) was collected from Mr. Collier's head at autopsy, along with 68 lead shotgun pellets. HPD officers also recovered an unfired .25 caliber cartridge at the scene. The bullet fragment, lead pellets, and a fired plastic shot carrier recovered from the crime scene were submitted to the Crime Lab on May 19, 1992. An HPD firearms examiner, Robert Baldwin, receive the evidence on May 19 and logged it in on May 20, 1992.

Donald Davis, an HPD Crime Lab firearms examiner, issued a report dated June 16, 1992 in which he concluded, based solely on a visual examination, that EB1 was a .25 caliber bullet. Mr. Davis also reported the plastic shot carrier was consistent with a 12 gauge shotgun but contained insufficient definite characteristics for identifying the firearm. The 68 fired shotgun pellets were found to be #6 birdshot. [footnote omitted].

On March 25, 1994, a bullet that was removed from Mr. Rasul's foot when he was treated at [the hospital] was retrieved from the hospital by an HPD investigator. The bullet is identified as EB2. [footnote omitted].

HPD firearms examiner C. E. Anderson examined items labeled as EB1 and EB2 on June 23, 1995. He reported that the two fired, jacketed lead bullets were partially mutilated but the land and groove measurements on both bullets indicated "that they could have been fired in a firearm of the same manufacturer." The caliber of EB1 and EB2 is not explicitly identified, but the report implies that Mr. Anderson concluded the bullets were .25 caliber because it states that "we are in concurrence with [Mr. Davis'] findings," which concluded that EB1 was a .25 caliber bullet. Mr. Anderson and Mr. Baldwin are identified as the firearms examiners responsible for this report."

b. Bromwich Report: "Trial Testimony Regarding the Crime Lab's Evaluation of Firearms Evidence"

"During the prosecution's direct examination, Mr. Baldwin was asked specific questions regarding the dates on which he examined State's Exhibit 21, which included EB1.

Q. Sir, what day did you actually examine State's Exhibit 21?

A. Well, actually there were two occasions that I had – there were two different occasions I had to examine this evidence. At the time Officer Horowitz initially submitted this evidence to the Firearms Laboratory I was responsible at that time for logging in evidence that had been submitted to the laboratory. That was the first occasion that I had to examine any of these items. And that was on May the 20th of 1992, I believe.

Q. Did you also examine it on a second date?

A. At a later time, based on a request from your office, there was a re-examination of the evidence by Mr. Anderson, Mr. C. E. Anderson, who is a senior firearms examiner, and myself; yes.

Q. On what date?

A. That was on June 23rd of 1995. [footnote omitted]

Responding to an IAD investigation that occurred eight years after Mr. Williams's trial, Mr. Baldwin acknowledged that the examination he performed when logging in the evidence on May 19, 1992 was cursory. Mr. Baldwin reported that his June 23, 1995 examination was more detailed and included a microscopic comparison of EB1 and EB2 to verify Mr. Anderson's identifications. During that comparison, Mr. Baldwin examined the individual characteristics of each bullet and evaluated the width of each bullet's lands and grooves. [footnote: Mr. Baldwin did not conduct a GRC [General Rifling Characteristic] analysis at that time, as the examination was performed by the "primary examiners", Mr. Davis and Mr. Anderson. During cross examination, Mr. Baldwin acknowledged that his testimony regarding possible manufacturer's of EB1 came from Mr. Davis's report.

Based on the examination he actually did conduct, Mr. Baldwin concluded that there were insufficient individual characteristics to relate EB1 and EB2 to each other, meaning that he was unable to conclude that the bullets were fired from the same weapon. Because both bullets exhibited rifling with a left twist and land and groove impressions of a similar width without any misalignment, Mr. Baldwin found no reason to disagree with the conclusions of Mr. Anderson (who was then head of the Firearms Section and an examiner with 20 years' experience) and Mr. Davis (also an experienced examiner). IAD nevertheless concluded that Mr. Baldwin violated HPD Internal Directives by testifying at Mr. Williams trial without conducting his own independent examination of the evidence.

As noted above, Mr. Guevara admitted on cross-examination that he told HPD investigators that he fired the .22 caliber derringer in the direction of Mr. Collier during the aborted drug deal. The State moved the derringer into evidence and elicited the following testimony from Mr. Baldwin to support its theory that it was not Mr. Guevara's gun that produced the bullet in Mr. Collier's head:

Q. Is there any way in the world, based on your training, your expertise and the examinations you made, that the bullet [EB1], which was part of the submission in State's Exhibit No. 21, was shot out of that Derringer, State's Exhibit No. 17?

A. No sir. It's the wrong caliber, plus the type of cartridge used in State's Exhibit 17 is a rim fire cartridge and the .22 automatics are center fire cartridges.

Mr. Baldwin then testified that EB1 and EB2 were both .25 caliber bullets and that neither could have been fired from the derringer:

Q. Before I go any further, show these to the jury, what type of bullet is the bullet that was submitted in State's Exhibit 22 [EB2]?

A. That's a .25 automatic full metal jacketed bullet.

Q. How can you [be] so certain that both of these are .25 caliber bullets?

A. Comparison. We compared both bullets to each other. The base diameters were consistent. We also compared the land and groove widths and the number of lands and grooves, and they were also consistent with each other.

Q. Now, these bullets don't look at all like each other. What makes you say they were the same type of bullet?

A. The reason they don't look the same is the fact that the one bullet recovered from the morgue is extremely mutilated. A very large portion of its mass is missing.

Q. In fact, sir, in addition to being the same diameter, are they the same make of ammunition?

A. Yes, sir.

Q. Is there any way the bullet in State's Exhibit 22 could have been fired out of that Derringer, which is State's exhibit 17, any way in this world?

A. No sir. As indicated, State's Exhibit No. 22 is a 25 automatic, and the cartridge for a .25 automatic is larger in diameter than this weapon would be chambered to handle. Also, the .22 is a rim fire cartridge and he – excuse me – the .25 automatic is a center fired cartridge. [footnote omitted]

When the prosecutor asked Mr. Baldwin to examine examples of .22 caliber cartridges, Mr. Baldwin confirmed that he had not previously examined the derringer:

> Q. Could a cartridge like that have been fired from that Derringer we have been speaking about or could it be used in that Derringer …?
>
> A. Yes, it could, but I don't know what the functional condition of that Derringer is. I have never checked it.

Neither Mr. Baldwin nor the investigators handling Mr. Williams case made arrangements to have the derringer examined by the Firearms Section after the trial."

c. Bromwich Report: "The Misidentification of EB1 is Discovered and Reported"

Lawyers handling Mr. Williams's appeal sought and eventually obtained an order that the firearms evidence be turned over to a defense expert. The Harris County District Attorney's Office asked Mr. Baldwin to test-fire the derringer before it was provided to the defense 'in order to preserve the integrity of the evidence.'\

On January 15, 1998, the Crime Lab test-fired the derringer for the first time. Based on a microscopic comparison of EB1 to the bullet obtained from the test-firing, Mr. Baldwin concluded that EB1 was actually a .22 Magnum caliber bullet that had been fired from the bottom barrel of the double-barreled derringer carried by Mr. Guevara.

Mr. Baldwin's results were verified by Michael Lyons, another HPD firearms examiner. Mr. Baldwin promptly issued a report with the new conclusions, and Mr. William's lawyers were notified of the results. The revised findings undermined a central aspect of the prosecution's case, which was based on the premise that Mr. Guevara's .22 caliber derringer could not have fired the bullet that was removed from Mr. Collier's head.

d. Bromwich Report: An excerpt from "State Court Proceedings":

"Ronald Singer, of the Tarrant County ME's Office, reviewed the firearms evidence for the defense. In an affidavit dated April 15, 1998, he stated that:

> Even in their "damaged" state, the .22 Magnum caliber bullet, State's exhibit EB1 and the .25 Auto caliber bullet, State's Exhibit EB2 are easily distinguishable from one another, particularly if examined with the aid of a comparison microscope, and should have presented no problem to a competent firearms examiner. Mr. Baldwin's testimony at trial that EB1 was a .25 caliber projectile that could have been fired from the same gun as the bullet EB2, recovered from another victim's foot, at best demonstrates extreme carelessness on his part, and at worst calls in to question his expertise. [footnote omitted] If the bullet had been correctly identified during at least one of the three times it was examined by the Houston Police department, the bullet could have been compared to the

Davis derringer prior to trial; this might have materially affected the outcome of the trial.

Mr. Singer was also quoted in the press, stating that '[t]here is enough of a difference between a .25 caliber bullet and a .22 Magnum that even a non-expert could look at them and tell the difference.'"

e. Bromwich Report: Summary of "The Investigative Team's Analysis of Work Performed by the HPD Crime Lab in the Williams Case"

The Bromwich Investigative Team conducted an independent evaluation of the fired bullet evidence submitted to the crime laboratory in the Williams case. The following is a summary of their findings.[1]

- Bullet base diameters vary slightly by manufacturer.

- EB1 was ultimately determined to be a .22 Magnum rimfire caliber copper jacketed lead bullet. When it is intact and undamaged, this bullet has a base diameter of approximately 0.225 inches.

- EB2 was determined to be a .25 caliber copper jacketed lead bullet, and when this bullet is undamaged and intact, it has a base diameter of approximately 0.250 inches.

- The difference in diameter between these bullets if intact and undamaged is approximately 0.025 inches. This difference in base diameters is small but visibly discernible.

- Mr. Singer's assertion that a non-expert can perceive the difference is true if the bullets being compared are in good condition.

- However, when a bullet strikes a hard surface, distortion and /or fragmentation typically occur and can alter the apparent base diameter of a bullet. It can also cause apparent rifling orientations to be different than actual (right twist can appear as left twist and vice versa).

- Identification of EB1 was not simple because only a bullet fragment was retrieved, and that fragment was extensively deformed.

- Caliber determinations are based on the assumption that the bullets fired from the same weapon show comparable base diameters. When a bullet has been fired, the diameter of the barrel bore is approximated by measuring the bullet's base diameter.

[1] [INSERT URL]

- In this case, the distortion of EB1 caused its bullet base diameter to be variable and some of those measurements were similar to those of a .25 caliber bullet.

- The distortion of EB1 seems obvious in a side-by-side comparison of EB1 and EB2, however, this distortion – and its effect on the determination of the bullet's caliber – would not be as obvious on an examination of EB1 alone (Mr. Davis did not have EB2 when he made his examination. EB2 was not submitted until more than a year later. This is especially true because Mr. Davis did not perform a microscopic examination).

- A firearms examiner should consider the possibility that the distortion of a bullet would affect its base diameter.

- Distinctions in the design and construction of an unfired .25 Magnum caliber bullet and am unfired .25 caliber bullet may be obvious, but the distinctions are not discernible in this case because of the extensive damage to and fragmentation of EB1. The only information regarding the design and construction that can be derived from EB1 is that it appears to be a full metal jacket design with a slightly concave base.

- To compare GRCs, a firearms examiner determines the number of lands and grooves on the bullets being compared and the direction of their twist. The width of the lands and grooves can also be measured, and the sum of those measurements should permit a calculation of the diameter and indicate the caliber.

- This step is not possible for distorted bullets or fragments like EB1. HPD examiners could only compare limited information from the fragment of EB1 to measurements from EB2.

- EB2 has six lands and grooves with a left twist. EB1 appears to have similar GRC's.

- The similarities in the rifling dimensions as well as in the gross features of the rifling impression themselves illustrate how an examiner could mistakenly conclude that EB1 and EB2 had similar class characteristics.

- The June 23, 1995 report issued by Mr. Anderson and co-signed by Mr. Baldwin states that the bullets "could have been fired in a firearm of the same manufacturer." Such a conclusion is appropriate where class characteristics are similar.

- Similar land and groove width dimensions can be found in bullets of two different calibers.

- Mr. Baldwin's testimony regarding his role in the May 1992 examination of EB1 may have given jurors the misleading impression that his review involved a more

substantive examination than the cursory inspection he described during a 2003 IAD investigation.

EXHIBIT C

2

Firearms and Ammunition: Physics, Manufacturing, and Sources of Variability

A firearm is a dynamic system for delivering maximum destructive energy to a target, in the form of a high-velocity bullet, with minimum delivery of energy to the shooter. To that end, the firing of a firearm and the subsequent generation of ballistic toolmarks are the end results of processes that are simultaneously characterized by high uniformity and great variability. Modern firearms and ammunition manufacture relies heavily on the uniformity and interchangeability of component parts, yet each step in the production cycle presents an opportunity for microscopically fine differences from part to part. Likewise, the firing of a gun depends on the rapid and repeated performance of numerous mechanical steps that is designed to produce combustion, done in a controlled manner yet still not creating exactly identical conditions in repeated firings.

In this chapter, we summarize the basic parts of firearms and ammunition (Section 2–A) and describe the physical processes that take place when a trigger is pulled and a gun is fired (2–B). These sections are not intended to be comprehensive examinations of the history and features of firearms and ammunition nor a complete catalogue of firearms products in current use. Rather, they provide context for the principal focus of this chapter: describing the types of toolmarks left on ballistics evidence by firing (2–C), particularly those that are typically imaged and input into ballistic image databases.[1] We close in Section 2–D with brief descriptions of concepts in the manufacture of both firearms and ammunition. A general understanding of manufacturing is essential not only for an appreciation

[1]More detailed information and images are available at http://www.firearmsid.com.

of the sources of variability in ballistic toolmarks, but also in assessing the feasibility of implementing technologies like wide-scale ballistic imaging or microstamping.

2–A ANATOMY OF FIREARMS AND AMMUNITION

2–A.1 Firearms

Firearms come in a wide array of designs and specific makes, and each represents a complex assemblage of numerous constituent parts. In this section we focus on the parts most central to the basic firing assembly since the interest is in toolmark creation. Due to their widespread use in crime, we also discuss some terminology in the specific context of handguns, as in differentiating between revolvers and pistols.

Barrels

Gun barrels are manufactured from solid pieces of steel whose composition is carefully selected for its chemical and metallurgical properties. A first step of the process, drilling, results in a comparatively rough hole of uniform diameter extending from one end of the barrel to the other. Next the barrel is bored with a reamer, designed to produce as smooth a surface as possible on the inside of the barrel. The interior surface or bore bears numerous scars and scratches from this drilling process; it is these random imperfections—more so than subsequent steps—that are said to account for individual characteristics on fired bullets (Heard, 1997:124–125).

Barrels are further subjected to a rifling process, creating a pattern of grooves on the inside the barrel. This rifling is essential to the firing accuracy of the weapon; as it is forced out of the barrel by gas pressure, the bullet impacts with the barrel rifling and is given a rotation—somewhat akin to the spin on a thrown football—that gives the bullet a more direct flight. Some weapons, typically shotguns, have no rifling ("smoothbore"). Most handguns and rifles have a spiral pattern of rifling to improve their accuracy. The rifling may be created by forcing a carbide button through the reamed barrel; it is the normal wear on this button, as many riflings are performed, that is said to impart individual microscopic variability in markings in the barrel (along with residual scars or imperfections from the original drilling). Additional steps in the process to finish a barrel include heat treating (to impart hardness) and cleaning.

Across manufacturers, barrels can vary in two fundamental features, each of which are basic class characteristics (see Section 3–B.1). The first is the *direction* in which the grooves in the barrel twist, whether left- or right-handed. Most U.S. makers use a right twist, although Colt revolvers

are known for their left twist (Rinker, 2004:128). The second is the number of *grooves* that are cut into the barrel—normally at a depth of 0.004–0.006 inch—to create the rifling, and, correspondingly, the number of raised *lands* between those grooves. Historically, "no standard was established and makers used, normally, six, seven, or eight grooves"; this remains the usual range, although firearms have been fielded with as few as 2 and as many as 24 grooves (Rinker, 2004:130, 131).

Barrels also vary in the degree of twist in the rifling, which affects how much rotation is put on bullets as they pass through the barrel and exit. Rinker (2004:127) observes that "few people agree on what is the proper twist. Some people want an over stabilized bullet from a fast twist. They claim best accuracy at all ranges. Other shooters believe a fast twist builds pressure and heat and they want a slow twist for minimum stability, and they have claims to back their theory."

Some firearms differ from conventional rifling with square-edged grooves, using *polygonal* rifling instead. "Polygonal rifling has no sharp edges," and instead the raised lands in the barrel have a smooth, "rounded profile which can be difficult to discern when looking down the barrel. This type of rifling is almost exclusively manufactured using the hammer or swage process" (Heard, 1997:123).

Chamber, Breech Face, and Firing Pin

The rear section (away from the muzzle) of the barrel bore is known as the chamber; it is designed and sized to fit a specific caliber of cartridge (see Section 2–A.2). The part of the firearm against which a cartridge sits when it is placed in the chamber is the breech, and the whole assembly may be referred to as the breechblock or breech bolt.

The specific surface of the breech that makes contact with the base of the cartridge is the breech face; Figure 2-1 depicts the breech faces of two firearms. The exact steps used to form the breech assembly can vary by manufacturer, and the breech face may vary in terms of the amount of filing or polishing done on it and whether any paint or other materials is applied to it. Basic filing can create gross striation marks in linear arrangements; in others, a rotary milling operation may be applied to the breech face surface, creating a pattern of concentric circles (American Institute of Applied Science, 1982:77). These steps are crucial to the theory of firearms identification as it is random imperfections created in these machining and filing processes that is said to make the surface (and the negative impressions of said surface, left on fired cartridge casings) unique.

A hole drilled through the breech assembly holds the firing pin, a very hard steel rod that can be forced to protrude from the breech to strike the primer of a cartridge seated in the chamber. While most firing pins have a

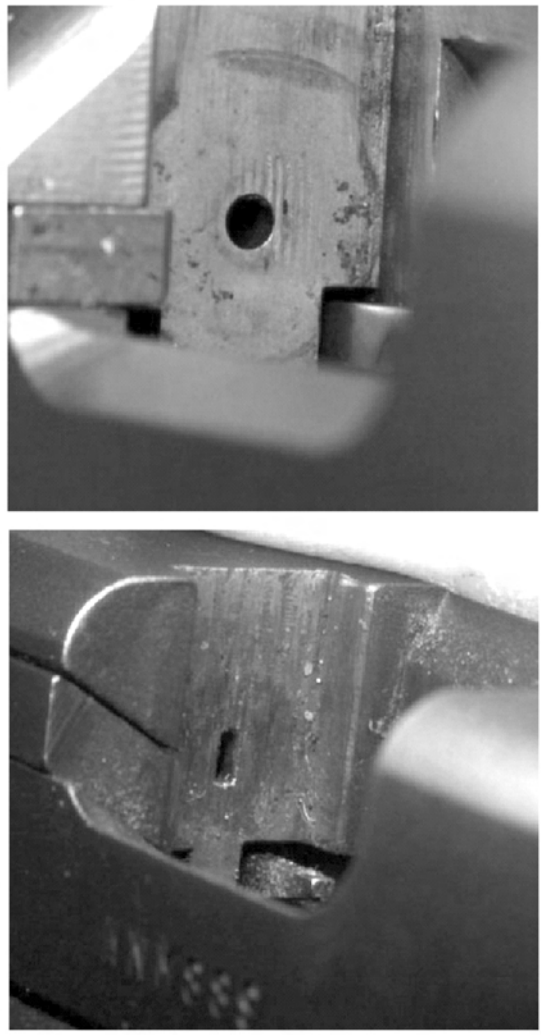


FIGURE 2-1 Breech faces with firing pin holes: Two firearms.
NOTES: The top image is the breech face of a Smith & Wesson firearm; the bottom image is the breech face of a Glock firearm. The shape of the firing pin hole for the Glock firearm indicates its characteristic rectangular firing pin.
SOURCE: Excerpted from Tulleners (2001:Fig. 3-3).

small rounded end or nose, some have more distinctive shapes; in particular, Glock firearms are known for a rectangular firing pin. Firing pins are generally made on a standard screw machine. Like the breech face, the tip of the firing pin is subject to machining and filing steps that impart microscopic imperfections.

Revolvers and Pistols

Handguns may be divided into two basic types—revolvers and pistols—by the manner in which ammunition is loaded and cycled through the firearm.

In a revolver, "the supply of ammunition is held in a cylinder at the rear of the barrel with each round having its own chamber," and a ratchet mechanism is then used to cycle the cylinder to the next position (Heard, 1997:18). Revolvers may be further subdivided by the manner in which this cycling is performed. In *single-action* revolvers, the shooter manually cocks the hammer, pulling it back and setting the ratchet action in motion. A trigger pull then causes the hammer to drop and commence the firing process. More complex—and more common—*double-action* revolvers save a step: "A long continuous pull on the trigger cocks the hammer, rotates the cylinder, then drops the hammer all in one operation" (Heard, 1997:18).

By comparison, pistols are self-loading, making use of ammunition "contained in a removable spring-loaded magazine housed within the grip frame." Pistols have a single chamber, and individual rounds of ammunition are cycled into the chamber by mechanical means; pulling back the slide rearward until the breech face is behind the top round in the magazine, and then releasing it, forces the round forward and into the chamber for firing. After firing, the spent cartridge case is ejected "through a port in the side, or occasionally top, of the slide. At the end of its rearward motion, the spring-loaded slide moves forward[,] stripping a fresh round off the top of the magazine and feeding it into the rear of the barrel" (Heard, 1997:19).

Pistols are often referred to as *semiautomatic* pistols (or semi-automatics); they are semiautomatic in that they are self-loading but require separate, distinct trigger pulls to fire different rounds. "Automatic" is used to describe "a weapon in which the action will continue to operate until the force is removed from the trigger or the magazine is empty." Though a few fully automatic pistols have been marketed, they are rare "due to the near impossibility of controlling such a weapon [for accurate shots]. . . . Each shot causes the barrel to rise during recoil and before the firer has time to reacquire the target within the sights, the next round has fired"; consequently, "even at close range it is unusual for more than two shots to hit a man-sized target" (Heard, 1997:17, 18).

For the objective of the recovery of ballistics evidence and imaging

thereof, the distinction between revolvers and pistols is vital: while pistols forcibly eject spent rounds, revolvers do not. Hence, casings may only be recovered at a crime scene involving a revolver if they are specifically emptied by a shooter (e.g., for reloading).

Extractor and Ejector

Both revolvers and pistols make use of an extractor, typically a small arm that fits over the rim of the cartridge. As the name implies, the extractor serves to pull a spent cartridge from the chamber so that a new cartridge can take its place. In a revolver, the extractor—which can remove all cartridges simultaneously by depressing the ejection rod (or extractor rod)—also has ratchet notches that advance the cylinder to the next chamber. In a semiautomatic pistol, however, the extractor removes the cartridge so that it makes contact with the ejector, typically a fixed protuberance that strikes the rim of the cartridge. Because these steps are performed very quickly, and with some speed and force, both the extractor and ejector mechanisms can leave marks on expended cartridge casings.

2–A.2 Ammunition

Modern ammunition takes the form of integrated, self-contained *cartridges*, integrating three key elements in one unit:

- a *bullet*, the actual projectile that is expelled from the firearm's barrel;
- *propellant*, which generates the force and pressure needed to put the bullet in motion and into flight; and
- a *primer*, which in modern usage is a volatile and pressure-sensitive chemical mixture that is responsible for igniting the propellant.

Historically, with firearms of the 18th century, shooters had to assemble these components manually in order to reload, inserting black gunpowder, wadding, and a spherical lead ball into the gun's barrel. With the intent of making reloading faster, early cartridges featured premeasured and prepackaged charges of powder, in small bags, but they still required an external source to provide a thermal "flash" to ignite the powder and fire the projectile. The innovation of the breechloader, by which the ammunition is loaded at the rear of the gun's barrel, made modern integrated ammunition possible. Modern ammunition links these three components together, placing them inside an outer *case*.

Ammunition is commonly identified based on the diameter of its bullet, for proper fitting with firearms barrels. The original designation of ammu-

nition size was by caliber: The unit of measurement was hundredths of an inch (e.g., .38 caliber corresponding to a bullet with diameter 0.38 inches). However, such caliber labels are only approximations, for example, a .38 caliber is actually 0.357 inches in diameter and a .40 caliber is actually 0.429 inches in diameter. Ammunition (and corresponding gun barrels) are also now identified using the metric system, such as 9mm or 10mm.[2]

Ammunition cartridges are primarily divided into two categories—rimfire and centerfire—depending on where the primer is located (and, correspondingly, where the gun's firing pin strikes the cartridge during firing). We explain the distinction in the next section.

Primer

The use of a chemical primer to ignite the propellant dates back to the development of the percussion cap in the early 1800s, when it was discovered that striking a cap containing fulminate of mercury created a flame that could then move into the main charge of powder. Today, the exact chemical composition of primer mixtures can vary and remains proprietary. "Lead styphnate is the main ingredient," generally, although individual primers may also include some of the following: "[trinitrotoluene (TNT)], lead or copper sulphocyanide, lead peroxide, sulfur, tetryl, barium peroxide, and barium nitrate" (Rinker, 2004:19). Ground glass may also be added as a "sensitizer," to create friction when impacted by the firing pin (Matty, 1987:10). A primer mixture is a high explosive; working with it and placing the primer in the case are extremely sensitive parts of the ammunition manufacture process.

Rimfire cartridges were first developed in the 1800s, and rimfire ammunition remains in heavy usage in .22 caliber cartridges. As the name implies, "the primer composition is spun into the rim of the cartridge case," putting it in immediate contact with the powder propellant (Rinker, 2004:19–20). By comparison, *centerfire* ammunition has a cylindrical cap seated in the cartridge head that contains the primer mixture. The cap consists of a cup-

[2]Care is needed with the use of the word "caliber." Here, "caliber" is shorthand for the *nominal caliber* of the ammunition, which refers specifically to the diameter of the bullet. However, *specific caliber* of ammunition "refers to a name given to a cartridge representing the entire design of the cartridge as intended by the manufacturer, [including not only] the diameter of the bullet but the entire shape and size of the cartridge" (Moran, 2000:235). That is, a nominal-caliber ammunition group may include a wide variety of specific varieties that can vary significantly in their length, case design, powder charge, and so forth. Both "nominal caliber" and "specific caliber" are used to describe and label firearms as well, referring to the "group of firearms which share the same bore diameter" and the "name given to a firearm representing the specifically designed cartridge which will fit into the firearm," respectively (Moran, 2000:235).

and-anvil combination and a pellet of primer mixture. During firing, the firing pin "compresses the primer composition between the cup and anvil," causing a flame that passes through a hole or vent to ignite the propellant charge (Rinker, 2004:19). Practically, the development of the centerfire system "was the great milestone in weapon and ammunition development;" with it, "only the primer cup needed to be soft enough to be crushed by the firing pin," freeing the main body of the cartridge case to be harder, providing "a gas seal for much higher pressures than could be obtained with rimfire ammunition" (Heard, 1997:11). Centerfire cartridges also developed, in part, due to the desire to reuse "the most expensive part of the cartridge, the case"; the centerfire configuration permits new primer assemblies to be inserted into expended casings (Matty, 1987:8).

Given its purpose, the primer assembly must meet specific criteria. The primer mixture "must always have a uniform flash that is hot enough without being too violent. In other words, it must always consistently produce the proper amount of heat" (Rinker, 2004:20). Likewise, the material holding the primer—either the cartridge brass of the rim in a rimfire cartridge or the cup in a centerfire primer—must withstand the impact of the firing pin, the detonation of the primer, and the expansion of gas from the ignited propellant without rupturing. Centerfire primer cups are typically brass or nickel.

Propellant

Though it derives from centuries of development, a critical part of ammunition is subject to popular misunderstandings and mislabelings. It is commonly referred to as *powder*, tracing from ancient formulations of black powder and more modern incarnations of smokeless gunpowder. As Hatcher (1935:96) observes, powder "originally meant, and still does mean, fine dust; but at the present time we find substances called powder which do not in any manner resemble dust and which are not even finely divided." *Propellant* is a more generic and more apt term for the substance used in modern ammunition. The individual particles of propellant may still be referred to as grains, even though they may not have a gritty or granular texture; however, the common use of *grains* to describe the exact quantity or charge of propellant in a cartridge has nothing to do with texture (a grain is a measured weight equal to 0.0648 grams).

Fundamentally, a propellant is not devised to *explode* violently: It is designed to *burn*, and burn rapidly. As Rinker (2004:21) summarizes, "all gunpowder produces the force to move a projectile as the result of 3 things. (1) When it burns, it produces a huge quantity of gas. (2) As it burns, it produces a huge amount of heat. (3) After ignition, it creates its own oxygen and needs no outside air. All three are required. At first, the need

for heat may not be as obvious as the other two, but hot gas expands and requires more space then cold gas," heightening the buildup of pressure in the gun's chamber.

Modern propellants are a form of nitrocellulose, first discovered in 1846 when cotton, nitric acid, and sulfuric acid were mixed. One pound of nitrocellulose-based powder contains 1.2–1.5 million foot pounds of stored chemical energy, in comparison with about 600,000 foot pounds of stored energy in one pound of the traditional saltpeter, charcoal, and sulfur combination of black powder (Rinker, 2004:23). "If ignited in an unconfined space," nitrocellulose propellant will burn gently; if, however, combustion occurs in a confined space—as in a cartridge—"the heat and pressure built up will accelerate the rate of combustion exponentially" (Heard, 1997:76). The charge of propellant utilized in cartridges is carefully tuned to the caliber, bullet weight, barrel length, and desired performance of the ammunition. Chemical "moderating" agents or other additives (e.g., graphite or barium nitrate) are often used to control the burn rate of the propellant, and the mixes used in final propellants are "very tightly-controlled trade secrets" (Heard, 1997:59).

Cartridge Cases

Cartridge cases have traditionally been manufactured from brass, an alloy of copper and zinc, although other materials have been used; in particular, steel casings (coated with copper or a lacquer) were developed during World War II due to brass shortages, and steel cases remain in use in some countries because of their lower cost. Cartridge brass is almost universally of the same composition: a 70-to-30 or 75-to-25 alloy (in percentage of weight) of copper and zinc, respectively. This combination was developed, along with methods for working with it, as a result of the physical demands put on the case during the firing of a gun. As described below, a cartridge case expands during firing, pressing against the chamber walls to create a seal and containing the high-pressure gases created in firing. To accomplish this *in situ* deformation, the hardness of the cartridge brass must be precise so that the case retains its original shape and can be readily extracted from the breech. Too hard a starting brass and the case may crack during firing; too soft and it will expand and deform too much and be difficult to extract. Although there are a number of manufacturing processes currently used to produce cartridges, the salient features of the general manufacturing process are similar. Within the same case, thickness must also vary in particular ways, tailored to suit various tasks: maximum hardness in the rim (of a centerfire cartridge) in which the primer cap is seated, medium hardness with good elasticity in the central walls of the case, and softest at the neck or mouth end where the bullet is seated.

One modern manufacturing process for producing a centerfire case starts with brass rod or wire, in coils. A machine called a cold header, similar to the one used to make common nails, feeds in the rod or wire, cuts off a piece large enough to make one case, and transfers it to a cavity in the machine, where it is struck by a punch. This process forms the irregularly shaped cylindrical piece into a precise sort of button shape. The button is annealed (heated and then cooled) to reduce its hardness, and is then fed into a two-stage transfer press that transforms the cartridge blank into a low, wide cup. The half-formed cup is next pushed through a die or series of dies that draw the blank to its final shape and dimensions. Additional annealing, cleaning, and forming steps are done sequentially until the blank is in the final shape of the cartridge case.

Bullets

The last major component of the cartridge is the bullet or projectile. Bullets in modern ammunition can consist of a variety of metals. There are bullets made entirely of aluminum, steel, and sometimes brass; nonmetallic substances like rubber and wood have also been used to make bullets. However, to provide the needed weight for improved accuracy and performance, bullets most often contain some amount of lead.

Bullets are designed for two basic purposes—penetration on impact with a target and perforation and expansion to increase damage—and the exact composition and construction of bullets are tailored to those purposes. An all-lead bullet is very soft and therefore expands rapidly on striking a target. Indeed, "pure lead is not used for lead bullets" precisely because "it is too soft [and] damages too easily in handling and loading"; antimony is most commonly added to lead as a hardening agent, though tin has also been used (Frost, 1990:27). Better penetration power at greater distances and accuracy can be attained by covering a lead core with a full jacket or partial jacket composed of a copper alloy. High-velocity, fully jacketed bullets are designed to penetrate deeply, while lower velocity jacketed bullets may tumble within the target and cause additional damage due to expansion. Mushrooming or expanding bullets, such as hollowpoints, are designed to transfer a maximum amount of energy to the target and to penetrate but not exit. The composition and design of bullets—along with what materials they do or do not strike—are important to forensic ballistics analysis as they affect what condition a recovered bullet will be in and hence how difficult it is to match to other evidence.

A lubricant is applied to bullets before they are seated in cartridge casings; it acts to cut down on metal fouling of the bore, the deposition of particles or residues from the bullet (Frost, 1991:31). In centerfire cartridges, where "grease grooves" are created in the case by knurling, the

lubricant is usually a wax or heavy grease type; due to its placement, it must be a substance that will neither contaminate the powder nor react with lead or copper plating.

2–B THE FIRING OF A WEAPON: INTERNAL BALLISTICS

The general concept of "ballistics" can be divided into separate stages; see Box 1-1. External ballistics (the flight path and behavior of the bullet between its exit from the barrel and its arrival at its target) and terminal ballistics (behavior of the bullet on striking a target) are both critical to complete firearms investigations.

Our primary focus is on internal ballistics—the actions that occur between the pulling of the trigger and the bullet's exit from the barrel of a firearm. Internal ballistics is "a series of actions or operations that every firearm must go through, whether .22 caliber revolver or a .50 caliber machine gun," all of which occur in a time span on the order of 0.003 seconds (Rinker, 2004:1, 2). The trigger pull starts the mechanical process of allowing the firing pin to strike the primer of the chambered cartridge. The pressure from the firing pin creates a dent in the primer surface of the cartridge; more significantly, it causes a small explosion, the heat from which passes through the hole in the primer cap and into the main body of the cartridge. There, the charge of powder burns rapidly in a confined space, converting from a solid to a gas and exerting great pressure against all surfaces. "When the pressure has built up to a sufficient level, known as *short shot*, the bullet will start to move because the pressure is greater than the holding force of the case neck." As the powder burn continues, "the pressure increases and the neck and body walls of the case expand to meet and grasp the inside chamber walls," creating a seal and increasing the pressure acting on the bullet's base, propelling it forward (Rinker, 2004:1). The bullet, being slightly larger than the barrel diameter, is forced to seat into the rifling (the lands and grooves) on the bore of the barrel, picking up rotation as it passes down the length of the barrel.

While this sequence of events drives the bullet through the barrel and out of the firearm, forces are also at work on the head of the cartridge. Hatcher (1935:270, 272) describes the processes for a centerfire cartridge:

> When a primer is struck by the firing pin, the very brusque and powerful mixture that it contains explodes with violence, [causing the flame that ignites the powder charge]. But the explosion of the primer mixture also reacts in a backward direction onto the primer cup itself, and blows it part way out of the primer pocket, unless the primer is strongly crimped in place, as is done with some kinds of rifle ammunition. Then when the main charge ignites, the powder pressure inside the case forces the case

> back sharply against the breech face or recoil plate, and this action seats the primer again. . . .
>
> When the material of the primer is very soft, or the breech pressure is very high, or more particularly if the soft primer has a very strong mixture in it and the vent hole is small, the metal forming the surface of the primer cup often is forced back more or less into the firing pin hole in the breech block, thus leaving a raised rim all around the firing pin impression.

The firing pin is often not fully retracted, and so it may impact the casing multiple times (Krivosta, 2006:42). Likewise, the firing pin may scrape or drag somewhat against the edge of the surface.

Also emitted from the barrel as a result of firing is gunshot residue, a mixture of partially burned and unburned particles of propellant, leftover primer mixture, and particles of metal and lubricant from the release of the bullet and its passage through the barrel. Some residue may also remain in the barrel and possibly on other internal surfaces of the gun; with time, and in the absence of cleaning, these residues can build up and alter the surface to which the bullet and cartridge case are exposed during firing.

2–C BASIC TOOLMARKS ON BALLISTICS EVIDENCE

2–C.1 Cartridge Case Markings

Breech Face Marks

Gas pressure created during the firing process exerts pressure in all directions, including forcing the head of the cartridge against the breech face. Hence, the surface area of the cartridge head may pick up negative impressions of any linear striations or other features left on the breech face when it is filed and machined. Some of these marks may register on the relatively hard cartridge brass that forms the outer ring (head stamp area) of the cartridge case, but most of the features show up in the softer surface of the primer cap. Hence, what is known as the breech face mark is the pattern of linear striations and other textural features on the surface of the primer, surrounding the indentation of the firing pin impression. Figure 2-2 illustrates the breech face marks and firing pin impression for two different firearms, one Glock and one Smith & Wesson.

Hatcher (1935:265–266) provided an early description of the breech face mark and recognized the mark's importance as a potentially identifiable feature:

> In both [semi]automatic pistols and revolvers there are certain fine tool marks or scratches left on the breech face or the metal against which the

FIGURE 2-2 Breech face markings and firing pin impressions for three ammunition types and two firearm brands.
NOTE: S & W = Smith & Wesson.
SOURCE: Adapted from Tulleners (2001:Fig. 3-4).

> cartridge presses when it is being fired. These marks are quite pronounced on metal surfaces that have been finished by a file as is commonly done on the breech face of the average [semi]automatic pistol or revolver. Examined under a microscope this surface appears to consist of a number of ridges or scratches, and when the cartridge is fired, the primer, being of copper or brass, which is much softer than the steel of the breech face, will take the impression of these fine ridges.

In gross appearance, features in the breech face impression may fall into some general categories depending on the specific filing or polishing steps used by the manufacturer. Straight filing creates linear features; other breech face impressions may feature cross-hatching or circular patterns. For example, Kennington (1995) documents the class of 9mm pistols for which the rotary cutting tool used in milling the breech face not only leaves distinctive arched markings that are impressed on the primer surface, but

may also be evident elsewhere on the cartridge head. Kennington suggests that the rifling characteristics from bullet evidence at a crime scene can be combined with evidence of arched markings on cartridge casings to rapidly identify the pistol make in question.[3]

Because breech face impressions are created by the pressure of firing, Tulleners (2001:3-2) notes that their detail "is dependent on cartridge chamber pressure and the type of breech face manufacture/condition. [Chamber pressure varies within caliber and depends on such factors as the bullet size and weight and the powder charge contained in the cartridge.] Lower pressure cartridges are not expected to consistently produce decent breech face impressions." He adds that cartridge chamber pressure, bullet weight, and primer hardness "can vary to such an extent that an examiner will not be able to identify test 1 to test 2 when different ammunition is used in the same gun;" hence, "one of the cardinal rules in firearm examination is to test fire the gun with similar ammunition as the evidence ammunition if at all possible" (Tulleners, 2001:3-3).

Firing Pin Impressions

The firing pin impression on the surface of the primer provides important information on the general class of the firearm that discharged a casing. The shape of the "pit" marking the firing pin's strike indicates the shape of the firing pin in the firearm (e.g., round, elliptical, rectangular). The firing pin impression will also bear the marks created by filing or smoothing the tip of the firing pin. "The point of the firing pin will have small ridges, and no two . . . firing pin points will be exactly alike," conjectured Hatcher (1935:266). However, Burrard (1962:113) notes that "great caution is necessary" in distinguishing individual markings from grosser features of firing pin marks, which "often take the form of a number of small concentric rings." Yet individual imperfections on the tip of the firing pin can be telltale: "Another by no means rare feature of a [firing pin] is the presence of a small 'pimple' on the extreme end," and so the presence of a corresponding mark on one cartridge and the absence on another "would be proof positive that the [second] cartridge could *not* have been fired" from the same weapon as the first.

For some guns and some firings, the firing pin impression may not be a clearly defined indentation on an otherwise flat surface. Instead, primer "flowback" may occur: a larger crater is created as the primer material

[3]However, he cautions that "the arch-producing machine process . . . may not be the final breechface treatment at the factory. The breechface can still be broached, filed, sandblasted, tumbled and/or plated," and residue buildup as a result of firing can obscure the arch markings.

around the pit is forced outward by gas pressure, partially flowing into the aperture in the breech from which the firing pin emerges. Though "flow-back" is commonly attributed to firearms in which excessive pressure can build during firing, Kreiser (1995) suggests other explanations that also correspond to characteristics of the particular make of firearm. Among these is the diameter of the firing pin aperture: the wider the aperture, the more primer surface is unsupported (not positioned directly against another object) during firing and hence more likely to crater outward.

In some firings, the firing pin may scrape against the surface of the primer as it is withdrawn. In these cases, the firing pin impression is not purely a mirror of the shape of the firing pin (e.g., circular) but has a drag mark trailing away from the main impression. Because drag marks may be repeated—that is, they may be a function of the behavior of the firing pin in a particular gun—they become important landmarks for traditional firearms identification and ballistic imaging alike, providing a benchmark to orient casings consistently. It is also important to note that the mechanics of firing is such that there is variability in the exact position where the firing pin impacts the cartridge across different firings; the pin may wobble slightly and strike at slightly different points and angles.[4]

In rimfire weapons, the firing pin strikes the brass of the outer rim of the cartridge head. As Hatcher (1935:68) observed, "[rimfire ammunition] takes a good impression showing the shape of the firing pin, but it does not often take a clear impression of the fine file marks and other irregular scratches on the breech block, which form the 'finger-prints' of the gun." Accordingly, he noted that "when an empty rim fire cartridge is found at the scene of a shooting, it is often easy to say what type of arm was used; but it is seldom possible to identify a rimfire cartridge to a definite individual gun by the impression of the file marks it left on the head, as is so often done in the case of a center-fire cartridge."

Ejector Marks

The ejector arms in automatic or semiautomatic firearms can vary in shape (e.g., rectangular, round, or triangular) and size; the footprint of the ejector determines the size and shape of the mark left by the ejector on the rim of the spent casing. Ejector marks can vary from tiny divots to

[4]Fadal (1995) provides an unusual but vivid example of the difference that placement and angle of the firing pin strike can have on the resulting marks. The Hi-Standard Model DM-101 is a .22 caliber derringer handgun that is double-barreled; however, the same rectangular firing pin is used to initiate the firing in each of the two barrels. The difference in the way the same pin hits the (rimfire) casings in the two barrels—one using the top part of the pin and the other the lower—is sufficiently large that an examiner cannot match firings from one barrel to firings from the second barrel on the firing pin marks alone.

more substantial indents on the cartridge head near the rim. Analysis of ejector marks can be made more difficult by the fact that the rim of the cartridge head is also where ammunition makers put their headstamp (brand identifier) and information on the size and caliber of the cartridge. These heavy-set alphanumeric characters are inscribed on the cartridge brass and—depending on where the ejector happens to hit—parts of the stamp may bleed into the ejector mark.

In addition to the shape of the ejector mark and any individual scrapes or textures therein, ejector marks also serve the same important purpose as a firing pin drag mark: They provide a point of reference for proper orientation of cartridge cases relative to each other in comparison.

Other Markings

During the firing process, gas pressure works on all surfaces, forcing the material of the cartridge against the chamber of the weapon; particularly in semiautomatic weapons, other firearms parts are used to circulate ammunition through the weapon and eject spent casings. These actions and parts can lead to a host of marks on the cartridge case that—though not imaged using current techniques—are sometimes used by examiners studying matches between pieces of evidence.

Chamber marks are parallel striated marks along the outer walls of the cartridge case, impressions from the scraping used to bore or ream the chamber (along with the rest of the barrel) from a solid piece of alloy. The extractor in a pistol that helps move a spent cartridge out of the chamber is typically a small arm that fits over the rim of the casing, holding it as the breech assembly slides backward. Accordingly, the extractor can leave marks where it makes contact, either on the edge of the rim of the cartridge head or on the neck separating the head from the main body. The slide that moves back and forth in semiautomatic pistols, allowing ejected casings to move away from the weapon, may leave a scuff mark on the edge of the cartridge head and a rough drag mark along the cartridge wall. As individual cartridges move from a magazine into chamber, a mark on the outer wall of the case may be caused by the magazine lip.

2–C.2 Bullet Markings

Hatcher's (1935:255) seminal text on firearms identification referred to "the fine ridges and grooves on the surface of the bullet, parallel to the rifling marks," as "the most important individual characteristics which are used" in the field. These marks on the bullet—known as striations or striae—"are caused by its passage over surface irregularities and rough spots on the interior of the gun barrel that got there principally during

the machining operations of reaming the bore and rifling the grooves. Any such machining operation will leave the bore at least slightly rough, and each rough spot will leave a mark on the bullet during its passage through the bore."

The rifling carved into the barrel takes the form of grooves separated by raised areas, known as lands. These lands and grooves create corresponding engraved areas—dubbed land engraved areas and groove engraved areas (and commonly abbreviated as LEAs and GEAs)—on the bullet surface, separated by shoulders. The land engraved areas, being the part of the bullets that scrape against the raised lands on the barrel, are the principal areas of interest for observing striations.

The pattern of land and groove engraved areas on recovered bullets can be used to determine basic information about the rifling characteristics of the gun that fired them, in order to identify a class of guns from which it came. Specifically, the number of lands is an important class characteristic, as is the direction of twist evident from a side view of the bullet. Bullets (and corresponding rifling characteristics) are commonly labeled by these two pieces of information—e.g., 5R for five lands and a right-hand twist. A recovered bullet can also be measured to suggest the caliber of the ammunition and weapon. However, this is not always possible—nor is a full analysis of striation marks—due to the condition of some bullets recovered from crime scenes (and victims).

Bullets fired through weapons using polygonal rifling create special difficulties. Compared to conventional, square-edged rifling, polygonal rifling has key advantages: it reduces metal fouling, and it increases bullet velocity by reducing friction as the bullet passes through the barrel (Heard, 1997:123). However, the smoothness and subtlety of polygonal rifling can make it difficult to discern even gross features on recovered bullets—the shoulders defining lands and grooves—much less fine individual detail. Heard (1997:131) concludes that "generally speaking it is possible, although extremely difficult, to match bullets from polygonally rifled barrels."

2–D THE MANUFACTURING OF FIREARMS AND AMMUNITION

The underlying theory of firearms identification depends critically on manufacturing processes, positing that the tools used to form component parts wear with use so that each part may share the same gross features yet differ in microscopic (and, presumably, uniquely individual) ways. Manufacturing processes are also essential to consider in assessing the costs and benefits of wide-scale ballistic imaging or alternatives such as microstamping. Introducing stages to the process of producing firearms or ammunition—for example, systematic test-firing to produce exhibit cases, imaging of exhibits in large batches, or laser-etching a unique mark on the

base of a bullet—can have major impacts on the cost of production and, perhaps, the feasibility of compliance with proposed changes.

We have already touched on some aspects of manufacturing in describing the anatomy of firearms and ammunition earlier in this chapter, and aspects of manufacture will arise in Chapter 3 as well (particularly in discussing challenging issues for firearms identification, generally). This section introduces basic issues but is not a comprehensive discussion.

2–D.1 Firearms

The manufacturing of most guns is highly automated and generally efficient, and as many as 5 million new firearms (domestic and foreign) enter the U.S. market each year. Befitting its historical development, dating to Samuel Colt's popularization of interchangeable parts and production line assemblies, the modern firearms industry remains one that is characterized by solid process control. That is, the process of mass-producing firearms is one that can be well partitioned: constituent parts of a new firearm can be drawn from large bins of fairly standardized parts and automatically fitted together with low yield loss, resulting in weapons of reasonably identical properties in terms of size, weight, and performance.

Yet individual manufacturers differ on the exact steps used in machining and assembling firearms, and choices on the amount of filing or polishing to do on firing pins or whether to apply paint to the breech face can have an impact on the resulting toolmarks. In addition, some manufacturing techniques affect the type and quality of marks created in firing. Champod et al. (2003:307) argue that "machining marks made by grinding, filing and some other machining methods are random and hence we expect no repeatability between tools." In comparison, "machining marks made by stamping, some cutting processes such as broaching, and some forging processes may be repeatable."

Various manufacturing techniques used by Lorcin Engineering drew interest in the 1990s, as firearms produced by the firm became more widely used in crimes;[5] they serve as useful illustrative examples. Thompson (1996:95) found two Lorcin L9MM semiautomatic pistols, bought at the same time, that produced sufficiently similar breech face markings that a match could be made to either weapon on that mark alone; they could, however, be distinguished by sidewall and extractor marks. Similarly, Matty

[5]In 2000, the Lorcin L380 semiautomatic pistol was the most traced firearm after recovery from juvenile possessors, and a Lorcin .25 caliber pistol ranked seventh. The L380 was also traced with high frequency after recovery from older offenders, ranked second among firearms recovered from 18–24-year-olds, and ranked third among firearms recovered from adults aged 25 and older (U.S. Bureau of Alcohol, Tobacco, and Firearms, 2002:15–16).

(1999:134) reports on a case where a search on a DRUGFIRE database—an initial competitor to the current Integrated Ballistics Identification System (IBIS) for ballistic imaging (described in Chapter 4)—suggested enough similarity to cause the physical evidence (both test-fired cartridge casings and the recovered Lorcin L9MM that produced them) to be retrieved from storage. On more detailed examination, "the breech face signatures were similar, but there was insufficient detail for an identification"; however, chamber and extractor marks failed to coincide at all.

"The heavy black 'paint' that adhered to the breech face" was originally believed to be a cause of this phenomenon (Thompson, 1996:95).[6] Ultimately, though, it was attributed to the fact that the breech faces for that model being formed by stamping, with no further grinding. In earlier Lorcin models, "the breechface area would become battered during firing as [a relatively soft alloy slide] hit the rim of a cartridge in the magazine as it fed the cartridge into the chamber"; this caused the breech face markings to be unstable and to change from firing to firing (Matty, 1999:135). Lorcin revised its process—in newer models, "a solid stamped steel insert is placed into a non-ferrous alloy slide"—but this stamped steel insert is prone to have marks that "can carry over from one steel insert to another" (Tulleners, 2001:3-4). (This phenomenon is an example of subclass carryover, discussed in fuller detail in Section 3–B.1.)

More generally, Collins (1997:498) observed that "the bullets and casings of the [Lorcin] L380 [.380 caliber semiautomatic] pistol are easy to characterize. The bullets exhibit slippage[7] and/or extremely shallow land impressions that often make even shoulder location difficult to determine," and even "breech face marks are either non-existent or change from shot to shot." Collins' specific inquiry into the manufacturing pistol was based on attempting (unsuccessfully) to replicate crescent shaped marks observed in some firings, imprinted directly below the firing pin impression and believed to be caused by peening of the breech face surface under repeated firings.

Another example of manufacturing processes that can directly affect the marks left by firearms and the ability to match them is the button rifling technique used by some manufacturers, notably Hi-Point (Roberge and Beauchamp, 2006:166):

[6]A thick coat of black paint was also judged to be the probable cause of highly similar breech face marks produced by two different 45 ACP Haskell semiautomatic pistols; individual characteristics would emerge on the breech face marks for each gun with repeated firings, as the paint chipped and wore off (Tulleners, 2001:3-4).

[7]"Slippage" means that a bullet does not fully grip the rifling on the barrel interior; hence, it can wobble and shift, rather than following the clear path of the rifling (and having marks carved into the side of the bullet as it passes through).

> This process creates the grooves in the barrel by compressing rather than removing the excess material resulting in a relatively shallow barrel groove. Another distinct characteristic of the Hi-Point barrels is the metal tailings left along the shoulder of the groove. The combination of button rifling and metal tailings creates a relatively smooth barrel with very coarse shoulders. With each shot fired, all or part of the metal tailings break off changing the coarse stria on the fired bullet. The shallow rifling also allows a great deal of slippage to occur. Furthermore, the crowning[8] of these barrels can add additional subclass characteristics.

All newly manufactured firearms are required to bear a unique serial number, and this number may be stamped or etched on various parts of the firearm frame and assembly. However, guns with consecutive serial numbers are generally not consecutively manufactured in full. Production of firearms is typically an assembly line process, drawing various preconstructed parts from large bins for assembly into a finished weapon. Hence, two firearms that bear consecutive serial numbers may have rolled off the line in sequence, but their frames, barrels, firing pins, and so forth need not have been manufactured right after each other. There are some exceptions to this rule; for instance, Lardizabal (1995) found that consecutive serial numbers in a set of Hechler & Koch 9mm USP semiautomatic pistols meant that the slide for these weapons had in fact been consecutively manufactured.[9]

2–D.2 Ammunition

Like firearms, ammunition cartridges are the result of numerous tooling and machining operations, and individual manufacturers vary in the specific techniques they use. It is standard practice for manufacturers to apply a head stamp, engraved on the rim of the cartridge head, to identify the manufacturers and perhaps the specific make of the ammunition; they may also use colored paints or other indicia to differentiate between specific makes and calibers. Ammunition manufacturers also vary in some post-processing steps, such as the application of a lacquer sealant to the primer surface. "Primer sealants are routinely applied to centerfire cartridges to increase the power and reliability of the ammunition," "placed at the junction between the primer and the primer cup [to] create a water and airtight

[8] "Crowning" is a finishing step on the muzzle or discharge end of a barrel, rounding or grinding the mouth so that it is flush or recessed slightly and thus providing no obstacle to the bullet's exit.

[9] Lardizabal (1995:50) found that firings from a set of these pistols with similar serial numbers could not be distinguished from each other by any mark, and this "persistence of detail" continued through 250 firings. A pattern of striations was observed on the breech face itself, above the firing pin hole; this mark appeared to have been created after a chemical finishing process.

seal [and prevent] oil and other foreign matter from entering the cartridge." The sealant also makes the cartridge resistant to moisture. However, while "most ammunition manufacturers limit the application of the sealant to the junction of the primer and primer cup," some (primarily European) manufacturers "apply the sealant so that it extends across the entire surface of the primer." The Czech-made Sellier and Bellot ammunition, in particular, is known for a red lacquer sealant over the entire primer (Hayes et al., 2004:139). The lacquer can act as a cushion, "absorb[ing] and dissipat[ing] a greater amount of energy" when involved in a collision (compared with metals), and consequently "reduc[ing] the amount of energy that reaches the metal surface of the primer" (Hayes et al., 2004:142).

The specific techniques of a manufacturer can combine with more ornamental and postprocessing steps to leave distinctive marks on the cartridge. Box 2-1 reviews these nonfiring manufacturing marks—features that are present on the cartridge before firing and traces of which may endure after firing. In comparing exhibits, firearms examiners must compensate for the presence of these nonfiring marks, lest they lead to a false identification or exclusion. While many of these nonfiring marks are deliberate design choices, others arise inadvertently due to other steps in manufacture. Yborra and McClary (2004) report finding distinct striated markings near the edge of the primer surface on a batch of 115 grain Remington 9mm Luger ammunition. The marks appeared to be due to manufacturing and not firing: when a pair of casings was rotated so that identifying marks in the firing pin impression were in the same orientation, the extractor marks on the cartridges also lined up but the newly found striated marks on the primer surface were 90 degrees out of alignment. Remington managers indicated that they had never previously experienced such a phenomenon but suggested that a possible cause might be the way the primer is seated in the cartridge. Two separate punches drive the primer to its final position about 0.002 to 0.005 inches below the level of the cartridge head; "a misalignment or damage to one of these punches MAY have caused the observed [marks], and being machine-based, would be consistent" (Yborra and McClary, 2004:309). But no such defect could be found; nor could similar marks be detected on other boxes of ammunition from the same lot. The punches used in primer seating were also suspected of causing parallel markings near the edge of the primer on some Winchester 9mm ammunition (Flater, 2002:315); it was also suggested that the die used to flatten the surface of the primer cup could also have impressed such a mark.

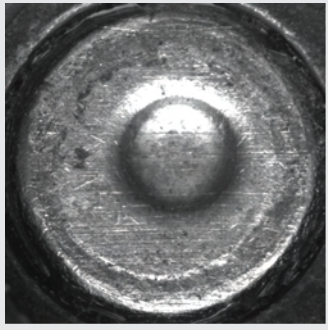
BOX 2-1
Nonfiring Manufacturing Marks

Nonfiring manufacturing marks on ammunition are features created by individual firms' manufacturing processes. They are not defects, in that they do not diminish the ammunition's performance or otherwise detract from the ammunition's quality. However, they may be mistaken for textures or striations created by the firing of a gun or that may complicate the determination of a pattern match between exhibits. Amassing knowledge of these marks—and developing the skill to adjust for their presence—is an important part of the experience of a firearms examiner.

Cataloging these nonfiring manufacturing marks, Tam (2001) suggests a rough typology based on their impact on the determination of a match between evidence: (1) marks that are not expected to cause a problem for identification (or exclusion); (2) marks that may cause problems but can be compensated for with some effort; and (3) marks that are problematic for comparison and difficult to analyze.

In the first class, there are marks that would easily be overwritten by firing-related marks, as in extremely fine pre-existing parallel marks on the primer surface. Other marks—being relatively simple and known in advance—are not problematic because the examiner can mentally compensate for their presence (e.g., a V-shaped or other stamped mark on the primer surface used to indicate certain brands). Other marks that may fall into this category are those that are on areas of the cartridge not typically considered for ballistic imaging or routine analysis, such as unique marks on the rim of the cartridge.

For the second class, manufacturing marks that may cause problems, Tam (2001) suggests that these features can be overcome by simple procedures. Marks in this class include thick striation-like parallel marks across the primer surface; these may obscure texture patterns in the breech face impression and may extend into the firing pin impression. Russian-made Wolf ammunition is well known for these marks, which have also been observed in other ammunition types. An IBIS image (using side light) of a fired round of Wolf ammunition is shown below; most of the visible horizontal parallel marks on the primer surface existed prior to firing.



Side light IBIS image of fixed casing using Wolf ammunition; heavy horizontal lines are preexisting manufacturing marks.

continued

BOX 2-1 Continued

Reitz (1975:103) observed "matchable striations on unfired primers of [exhibits from particular lots of] Winchester, .38 special cartridges." These marks were attributed to a particular punch used during the primer seating process, which had not been produced to the same smoothness as is typically the norm. "These markings remained prevalent even after firing, which could be perilous to comparison examinations by unwary examiners." Similarly, Robinson (1996:164) observed Russian-made ammunition with primers that, before firing, "had parallel marks like one might find as a result of breechface impressions." Finding that "the marks continue around the curve of the primer into the sides which were not visible," he concluded that "the only way that marks could have gotten there was by the rollers in the brass mill where the sheets of brass were made."

The third class of marks, those that are problematic for comparison, include ammunition types with existing distinct parallel and cross marks on the primer surface, making it difficult to discern which textural features were created by firing. Murray (2004:314) reports on toolmarks on the primer surface of Fiocchi .25 Auto ammunition whose cause is unknown; the manufacturer suggested that they might be attributed to a rare, imperfect configuration of the feeder during the process in which the primer is seated in the empty shell. The marks were problematic because they were not consistently prominent across the whole primer surface. When, as in the Wolf ammunition toolmarks, the markings span the whole primer, an examiner can compensate for them because they can be traced from the face of the primer into the pit of the firing pin impression. Maruoka (1994a; see also Maruoka and Ball, 1995) had previously noted parallel marks on the primer surface of some Fiocchi ammunition, but those marks did span the entire surface. But these inconsistent marks offer no such traceability, so that "differentiating these marks from breech face marks would be very difficult, if not impossible" (Murray, 2004:314). Some ammunition may also bear random marks on the rim of the cartridge that could be mistaken for ejector marks.

EXHIBIT D

TOOLMARK-COMPARISON TESTIMONY: A REPORT TO THE TEXAS FORENSIC SCIENCE COMMISSION

The Yale Law School Forensic Science Standards Practicum

May 2, 2022

TABLE OF CONTENTS

Introduction

On October 6, 2021, the Innocence Project filed a complaint with the Texas Forensic Science Commission requesting that the Commission "investigate and report 'the integrity and reliability' of toolmark and firearms analysis . . . as used in criminal proceedings."[1] The Project's "request is that this Commission set appropriate limits on the conclusions of FTE [Firearm Toolmark Evidence] examiners in traditional bullet-to-firearm matching testimony, and determine what conclusions—if any—can be proffered in other toolmark assays."[2] Following a discussion of the complaint with counsel to the Commission, the members of the Yale Law School Forensic Science Standards Practicum reviewed case law on the admissibility of toolmark-comparison testimony; forensic-science standards; legal, statistical, and scientific literature on this type of pattern and impression evidence and on the presentation of forensic-science identification evidence generally.[3]

This report describes the range of approaches that courts, legal commentators, and scientists have proposed for presenting toolmark-comparison evidence in trial settings. The report focuses on "identifications at crime labs … still performed manually by experts in optical (side-by-side) comparison microscopes."[4] Part I describes a line of cases imposing limitations on how expert witnesses may present the conclusions they have reached about the source of an item of physical evidence that contains toolmarks. Courts have issued limiting rulings of two kinds. Some simply forbid the use of certain expressions or types of statements. Other rulings specify the strongest form that source attributions can take to be consistent with the rules of evidence. We collect these rulings and summarize the reasoning behind them.

Part II describes the most applicable voluntary standards on FTE testimony that have been adopted or are under development. These standards presuppose that the expert should try to reach an opinion on whether the compared items originate from the same tool, but they do not state how an expert should address the issue of quantifying the uncertainty in these classifications.

Part III turns to an aspect of this last issue. It argues that, when computing false-positive "error rates"[5] from experiments that investigate the ability of FTE examiners to make accurate

---

[1] Innocence Project, Email to Texas Forensic Science Comm'n, Oct. 6, 2021.

[2] *Id*.

[3] The Yale Law School Forensic Science Standards Practicum is an experiential learning course offered in the spring semester of 2022. The participants in the practicum are Gregory Antill, Ph.D., Elaine Emmerich, B.A., R. Charlotte Ishida, M.A., Marnie Lowe, B.A., Rachel Perler, M.P.H., and Visiting Professor David Kaye, J.D., A.M. We thank Thomas Busey for meeting with us to present the study described in Part III(B) and Jonathan J. Koehler for sharing ideas on "error rates" at another session of the class.

[4] T.V. Vorburger et al., *Topography Measurements and Applications in Ballistics and Tool Mark Identifications*, 4 Surface Topography: Metrology & Properties 013002, at 2 (2016) available at https://iopscience.iop.org/article/10.1088/2051-672X/4/1/013002/pdf.

[5] In forensic statistics, the phrase "error rate" commonly designates the proportion of cases in which an error occurs in an experiment or in casework. Some statisticians use the word "rate" to mean "the frequency with which an event occurs in a defined population over a specified period of time." Centers for Disease Control and Prevention, Principles of Epidemiology in Public Health Practice 3-7 (3d ed. 2006, updated 2012), available at https://www.cdc.gov/csels/dsepd/ss1978/SS1978.pdf. A proportion is "a portion or part in its relation to the whole." OECD, Glossary of Statistical Terms (Dec. 1, 2005), https://stats.oecd.org/glossary/detail.asp?ID=6689; *see also* Joseph L. Fleiss et al., Statistical Methods for Rates and Proportions (3d ed. 2003). This report uses the term "rate" and "proportion" interchangeably.

source attributions for evidentiary use, judgments of "inconclusive" should be excluded from the analysis (although there is value in studying inconclusive results for improving the yield of definitive results). Part III also analyzes data from a new study that sheds some light on how FTE examiners might perform in non-firearms cases with a richer set of reporting categories than the conventional yes-no (or inconclusive) scale.

Finally, Part IV broadens the discussion by pointing out other alternatives to the conventional scale. These include "evidence-centric" methods in which examiners do not try to render opinions on claims about the source of the FTE. They also include additional "conclusion-centric" methods and possible enhancements to the presentation of source attribution conclusions that could better convey the general accuracy of toolmark-comparison conclusions to the factfinder.

The report does not review and synthesize the scientific studies of the accuracy, repeatability, and reproducibility of source attributions from largely subjective microscopic comparisons of FTE. Consequently, it reaches no conclusions on the scientific validity required for FTE testimony to be admissible under the Texas and federal rules of evidence. Its goal is to assist the Commission to understand the present landscape of toolmark-comparison testimony and to improve the utility and integrity of this testimony in Texas, should the Commission conclude that scientific research warrants the introduction of FTE examination findings in trials. Our major conclusion is that the existing form of FTE opinion testimony needs to be improved, and we make multiple suggestions for doing just that.

I. LIMITATIONS ON TOOLMARK-COMPARISON TESTIMONY IMPOSED BY COURTS

Firearms-and-toolmark-examiner testimony linking guns and other tools to items associated with crimes via the impressions left on the recovered items has "been 'almost uniformly accepted by federal courts.'"[6] Indeed, it is fair to say that in all jurisdictions, "long-standing tradition allow[ed] the unfettered testimony of qualified [toolmark] experts."[7] Starting in the early 2000s, however, courts haltingly began to impose limits on this testimony. Today, FTE testimony has become "an evidentiary issue of increasing interest and controversy . . . resulting in a heightened apprehension in the scientific reliability and admission of this evidence."[8] According to one federal district court, "[l]imitations restricting the degree of certainty that may be expressed on firearm and toolmark expert testimony are not uncommon."[9] In this Part, we summarize the concerns articulated in the limiting opinions and the resulting evidentiary restrictions.[10]

In focusing on these opinions, we do not claim that any one approach within the limiting opinions represents a majority rule for the admission of FTE testimony. Neither do we maintain that the case law in the majority of jurisdictions has yet departed from the traditional

[6] United States v. Brown, 973 F.3d 667, 704 (7th Cir. 2020) (quoting United States v. Cazares, 788 F.3d 956, 989 (9th Cir. 2015)).

[7] United States v. Casey, 928 F. Supp. 2d 397, 400 (D.P.R. 2013) (referring to "ballistics experts").

[8] United States v. Davis, No. 4:18-cr-000112, 019 WL 4306971, at *3 (W.D. Va. 2019).

[9] United States v. Harris, 502 F.Supp.3d 28, 44 (D.D.C. 2020).

[10] We use the phrase "limiting opinions" to refer to the ones that prohibit or otherwise constrain expert firearm or other toolmark source attributions. We use "firearm-marks" to denote toolmarks resulting from the operation of firearms.

acquiescence to "unfettered testimony."[11] There are relatively few modern appellate court opinions on firearm toolmark identification and even fewer on other toolmark testimony. Most of the small set of recent appellate cases do not seriously engage with the validity and reliability of firearm and toolmark evidence. This makes it difficult to state definitively what the law is or is becoming. The most penetrating and thorough opinions have come from trial courts and hence are not binding precedent. Nevertheless, these opinions can influence other courts and tend to be cited across jurisdictions.[12] The traditional expectation of unfettered testimony may be changing. And even if the rules of evidence do not *compel* all courts to adopt the procedures introduced in the innovative cases, trial courts possess the *discretion* to channel or respond to the presentation of admissible evidence so as to reduce the risk that jurors will overvalue expert evidence.[13]

A. The Outer Limits Dictated in the Limiting Opinions on Firearms-mark Testimony

The first modern opinion to block testimony that a given tool was definitely the source of a set of impressions, or that no other source was scientifically or practically possible, came from the federal district court for the District of Massachusetts in 2005.[14] In view of the seemingly standardless subjectivity of the Association of Firearms and Toolmark Examiners' (AFTE) theory of identification, the potential for bias, and the lack of data on error rates, the district court perceived "no accurate way of evaluating the testimony"[15] and restricted the examiner "to testify to similarities . . . but not to testify to the ultimate conclusion that two samples matched."[16] This case, *United States v. Green*, was followed by a growing line of rulings adopting various measures to curtail expert traditional firearms-impression testimony. Table 1 enumerates the formulations for source attributions that courts have ruled in as the strongest that are permissible.

---

[11] *Casey,* 928 F. Supp. 2d at 400.

[12] *But see* Katie Kronick, *Forensic Science and the Judicial Conformity Problem*, 51 Seton Hall L. Rev. 589 (2020).

[13] *See* United States v. Green, 405 F.Supp.2d 104, 123 (D. Mass. 2005) (interpreting United States v. Mooney, 315 F.3d 54, 63 (1st Cir. 2002), which upheld the admission of handwriting-identification testimony, as "suggest[ing] that the trial court has discretion to either include or exclude [aspects of] expert testimony in this context:); State v. Raynor, 254 A.3d 874 (Conn. 2020) (although it was not an abuse of discretion for the trial court to reject defendant's request to restrict testimony to a "more likely than not" statement, "our decision does not preclude trial courts from imposing appropriate limits on such expert testimony"); *see generally* 2 McCormick on Evidence § 185 (Robert Mosteller ed., 8th ed. 2020).

[14] *Green*, 405 F. Supp. 2d 104. However, *Green* was not the first time an expert was limited to commenting on the similarities of toolmarks and prohibited from expressing an opinion that a mark resulted from a particular tool. *See* David H. Kaye et al., The New Wigmore on Evidence: Expert Evidence § 15.3 (2d ed. 2010) (descripting rulings in the federal trials resulting from the 1995 Oklahoma City bombing of a federal office building that limited an FBI toolmark expert who was certain that a particular drill bit was used to open a padlock at a quarry from which explosives were stolen "to show[ing] what he saw through the . . . comparison microscope and then with his experience and training the similarities . . . .").

[15] *Id*. at 121.

[16] *Id*. at 124.

Table 1. Explicit Upper Bounds Placed on Testimony Associating Firearms with Ammunition Components

Strongest source-attribution statement allowed	*Opinions*
Pointing out features and their similarities	United States v. Green, 405 F.Supp.2d 104 (D. Mass. 2005) United States v. Adams, 444 F.Supp.3d 1248 (D. Or. 2020)[17] People v. Ross, 129 N.Y.S.3d 629 (N.Y. Sup. Ct., Bronx Cnty. 2020) (as to nonclass characteristics)[18]
Features are consistent with or fail to exclude the tool as the source of the impressions	United States v. Shipp, 422 F.Supp.3d 762 (E.D.N.Y. 2019)[19] United States v. Davis, No. 4:18-cr-00011, 2019 WL 4306971 (W.D. Va. Sept. 2019)[20] United States v. Tibbs, No. 2016 CF1 19431, 2019 D.C. Super. LEXIS 9, 2019 WL 4359486 (D.C. Super. Ct., Sept. 5, 2019)[21]

[17] The *Adams* court enumerated the similarities that could be presented as follows: "expert testimony is limited to the following observational evidence: (1) the Taurus pistol recovered in the crawlspace of Mr. Adams's home is a 40 caliber, semi-automatic pistol with a hemispheric-tipped firing pin, barrel with six lands/grooves and right twist; (2) that the casings test fired from the Taurus showed 40 caliber, hemispheric firing pin impression; (3) the casings seized from outside the shooting scene were 40 caliber, with hemispheric firing pin impressions; and (4) the bullet recovered from gold Oldsmobile at the scene of the shooting were 40/10mm caliber, with six lands/groves and a right twist." *Id*. at 1367. The court insisted that "[n]o evidence relating to [the expert's] methodology or conclusions relating to whether the shell casings matched the Taurus will be admitted at trial." *Id*. At the same time, the opinion emphasized that the ruling "is not an indictment of forensic evidence or toolmark comparison analysis writ large" because

> Even at its worst, comparison analysis has a very low rate of error and yields results that cannot be random. But it is not clear that those results are the product of a scientific inquiry. Nothing in Mr. Gover's testimony explains how or why he reached his conclusion in any quantifiable, replicable way. It is possible that the AFTE method could be expressed in scientific terms, but I have not seen it done in this case, nor elsewhere.

444 F. Supp. 3d at 1266-67. An accompanying footnote explained that "scientific terms" would have to entail "a more quantitative measure of sufficient agreement" than a "trained examiner['s] . . . impression—call it a hunch—that it is actually a match." *Id*. at 1267 n.9. "To be admissible . . . as scientific evidence, AFTE will have to shift away from hunches to numbers." *Id*.

The *Adams* court also intimated that it might have been more receptive to more pointed testimony from a firearms-toolmark expert who did not claim to be "solving crimes [through] conclusive, scientific, forensic testing." *Id*. at 1257.

[18] The *Ross* court allowed testimony of a failure to exclude based on class characteristics, but "the examiner may not opine on the significance of any [other] marks." 129 N.Y.S.3d at 642.

[19] *Shipp* explained that the expert "may testify that the toolmarks on the recovered bullet fragment and shell casing are consistent with having been fired from the recovered firearm, and that the recovered firearm cannot be excluded as the source of the recovered bullet fragment and shell casing. However, [the expert] may not testify, to any degree of certainty, that the recovered firearm is the source of the recovered bullet fragment or the recovered shell casing." 422 F.Supp.3d at 783.

[20] The *Davis* court ruled that, at most, the experts may "render an opinion as to whether toolmarks on certain cartridge cases bear marks consistent with each other." 2019 WL 4306971 at *8.

[21] 2019 D.C. Super. LEXIS at 80-81 (the "expert may testify that based on his examination, the recovered firearm cannot be excluded as the source of the cartridge casing found on the scene of the alleged shooting. . . . Any statements by the expert involving more certainty . . . would stray into territory not presently supported by reliable principles and methodology.").

Attribution opinion "more likely than not"	United States v. Glynn, 578 F. Supp. 2d 567 (S.D.N.Y. 2008)
Attribution opinion "to a reasonable degree of ballistic certainty" or close variants	United States v. Diaz, No. CR 05-00167 WHA, 2007 WL 485967 (N.D. Cal. Feb. 12, 2007)[22] United States v. Cazares, 788 F.3d 956 (9th Cir. 2015)[23] United States v. Monteiro, 407 F.Supp.2d 351 (D.Mass. 2006)[24] United States v. Taylor, 663 F. Supp. 2d 1170 (D.N.M. 2009)[25] United States v. Ashburn, 88 F. Supp. 3d 239 (E.D.N.Y. 2015)[26] United States v. Hunt, 464 F. Supp. 3d 1252 (W.D. Okla. 2020)[27] Commonwealth v. Pytou Heang, 942 N.E.2d 927 (Mass. 2011)
Attribution opinion with no statement of a degree of certainty	United States v. Willock, 696 F. Supp. 2d 536 (D. Md. 2010), aff'd sub nom. United States v. Mouzone, 687 F.3d 207 (4th Cir. 2012)
Attribution to any degree of confidence–may be "to the exclusion of" or to "a practical impossibility" for any other possible source	The traditional rule for most of the 1900s.[28]

Most of the opinions in Table 1, as well as a substantial number of others, explicitly condemn certain opinions as overstated and unacceptable. Table 2 lists particular phrases that courts have ruled are inadmissible. The 2021 briefing materials for the Advisory Committee on the Rules of Evidence refer to this line of cases[29] as a reason for the newly proposed amendment to Federal

[22] 2007 WL 485967 at *14 ("reasonable degree of certainty in the ballistics field").

[23] 788 F. 3d at 989 (dictum that "'a reasonable degree of certainty in the ballistics field' is the proper expert characterization of toolmark identification'").

[24] The *Monteiro* court ruled that "[t]he government must ensure that its proffered firearms identification testimony comports with the established standards in the field for peer review and documentation. If the expert opinion meets these standards, the expert may testify that the cartridge cases were fired from a particular firearm to a reasonable degree of ballistic certainty." 407 F.Supp.2d at 355.

[25] 663 F.Supp.2d at 1180 ("may only testify that, in his opinion, the bullet came from the suspect rifle to within a reasonable degree of certainty in the firearms examination field")

[26] 88 F. Supp. 3d at 249 ("[T]he court will limit [the expert] to stating that his conclusions were reached to a "reasonable degree of ballistics certainty" or a "reasonable degree of certainty in the ballistics field.")

[27] 464 F.Supp.3d at 1262 ("[T]he Court will permit the Government's experts to testify that their conclusions were reached to a reasonable degree of ballistic certainty, a reasonable degree of certainty in the field of firearm toolmark identification, or any other version of that standard.").

[28] On the shift from an early view that toolmarks from firearms were admissible, but expert statements source attributions as "facts" were not, to the unfettered regime noted in United States v. Casey, 928 F. Supp. 2d 397, 400 (D.P.R. 2013), *see* David H. Kaye, *Firearm-Mark Evidence: Looking Back and Looking Ahead*, 68 Case W. Res. L. Rev. 723, 724-25 (2018); *see also* 4 David L. Faigman et al., Modern Scientific Evidence § 34:2 (2021-2022) (non-firearms toolmarks); *id*. § 34:3 (firearm-marks).

[29] Advisory Comm. on Evid. Rules, Agenda for Comm. Meeting, Apr. 30, 2021 (Tab 2B, Daniel J. Capra, Forensic Case Digest 2008-Present, in Advisory Committee on Evidence Rules April 30, 2021, pp. 111-202).

Rule of Evidence 702 to clarify that the rule "does not permit the expert to make extravagant claims that are unsupported by the expert's basis and methodology."[30]

Table 2. Expressions for or About
Source Attributions Deemed Inadmissible

Inadmissible Expressions
Same-source expressions ● "reflect a 'signature'"[31] or "unique"[32] ● "were fired by the same gun"[33] ● "a 'match' to other cartridge cases or firearms"[34] ● "subjective terms such as 'sufficient agreement' or 'consistent with'" for nonclass characteristics[35]
Expressions of certainty ● "to the exclusion of all other firearms in the world"[36]; "practical exclusion of all other guns"[37] ● "certain" or "100% sure"[38]; "absolute or 100% certainty"[39] ● "practical impossibility'"[40]; "practical certainty"[41] ● "probability . . . is so small it is negligible"[42] ● "scientific certainty"[43] or "reasonable degree of scientific certainty"[44] ● "a match to an exact statistical certainty"[45]

[30] Fed. R. Evid. 702 advisory committee note to 2021 proposed amendment.

[31] United States v. Davis, No. 4:18-cr-00011, 2019 WL 4306971 (W.D. Va. Sept. 2019).

[32] Williams v. United States, 210 A.3d 734, 738 (D.C. 2019).

[33] *Davis*, 2019 WL 4306971.

[34] *Id*.

[35] People v. Ross, 129 N.Y.S.3d 629, 642 (N.Y. Sup. Ct., Bronx Cnty. 2020).

[36] United States v. Cazares, 788 F.3d 956, 989 (9th Cir. 2015); United States v. Taylor, 663 F. Supp. 2d 1170 (D.N.M. 2009) ("a match to the exclusion, either practical or absolute, of all other guns"); United States v. Ashburn, 88 F. Supp. 3d 239, 249 (E.D.N.Y. 2015).

[37] People v. Azcona, 272 Cal.Rptr.3d 471, 480 (Cal. Ct. App. 2020).

[38] United States v. Parker, 871 F.3d 590 (8th Cir. 2017); *Ashburn*.

[39] Gardner v. United States, 140 A.3d 1172, 1183 (D.C. 2016); Commonwealth v. Pytou Heang, 942 N.E.2d 927, 946 (Mass. 2011).

[40] *Davis*, 2019 WL 4306971; *Ashburn*, 88 F. Supp. 3d 239; *Pytou Heang*, 942 N.E.2d at 946; State v. Terrell, No. CR170179563. 2019 WL 2093108 (Conn. Super. Ct. Mar. 21, 2019).

[41] United States v. Jackson, No. 1:11-CR-411-WSD, 2012 WL 2513499 (N.D. Ga. July 25, 2012).

[42] United States v. Hunt, 464 F.Supp.3d 1252, 1262 (W.D. Okla. 2020).

[43] United States v. Taylor, 663 F. Supp. 2d 1170, 1180 (D.N.M. 2009).

[44] *Hunt*, 464 F.Supp.3d at 1261.

[45] United States v. Monteiro, 407 F.Supp.2d 351, 355 (D.Mass. 2006).

B. Reasons for Limitations

The opinions articulating an upper limit on the testimony that will be allowed (Table 1) or rejecting specific phrases as overstatements (Table 2) justify the limitations on various grounds. In applying Rule 702 and *Daubert v. Merrell Dow Pharmaceuticals*[46] and its progeny (or their state rules and cases on scientific evidence generally) to the question of admissibility, they have expressed the greatest dissatisfaction with the highly subjective nature of microscopic comparisons conducted under the AFTE method.[47] The absence of any controlling standard, outside the mind of the examiner looking for "sufficient agreement" in pairs of impressions, leads the limiting courts to denigrate the method of human judgment as tautological, circular, vague, and obscure.[48] Even some courts that dismiss requests "to exclude . . . testimony wholesale [as] unprecedented" find that "the lack of objective criteria governing the application of the AFTE method" is the most compelling argument for exclusion.[49]

But the inability of toolmark examiners to articulate how they decide when "sufficient agreement" is present—beyond the statement that they have acquired an internalized sense of the variations that occur with toolmarks from the same tool—does not prove that their training and experience is for naught. Many of the limiting opinions discuss the extent to which experiments have demonstrated that the process of human judgment of the similarities on the part of trained examiners is valid and reliable. Most opinions list error rates for source attributions in various studies and characterize them as low.[50] This conclusion leads courts to deny motions to exclude toolmark evidence entirely, but several courts have not found the research presented to them totally reassuring. The 2016 report of the President's Council of Advisors on Science and Technology (PCAST) dismissed the vast bulk of studies of the performance of FTE examiners as biased toward unrealistically low error rates,[51] and this critique has figured prominently in a few

---

[46] 509 U.S. 579 (1993).

[47] The AFTE "Theory of Identification as it Relates to Toolmarks" is quoted *infra* Part II.

[48] *E.g.*, United States v. Cloud, No. 1:19-cr-02032-SMJ-1, 1:19-cr-02032-SMJ-2, 2021 WL 7184484, at *7 (E.D. Wash. Dec. 17, 2021) ("The most troubling aspect of the methodology is the tautology at its heart."); United States v. Adams, 444 F.Supp.3d 1248, 1263 (D. Or. 2020) ("a tautology that doesn't mean anything"); *id*. at 1258 ("almost entirely subjective and inscrutable"); United States v. Shipp, No. 19-cr-029-NGG, 2019 WL 6329658, at *13 (E.D.N.Y. Nov. 26, 2019) ("circular and subjective").

[49] United States v. Romero-Lobato, 379 F.Supp.3d 1111, 1121-22 (D. Nev. 2019) ("With the AFTE method, matching two tool marks essentially comes down to the examiner's subjective judgment."); *see also* United States v. Chavez, No. 15-CR-00285-LHK-1, 2021 WL 5882466, at *4-*5 (N.D. Cal. Dec. 13, 2021) (that "the AFTE methodology does not have an objective standard" is the only *Daubert* factor that "weighs against a finding of reliability.").

[50] *E.g.*, United States v. Harris, 502 F.Supp.3d 28, 39 (D.D.C. 2020) ("the evidence shows that error rates for false identifications made by trained examiners is low."); United States v. Hunt, 464 F.Supp.3d 1252, 1258 (W.D. Okla. 2020) ("Other federal courts examining the AFTE method's rate of error have likewise found it to be low."); United States v. Tibbs, No. 2016 CF1 19431, 2019 D.C. Super. LEXIS 9, at *38 (D.C. Super. Ct., Sept. 5, 2019) ("The vast majority of courts have nonetheless accepted the notion that existing studies support the conclusion that the discipline's error rate is quite low—between one and two percent."). A few courts have relied on the fact that the false-positive probability for repeated, independent tests is the product of the false-positive probabilities for a single test. *E.g.*, United States v. Chavez, No. 15-CR-00285-LHK-1, 2021 WL 5882466, at *4 (N.D. Cal. Dec. 13, 2021) (assuming a false-positive probability of 2.2% per test, "[w]ith just another independent examiner, the cumulative probability of a false positive rate could be as low as 0.05%."). However, this simple multiplication is correct only when the examiners conducting subsequent tests are blinded to the outcome of the earlier tests. *Id*. at *14 n.2.

[51] Exec. Office of the President, President's Counsel of Advisors on Sci. & Tech., Report to the President: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods (2016).

limiting opinions. The district court in *United States v. Cloud*,[52] for example, agreed with PCAST that "multiple, appropriately designed, 'black-box' studies [are] essential."[53] It therefore focused on four of the most recent studies. The reported error rates were 2% or less, "[b]ut even these 'open' studies," the court noted, "have their drawbacks."[54] For this court,

> The most troubling concern . . . is that the studies allow the examiners to answer "inconclusive," even though the examiners know they are being tested. To be sure, incorrectly selecting "inconclusive" (a false-negative or Type-II error) in the field has little to no implications on the *Daubert* analysis: incorrect exclusions increase the likelihood that a guilty party is acquitted and therefore typically benefit the party objecting to proffered expert testimony. But providing examiners in the study setting the option to essentially "pass" on a question, when the reality is that there is a correct answer—the casing either was or was not fired from the reference firearm—fundamentally undermines the study's analysis of the methodology's foundational validity and that of the error rate.[55]

We discuss the implications of the "inconclusive" category in studies in Part III.

The district court in *United States v. Adams* [56] raised an issue of external or ecological validity: would examiners behave differently in an experiment than in real casework? The court connected this issue to the "inconclusive" option, writing that:

> The incentive structure for the testing process is also concerning. It appears to be the case that the only way to do poorly on a test of the AFTE method is to record a false positive. There seems to be no real negative consequence for reaching an answer of inconclusive. Since the test takers know this, and know they are being tested, it at least incentivizes a rate of false positives that is lower than real world results. This may mean the error rate is lower from testing than in real world examinations.[57]

As such, the court mused that "[i]t is hard to know exactly what to make of these results. It is possible that the error rate for toolmark testing is very low, but it is more likely that it is not."[58]

---

[52] No. 1:19-cr-02032-SMJ-1, 1:19-cr-02032-SMJ-2, 2021 WL 7184484 (E.D. Wash. Dec. 17, 2021).

[53] *Id*. at *9; *see also Tibbs*, 2019 D.C. Super. LEXIS 9, at *36-*37 (emphasizing that "determining the error rate for a particular methodology appears essential to determining its ultimate reliability" and "agree[ing] with one of the essential premises of the 2016 PCAST Report.").

[54] *Cloud*, 2021 WL 7184484, at *10.

[55] *Id*. The court added that

> Given that the items to be examined are recovered in a laboratory setting, the high "inconclusive" response rate is at least notable, even if some samples might in fact have an insufficient number of features from which a conclusive determination can be made. Whether or not inconclusive responses should be permitted in a study is thus a close question, and one researchers should consider revisiting upon designing a new study.

Id. Yet, the court wrote that if the examiner were to go beyond a statement that the gun was not excluded, it would instruct the jury on the upper confidence limits on the false-positive error rates measured in three of the experimental studies. *Id*.

[56] 444 F.Supp.3d 1248 (D. Or. 2020).

[57] *Id*. at 1265 (note omitted).

[58] *Id*. The courts in both *Adams* and *United States v. Shipp*, 422 F.Supp.3d 762, 779 (E.D.N.Y. 2019), suggested that false-positive error rates of 2% or more were disturbingly large. There are indications in the opinions that the courts regarded the false-positive probability as if it were the probability of a non-source given a reported source

The relative lack of external indicia of the quality of traditional FTE examinations has also reduced courts' confidence in the method. Courts have reservations about how "scientific" FTE techniques are when such techniques were produced by panels whose members include research scientists from outside the forensic-science establishment.[59] Another source of doubt is the status of the journal in which most FTE research is published. Judges who conclude that much of the research proffered in support of the practice is not published in mainstream scientific journals with a history of rigorous pre-publication review and widespread post-publication scrutiny[60] are more reticent to allow strong claims of accuracy from expert witnesses.

C. Other Toolmark-comparison Testimony

Recent court opinions restricting expert testimony are most common for firearms-mark testimony, but not because courts perceive the examination process to have a weaker scientific foundation for firearms marks than for other toolmarks. The method is the same, and, if anything, there is less research into the accuracy of associating impressions from tools such as screwdrivers,[61] crowbars,[62] knives,[63] and even fingernails.[64] There are fewer limiting opinions involving source attribution to other tools, probably because fewer of these examinations are performed, and fewer reports bubble up to the courts.

These opinions do not discuss the possibility of limiting statements about source attribution as has occurred with firearms, but a toolmark examiner's attribution of marks on the cartilage of a homicide victim to the defendant's knife in *Ramirez v. State*[65] led to three successive reversals by the Florida Supreme Court. The court was taken aback by "the extraordinarily precise claims of identification":[66]

> [The] police crime technician . . . made the extraordinary claim that his newly formulated knife mark identification procedure was infallible. He contended that he could identify the murder weapon to the exclusion of every other knife in the world—even if there had been two million consecutively produced knives of the

---

association. This understanding transposes the arguments in the conditional probability. The false-positive probability is the probability of reporting a source association given that there is none.

[59] *E.g.*, *Adams*, 444 F.Supp.3d 1248.

[60] *Adams*, 444 F. Supp. 3d at 1265 ("[T]he *AFTE Journal* is a trade publication, meant only for industry insiders, not the scientific community."); *Tibbs*, 2019 D.C. Super. LEXIS 9, at *33 ("The *AFTE Journal* is thus, in a sense, 'comparable to talk within congregations of true believers' rather than an example of 'the desired scientific practice of critical review and debate mentioned in *Daubert*.") (quoting David H. Kaye, *How* Daubert *and its Progeny Have Failed Criminalistics Evidence and a Few Things the Judiciary Could Do About It*, 86 Fordham L. Rev. 1639, 1645 (2018)).

[61] Fletcher v. Lane, 446 F. Supp. 729 (S.D. Ill. 1978) (rejecting a federal habeas corpus petition regarding the state trial court's admission of expert testimony that prymarks on victim's door conclusively and uniquely matched a screwdriver found in defendant's home).

[62] State v. Raines, 224 S.E.2d 232, 234 (N.C. Ct. App. 1976).

[63] United States v. Smallwood, 456 Fed.Appx. 563, 2012 WL 171402 (6th Cir. Jan. 23, 2012); State v Churchill, 646 P2d 1049 (Kan. 1982).

[64] Commonwealth v. Graves, 456 A.2d 561 (Pa. Super. Ct. 1983) (toolmark examiner's testimony that fingernail marks on victim's neck were a "high probability" match to defendant's fingernail was admissible, as was "practical impossibility" testimony from a forensic odontologist).

[65] 810 So.2d 836 (Fla. 2001).

[66] *Id*. at 849.

> same type—based on a striation "signature" arising from microscopic imperfections in the steel of the blade.[67]

Notwithstanding publications in journals of forensic science and medicine, the court deemed such extreme claims devoid of support in the scientific literature.[68] It noted the absence of any standards for the judgment of sufficient agreement[69] and the absence of any quantified error rate.[70] Nevertheless, the court did not question the admissibility of "traditional knife mark identification theory"[71] or testimony "that a victim's wounds were caused either by a particular knife or a knife similar thereto."[72] Thus, *Ramirez* requires trial courts to exclude the most extreme form of source attribution testimony, at least for wound marks from knives, but it did not directly address the question of what alternative testimony should be allowed.

In addition to arriving at a permissible or optimal upper limit on the nature of toolmark-identification testimony with tools other than firearms, the Commission may wish to consider the extent to which findings from firearms studies can be generalized to different types of tools and impressed materials. The method of ascertaining when patterns are sufficiently similar to make a positive association is an individual cognition based on training and experience. It relies on the toolmark examiner's personal understanding of how much variation occurs when a given tool makes repeated marks on the same material. Consequently, expertise in making source attribution from firearm marks does not necessarily carry over to other marks from other tools. In *United States v. Smallwood*,[73] the district court found that an FTE examiner with "significant experience with toolmarks generally" was not qualified to opine on the association between a knife and punctures made to an automobile's tires. The government appealed the ruling, arguing that, "although the AFTE theory lacks an objective standard, competent firearms toolmark examiners still operate under standards controlling their profession."[74] The court of appeals rejected any suggestion that one size fits all. It described the limited amount of training and experience the examiner had with knife marks and concluded that the "opinion that there is 'sufficient agreement' between her test marks and the puncture marks found in the tires of the . . . vehicle is unreliable under the AFTE's own standard because she has virtually no basis for concluding that the alleged match exceeds the best agreement demonstrated between tool marks

---

[67] *Id*. at 853.

[68] The court determined that:

> [T]he record does not show . . . meaningful peer review or publication . . . . At the *Frye* hearing below, the court reviewed two groups of published articles addressing knife mark evidence—one group North American, the other European. The North American articles were written by law enforcement technicians and . . . none undertakes the kind of searching, critical review that is the sine qua non of scientific acceptance. The European articles, on the other hand, were written by medical doctors and professors and are far more discerning; they delineate general studies and contain extensive analyses. The articles in that group, however, address only traditional knife mark theory relative to striation signatures. None address . . . the absolute certainty of identification deduced from such a test.

Id. at 849-50.

[69] *Id*. at 847.

[70] *Id*. at 851.

[71] *Id*. at 845.

[72] *Id*. at 846.

[73] 456 Fed.Appx. 563, 2012 WL 171402 (6th Cir. Jan. 23, 2012).

[74] *Id*. at 565.

known to have been produced by different tools."[75] Studies across tool types would help to reveal whether validity varies according to the type of the tool that leaves the marks.[76]

II. VOLUNTARY STANDARDS GOVERNING TOOLMARK COMPARISONS AND TESTIMONY

A. AFTE Documents

Standards in the form of guidelines or requirements for laboratories to adopt can come from professional associations or consensus-based standard-developing organizations (SDOs). The Association of Firearm and Toolmark Examiners (AFTE) published a three-paragraph "theory of identification" that assumes the goal of the comparative analysis is to form an opinion as to the source of a toolmark.[77] A related AFTE Range of Conclusions document for expressing source opinions has two categories for such opinions: "identification" and "elimination."[78] An "identification" is a finding of "[a]greement of a combination of individual characteristics and all discernible class characteristics where the extent of agreement exceeds that which can occur in the comparison of toolmarks made by different tools and is consistent with the agreement demonstrated by toolmarks known to have been produced by the same tool."[79] An "elimination" occurs when the examiner believes there is "[s]ignificant disagreement of discernible class characteristics and/or individual characteristics."[80] Of course, an FTE examiner might be unable or unwilling to reach one of these categorical conclusions, either because the examination could

---

[75] *Id.*

[76] One such study is described *infra* Part III(B).

[77] AFTE Theory of Identification as it Relates to Toolmarks (undated), https://afte.org/about-us/what-is-afte/afte-theory-of-identification, last visited Apr. 15, 2022, The theory is as follows:

> 1. The theory of identification as it pertains to the comparison of toolmarks enables opinions of common origin to be made when the unique surface contours of two toolmarks are in "sufficient agreement".
>
> 2. This "sufficient agreement" is related to the significant duplication of random toolmarks as evidence by the correspondence of a pattern or combination of patterns of surface contours. Significance is determined by the comparative examination of two or more sets of surface contour patterns comprised of individual peaks, ridges and furrows. Specifically, the relative height or depth, width, curvature and spatial relationship of the individual peaks, ridges and furrows within one set of surface contours are defined and compared to the corresponding features in the second set of surface contours. Agreement is significant when the agreement in individual characteristics exceeds the best agreement demonstrated between toolmarks known to have been produced by different tools and is consistent with agreement demonstrated by toolmarks known to have been produced by the same tool. The statement that "sufficient agreement" exists between two toolmarks means that the agreement of individual characteristics is of a quantity and quality that the likelihood another tool could have made the mark is so remote as to be considered a practical impossibility.
>
> 3. Currently the interpretation of individualization/identification is subjective in nature, founded on scientific principles and based on the examiner's training and experience

[78] AFTE Range of Conclusions (undated), https://afte.org/about-us/what-is-afte/afte-range-of-conclusions (last visited Apr. 15, 2022). The conclusion scale was "adopted by the Association membership at its annual business meeting in April 1992 (and published in the AFTE Journal Volume 24, Number 3)." *Id.*

[79] *Id.*

[80] *Id.*

not be conducted (because the materials were "unsuitable for examination") or because further analysis turned out to be "inconclusive."[81]

B. SWGGUN Standards

The former Scientific Working Group on Firearms and Toolmark (SWGGUN) described itself and other working groups as "tasked by the National Academy of Sciences (NAS) . . . with directing the forensic science communities in the establishment of standardized procedures and protocols."[82] To this end, SWGGUN released a document entitled "SWGGUN Systemic Requirements/Recommendations for the Forensic Firearm and Toolmark Laboratory."[83] The document simply delegates to individual laboratories the task of standardizing the microscopic comparisons.[84] The only other SWGGUN standard on interpreting comparative observations is titled "Guidelines: Criteria for Identification."[85] The document calls on laboratories to "include in their protocol a Criteria for Identification that is generally accepted by members of the forensic firearms community";[86] it then endorses the AFTE theory of identification as the protocol.[87]

C. ASB Standards

Following the creation of the Organization of Scientific Area Committees for Forensic Science through NIST in 2014, SWGGUN dissolved, and the Firearms and Toolmark Subcommittee of the Scientific Area Committee on Physics/Pattern Interpretation continued to draft standards and submit them for further modification and publication by an SDO—namely, the American National Standards Institute-Academy Standards Board (ASB). None of the eight standards on firearms and toolmarks that ASB published contains guidelines or requirements for the microscopic comparison process.

According to one OSAC webpage,[88] ASB is considering an OSAC Proposed Standard for Verification of Source Conclusions in Toolmark Examinations from the OSAC subcommittee.[89]

---

[81] The definition of "inconclusive" is either "[s]ome agreement of individual characteristics and all discernible class characteristics, but insufficient for an identification"; [a]greement of all discernible class characteristics without agreement or disagreement of individual characteristics due to an absence, insufficiency, or lack of reproducibility"; or "[a]greement of all discernible class characteristics and disagreement of individual characteristics, but insufficient for an elimination." *Id*.

[82] SWGGUN Systemic Requirements/Recommendations for the Forensic Firearm and Toolmark Laboratory (undated), available at https://www.nist.gov/system/files/documents/2016/11/28/swggun_systemic_report.pdf.

[83] *Id*.

[84] It requires that "[s]tandardized procedures will be developed by a laboratory to provide guidance in the examination, documentation and reporting of firearm and toolmark related evidence. Part of the standardized procedures will include a verification process . . . ." *Id*. at 4,

[85] SWGGUN, Guidelines: Criteria for Identification (undated), available at https://www.nist.gov/system/files/documents/2016/11/28/guidelines_for_criteria_for_identification.pdf.

[86] *Id*. § 2.1.

[87] *Id*. § 2.2.

[88] OSAC, Firearms & Toolmarks Subcommittee (Mar. 1, 2022), https://www.nist.gov/osac/firearms-toolmarks-subcommittee (links to "Standards at an SDO"). However, the page OSAC Registry, https://www.nist.gov/organization-scientific-area-committees-forensic-science/osac-registry (Apr. 5, 2022), lists no OSAC Proposed Standards coming from the subcommittee.

It requires that a second examiner compare the items as a quality assurance measure. It does not address the interpretive process itself. ASB also is considering a standard from the OSAC subcommittee[90] that would expand the AFTE scale for reporting conclusions. This "Standard Scale of Source Conclusions and Criteria for Toolmark Examinations" would replace the two AFTE categories (or three, if one thinks of "inconclusive" as a conclusion) with identification, insufficient support for identification, insufficient support for either exclusion or identification, insufficient support for exclusion, and exclusion.[91] The draft standard gives some explanation of what belongs in each category, but it relies on the AFTE theory, which itself relies on the examiner's impression that variations are within the range that occur for all same-source pairs and are outside the range for different-source pairs.[92] The draft standard does not supply the kind of standardized procedures sought in some of the limiting judicial opinions.[93]

The draft standard also includes a section on "qualifications and limitations." Drawn from a longer list in the first edition of the Department of Justice's Uniform Language for Testimony and Reports [ULTRs] for the Forensic Firearms/Toolmarks Discipline – Pattern Match Examination,[94] they resemble some of the denigrated expressions in Table 2.

D. Department of Justice ULTR

Because several of the opinions introduced in Part II treated conformity with the ULTR as a sufficent constraint on testimony, the current ULTR merits discussion in its own right. The ULTR applies only to FTE examiners within the Department of Justice. It confines them to the two AFTE categories for a conclusion and the inconclusive category, using the labels of "source

---

[89] OSAC Proposed Standard for Verification of Source Conclusions in Toolmark Examinations (Aug. 2020), https://www.nist.gov/system/files/documents/2020/08/06/Standard%20for%20Verification%20of%20Source%20Conclusions%20in%20Toolmark%20Examinations.pdf.

[90] OSAC, Firearms & Toolmarks Subcommittee (Mar. 1, 2022) https://www.nist.gov/osac/firearms-toolmarks-subcommittee (links to "Standards at an SDO").

[91] OSAC Firearms & Toolmarks Subcomm., Standard Scale of Source Conclusions and Criteria for Toolmark Examinations (ver. 1.0, undated), https://www.nist.gov/system/files/documents/2020/03/24/100_fatm_roc_and_criteria_standard_asb_mar2019_OSAC%20Proposed.pdf, For a small-scale study of the impact the expanded scale might have on how examiners classify FTE, see Thomas Busey et al., Validating Strength-of-support Conclusion Scales for Fingerprint, Footwear, and Toolmark Impressions, J. Forensic Sci. (forthcoming 2022), https://doi.org/10.1111/1556-4029.15019. Despite its title, this paper does not purport to demonstrate that examiners will use the new scales accurately. *See infra* Part III(B).

[92] *Id*. § 4.2.5.1.1 ("An Identification conclusion is based on an examiner's determination that all discernible class and individual characteristics agree such that the extent of agreement exceeds that which has been demonstrated by toolmarks made by different tools (KNM) and is consistent with the agreement demonstrated by toolmarks known to have been made by the same tool (KM).") (note omitted).

[93] *E.g.*, United States v. Adams, 444 F.Supp.3d 1248, 1267 n.9 (D. Or. 2020) ("It seems equally likely that a more quantitative measure of sufficient agreement would result in a finding of inconclusive in cases that currently result in a match. Mr. Gover and his peers seem reluctant to impose quantitative restrictions on their methodology because it would fail to justify a match in those cases where the numerical standard isn't met."); United States v. Cloud, No. 1:19-cr-02032-SMJ-1, 1:19-cr-02032-SMJ-2, 2021 WL 7184484, at *7 (E.D. Wash. Dec. 17, 2021) ("a vague standard that ultimately comes down to their training and experience … is truly an unforced error, and forensic science would be improved by a concerted effort to reframe the methodology's procedures in concrete, objective terms.").

[94] A comparison of the two versions can be found in Box 1 of David H. Kaye, *Mysteries of the Department of Justice's ULTR for Firearm-toolmark Pattern Examinations*, Forensic Sci., Stat. & L., Nov. 27, 2020, http://for-sci-law.blogspot.com/2020/11/mysteries-of-department-of-justices.html.

identification (i.e., identified)," "source exclusion (i.e., excluded)," and "inconclusive."[95] The ULTR defines "identified" in three ways that do not necessarily cohere. Vacillating between traditional "conclusion-centric" statements and purely "evidence-centric" assessments (*see infra* Part V), it permits source attribution as long as:

1) "all observed class characteristics are in agreement and . . . the examiner would not expect to find that same combination of individual characteristics repeated in another source and has found insufficient disagreement of individual characteristics to conclude they originated from different sources";
2) "the observed class characteristics and corresponding individual characteristics provide extremely strong support for the proposition that the two toolmarks originated from the same source and extremely weak support for the proposition that the two toolmarks originated from different sources"; or
3) the examiner believes that "the probability that the two toolmarks were made by different sources is so small that it is negligible."[96]

To make an identification under the AFTE range of conclusions, an examiner must determine that "the extent of agreement exceeds that which can occur in the comparison of toolmarks made by different tools and is consistent with the agreement demonstrated by toolmarks known to have been produced by the same tool."[97] The first ULTR condition is merely that "the examiner would not expect to find" the observed agreement from any other source. Arguably, the latter is a weaker standard.

The ULTR is clearer in listing things that an examiner cannot do than in defining the conclusions an examiner can reach. An FTE examiner "shall not" do any of the following:

- assert that a 'source identification' or a 'source exclusion' conclusion is based on the 'uniqueness' of an item of evidence.
- use the terms 'individualize' or 'individualization' when describing a source conclusion.
- assert that two toolmarks originated from the same source to the exclusion of all other sources.
- assert that examinations . . . are infallible or have a zero error rate.
- provide a conclusion that includes a statistic or numerical degree of probability except when based on relevant and appropriate data.
- cite the number of examinations conducted in the forensic firearms/toolmarks discipline performed in his or her career as a direct measure for the accuracy of a conclusion provided.
- assert that two toolmarks originated from the same source with absolute or 100% certainty, or use the expressions 'reasonable degree of scientific certainty,'

---

[95] U.S. Dept. of Justice, *Uniform Language for Testimony and Reports for the Forensic Firearms/toolmarks Discipline Pattern Examination*, Aug. 15, 2020, at 2, available at https://www.justice.gov/olp/page/file/1284766/download.

[96] *Id*. at 2. These conditions appear in separate paragraphs. Presumably, they are meant to be equivalent. For criticism of this language and related statements in the ULTR is presented in David H. Kaye, *Mysteries of the Department of Justice's ULTR for Firearm-toolmark Pattern Examinations*, Forensic Sci., Stat. & L., Nov. 27, 2020, http://for-sci-law.blogspot.com/2020/11/mysteries-of-department-of-justices.htm,

[97] AFTE Range of Conclusions (undated), https://afte.org/about-us/what-is-afte/afte-range-of-conclusions (last visited Apr. 15, 2022).

'reasonable scientific certainty,' or similar assertions of reasonable certainty in either reports or testimony unless required to do so by a judge or applicable law.[98]

III. ERROR RATES FOR ASSESSING VALIDITY: INCONCLUSIVES AND A MORE GRADUATED REPORTING SCALE

A. The Place of Inconclusives When Computing the False-alarm Proportion

As noted in Part I, courts have discussed "error rates" in experiments in which examiners make pairwise comparisons with ammunition components from both the same and different firearms, blinded to which pairs are from the same source ("true pairs") and which ones are from different sources ("false pairs").[99] They also have relied on experiments with other designs. In presenting "error rates" from studies with "ground truth" known to the researchers, one might think that judgments of "inconclusive" or "not suitable for comparison" are irrelevant to the legal question of the probative value of a source attribution.[100] The Innocence Project petition, however, maintains that:

> [N]early every study involving toolmark examination mishandles the treatment of "inconclusive" results, failing to count these conclusions as incorrect, even though there exists a ground truth answer of match or non-match. The failure to treat inconclusives as mistakes results in misleadingly low error rates that do not reflect reality. This treatment of inconclusives is akin to giving a student 90% on a test where he answers 10% percent of the questions incorrectly and skips the rest. An inconclusive result that does not match ground truth must be considered: it is an error and must be counted as such.[101]

The issue is not so simple. A better analogy is a true-false test in which the student also can answer "don't know." If the student *should* know the answer based on the course content but does not, then it is reasonable to score a "don't know" as an error when assessing the student's mastery of the material. But suppose one would not expect the student to be able to answer the question with the information provided in the course. Then it is less reasonable to score "don't know" as a wrong answer. Being unable to give a yes-or-no answer is a limitation, but it is not always a mistake.[102]

---

[98] *Id*. at 3.

[99] We use the word "pair" for convenience even though the evidence being examined in an experiment could consist of one questioned impression and several impressions made by a known tool (such as three impact marks on the shell casings from three test firings of the suspect's pistol. The examiner is asked whether the known impressions (from the test firings), paired with questioned impression, all emanated from the same known source.

[100] *See, e.g.*, Expert Working Grp. on Human Factors in Latent Print Analysis, Nat'l Inst. of Standards & Tech., Latent Print Examination and Human Factors: Improving the Practice Through a Systems Approach 30 (David H. Kaye, ed. 2012).

[101] Innocence Project, *supra* note 1 (footnotes omitted).

[102] For discussions of the meanings of "error" in pattern-matching tasks, see Alex Biedermann & Kyriakos N. Kotsoglou, Forensic Science and the Principle of Excluded Middle: "Inconclusive" Decisions and the Structure of Error Rate Studies, 3 Forensic Sci. Int'l: Synergy 100147 (2021); Expert Working Grp. on Human Factors in Latent Print Analysis, Nat'l Inst. of Standards & Tech., Latent Print Examination and Human Factors: Improving the Practice Through a Systems Approach 12-13, 24-31 (David H. Kaye, ed. 2012) (distinguishing between "outcome errors" and "process errors").

When examiners often do not come to definitive conclusions, the simple dichotomy of false-positives versus false-negatives is insufficient to describe the kinds of errors that can occur. Some terminology from signal-detection theory may be helpful. The simplest situation occurs when one has to decide whether input contains a signal or whether it is just white noise. If the system classifies the input as a signal, it sounds an alarm; if it classifies it as noise, it stays silent. Consequently, there are two types of possible errors: false alarms and missed signals. Only one type of false alarm is possible—the false positive—and all missed signals are false negatives.

With FTE evidence, the examiner can (1) sound an alarm by reporting that a true pair is present (positively associating the defendant with the toolmarks); (2) sound an alarm by reporting that the pair is false (eliminating the defendant's tool as a plausible source); or (3) do neither, and report that he or she cannot tell what the input is. Consequently, there are two types of false alarms. False-source-attribution alarms occur when an examiner reports a true pair when given a false pair. False-elimination alarms arise when an examiner reports a false pair given a true pair. But what about misses? The false alarms are all misses, but the third reporting option also produces missed signals of two types—those arising for false pairs (false-pair misses) and those for true pairs (true-pair misses). There are also two types of false inconclusives—false-pair inconclusives miss the false pairs, and true-pair inconclusives miss the true pairs.[103] The crucial point behind these distinctions is that the inconclusives do not enter into either of the false-alarm rates. They only enter into two false-inconclusive rates, and through those, into the two missed-signal rates.[104]

These statistics on errors of various kinds do not exist in a vacuum. They help answer specific research or policy questions. What, then, should the Commission make of this array of error proportions when evaluating the data on FTE examiner-performance in validation studies? Large false-inconclusive proportions might indicate an opportunity to make examiners more sensitive to false pairs (making the system more protective of falsely accused defendants) or to true pairs (making the system better able to produce evidence against criminals). These statistics

---

[103] *Cf.* Heike Hofmann & Alicia Carriquiry, *Treatment of Inconclusives in the AFTE Range of Conclusions*, 19 Law, Probability & Risk 317 (2020) (diagramming the different types of misses and alarms).

[104] Let $N(TPA_{FP})$ be the number of true-pair alarms (*TPA*s) for false pairs (*FP*s); let $N(Inc_{FP})$ be the number of inconclusives for false pairs; let $N(E_{FP})$ is the number of eliminations for false pairs; and let $N(FP)$ be the total number of false pairs. Notice that $N(FP) = N(TPA_{FP}) + N(Inc_{FP}) + N(E_{FP})$. The proportion of missed *FP*s relative to all *FP*s is

$$\begin{aligned} \text{Prop}(Miss \mid FP) &= N(TPA_{FP}) + N(Inc_{FP}) \,/\, N(FP) \\ &= N(TPA_{FP}) \,/\, N(FP) + N(Inc_{F}) \,/\, N(FP). \end{aligned}$$

This total missed-false-pair proportion, $\text{Prop}(Miss \mid FP)$, thus has two components: $\text{Prop}(TPA \mid FP) = N(TPA_{FP}) / N(FP)$, and $\text{Prop}(Inc \mid FP) = N(Inc_{FP}) / N(FP)$. The first term is a false-alarm proportion that indicates how often analysts are missing false pairs by sounding source-attribution alarms when presented with false pairs. So *FP*s leading to inconclusives must be included in the denominator but not the numerator. Similarly, the second term reveals how often analysts are missing false pairs, but this time by opting out with an inconclusive. Once again, *FP*s leading to inconclusives must be included in the denominator, but they also comprise the numerator.

An equivalent decomposition applies to the two types of "wrong" responses to true pairs (*TP*s). The total missed-true-pair proportion is $\text{Prop}(FPA \mid TP) = \text{Prop}(Inc \mid TP) = N(FPA_{TP}) / N(TP) + N(Inc_{TP}) / N(TP)$. The first term of the sum is a false-alarm proportion that indicates how often analysts are missing true pairs by sounding exclusion alarms when presented with true pairs. So *TP*s leading to inconclusives must be included in the denominator but not the numerator. Similarly, the second term reveals how often analysts are missing true pairs by opting out with an inconclusive. Inconclusives are counted in the numerator and the denominator of this true-pair inconclusive proportion.

are thus of system-wide interest and should be monitored by laboratory managers and other actors concerned with increasing the utility of FTE.[105] Inconclusives are important here.

However, the most salient statistics are different when the purpose is simply to ascertain the risk that scientific evidence, as currently produced, will serve to falsely accuse a suspect or convict a defendant. The law's fascination with "error rates" has two dimensions. One is the legal question of admissibility. When the probability of error with a scientific method is too large, the evidence may be kept from the factfinder. The other reason to have good estimates of error rates is not to exclude source attributions, but to give factfinders statistics that would help them assess the probative value in a given case.[106] For these purposes, inconclusives do not affect the relevant error statistics. As one decision theorist wrote with regard to latent print examinations:

> When an examiner offers an "inconclusive" opinion about whether two prints match, there is a sense in which he has erred. After all, he did not get the answer right, and the consequences of this failure may be serious (e.g., missed opportunity to exonerate a suspect). However, in the more usual sense of the meaning of error, an inconclusive is not an error. It is a pass. An inconclusive means that the examiner offers no judgment about whether two prints do or do not share a common source.[107]

As such, the proportions of missed true or false pairs are beside the point. The system for generating alarms and exclusions may not be functioning efficiently, as it may be allowing too many examiners to abstain in hard cases. However, the appropriate summary of the data from the experiment—for the purpose of judging whether the risk of false alarms is too high to allow examiners to make source attributions—is the proportion of false source attributions without regard to inconclusives. Contrary to what the petition to the Commission suggests, the mere fact that inconclusives are not reports of the true state of affairs supplies no mathematical or logical justification for adding a number of inconclusives to the numerator of an observed error

---

[105] *See* Heike Hofmann & Alicia Carriquiry, *Treatment of Inconclusives in the AFTE Range of Conclusions*, 19 Law, Probability & Risk 317, 330-31 (2020). So too, one would include inconclusives in some statistics when the research question is whether examiners are more averse to the risk of a false source attribution than a false exclusion. *See also* Expert Working Grp. on Human Factors in Latent Print Analysis, Nat'l Inst. of Standards & Tech., Latent Print Examination and Human Factors: Improving the Practice Through a Systems Approach 30 (David H. Kaye, ed. 2012):

> [F]rom the perspective of the police, what matters is all the cases that an examiner considers rather than just those in which the examiner ultimately might testify. In terms of improving the contribution of the examiner to the investigative process, it is appropriate to regard the failure to identify or exclude when the latent print contains adequate information as a potentially correctable error. Likewise, deciding that the latent print is of sufficient quality, but concluding that the comparison is inconclusive when, in fact, the similarities (or differences) are distinct and extensive, also is an error. Whether one regards such errors as outcome or process errors (or both), they are important because they might warrant a change in training or operational procedures to take fuller advantage of the latent friction ridge data.

[106] We discuss this issue further *infra* Part IV.

[107] Jonathan J. Koehler, *Fingerprint Error Rates And Proficiency Tests: What They Are And Why They Matter*, 59 Hastings L.J. 1077 (2008) (note omitted). When presented to the jury, however, the probabilities of an inconclusive conditional on the source hypotheses can be important. Imagine an examiner called to rebut an identification by testifying that the evidence does not support that opinion because it is inconclusive. The jury could benefit from knowing the probability of this examiner's opting for "inconclusive" when the tool is and is not a true source.

proportion.[108] But neither is it appropriate to add some number of inconclusives to the denominator to water down the observed error proportion.[109] As a NIST Expert Working Group explained in 2012:

> The view that inconclusives should not count is appropriate from the perspective of a judge or juror who might consider error rates or probabilities to assess the probative value of an identification or an exclusion. For that purpose, it does not matter how often the examiner refrains from reaching a categorical conclusion. What matters is accuracy in those cases in which the examiner does offer an opinion on identification or exclusion. These are the only cases in which an examiner's testimony might lead the jury astray. Testimony that the latent print contained inadequate information to reach any conclusion as to the origin of the print occurs less often and should not propel the jury in any particular direction. Therefore, any calculated error rate presented in a trial involving an identification or an exclusion should be based upon the subset of cases in which examiners actually make an identification or an exclusion.[110]

In statistical jargon, it is appropriate to condition on the fact that the examiner has chosen one of the binary alternatives to "inconclusive" and to use the resulting "decision-specific" proportion in assessing the legal value of the evidence.[111] On this basis, the false-positive error proportion in the 2014 Ames study made prominent in the PCAST report was 22/1443 = 1.5%.[112]

All the same, the inconclusive category is relevant to the validity of source attributions, but not because the number of inconclusives belongs in the proportion. There is the possibility, noted in Part I(B), that examiners who are not blind to the fact that they are experimental

---

[108] It is important that "inconclusives" are presented and understood as providing no information one way or the other about the truth of a source hypothesis. Testimony of an "inconclusive" outcome should be complete enough to make it clear that "inconclusive" is indeed a pass. Likewise, lawyers should nor argue that "not excluded" is probative of the same-source hypothesis when the examiners consider the level of matching to be inconclusive.

[109] Jonathan J. Koehler, *Forensics or Fauxrensics? Ascertaining Accuracy in the Forensic Sciences,* 49 Ariz. St. L.J. 1369, 1140 n.180 (2017). The effect of including inconclusives in only the denominator can be substantial (or not). Heike Hofmann & Alicia Carriquiry, *Treatment of Inconclusives in the AFTE Range of Conclusions,* 19 Law, Probability & Risk 317, 330 (2020).

[110] Expert Working Grp. on Human Factors in Latent Print Analysis, Nat'l Institute of Standards & Tech., Latent Print Examination and Human Factors: Improving the Practice Through a Systems Approach 29-30 (David H. Kaye, ed. 2012).

[111] Heike Hofmann & Alicia Carriquiry, *Treatment of Inconclusives in the Afte Range of Conclusions*, 19 Law, Probability & Risk 317, 328-29 (2020); *see also* J.J. Koehler, *Proficiency Tests to Estimate Error Rates in the Forensic Sciences,* 12 Law, Probability & Risk 89, 95 (2013). For related arguments, see Alex Biedermann & Kyriakos N. Kotsoglou, *Forensic Science and the Principle of Excluded Middle: "Inconclusive" Decisions and the Structure of Error Rate Studies*, 3 Forensic Sci. Int'l: Synergy 100147 (2021); Itiel E. Dror, Glenn Langenburg, *"Cannot Decide": The Fine Line Between Appropriate Inconclusive Determinations VS. Unjustifiably Deciding Not to Decide*, 64 J. Forensic Sci. 1e15 (2019), https://doi.org/10.1111/1556-4029.13854; Itiel E. Dror & Nicholas Scurich, *(Mis)use of Scientific Measurements in Forensic Science*, 2 Forensic Sci. Int'l, 333 (2020); Itiel E. Dror & Nicholas Scurich, *Continued Confusion About Inconclusives and Error Rates: Reply to Weller & Morris*, 2 Forensic Sci. Int'l: Synergy 703e704 (2020); Todd J. Weller & M.D. Morris, *Commentary on I. Dror, N Scurich "(Mis)use of Scientific Measurements in Forensic Science,"* 2 Forensic Sci. Int.: Synergy 701 (2020).

[112] The false-negative proportion was 4/1079 = 0.37%. David P. Baldwin et al., Ames Laboratory, Dep't of Energy, A Study of False-Positive and False-Negative Error Rates in Cartridge Case Comparisons, Technical Report #IS-5207 (2014), https://afte.org/uploads/documents/swggun-false-postive-false-negative-usdoe.pdf [https://perma.cc/4VWZ-CPHK].

subjects will be more cautious to avoid making the serious error of false source attributions and false exclusions. To the extent that this heightened risk aversion occurs, the proportion from the experiment will tend to underestimate the false-alarm proportion that would occur in practice. Some of the inconclusives in the experiment would become false alarms in real cases.

There are at least two ways to address the risks of generalizing from the experiment to the real world. First, one could consider whether the inconclusive proportion is higher in actual cases than in the experiments. If the experimental pairings are similar to those in case work, finding that there is no large difference in the proportion of inconclusives would allay the fear that the experimental subjects are manipulating (unconsciously or otherwise) the inconclusive category. Second, and even better, the experimental subjects could be blinded to the cases in which their performance is being checked against ground truth,[113] as the Houston Forensic Science Center has done.[114]

B. Validating a More Finely Grained Reporting Scale

The question of false-alarm proportions is more complicated when the binary reporting scale is expanded to allow firm inclusions and weak exclusions to be reported separately, as in the draft ASB standard mentioned in Part II. A more finely grained reporting scale could theoretically provide some significant benefits over the binary scale. One would expect some of the previous monolithic source attributions to shift down to the weaker source-attribution category (a one-bell alarm instead of a two-or-more-bell alarm, so to speak). Likewise, some formerly monolithic inconclusives could become informative one-bell alarms. To assess the relative merits of an expanded scale, however, it is important to determine how examiners using the expanded scale actually would perform.

We have located no research directly validating the use of an expanded scale. A very recent study entitled "Validating Strength-of-support Conclusion Scales for Fingerprint, Footwear, and Toolmark Impressions" compares examiners' classifications made using on a scale that resembles the AFTE one of "identification," "exclusion," and "inconclusive," but adds two additional intermediate categories," making it a five-point scale, with classifications made using a corresponding five-point scale for degrees of evidentiary support rather than categorical conclusions.[115] Both scales differ from the traditional AFTE one and from the newly proposed ASB standard. Moreover, the study does not show that the examiners accurately grade the degrees of similarity or their consequences with either of the scales it had the examiners use. The validity of expanded scales therefore remains an open question.

That said, because it uses pairings of known ground truth, the study does supply data on the accuracy of toolmark identifications for certain non-firearm toolmarks.[116] The researchers

---

[113] *E.g.*, Jonathan J. Koehler, *Proficiency Tests to Estimate Error Rates in the Forensic Sciences,* 12 Law, Probability & Risk 89 (2013); Jonathan J. Koehler, *Fingerprint Error Rates and Proficiency Tests: What They Are and Why They Matter*, 59 Hastings L.J. 1077 (2008).

[114] Maddisen Neuman et al., *Blind Testing in Firearms: Preliminary Results from a Blind Quality Control Program*, J Forensic Sci. (forthcoming 2022); DOI: 10.1111/1556-4029.15031.

[115] Thomas Busey, Morgan Klutzke, Alyssa Nuzzi & John Vanderkolk, *Validating Strength-of-support Conclusion Scales for Fingerprint, Footwear, and Toolmark Impressions,* J. Forensic Sci. (2022), https://doi.org/10.1111/1556-4029.15019.

[116] Various studies on the accuracy of microscopic comparisons of firearms toolmarks are reviewed in Heike Hofmann & Alicia Carriquiry, *Treatment of Inconclusives in the AFTE Range of Conclusions*, 19 Law, Probability

created striated toolmarks from 15 quarter-inch screwdrivers and wood chisels on heavy-duty aluminum foil at 20-and 35-degree angles (three scrapings per angle) and photographed the toolmarks. They organized the photographs into true and false pairings[117] and recruited 20 toolmarks examiners from local, county, and state agencies[118] to compare the various pairs. One randomly selected group of examiners used the expanded AFTE categories, and the other group used the newer strength-of-support categories (a concept discussed further in Part IV). An abridged version of the authors' Table 4, juxtaposing the scales and adding our own abbreviations, follows:

Table 3. Labels for the Categories of the Conclusion-based and the Strength-of-evidence Scales in Busey et al. (2022)

Identification (ID)	Extremely Strong Support for Common Source (ES)
Insufficient for Identification (weak ID)	Support for Common Source (SS)
Inconclusive (Inc)	Inconclusive (Inc)
Insufficient for Elimination (weak Elim)	Support for Different Sources (SD)
Elimination (Elim)	Extremely Strong Support for Different Sources (ED)

The top two rows label an examiner's view that the paired toolmark photographs are, at the very least, indicative of the same source (S). The bottom two label an examiner's view that they are indicative of two different tools (D). Inconclusives express no opinion one way or the other. False-positive errors occur when examiners report ID, weak ID, ES, or SS for false pairs (D). False-negative errors occur when they report EX, weak EX, ED, or SD for true pairs (S). The 20 examiners in the study had the following pattern of correct and incorrect responses:

& Risk 317 (2020); *see also* L. Scott Chumbley et al., *Accuracy, Repeatability, and Reproducibility of Firearm Comparisons Part 1: Accuracy* (July 30, 2021), https://arxiv.org/abs/2108.04030; Erwin J.A.T. Mattijssen et al., *Validity and Reliability of Forensic Firearm Examiners*, 307 Forensic Sci. Int'l 110112 (2020) (discussed in David H. Kaye, *"Quite High" Accuracy for Firearms-mark Comparisons,* Forensic Sci., Stat. & L., Aug. 18, 2020, http://for-sci-law.blogspot.com/2020/02/a-new-validity-and-reliability-study-of.html); Erwin J.A.T. Mattijssen, *Interpol Review of Forensic Firearm Examination 2016-2019,* 2 Forensic Sci. Int'l: Synergy 389 (2020).

[117] True pairs consisted of one scraping from each tool at either angle and the images from the same tool from a different scraping at both angles. False pairs were similar-looking scrapings from different tools. The authors noted that "it is difficult to determine whether the task difficulty was comparable to typical casework." In another section of the article, they state that "It is difficult to establish the task difficulty of these comparisons relative to casework, although the fact that the toolmarks were created by tools of the same make and model does make this a particularly challenging task."

[118] The article does not describe the method of recruitment and does not provide data on how representative the volunteers may have been of toolmark examiners generally. There was dropout "due to the Covid-19 pandemic."

Table 4: Examiners' Classifications Reported in Table 12 of Busey et al. (2022)

	Elim + ED	Weak Elim + SD	Weak ID + SS	ID + ES
D	84 + 45	68 + 119	24 + 13	2 + 2
S	8 + 11	20 + 19	67 + 84	177 + 136

Conclusion-based Scale

False-positive proportions
Prop(ID | D) = 2/(84+68+24+2) = 2/178 = 1.1%
Prop(Weak ID | D) = 24/178 = 13.5%
Prop(Some ID | D) = 14.6%

False-negative proportions
Prop(Weak Elim | S) = 20/(8+20+67+177) = 20/252 = 7.9%
Prop(Elim | S) = 8/252 = 3.1%
Prop(Some Elim | S) = 11.1%

Strength-of-evidence Scale

False-positive proportions
Prop(ES | D) = 2/(45+119+13+2) = 2/179 = 1.1%
Prop(SS | D) = 13/179 = 7.3%
Prop(Some Support for S | D) = 8.4%

False-negative proportions
Prop(ED | S) = 11/(11+19+84+136) = 11/250 = 4.4%
Prop(SD | S) = 19/250 = 7.6%
Prop(Some Support for D | S) = 12%

Performance at the extreme ends of the scales is respectable. The false-positive proportion is 1.1%, and the false-negative proportion is 3.1% for the conclusion-based scale. For the strength-of-evidence scale, they are 1.1% and 4.4%, respectively. As previously noted, it is not clear that the examiners are using the ends of the expanded scales just as they use the "identification" and "elimination" labels with the AFTE system. But if one were to assume some rough equivalence, then these false-positive and false-negative proportions could be compared with those from the firearms studies (computed, as here, without regard to inconclusives).[119] Despite the small number of examiners in the study and the authors' suggestion that the toolmark evidence they manufactured may be more difficult than that encountered in practice, we have presented these results because there are relatively few validity studies for non-firearms toolmarks.

IV. POSSIBLE MODES OF TESTIMONY

In this Part, we place the limiting formulations introduced in Part I within a broader range of possible forms of expert testimony about FTE. Rather than just describing what has been

[119] Table 12 of Busey et al. reports 113 and 57 inconclusives for false pairs and true pairs, respectively, using the conclusion-based scale. It reports corresponding counts of 110 and 61 using the strength-of-evidence scale. For some purposes, one might want to compute proportions that include these numbers in some manner. *See supra* Part III.

legally admissible to date, we consider what might be best for assisting the jury in understanding the toolmark evidence and its implications for determining the source of a questioned item. The limitations in Table 1 presuppose that the expert makes judgments about the truth or probability of hypotheses as to the source of the toolmarks. These opinions are not the only way to present the findings, and they may not be the best way. This Part therefore outlines a more extensive range of presentations of toolmark comparisons and offers ideas as to which ones are most suitable.

The possibilities can be arrayed in various ways. The most fundamental distinction between types of FTE evidence presentation is between statements about source hypotheses and statements about the strength of the toolmark evidence. An example of a source-hypothesis statement is an assertion that there is a negligible probability that the marks came from any other tool than the known tool. It expresses the degree of the examiner's belief about the different-source hypothesis being true. In contrast, an assertion that "the similarities in the examined items strongly support the hypothesis that the marks came from the known tool" would be an example of a statement about the strength of the evidence. Such a statement is not about the probability of the hypothesis and does not express any degree of belief about a source conclusion. It is just a statement that the information discerned in the toolmark evidence is powerful evidence for a source attribution.

The degree of belief in a source hypothesis is a function of the strength of the toolmark evidence *and* the degree to which, before making any comparisons, the examiner believed that the known tool was used. Forensic-science and evidence theorists often refer to the pre-test degree of belief as a prior probability that is adjusted according to the strength of the toolmark evidence to arrive at a posterior probability.[120] Bayes' rule can serve as a model for this belief-revision process. It prescribes that the posterior odds are simply the prior odds multiplied by a quantity known as the Bayes' factor or, in this context, the likelihood ratio.[121] The numerator of the likelihood ratio is the probability of observing the processed pair of toolmarks given the assumption that the marks come from the same source. The denominator is the probability of observing those toolmarks given a different hypothesis—usually, that the marks come from an unspecified different tool. The likelihood ratio thus states how many times more probable the processed toolmark evidence is under one source hypothesis versus another. Because this factor reveals how the odds change based on the examiner's understanding of the toolmarks, it expresses the strength of the evidence. Large likelihood ratios mean that the evidence strongly supports the same-source hypothesis in the numerator. It gives a large boost to the odds in favor of that hypothesis. Small likelihood ratios mean that the evidence has only a minor impact on the degree of belief.

In short, this Bayesian model of how beliefs should change with new evidence neatly separates conclusions from evidence, at least in principle.[122] An examiner who testifies to an

[120] The Bayesian perspective dominates an extensive and longstanding body of academic writing on inference in forensic science. A comprehensive textbook is Colin G.G. Aitken, Franco Taroni & Silvia Bozza, Statistics and the Evaluation of Evidence for Forensic Scientists (3d ed. 2021).

[121] If the odds on an outcome are a to b, the probability is $a/(a+b)$, For example, odds of 1:5 correspond to a probability of ⅙.

[122] This is not to say that jurors actually process evidence according to Bayes' rule. *See, e.g.,* Jonathan J. Koehler, *On Conveying the Probative Value of DNA Evidence: Frequencies, Likelihood Ratios, and Error Rates*, 67 U. Colo. L. Rev. 859 (1996) (arguing that jurors tend to lack a sophisticated statistical understanding and may not be practically influenced by differences between evidence- and conclusion-centric testimony).

opinion about the true source of the marks must be applying a likelihood ratio to prior probabilities for the source hypotheses.[123] An examiner who does not offer a source conclusion but merely testifies to how strongly evidence points to one hypothesis as opposed to another does not consider the prior probabilities. To capture this distinction, we can say that strength-of-evidence statements are evidence-centric testimony, whereas assertions about the truth of the source hypotheses are conclusion-centric testimony.[124]

The second major distinction between types of FTE presentation is between classifications and quantifications. The strength, or probative value, of the evidence as expressed in the likelihood ratio is numerical when the conditional probabilities are numbers on the usual 0-1 scale. But the probative value of the evidence also can be described by less refined categories with labels such as "strong support" or "weak support." We can describe the categories for these verbal likelihood ratios as classifications. Whether or not an examiner starts with a numerical estimate for a likelihood ratio, he or she can characterize the perceived likelihood ratio on a verbal scale that uses these classifications.

Similarly, the conclusion-centric posterior probabilities can be expressed numerically (as subjective probabilities) or categorically. From that Bayesian perspective, the latter classifications are just named intervals for sorting the underlying numbers into somewhat arbitrary categories. Categorical statements of association (for example, "identification" or, dropping down one notch in the draft ASB standard mentioned in Part II, "insufficient support for identification") are two conceivable classifications. The first classification is same-source pairs; the second is probably-but-not-so-clearly-same-source pairs.[125] Because the classifications pertain to the asserted status of the pairs, they are conclusion-centric classifications.

With these two distinctions (evidence-centric vs. conclusion-centric and classification vs. quantification) in mind, we can organize and comment on the possible approaches to toolmark-comparison testimony. We begin with an extreme version of evidence-centric testimony that limits the examiner to explaining the features of interest and the extent to which they are similar or different. We then consider other evidence-centric approaches. Finally, we return to the conclusion-centric approaches that are the status quo for testimony in the United States.

A. Features-only Testimony

The most restrictive mode of testimony confines the expert to displaying or describing potentially distinguishing features and leaving it at that. This austere mode of presentation requires a judge or jury unable to discern the similarities and differences in pairs of impressions, largely eliminating the expert part of the testimony. A slightly less restrictive and more informative mode allows the expert to characterize particular features as similar or different. As

---

[123] *E.g.*, William C. Thompson et al., *Perceived Strength of Forensic Scientists' Reporting Statements About Source Conclusions,* 17 Law, Probability & Risk 133, 134 (2018) ("Forensic scientists cannot logically draw conclusion about source probabilities without taking a position on the prior odds that the items in question have the same source. Doing that, however, requires the forensic scientist to delve into matters outside their scientific expertise.").

[124] David H. Kaye, *The Nikumaroro Bones: How Can Forensic Scientists Assist Factfinders?*, 6 Va. J. Crim. L. (2018), available at https://papers.ssrn.com/abstract_id=3177752.

[125] The wording of the classification ("insufficient support for … ") does not capture its meaning and place of the category in the scale that descends from "identification" through "exclusion." Despite the word "support" in the label, the classification comes from the examiner's posterior subjective probability.

a response to arguments about a lack of scientific validity, a few courts have applied this approach to testimony on latent fingerprints and handwriting.[126]

It is reasonable to eschew allegedly expert evaluations when presenting some types of evidence. For example, fracture matches do not necessarily need expert characterizations to convey their meaning. Once the bits of a broken windshield are reassembled into their original configuration, for example, the gloss of an expert opinion that the fragments came from the same source adds little to a juror's appreciation of the evidence. But the fact that two small fragments align nicely at their edges under magnification can be harder for anyone, expert or otherwise, to evaluate. The evidence seems probative (to some degree), and admitting it may be worth the risk of it being overvalued, but the features-only presentation is an awkward compromise for esoteric distinguishing features such as those used in toolmark comparisons.[127] For microscopic toolmark comparisons, a features-only presentation threatens to be unhelpful to the factfinder by providing *too little* expert opinion, leaving the factfinder to surmise the strength of the evidence unguided.

B. Perceived Strength-of-evidence Testimony

Rather than confining testimony to the underlying features, experts could be permitted to give their estimates of the likelihood ratio for the data. These could come from traditional visual comparisons in which examiners have an implicit sense of the variations in true pairs and false pairs that they encounter.[128] Whereas "[b]y their very nature, posterior odds in the firearm–toolmark discipline incorporate non-scientific contextual information that is contained in the prior odds,"[129] "the likelihood ratio expresses the strength of the obtained evidence irrespective of the prior odds."[130] An examiner could describe the strength of evidence to the factfinder with a statement such as, "I believe the similarity in the microscopic characteristics I have compared are *x* times more likely to arise if the defendant's tool made the marks than if a randomly selected tool did," and then explain the basis for that judgment.[131] More objective, statistical or machine-learning methods also have been developed to give "appropriate numbers for each individual case . . . a quantitative measure for the weight of the evidence" so that "the judge or jury decides whether to accept the evidence and what weight to assign to it."[132] For these

---

[126] *See* David H. Kaye et al., The New Wigmore on Evidence: Expert Evidence § 15.3 (2d ed. 2010) (citing cases).

[127] *Id*.; Jennifer L. Mnookin, *The Courts, the NAS, and the Future of Forensic Science,* 75 Brook. L. Rev. 1209 (2010).

[128] *See* Wim Kerkhoff et al., *The Likelihood Ratio Approach in Cartridge Case and Bullet Comparison,* 45 AFTE J. 284 (2013) (summarizing the discussion that led to the implementation of the likelihood ratio approach to firearms identification by the Firearms Section of the Netherlands Forensic Institute.).

[129] Stephen Bunch & Gerhard Wevers, *Application of Likelihood Ratios for Firearm and Toolmark Analysis*, 53 Sci. & Just. 223, 227 (2013).

[130] J. Song et al., *Estimating Error Rates for Firearm Evidence Identifications in Forensic Science*, 284 Forensic Sci. Int'l 15 (2018), https://doi.org/10.1016/j.forsciint.2017.12.013.

[131] *See* David H. Kaye, *Likelihoodism, Bayesianism, and a Pair of Shoes*, 53 Jurimetrics J. 1 (2012) (discussing footwear-mark testimony). For largely subjective likelihood ratios, the explanation would have to acknowledge the imprecision of the specific values to avoid the criticism that "numbers may lend a false air of precision to a subjective approximation." William C. Thompson, *How Should Forensic Scientists Present Source Conclusions?*, 48 Seton Hall L. Rev. 773, 780 (2018).

[132] J. Song et al., *Evaluating Likelihood Ratio (LR) for Firearm Evidence Identifications in Forensic Science Based on the Congruent Matching Cells (CMC) Method*, 317 Forensic Sci. Int'l 110502, at 2(2020),

reasons, having experts address only the likelihood ratio (or its components) has won widespread endorsement from statistical[133] and scientific or laboratory associations[134] and agencies[135] as well as from scholars of law and statistics.[136]

Although giving a numerical likelihood ratio may be conceptually sound, judges and juries may have difficulty understanding exactly what they mean. To address this difficulty, a testifying expert could explain the likelihood ratio with an analogy to a diagnostic medical test—for example, a test for SARS-CoV-2 that is 100 times as likely to come back positive when the virus is present than when it is not corresponds to FTE with a likelihood ratio of 100. The expert might also use a frequency figure. An estimated likelihood ratio of 100 is equivalent to a match on a set of marks in a population in which it is estimated that 1 in 100 guns would produce the

---

https://doi.org/10.1016/j.forsciint.2020.110502. The validation and admissibility of the CMC method and other computerized methods are beyond the scope of this report.

[133] *Am. Stat. Ass'n Position on Statistical Statements for Forensic Evidence*, Am. Stat. Ass'n 1, 2-4 (Jan. 2, 2019), https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf [https://per ma.cc/X4AM-AVBU]:

> To evaluate the weight of any set of observations made on questioned and control samples, it is necessary to relate the probability of making these observations if the samples came from the same source to the probability of making these observations if the questioned sample came from another source in a relevant population of potential sources. . . . We . . . strongly advise forensic science practitioners to confine their evaluative statements to expressions of support for stated hypotheses: e.g., the support for the hypothesis that the samples originate from a common source and support for the hypothesis that they originate from different sources.

[134] Colin Aitken et al., *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses* (2010), http://www.rss.org.uk/Images/PDF/influencing-change/rss-fundamentals-probability-statistical-evidence.pdf [https://perma.cc/NV7K-VJ9C] (committee of the Royal Statistical Society); Ass'n of Forensic Sci. Providers, *Standards for the Formulation of Evaluative Forensic Science Expert Opinion*, 49 Sci. & Just. 161 (2009); Eur. Network of Forensic Sci. Insts., *ENFSI Guideline for Evaluative Reporting in Forensic Science* 10 (2015), http://enfsi.eu/wp-content/uploads/2016/09/m1_guide line.pdf [https://perma.cc/H296-YKML] ("Evaluative reports should address the probability of the findings given the propositions and relevant background information and not the probability of the propositions given the findings and background information."); *cf.* Royal Society, Forensic DNA Analysis: A Primer for Courts 36 (2017) ("Likelihood ratios are generally accepted as being the most appropriate method for evaluating the evidential strength of DNA profiles.").

[135] Subcomm. on Reporting and Testifying of the National Commission on Forensic Science. Nat'l Comm'n on Forensic Sci., *Views of the Commission: Statistical Statements in Forensic Testimony*, U.S. Dep't Justice (Feb. 9, 2017), https://www.justice.gov/archives/ncfs/page/file/965931/download [https://perma.cc/3WU L-3N2R] ("Forensic science practitioners should confine their evaluative statements to the support that the findings provide for the claim linked to the forensic evidence."); Nat'l Inst. of Forensic Sci. Austl. N.Z., *An Introductory Guide to Evaluative Reporting* 6 (2017), available at https://www.anzpaa.org.au/forensic-science/our-work/products/publications:

> The fundamental principles of evaluative reporting or interpretation are . . . (iii) that the role of the expert is to comment on the probability of their findings, given these propositions and not on the propositions themselves. It is this last principle that allows the fact-finders to combine aspects of evidence they hear during the course of the trial with their judgement in their deliberations. This framework of evidence evaluation is commonly referred to as evaluative reporting, but may also be referred to as the likelihood ratio approach, logical thinking or Bayesian inference.

[136] *E.g.*, Edward K. Cheng, *The Burden of Proof and the Presentation of Forensic Results*, 130 Harv. L. Rev. F. 154, 161-62 (2017) ("Scholars have long argued in favor of presenting forensic results using likelihood ratios, and indeed some forensic communities in Europe have embraced them . . . . The key is that likelihood ratios present a clear path to improving the use of forensics testimony in court.") (footnotes omitted); Colin G.G Aitken & 30 co-authors, *Expressing Evaluative Opinions: A Position Statement*, 51 Sci. & Just. 1 (2011), http://dx.doi.org/10.1016/j.scijus.2011.01.002.

matching marks. People may be better able to use "1 in N" statements than a ratio of conditional probabilities.

Psychological research indicates that individuals tend to alter their prior beliefs in response to numerical probabilities, frequencies, or likelihood ratios *less* than Bayes' rule would support.[137] In lay terms, this indicates that juror perceptions may be more tied to prior perceptions and beliefs than to these probabilities and statistics. However, the degree of conservatism might vary with the type of forensic-science evidence,[138] the magnitude of the likelihood ratios,[139] and other factors.[140] Of course, other modes of presentation raise similar questions, and it has been said that "whether jurors can understand more complex statistical terms such as likelihood ratios and random match probabilities is an empirical question with no clear answer in the literature."[141] In addition, examiners used to thinking in terms of an AFTE-like conclusion-centric scale would need to be retrained to articulate personal likelihood ratios. Furthermore, in light of the unstructured and unstandardized process for formulating likelihood ratios, they would need to be studied for accuracy and reliability within and across examiners.

Instead of forming (or reporting) numerical likelihood ratios, examiners could testify to the probative value of the evidence qualitatively, with statements such as, "[I]t is far more probable that this degree of similarity in features would occur when comparing [the questioned impressions] with the defendant's [tool] than with [some other tool]."[142] Or the testimony could draw on a standard table of expressions for the degree to which the evidence supports a source conclusion, as recommended in the 2009 National Academies report[143] and implemented in scales used in other countries[144] and for DNA evidence.[145] The ASB draft standard (see Part II)

---

[137] *E.g.*, Dale A. Nance & Scott B. Morris, *An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Large and Quantifiable Random Match Probability*, 42 Jurimetrics J. 403 (2002).

[138] *See* William C. Thompson & Eryn J. Newman, *Lay Understanding of Forensic Statistics: Evaluation of Random Match Probabilities, Likelihood Ratios, and Verbal Equivalents*, 39 Law & Hum. Behav. 332 (2015) (finding good adherence to Bayesian norms for DNA evidence and underutilization of a moderate LR for shoeprint evidence).

[139] *Id*. (finding large likelihood ratios to produce appropriate changes in beliefs).

[140] Kristy A. Martire & Gary Edmund, *How Well Do Lay People Comprehend Statistical Statements from Forensic Scientists?*, in Handbook of Forensic Statistics 201, 215 (David Banks et al. 2021).

[141] Thomas Busey et al., *Validating Strength-of-support Conclusion Scales for Fingerprint, Footwear, and Toolmark Impressions*, J. Forensic Sci. (forthcoming 2022), https://doi.org/10.1111/1556-4029.15019; *accord*, Edward K. Cheng, *The Burden of Proof and the Presentation of Forensic Results*, 130 Harv. L. Rev. F. 154, 161 (2017) ("An increasingly complex literature has emerged on lay understanding of likelihood ratios and how such quantitative information is best presented. Research thus far has yielded no easy answers . . . .").

[142] NIST Expert Working Group on Human Factors in Latent Print Analysis, *Latent Print Examination and Human Factors: Improving the Practice Through a Systems Approach* 134 (David H. Kaye ed. 2012); *cf*. David H. Kaye, *Likelihoodism, Bayesianism, and a Pair of Shoes*, 53 Jurimetrics J. 1 (2012) (discussing footwear-mark testimony).

[143] Comm. on Identifying the Needs of the Forensic Science Community, Nat'l Research Council, *Strengthening Forensic Science in the United States: A Path Forward* 186 (2009) (referring to gradations such as limited, moderate, moderately strong, strong, and very strong evidence to support a conclusion).

[144] Ass'n of Forensic Sci. Providers, *Standards for the Formulation of Evaluative Forensic Science Expert Opinion*, 49 Sci. & Just. 161 (2009); Eur. Network of Forensic Sci. Insts., *ENFSI Guideline for Evaluative Reporting in Forensic Science* 10 (2015), http://enfsi.eu/wp-content/uploads/2016/09/m1_guide line.pdf [https://perma.cc/H296-YKML].

uses the notion of evidentiary support in the intermediate categories, but the labels for the highest and lowest groupings ("identification" and "exclusion") are conclusions rather than degrees of support, depriving it of the advantages of a fully evidence-centric scale. Limited psychological research on such scales has been done to investigate how forensic-science practitioners understand terms such as "moderate support" and "strong support,"[146] and how lay individuals use them.[147]

Strength-of-evidence testimony does not require experts to draw a sharp line between the overall similarity of paired samples that establishes (in the mind of the examiner) that the pair originated from the same tool. Both qualitative and quantitative likelihood ratios range from marking evidence as highly supportive for one source hypothesis to depicting evidence as highly supportive of the alternative source hypothesis.

C. Source-probability Testimony

Beliefs in a source hypothesis can also be presented on quantitative and qualitative scales of personal or subjective probabilities. The FTE community seems to be in general consensus that limiting conclusions to "identification," "exclusion" and "inconclusive" is a best practice. However, other systems for characterizing data on a spectrum exist and are widely accepted in other fields. As noted at the outset of this Part, posterior probabilities can be elicited in overtly numerical terms. As a compromise between a full probability scale and the three-category scale, an ordinal scale using set terms to indicate specific ranges of probability could be used. For example, the U.S. Intelligence Community has a clearly articulated range of expressions to describe probability, corresponding to specific percentage ranges. These series, ranging from "almost no chance" and "remote" for 1-5% to "almost certain" for 95-99%, use ordinary words to verbalize the numerical probability.[148] The full table is presented in Table 5 below.

---

[145] Dep't of Justice, *Uniform Language for Testimony and Reports for Forensic Autosomal DNA Examinations Using Probabilistic Genotyping Systems*, Sept. 18, 2018, at 4 (permitting a numerical likelihood ratio to be accompanied by a "verbal qualifier" of uninformative, limited support, moderate support, strong support, or very strong support, depending on the order of magnitude of the computed likelihood ratio).

[146] Thomas Busey et al., *Validating Strength-of-support Conclusion Scales for Fingerprint, Footwear, and Toolmark Impressions*, J. Forensic Sci. (forthcoming 2022), https://doi.org/10.1111/1556-4029.15019; Elmarije K.van Straalenab et al., *The Interpretation of Forensic Conclusions by Criminal Justice Professionals: The Same Evidence Interpreted Differently*, 313 Forensic Sci. Int'l (2020).

[147] Eleanor Arscott et al., *Understanding Forensic Expert Evaluative Evidence: A Study of the Perception of Verbal Expressions of the Strength of Evidence*, 57 Sci. & Just. 221 (2017); Kristy A. Martire & Gary Edmund, *How Well Do Lay People Comprehend Statistical Statements from Forensic Scientists?,* in Handbook of Forensic Statistics 201 (David Banks et al. 2021); Kristy A. Martire & Ian Watkins, *Perception Problems of the Verbal Scale: A Reanalysis and Application of a Membership Function Approach*, 44 Sci. & Just. 264 (2015); Kristy A Martire et al., *On the Interpretation of Likelihood Ratios in Forensic Science Evidence: Presentation Formats and the Weak Evidence Effect*, 240 Forensic Sci. Int'l 61 (2014); Kristy A. Martire et al., *The Expression and Interpretation of Uncertain Forensic Science Evidence: Verbal Equivalence, Evidence Strength, and the Weak Evidence Effect*, 37 Law & Hum. Behav. 187 (2013); W.C. Thompson et al., *Perceived Strength of Forensic Scientists' Reporting Statements About Source Conclusions*, 17 Law, Probability & Risk 133 (2018), http://doi.org/10.1093/lpr/mgy012; William C. Thompson & Eryn J. Newman, *Lay Understanding of Forensic Statistics: Evaluation of Random Match Probabilities, Likelihood Ratios, and Verbal Equivalents*, 39 Law & Hum. Behav. 332 (2015).

[148] Intelligence Community Directive 203: Analytic Standards, Jan. 2, 2015, https://irp.fas.org/dni/icd/icd-203.pdf (last visited Apr. 25, 2022). In the table, "likelihood" has its ordinary-language meaning. It is not a "likelihood" in the sense of the probability of data given a hypothesis.

Table 5. Intelligence Community Expressions of Likelihood and Probability[149]

Percentage	01-05 %	05-20 %	20-45 %	45-55 %	55-80 %	80-95 %	95-99 %
Likelihood	Almost no chance	Very unlikely	Unlikely	Roughly even chance	Likely	Very likely	Almost certainly
Probability	Remote	Highly improbable	Improbable	Roughly even odds	Probable	Highly probable	Nearly certain

One might think that presenting source conclusions as personal probabilities is mere honesty, but the Department of Justice ULTR and the draft ASB standard squarely oppose this manner for expressing uncertainty. Both documents state 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."[150] Neither document explains what "relevant and appropriate data" would be, but presumably the fear is that personal probabilities would be misconstrued as being better founded than they actually are. Yet, it is not clear why the foundation for a statement like "I judge that the probability that the bullet with the marks came from the gun is 90%" cannot be presented to the judge or jury to make it clear that it is not the product of a formal, statistical analysis.

The more fundamental problem with posterior-probability testimony is that, as with all conclusion-centric testimony, it incorporates some normally unstated prior probability. Two experts with different views of the strength of the evidence might come to the same source-probability because they began with different prior probabilities. If the factfinder has no idea what prior probability the expert used, it is in no position to use the expert's testimony to modify *its* prior, which should be based on the non-toolmark evidence in the case. Thus, it has been argued that giving source-probability opinions makes experts go "where the science runs out" and "beyond [their] expertise."[151]

But regardless of how examiners arrive at subjective probabilities, studies could show whether their assessments are well calibrated.[152] Weather forecasters give predictions in the form of probabilities when they report that the chance of rain tomorrow is, say, 90%. A well calibrated predictor would be correct in about 90% of these forecasts—that is, of all the days when rain is forecast with a probability of 90%, it rains on 90% of them. However, it is not standard practice for FTE examiners to make probability assessments, and we know of no studies to confirm that

[149] *Id.* at 3.

[150] U.S. Dept. of Justice, *Uniform Language for Testimony and Reports for the Forensic Firearms/toolmarks Discipline Pattern Examination*, Aug. 15, 2020, at 3, available at https://www.justice.gov/olp/page/file/1284766/download; OSAC Firearms & Toolmarks Subcomm., Standard Scale of Source Conclusions and Criteria for Toolmark Examinations § 5.3 (ver. 1.0, undated), https://www.nist.gov/system/files/documents/2020/03/24/100_fatm_roc_and_criteria_standard_asb_mar2019_OSAC%20Proposed.pdf.

[151] David H. Kaye, *The Nikumaroro Bones: How Can Forensic Scientists Assist Factfinders?*, 6 Va. J. Crim. L. 101, 105 (2018); *see also* William C. Thompson, *How Should Forensic Scientists Present Source Conclusions?*, 48 Seton Hall L. Rev. 773 (2018).

[152] On this type of calibration, see, for example, A. Philip Dawid, *The Well-calibrated Bayesian*, 77 J. Am. Stat. Ass'n 605 (1982); Morris H. DeGroot & Stephen E. Fienberg, *The Comparison and Evaluation of Forecasters*, 32J. Royal Stat. Soc'y: Series D (The Statistician) 12 (1983).

their personal probabilities are well calibrated, weighing against the immediate adoption of this presentation mode.

D. Source-category Testimony

Binary classifications such as "identified" and "excluded" are, of course, the norm in FTE testimony. These classifications are subject to similar criticism relating to prior probabilities, and they also suffer from a loss-of-information problem in that all cases with extremely large personal probabilities are clumped together as "identifications," and all those with extremely small probabilities are "exclusions." In between, the examiner must be agnostic and uninformative. The use of only two categories means that two cases that are barely distinguishable because the overall degree of similarity lies just above and just below the borderlines are treated as radically different in their implications.[153] A richer categorical scale could mitigate the loss of information but still would draw arbitrary lines between the categories.

Table 1 delineated ways devised by courts to allow experts to present categorical judgments by tinkering with their wording or by adding statements of how certain it is that the source attribution is correct. The modes of testimony are as follows:

1) The features are consistent with or fail to exclude the tool as the source of the impressions;
2) "More likely than not," the questioned impression was made by the known tool;
3) "To a reasonable degree of ballistic certainty" (or close variants), the questioned impression was made by the known tool;
4) The questioned impression was made by the known tool (with no statement of a degree of certainty, strong or otherwise).

Each approach has some drawback. "Consistent with" can be misunderstood as a definitive source attribution. "Not excluded" is perhaps less likely to be overvalued by a lay juror, but it too supplies no information on what to make of the inability to exclude. "More likely than not" may connote more uncertainty than is necessary. Phrases like "a reasonable degree of ballistic certainty" are odd and impenetrable. Finally, allowing the ultimate opinion without permitting inquiry into how certain it is makes it difficult for the jury to know what weight to give to the opinion. Thus, legal commentators have been decidedly unenthusiastic about these measures.[154] What else can be done to enhance the usefulness of the traditional binary classifications?

1. Sensitivity, Specificity, and Conditional Error Probabilities for Binary Classifications

Instead of rewording binary classifications, courts could require evidence as to the relative frequency with which an examiner applies that classification correctly. Knowing how frequently examiners make correct and incorrect classifications of true and false pairs of marks not only can validate the claim of expertise at this task,[155] but the calibration also gives the

[153] Artificial boundaries also mean that overall false-positive and false-negative error rates seen in the experiments with FTE examiners underestimate these conditional error probabilities near the boundaries, at least for examiners who rarely opt out by declaring that the evidence is inconclusive.

[154] David H. Kaye, *Firearm-Mark Evidence: Looking Back and Looking Ahead*, 68 Case W. Res. L. Rev. 723, 734-35 (2018) (quoting negative reactions from law professors).

[155] Demonstrating expertise also requires proof that the experts' "operating characteristics" of sensitivity and specificity are higher than for non-experts who are given the same data to evaluate. *See, e.g.,* Roger C. Park, *Signature Identification in the Light of Science and Experience*, 59 Hastings L.J. 1101 (2008).

factfinder a track record it can use in weighing the expert's testimony. Data on the performance of the particular examiner with evidence of the same level of difficulty as that in the case at bar would give the best scientifically defensible estimates of the examiner's false alarm rates (or the complementary quantities of sensitivity and specificity).[156] Lacking such case-specific data, one can only look at the success or failure rates for similar comparisons in the toolmark field generally. The factfinder would use these statistics for general guidance, conceivably adjusting them to account for factors that plausibly would make the examiner's conclusions more (or less) likely to be correct.

This is essentially the form of presentation recommended in the recent 2016 PCAST report.[157] Specifically, the report proposes that juries be informed of the upper 95%, one-sided confidence limit for the conditional false-positive proportions found in experiments in which blinded examiners respond to true and false pairings.[158] The district court in *United States v. Cloud*[159] raised this false-positive-probability strategy when it promised the defendant that:

> [I]f the examiner intends to go beyond testimony that merely notes the recovered cartridge casings could not be excluded as having been fired from the recovered firearm, the Court will inform the jury that: (1) only two studies that meet the minimum design standard have attempted to measure the accuracy of fingerprint comparison and (2) these studies found false positive rates that could be as high as 1 in 306 in one study, 1 in 18 in the second study.[160]

Framing a false-alarm rate as "1 in *N*" is a useful way to call the factfinder's attention to the possibility of other sources for the toolmark evidence. But the details of the method for computing the false-positive probabilities presented in the PCAST report have been questioned,[161] and judges and experts must take care not to present experiments with small sample sizes as proof of high error probabilities (as might have occurred in *Cloud*).[162] Testimony or instructions also should be structured to avoid the misperception that an estimated false-positive proportion, often referred to as the "error rate," is the proportion of cases in which examiners make false source attributions. The false-positive proportion is the proportion of cases in which examiners confronted with false pairs make source attributions, not the proportion of all source attributions that are false. The tendency to equate the latter probability with the former is a version of the "transposition fallacy" that is a worry with many of the ways to present

---

[156] Sensitivity is the probability of a source attribution for a true pair, Specificity is the probability of a source exclusion for a false pair.

[157] 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.pdf [https://perma.cc/R76Y-7VU].

[158] For discussion of computing the proportions in studies with large numbers of "inconclusives," see *supra* Part III(A).

[159] No. 1:19-cr-02032-SMJ-1, 1:19-cr-02032-SMJ-2, 2021 WL 7184484 (E.D. Wash. Dec. 17, 2021).

[160] *Id*. at *15.

[161] David H. Kaye, *PCAST's Sampling Errors (Part I)*, Forensic Sci., Stat. & L. (Oct. 24, 2016); David H. Kaye, *PCAST's Sampling Errors (Part II: Getting More Technical)*, Forensic Sci., Stat. & L. (Dec. 11, 2016).

[162] Listing the upper ends of confidence intervals and ignoring the observed proportions themselves can be misleading. An upper confidence limit reflects both sampling uncertainty and the magnitude of the observed value. A small observed error proportion could have a high confidence limit associated with it just because the sample size is small. In that case, the study does not prove that the long-term error proportion is large. It means that the study does not provide a precise estimate of the true proportion (the one that would arise from averaging the observed proportions from a great many repeated experiments).

probabilities or proportions in forensic-identification science.[163] Still, false-positive probability estimates from pertinent studies are a reasonable way to inform the jury of the limitations of examiner classifications.

2. Likelihood Ratios for Binary Classifications

Courts sometimes suggest that false-negative probabilities are irrelevant in cases of source attribution. This is an oversimplification. The probative value, or evidentiary strength, of the examiner's classification of the paired material is only partly determined by the false-positive probability. If the judge or jury is to receive guidance on how much confidence it can have in the examiner's source attribution, it becomes necessary to regard the examiner as a test instrument. The human instrument responds with some reading on the scale for categorical conclusions about the source hypotheses. The goal is to inform the factfinder of the probative value of the examiner's readings on the binary AFTE scale. A likelihood ratio indicates the evidentiary strength of the report on this scale, but it is slightly different from the likelihood ratio discussed in Part IV(B). There, we spoke of the likelihood ratio with respect to the examiner's perceived degree of similarity between the paired items. For convenience, assume similarity could be represented as a single number—a similarity score. The likelihood ratio discussed above summarized the value of a particular similarity score by presenting the probability of that score given a true pair divided by the probability of the same score given a false pair. But assessing the examiner's performance as a binary test instrument requires a likelihood ratio that pertains to more than this single similarity score. The report is not the score itself. It is that the unspecified score falls into an acceptance region for the hypothesis that the pair of impressions are from the same source. The question for evaluating the value of such a positive classification is how much more likely it is for the examiner to make these attributions for same-source pairs than for different-source pairs.

Dividing the probability of the former (the sensitivity) by the probability of the latter (the false-alarm probability) answers this question. It yields an "alarm likelihood ratio" that would help the factfinder make such an evaluation. One can estimate this alarm likelihood ratio objectively, from appropriate experiments that check on the performance of examiners relative to the true state of affairs.[164]

Judges and jurors could receive testimony on the alarm likelihood ratio just as they could hear testimony quoting only the false-alarm proportions from experiments. As with the examiner's personal likelihood ratio of Part IV(B), however, questions about juror comprehension arise. Once again, the qualitative "verbal scales" could be used to supplement a purely numerical presentation. As noted in Part IV(B), the testifying expert could make analogies

[163] *See, e.g.*, Kristy A. Martire & Gary Edmund, *How Well Do Lay People Comprehend Statistical Statements from Forensic Scientists?*, in Handbook of Forensic Statistics 201(David Banks et al. 2021).

[164] Although the alarm likelihood ratio denominator is a kind of random-match probability, it differs from a random-match probability derived from a probability model for generating a population of patterns..Probability models for the generation of toolmarks have been proposed to compute random-match probabilities, and "substantial efforts" show that 3D patterns from tools can be reduced to one-dimensional similarity scores that produce relatively few matches (as defined by computerized algorithms) among the many false pairs that can be formed from particular datasets. John E. Murdock et al., *The Development and Application of Random Match Probabilities to Firearm and Toolmark Identification*, 62 J. Forensic Sci. 619 (2017). Although these studies indicate that there is considerable distinguishing information in the impressions, they do not provide direct estimates for a random-match probability as matches are judged by toolmark examiners. *Id.*

to clinical medical test outcomes or offer a frequency figure. Still, psychological research does not yet establish which of the various modes of presentation of the likelihoods, their ratio, corresponding random-match probabilities, and qualitative categories—or which mixture of these concepts—is best for juror comprehension.

Summary and Conclusions

In response to doubts about the extent to which research establishes the validity and reliability of visual, microscopic comparison of toolmark impressions, some courts have eschewed certain phrases and placing limits on how strongly an FTE examiner can testify to a source attribution. Many of these rulings, cataloged and explained in Part I, prevent the examiner from expressing a subjective degree of confidence in the conclusion. Although certain wording changes may represent improvements over the most extreme testimony, the limitations leave legal finders of fact without the knowledge required to appreciate the probative value of the limited conclusions. The voluntary, consensus standards from the forensic-science community, described in Part II, have not filled this gap.

If testimony includes source attributions, then estimated measures of accuracy derived from pertinent studies of examiner performance should accompany them, with the recognition that these figures are averages across examiners and across the toolmarks compared in the studies. Technically, the best measure of probative value is not the false-positive probability, as courts guided by parts of the PCAST report have assumed. Rather, it is the likelihood ratio involving the true and the false-positive proportions.[165] Part IV(D) made suggestions on how such numbers could be presented and perhaps supplemented by particular words.

But the assumption that FTE examiners should classify the pairings of impressions under comparison is questionable. As explained in Part IV(B), there are cogent and longstanding arguments for shifting the focus of the examination and testimony away from conclusions about the truth of the same-source hypothesis to direct statements of support for this hypothesis. In the absence of validated ratios from automated systems for comparisons, FTE examiners could provide personal likelihood ratios, with suitable explanations of their basis and meaning. As noted in Part IV(B), they could grade the strength of the evidence categorically as well as (or even instead of) numerically by using a standardized terminology about the degree of support the examiner has found for a possible conclusion.

Forensic science best serves the legal system in cases involving toolmark evidence when FTE examiners supply the best available scientifically justified information in a manner that successfully conveys the expert's understanding of the evidence. None of the approaches canvassed here are panaceas, but this report has summarized viable alternatives to traditional firearms and toolmark comparison testimony and indicated some of the more detailed issues involved in implementing these alternatives. We hope that the information will assist the Commission in responding to the petition to improve FTE testimony in Texas.

[165] For the reasons given in Part III, "inconclusives" should not be included in the fractions.

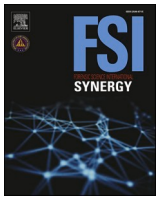
EXHIBIT E

Forensic Science International: Synergy 8 (2024) 100472

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: www.sciencedirect.com/journal/forensic-science-international-synergy



Inconclusive decisions and error rates in forensic science

H. Swofford *, S. Lund, H. Iyer, J. Butler, J. Soons, R. Thompson, V. Desiderio, J.P. Jones, R. Ramotowski

National Institute of Standards and Technology (NIST), USA

ARTICLE INFO

Keywords:
Forensic science
Error rates
Inconclusives
Likelihood ratio
Validation data
Bayesian reasoning
Black-box study

ABSTRACT

In recent years, there has been discussion and controversy relating to the treatment of inconclusive decisions in forensic feature comparison disciplines when considering the reliability of examination methods and results. In this article, we offer a brief review of the various viewpoints and suggestions that have been recently put forth, followed by a solution that we believe addresses the treatment of inconclusive decisions. We consider the issues in the context of *method conformance* and *method performance* as two distinct concepts, both of which are necessary for the determination of reliability. Method conformance relates to an assessment of whether the outcome of a method is the result of the analyst's adherence to the procedures that define the method. Method performance reflects the capacity of a method to discriminate between different propositions of interest (e.g., mated and non-mated comparisons). We then discuss implications of these issues for the forensic science community.

Disclaimers

These opinions, recommendations, findings, and conclusions do not necessarily reflect the views or policies of NIST or the United States Government.

One or more of the authors of this paper serve(s) as an Associate Editor/the Editor-in-Chief of this journal. The standard peer review process was followed and an editor who is not on the author panel has handled the review process for this paper. The authors had no influence over the peer review process. The final decision is made by an editor who is not on the author panel.

1. Introduction

The forensic science community faces scrutiny from legal and scientific scholars, who question (measures for) the reliability[1] of forensic examination methods, with particular emphasis on those that rely predominantly on visual observation and human judgment (e.g., feature comparison methods used in pattern evidence examination, such as friction ridge, firearms and toolmarks, footwear, tire tracks, handwriting) [1,2]. In the 1993 Supreme Court ruling in *Daubert v. Merrell Dow Pharmaceuticals, Inc.* [3], the Court declared that scientific evidence must be relevant and reliable, and provided examples of factors to consider when evaluating its admissibility, such as testability, peer review, error rates, standards, and acceptance in the scientific community. Largely in response to *Daubert*, error rates (e.g., false positive or false negative rates) began to receive increased attention as a key measure of performance.

In 2009, the National Research Council (NRC) report on forensic science renewed the call for determinations of error rates [1] and set in motion efforts to design and execute large-scale testing schemes to evaluate reliability across forensic science disciplines, with an initial emphasis on friction ridge and firearms analyses [4–10]. Likewise, the 2016 report by the President's Council of Advisors on Science and Technology (PCAST) emphasized the need for empirical measures of performance and appropriate determinations of error rates as factors underlying determinations of validity and reliability [2].

The focus on error rates as a primary measure of method performance is generally satisfactory when experts report results using a binary scale, such as identification or exclusion. In this context, the false

* Corresponding author.
E-mail address: henry.swofford@nist.gov (H. Swofford).

[1] In this paper, the term "reliable" is used as an all-encompassing term that relates to the extent to which a method can be relied upon to produce accurate and consistent results, and includes the concepts of "validity," "reproducibility," and "repeatability."

https://doi.org/10.1016/j.fsisyn.2024.100472
Received 14 March 2024; Received in revised form 17 April 2024; Accepted 18 April 2024
Available online 4 May 2024

Table 1a
A 2 × 3 table representing performance metrics relating to hypothetical Method 1 where all reported outcomes for both mated and non-mated comparisons are "inconclusive."

Method 1	Identification	Inconclusive	Exclusion
Mated Comparisons	0 %	100 %	0 %
Non-Mated Comparisons	0 %	100 %	0 %

Table 1b
A 2 × 3 table representing performance metrics relating to hypothetical Method 2 where all reported outcomes for mated comparisons are "Identification" and all non-mated comparisons are "Exclusion."

Method 2	Identification	Inconclusive	Exclusion
Mated Comparisons	100 %	0 %	0 %
Non-Mated Comparisons	0 %	0 %	100 %

positive rate is defined as the proportion of times the method results in an "identification" in non-mated comparisons (e.g., in a validation study) and the false negative rate is defined as the proportion of times the method results in an "exclusion" in mated comparisons.[2] However, few feature comparison disciplines operate using a binary scale. Most use a three-point (or more) scale, which is some variation of identification, inconclusive, or exclusion.[3] Even with the additional option of inconclusive, it might seem natural to apply the classical definitions of false positive rate and false negative rate. However, careful consideration quickly reveals that it is unsatisfactory to use error rates alone as the metric of performance for a method in these feature comparison disciplines.

Consider the following hyperbolic example to illustrate this point (Tables 1a and 1b).[4] Suppose we have two methods with the following outcomes for mated and non-mated comparisons.

In Tables 1a and 1b, we see that neither Method 1 nor Method 2 results in any identification decisions for non-mated comparisons or exclusion decisions for mated comparisons. Therefore, both methods have the ideal false positive and false negative rates of 0 % (or correspondingly, a seemingly ideal total combined error rate of 0 %). The usefulness of the two methods, however, could not be further apart.

The purpose of the forensic examination (e.g., in feature comparison disciplines) is to help others determine whether or not two patterns could have originated from the same source. Thus, a method's utility is characterized by how successfully the method's output distinguishes non-mated comparisons from mated comparisons. Method 1 leads to an inconclusive result for every comparison. This outcome means that Method 1 does not provide any information to help a user of the reported result (e.g., factfinder) determine whether a given comparison is non-mated or mated. Method 2, however, perfectly distinguishes all non-mated comparisons from mated comparisons. That is, a user of the reported result who inferred a comparison was non-mated if the result from Method 2 was exclusion and inferred a comparison was mated if the result was identification would have been correct every time. This example illustrates that, when a conclusion scale is not binary, false positive and false negative rates alone do not accurately convey how successfully one could use the method output to distinguish non-mated comparisons from mated comparisons and therefore do not adequately characterize method performance.[5]

Nevertheless, perhaps motivated by the fact that the term "error rates" is explicitly mentioned in the *Daubert* decision as well as the NRC and PCAST reports, the desire to represent method performance in terms of error rates has continued. Consequently, disagreements over the treatment of inconclusive decisions also remain. To avoid the misleading nature of classical definitions for false positive rate and false negative rate for non-binary conclusion scales, alternative definitions for false positive and false negative rates have been proposed—primarily manifesting in various ways of treating inconclusive outcomes. For example, the PCAST suggested omitting inconclusive decisions altogether so that (error) rate estimates are based on the proportion of conclusive examinations rather than the proportion of all examinations [2].

Although PCAST touched on this issue in 2016, controversy surrounding the treatment of inconclusive decisions began to surface in 2019 when Dror and Langenburg raised concern that there is a lack of transparency and accountability on the use of inconclusive decisions and recommended that the forensic science community establish criteria to know whether and when inconclusive decisions are "justifiable" [11]. This was followed by recommendations by Dror and Scurich in 2020 in which inconclusive decisions that did not conform to some established criteria ought to be counted as errors [12]. Not long after, several different articles were published expressing various viewpoints relating to the treatment of inconclusive decisions [13–18].

When deliberating on this issue, nearly every possible option has been proposed, including: inconclusive decisions be ignored altogether, inconclusive decisions always be considered correct, inconclusive decisions always be considered incorrect, inconclusive decisions be considered correct in some situations and incorrect in other situations, and inconclusive decisions be considered neither correct nor incorrect. Consequently, we are left with an array of proposed definitions of false positive and false negative rates that can lead to wildly different estimates of error rates, and, therefore, different representations and interpretations of the reliability of forensic science results, all with potential consequences regarding the admissibility of such evidence in judicial proceedings.

2. Discussion

When considering how inconclusive decisions should be treated (or any outcome for that matter), it is important to first take a step back and frame the context of the situation. There are two important things to consider:

First, in forensic casework, a particular issue might be disputed and the ground-truth of that issue (e.g., true source-origin of a particular set of compared items) is unknown and, oftentimes, unknowable. Further, items or impressions from crime scenes are often presented to analysts in

[2] Non-mated comparisons refer to items that were known to have been made by different sources. Mated comparisons refer to items that were known to have been made by the same source.

[3] We recognize there are different ways of conducting feature comparisons and communicating results (e.g., probabilistic, categoric). In this paper, we limit our discussion to the use of conclusion scales that include inconclusive as a legitimate response option since it represents traditional practices in many feature comparison disciplines. Further, we recognize conclusion scales vary in terms of the number of response options available (e.g., some might have multiple derivations of inconclusive, levels of support, or options to declare items "not suitable" for comparison). For simplicity, we focus our discussion on a single catch-all class of "inconclusive" response options that indicates a comparison outcome that is not an explicit assertion of the ground-truth state of the compared items (e.g., comparison outcomes other than "identification" or "exclusion").

[4] In this example, we use we use percentages of total response outcomes for mated and non-mated comparisons for illustrative purposes, but, in real studies, actual numbers should be provided to enable estimation of uncertainty.

[5] A similar example could be constructed to show that the alternative metrics of sensitivity (true positive rate) and specificity (true negative rate) also do not adequately characterize method performance. Such examples illustrate the perils of trying to summarize the performance of a method with a non-binary range of conclusions with the same number of parameters as a method with a binary range of conclusions. Two additional independent parameters or rates are required to fully characterize method performance for each element added to a binary conclusion scale.

Table 2
Brief description of recent articles discussing the treatment of inconclusive decisions in forensic science.

	Articles	Description of Viewpoints
1	*Dror and Langenburg (2019)* [11]	Called for greater transparency and accountability for the use of inconclusive decisions. An option of inconclusive should not be available when there is sufficient information to make a conclusive decision to avoid an "easy way out." They supported developing criteria to determine situations where fingerprint examiners would not be allowed to choose inconclusive and to use statistical models or qualified opinion scales that provide greater distinction of the perceived strength of evidence within the broad inconclusive category along with blind verification to assess appropriateness of an inconclusive decision.
2	*Dror and Scurich (2020)* [12]	Recognized the need for inconclusive decisions in some cases but claimed that these decisions ought to be considered correct or incorrect based on whether the evidence contains sufficient quantity and quality of information for a conclusive determination. They proposed either using a panel of independent experts or consensus data from a study to determine which comparisons should be deemed as inconclusive.
3	*Weller and Morris (2020)* [13]	Suggested that the rates of all decision types be reported as they relate to ground-truth with the recognition that there are two ground-truth states and three meaningful response categories. They expressed concerns with Dror and Scurich (2020) views of categorizing every result as correct or erroneous and representing measures of reproducibility as measures of accuracy.
4	*Hofmann* et al. *(2020)* [14]	Outlined and critiqued four approaches to address inconclusive decisions in calculating error rates, such that inconclusive decisions are: (1) ignored altogether, (2) considered as correct, (3) considered as incorrect, and (4) considered equivalent to an exclusion. They distinguished between "source-specific" and "decision-specific" metrics, suggesting they should be used for different purposes (method performance and court testimony).
5	*Biedermann and Kotsoglou (2021)* [15]	Argued that Dror and Scurich (2020) views conflate the ontological level of analysis (where ground-truth is fixed) with the epistemic level of analysis (where ground-truth remains uncertain). They warned against the artificial category of a "forensically correct" determination that does not have a ground-truth. They encouraged monitoring all response types as they relate to ground-truth so that the true limits of the method can be understood.
6	*Arkes and Koehler (2021)* [16]	Emphasized that inconclusive decisions are a statement about the insufficiency of available evidence and are neither correct nor incorrect as there is no applicable ground-truth. They proposed the use of signal detection theory as a framework for understanding the role inconclusive decisions play and opposed scoring inconclusives as either correct or incorrect when computing error rates.
7	*Dorfman and Valliant (2022)* [17]	Described an ideal "mechanical scheme" for establishing an objective basis to categorize inconclusive decisions as errors using objective measurements, statistical algorithms, and likelihood theory and illustrated how this could be used to assess overall error rates as described by Dror and Scurich (2020). Until such measures are available, they suggested blind testing schemes be employed to estimate error rates and that inconclusive decisions must be regarded as potential errors.
8	*Guyll* et al. *(2023)* [18]	Argued that inconclusive decisions are different because they forgo any assertion as to the ground-truth state of the evidence. They advocated for the rates of all decision types to be reported as they relate to ground-truth, conclusive and inconclusive alike, to make results useful for the widest range of purposes. They also suggested that the likelihood ratio of a decision (e.g., calculated in terms of "the proportion of all same-source comparisons that are given a particular decision divided by the proportion of all different-source comparisons that are given that same decision") be used as a metric for expressing its "probative value." They recognized, however, that evaluations of a technique for designating "decision correctness" (such as the use of a decision rule, consensus opinion, or similarity measure with cutoff criterion) may be useful in some contexts, such as training or determining appropriateness of examiners' decision in relation to evidence quality.

Table 3a
A 2 × 3 table representing performance metrics relating to hypothetical Method 3.

Method 3	Identification	Inconclusive	Exclusion
Mated Comparisons	89 %	10 %	1 %
Non-Mated Comparisons	1 %	40 %	59 %

Table 3b
A 2 × 3 table representing performance metrics relating to hypothetical Method 4.

Method 4	Identification	Inconclusive	Exclusion
Mated Comparisons	59 %	40 %	1 %
Non-Mated Comparisons	1 %	10 %	89 %

a partial, degraded, or low-quality state. Thus, it is certainly conceivable that forensic analysts will encounter situations where an examination does not yield sufficient information to support a conclusive opinion as to the potential source. Thus, an inconclusive determination is a possible, and sometimes necessary and important, outcome of the examination to ensure a binary decision (e.g., exclusion or identification) is not forced where it is not warranted and achievable. We recognize that this point is largely uncontroversial. What is contentious, however, is when inconclusive determinations might be warranted or justifiable and how inconclusive determinations should be treated when assessing the reliability of a method.

Second, users of forensic results (e.g., factfinders) are presented with the outcome of an examination conducted by a particular analyst and tasked with making inferences and decisions about the truth of various propositions in question (e.g., whether or not two patterns originated from the same source). Users of the reported result must therefore weigh the reliability of the result by considering at least three questions.

(1) What method did the analyst apply when conducting the forensic examination?
(2) How effective is that method at discriminating between the propositions of interest?
(3) How relevant is the data describing the discriminability (i.e., diagnostic capacity) of that method (generally) to the examination in the case at hand (specifically)?

To address these questions, information about whether the analyst conformed to a particular method as well as measures relating to the performance of that method are needed. In this context, we distinguish between two important concepts: *method conformance* and *method performance*.

- Method conformance relates to assessments of whether the outcome of a particular method is the result of the analyst's adherence to the procedures that define that method.
- Method performance relates to measures that reflect the extent to which the outcome of a particular method can effectively distinguish between different propositions of interest (e.g., between same-source and different-source comparisons).

Method performance includes information relating to both

discriminability and *reproducibility* of outcomes produced by the method.[6] Importantly, measures of reproducibility provide the gauge by which measures of discriminability (based on outcomes from multiple analysts generally) are relevant to an outcome by a particular analyst (specifically) as well as the adequacy of the procedures that define the method.[7] Further, while measures of method performance are the means by which methods are deemed "acceptable" for the intended application (e.g., from a validation study),[8] those measures of performance are only applicable to the extent that assessments of conformance are possible. Thus, determinations of reliability require consideration of results in the context of both method conformance *and* method performance.

In reviewing previously published viewpoints, we see several attempts to provide a better way of assessing the reliability of analysts' decisions. However, there are three general issues that we consider to have caused many of these prior viewpoints to be incomplete: (1) error rates alone (i.e., false positive and false negative rates) have been used as primary measures of method performance despite being unsuitable for non-binary conclusion frameworks, (2) measures of reproducibility (or other factors that do not consider decision outcomes in relation to ground-truth) have been conflated with measures of discriminability, and (3) assessments of method conformance have not been fully considered as a necessary factor for determinations of reliability for a particular case. A brief description of the viewpoints from eight different articles is provided in Table 2. A summary assessment of each article and a more detailed discussion of these three issues follows.[9]

Dror and Langenburg (2019) [11], Dror and Scurich (2020) [12], Hofmann et al. (2020) [14], and Dorfman and Valliant (2022) [17] focused predominantly on the use of error rates as primary measures of performance. In doing so, they offered multiple alternative definitions of error rates through different treatments of inconclusive responses. These alternative definitions conflate (explicitly or implicitly) measures of reproducibility (or other factors that do not consider decision outcomes in relation to ground truth) with measures of discriminability (i.e., suggesting that analysts' decisions that are not consistent with majority or expert panels, or do not conform to method-specific decision criteria, can be represented as erroneous outcomes). The decision-specific metrics discussed by Hofmann et al. [14] are affected by the prior odds of mated versus non-mated samples. For a performance study, this is determined by the arbitrary choice of the ratio of the respective comparisons. For court testimony, the evaluation of prior odds is typically outside the purview of the forensic evaluation. Thus, such decision-specific metrics do not provide clear information regarding a method's ability to discriminate between the propositions of interest. Arkes and Koehler (2021) [16] seemed to implicitly perpetuate the use of error rates as primary measures of performance. They did, however, touch on the concept of method conformance as distinct from method performance. Weller and Morris (2020) [13], Biedermann and Kotsoglou (2021) [15], and Guyll et al. (2023) [18] recognized the misleading and incomplete nature of error rates when used as sole measures of method performance for non-binary conclusion scales and instead advocated for presenting all decision outcomes when representing performance. Guyll et al. [18] touched on the concept of method conformance as distinct from method performance. However, framing conformance considerations as "decision correctness" conflates the concepts and may cause confusion. Guyll et al. [18] went further and proposed an alternative non-error rate metric—a likelihood ratio for each possible result—that can help convey how successfully one could use the method output to distinguish non-mated comparisons from mated comparisons.

2.1. Issue 1: focusing solely on two (error) rates

The first issue of concern is the focus on two (error) rates to represent method performance for non-binary conclusion scales. This approach overlooks important details about the performance of the method, and the array of proposals for different ways of computing false positive and false negative rates could be seen as a discussion of which details should be overlooked. That is, using two error rates as a sole measure of performance loses information relative to presenting the rate of each decision level (e.g., exclusion, inconclusive, identification) for non-mated comparisons and for mated comparisons (e.g., a 2×3 table, representing the two ground-truth states and three possible decision outcomes, as illustrated by Tables 1a and 1b). This is evident by noting that, regardless of what definitions are adopted for false positive rate and false negative rate, the full 2×3 table is not recoverable from these two numbers. For each of the proposed approaches for computing error rates, examples can be readily constructed of two methods that produce identical error rates but have different abilities to discriminate non-mated comparisons from mated comparisons or have different levels of reproducibility. Thus, for non-binary conclusion scales, error rates alone do not provide sufficient information for characterizing method performance (i.e., discriminability and reproducibility). This issue of losing information also extends to other summaries of performance where the full 2×3 table is not recoverable, such as the area under the receiver operator characteristic curve (AUC) or empirical cross entropy (ECE) [19].

Additionally, computing error rates raises the question of how to label inconclusive decisions. This has led to the various viewpoints summarized in Table 2 and some controversy because inconclusive decisions are not necessarily correct or incorrect. A "correct" decision is one that accurately represents the true source-origin state of items being compared. An "incorrect" decision is one that falsely represents the true source-origin state, resulting in an error (i.e., falsely asserting that two impressions originated from the same source or falsely asserting that two impressions originated from different sources). An inconclusive decision, on the other hand, is an outcome of the examination for which an assertion about the source-origin state of the items being compared was not explicitly made. Thus, an inconclusive decision is neither a correct nor erroneous representation of the true source-origin state. Other summaries, such as AUC or ECE offer an advantage in the sense that they do not require such binary labels; however, any summary from which the 2×3 table cannot be reconstructed is unsuitable for providing a complete characterization of a method's performance in discriminating between the propositions of interest.

Information regarding method performance should help others assess what weight to give to the method's result in a given case (for which ground-truth is not known). For instance, as noted by Guyll et al. [18], one could consider the "probative value" of the result by assessing the likelihood ratio for the analyst's decision using data collected under relevant conditions (e.g., approximated by calculating the portion of all mated comparisons for a particular decision divided by the portion of all non-mated comparisons for the same decision). This requires a complete and transparent representation of all possible outcomes as they relate to ground-truth of the compared items under specified conditions. Thus, when considering a more suitable way of conveying performance characteristics, we agree with the viewpoints and suggestions put forth by Weller and Morris [13], Biedermann and Kotsoglou [15], and Guyll et al. [18]—to provide the entire table of outputs representing all

[6] The term "discriminability" refers to the extent to which the outcomes of a method can accurately distinguish between non-mated and mated comparisons. The term "reproducibility" refers to the extent to which the outcomes of a method are consistently produced.

[7] This is important when analysts vary in their performance and measures of discriminability and reproducibility are based on aggregate outcomes from multiple analysts.

[8] The decision by a user or a group of users that a method is acceptable for its intended purpose does not obligate or constrain others (e.g., factfinders) to accept that determination when they are later tasked with evaluating the evidence in the context of a case.

[9] We do not claim this to be a comprehensive list. The eight articles presented here illustrate a range of viewpoints on the topic.

possible outcomes (e.g., a 2 × 3 table, such as that represented in Tables 1a, 1b, 3a, and 3b).[10] This provides greater transparency about the method's performance and enables users of the information to more effectively discriminate between propositions of interest (i.e., mated versus non-mated).

Consider the following 2 × 3 tables describing results of validation testing from hypothetical methods 3 and 4, reflected in Tables 3a and 3b.

There are several performance summaries for which methods 3 and 4 appear equivalent (e.g., error rates, AUC).[11] However, the complete tables reveal several important differences between the methods. Table 3a indicates that inconclusive decisions from method 3 occur at a rate among non-mated comparisons that is four times greater than the rate among mated comparisons. Table 3b, however, indicates that inconclusive decisions from method 4 occur at a rate among mated comparisons that is four times greater than the rate among non-mated comparisons. Thus, inconclusive decisions have different implications depending on whether they resulted from method 3 or method 4. The implied "probative value" of inconclusive decisions between methods 3 and 4 differ by a factor of 16. Differences also occur for identification and exclusion decisions. Decisions made by factfinders (or others within the criminal justice system, such as investigators, litigators, or judges) in response to an expert's opinion in a given case may depend on whether the expert applied method 3 or 4 (i.e., they may make different decisions depending on whether Table 3a or 3b is provided). This example illustrates the general fact that any summary of method performance from which the 2 × 3 table cannot be inferred risks losing information important for assessing what weight to give an expert's opinion in a given case.

Presenting the complete 2 × 3 table ensures that users of the information can make the best possible decision for the relevant conditions in the case. This is particularly true when inconclusive decisions are not symmetrically distributed between mated and non-mated comparisons. Excluding inconclusive decisions, combining them into a different category of decisions (for purposes of labeling them as correct or incorrect decisions),[12] or only representing incomplete summary statistics reflecting a subset of performance characteristics of the method (such that the 2 × 3 table cannot be reconstructed) prevents a meaningful interpretation of the performance of the method. Instead, such treatment of inconclusive decisions causes those performance characteristics to be represented in a distorted and potentially misleading way that can ultimately lead to fewer accurate factfinder decisions overall. Appendix I discusses this in more detail based on two pillars of statistical inference dealing with optimal decision making—Bayesian decision theory [20,21] and the Neyman-Pearson Lemma [22].

2.2. Issue 2: conflating reproducibility with discriminability

The second issue of concern is the suggestion that measures of reproducibility can be used as the basis for representing measures of discriminability of the method. Measures of reproducibility do not consider decision outcomes in relation to ground-truth; thus, they cannot provide a complete representation of the accuracy of an outcome or a method's utility in discriminating between non-mated and mated comparisons. At most, they provide limited information regarding discriminability (i.e., imperfect reproducibility indicates imperfect accuracy).

One approach to represent reproducibility data for a three-point conclusion scale is through a 3 × 3 table (e.g., Table 4). The data reflected in 3 × 3 tables provide an indication of the adequacy of the procedures that define the method. A 3 × 3 table formed using outcomes that have been assessed as properly conforming to the procedures that define a particular method reflects the extent to which the method can produce consistent results and the variability between laboratories or analysts for a given input and conditions. To the extent that measures of reproducibility among such decisions (i.e., variability among laboratories or analysts) are acceptable, the procedures that define the method and approaches for assessing conformance are adequate (i.e., the method is sufficiently well-defined and conformance to those procedures can be effectively demonstrated). However, if the measures of reproducibility among such decisions are such that it is common for different analysts to reach different decisions for a given input and conditions, or if the extent of the variability is otherwise unacceptable, then the procedures that define the method might be not be adequately specified (i.e., loosely defined) or the approaches for assessing conformance might not be sufficient (i.e., outcomes have been improperly assessed as conforming).

The data reflected in 3 × 3 tables also provide an indication of the extent to which aggregate measures of discriminability (reflected by a 2 × 3 table) across multiple analysts for a given method are relevant to a particular analyst's application of that method. While high measures of reproducibility indicate that analysts are performing with similar levels of discriminability, this is not necessarily true when measures of reproducibility are lower. Although lower measures of reproducibility will have some impact on aggregate measures of discriminability, it might not be clear whether that impact is due to some analysts performing poorly and other analysts performing well or due to all analysts performing mediocre. In other words, when measures of reproducibility are low, there could be substantial differences between assessments of performance based on the pooled 2 × 3 discrimination table and the corresponding table constructed using data for any given individual analyst. In that case, when presented with an outcome from a particular analyst for whom individual performance data is not available (as is often the case in practice), there will be no way to know where that analyst aligns in terms of the full range of performance among other analysts represented by the aggregate performance data. Thus, aggregate measures of reproducibility provide a gauge by which measures of discriminability (based on outcomes from multiple analysts generally) are relevant to an outcome by a particular analyst (specifically).

Measures of reproducibility (e.g., as reflected in 3 × 3 tables) can be obtained without knowing the ground-truth state (i.e., whether the comparisons are mated or non-mated), and can therefore be evaluated from actual casework data, at least conceptually. While these tables provide useful information, no summary from a 3 × 3 reproducibility table can provide the essential information contained in a 2 × 3

[10] For feature comparison disciplines, this can be accomplished using a 2 × 3 table or equivalent rate parameters reflecting the occurrence of identification, exclusion, and inconclusive decisions as they relate to ground-truth of the compared items. A 2 × 3 table is used in this discussion; however, this recommendation generalizes to a 2x*k* table, where *k* is the total number of possible outcomes that can be produced by the method, such as feature comparison disciplines that employ a 5-level scale, 7-level scale, 9-level scale, or another similar type of scale.

[11] Tables 3a and 3b lead to different ECE curves, which are reflections of each other about the vertical axis. However, permuting the column labels (i.e., identification, inconclusive, exclusion) in any 2 × 3 table will produce an identical ECE curve. This means that ECE curves also omit information relevant to assessing the weight of an expert opinion. See Appendix II for an example.

[12] For example, by calculating error rates after combining inconclusive decisions with identification decisions or exclusion decisions (i.e., treating all inconclusive decisions as if they were identification decisions or exclusion decisions), as was briefly discussed by Hofmann et al. [14] and Cuellar et al. (2024) [24]. Cuellar et al. [24] reference the Food and Drug Administration (FDA) Guidance for evaluating diagnostic testing when "equivocal" or "indeterminant" results are obtained [25]. While the FDA Guidance provides a means of representing a bounded range for possible error rates, the FDA recognize "[t]his may or may not be reasonable for [a given] situation" [25]. In the context of forensic science, we do not believe the FDA guidance is applicable or appropriate because it masks the actual outcomes produced by the method when tested, does not provide a complete representation of the performance of the method, and hinders the ability for a factfinder to assess the weight of a particular result.

Table 4
An example 3 × 3 table representing the reproducibility of decisions for a method. The table reflects the extent to which multiple applications of the same method between different laboratories or analysts produce consistent results. A well-defined method will yield a high proportion of consistent outcomes. Inconsistent outcomes reflect the extent of variability between laboratories or analysts and any ambiguity on what the method can be expected to produce for a given input and conditions.

Reproducibility	Identification	Inconclusive	Exclusion
Identification	*Consistent*	Inconsistent	Inconsistent
Inconclusive	Inconsistent	*Consistent*	Inconsistent
Exclusion	Inconsistent	Inconsistent	*Consistent*

discrimination table, such as those illustrated in Tables 1a, 1b, 3a, or 3b. The diagonal and off-diagonal elements of the 3 × 3 tables (labeled as "consistent" and "inconsistent" outcomes, respectively, in Table 4) are measures of (ir)reproducibility and must not be mistaken as suitable summaries of method discrimination.

This issue with using measures of reproducibility as a means of representing measures of discriminability also extends to the use of any other criteria or factors that do not consider results in relation to ground-truth (e.g., based on assessments of method conformance or comparing outcomes from one method to those from another method).[13]

2.3. Issue 3: lack of considerations for method conformance

The third issue of concern is the limited appreciation for the importance of method conformance when assessing or reporting measures of method performance. Method conformance is related to method performance. Performance data for one method is not relevant to a different method. If an analyst deviates from procedures for a particular comparison, they are no longer using the method specified by those procedures. Deviating from the procedures does not mean that an analyst is necessarily performing better or worse than those analysts following the procedures; however, it does mean that performance data for that method (i.e., from the other analysts who did follow the procedures, such as assessed during validation studies) might not adequately reflect the performance of the given analyst for the comparison in question, which could leave little or no information with which to assess the reliability of the outcome produced by the non-conforming analyst.

2.4. Evaluation of results

Taking into consideration these three issues, in the context of measuring *method performance*, we stress that the discriminability of analysts' decisions can only be assessed in terms of ground-truth, and because "inconclusive" decisions are not an assertion about the source-origin state of the items being compared, they are neither "correct" nor "incorrect." However, in the context of assessing *method conformance*, all analysts' decisions (including inconclusive decisions) should be assessed as "appropriate" or "inappropriate" in terms of whether they resulted from a proper application of a specified method. Thus, we agree with Dror and Langenburg [11] and Dror and Scurich [12], in the sense that one might wish to assess whether a particular decision, such as an inconclusive, is "justifiable." Whether a particular decision is "justifiable," however, depends on whether the outcome of the examination was "appropriate" (i.e., produced by proper conformance to the method procedures, including relevant decision criteria, if applicable) and whether empirical measures relating to the performance of that method (i.e., discriminability and reproducibility) under conditions relevant to a particular case have been deemed "acceptable." A result that is inappropriate does not mean it is incorrect; however, it does mean that there is likely little to no data with which the weight of the result can be assessed.

Consider the following two scenarios, for example, to elaborate on this point using a hypothetical method that includes explicit criteria to support decisions of identification or exclusion (e.g., specified minimum quality and quantity of corresponding or discordant features) and for which performance characteristics of the method have been deemed "acceptable" for use:

(1) When the criteria specified by a method to support a decision of identification or exclusion *have not been met:*
   a. Inconclusive decisions that are produced under this situation represent an outcome that is expected when procedures that define the method are adhered to. Such decisions reflect that the method has been applied in accordance with the scope of its validation and in a manner deemed acceptable for use. Therefore, in this situation, such decisions are *appropriate* as they relate to assessments of method conformance. Of course, the more often a method produces inconclusive outcomes, the less useful it would be and less likely the method might be deemed "acceptable" for operational use.
   b. Identification or exclusion decisions that are produced under this situation represent an outcome that is not expected when the procedures that define the method are adhered to. Such decisions reflect that the method has not been applied in accordance with the scope of its validation of what has been deemed to be acceptable. Therefore, in this situation, such decisions are *inappropriate* as they relate to assessments of method conformance. It is important to note that even if such decisions happen to be correct (based on ground-truth), they still represent an outcome that is not in conformance with the specified requirements, or criteria, deemed to be appropriate and acceptable for the intended use (i.e., the risk and consequences of producing errors when such conclusive decisions are made for a given input and conditions have been deemed to be too great).

(2) When the criteria specified by a method to support a decision of identification or exclusion *have been met:*
   a. Inconclusive decisions that are produced under this situation represent an outcome that is not expected when the procedures that define the method are adhered to. Such decisions reflect that the method has not been applied in accordance with the scope of its validation or in a manner deemed acceptable for use. Therefore, in this situation, such decisions

[13] For example, the 3 × 3 table in Fig. 1 by Dror and Scurich [12] reflects outcomes labeled as "correct" or "error" based on whether there is "sufficient information to justify such a decision," as determined by method-specific decision criteria (e.g., suggested by Dror and Langenburg [11]), consensus opinion or majority outcomes (suggested by Dror and Scurich [12]), or algorithmic assessments (suggested by Dorfman and Valliant [17]).

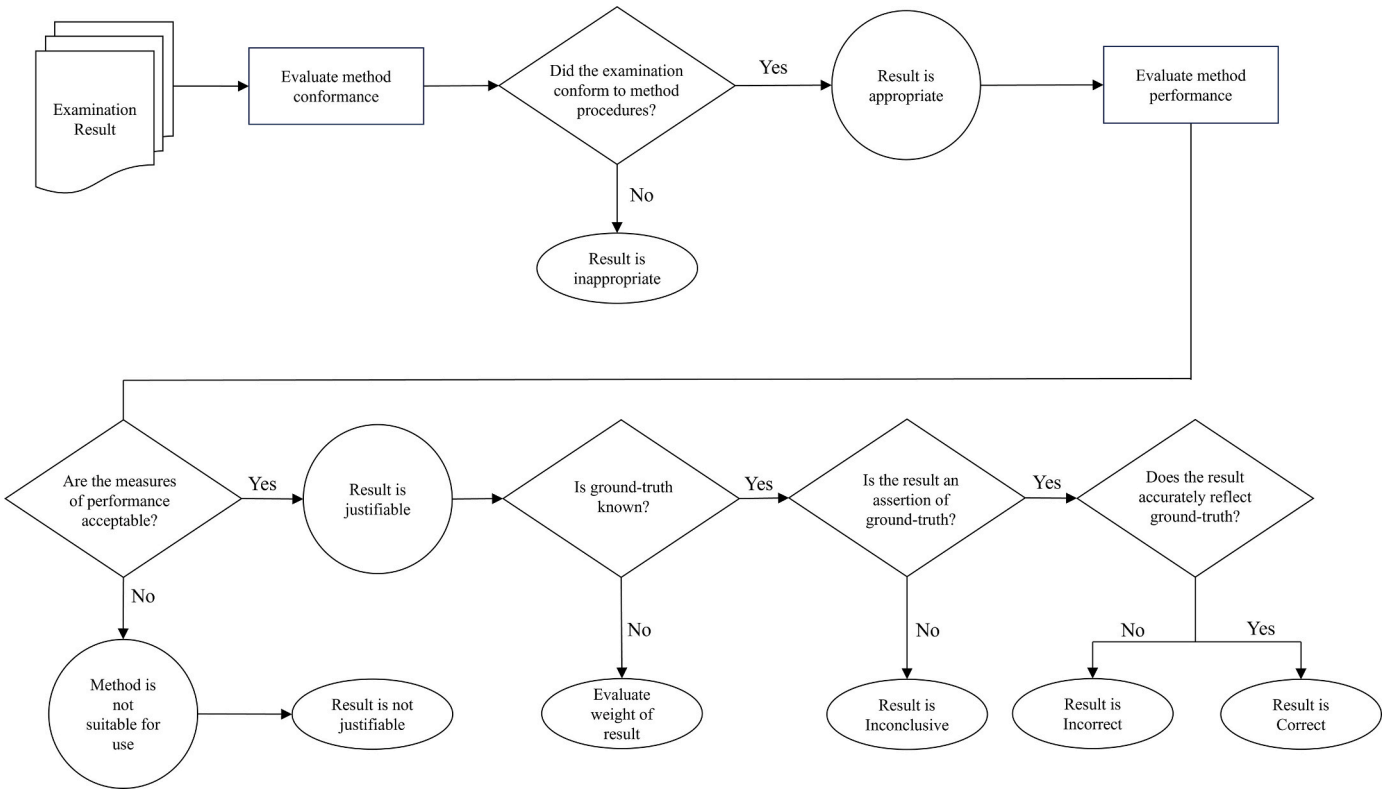


Fig. 1. Simplified flow diagram reflecting the process for evaluating examination results. The diagram illustrates the distinctions between results labeled as "appropriate" vs. "inappropriate," "justifiable" vs. "not justifiable," and "correct" vs. "incorrect."

are *inappropriate* as they relate to assessments of method conformance.

b. Identification or exclusion decisions meeting the relevant criteria that are produced under this situation represent an outcome that is expected when the procedures that define the method are adhered to. Such decisions reflect that the method has been applied in accordance with the scope of its validation of what has been deemed to be acceptable. Therefore, in this situation, such decisions (identification or exclusion, depending on the criteria relevant for each type of conclusive decision) are *appropriate* as they relate to assessments of method conformance. Like the counter-scenario described above (where an outcome might be correct yet inappropriate), it is important to note that even if such conclusive decisions provided under these circumstances happen to be incorrect, they still represent an outcome of the method that is in conformance with the specified requirements, or criteria, deemed to be appropriate and acceptable for the intended use. In other words, although there might be occasions where such decisions are incorrect, the tradeoff between correct and incorrect outcomes has been deemed acceptable to permit use of the method. Of course, the more often a method produces incorrect outcomes, the less useful it would be and less likely the method might be deemed "acceptable" for operational use.

While method conformance and method performance are both important aspects for determinations of reliability, care must be taken not to confuse or conflate the two. These two concepts are distinct, and both must be accounted for separately when considering the reliability of a particular method (e.g., during validation testing) or evaluating the weight of a particular result of a method (e.g., in a particular case). For method conformance, assessments must be based on an empirical demonstration that the established requirements and criteria inherent in the method have been satisfied (e.g., relating to analyses of quality, quantity, similarity, or rarity of comparison features and any relevant and applicable decision criteria).[14] For method performance, measures of discriminability must be assessed in terms of ground-truth (i.e., mated or non-mated comparisons) and measures of reproducibility must be assessed in terms of the consistency of decisions for a given input and conditions when the same method is applied by different analysts. Importantly, while measures of reproducibility provide an indication of the adequacy of the procedures that define the method (i.e., well-defined procedures produce more consistent results), demonstrating consistency of outcomes (e.g., agreement between analysts) post hoc is not sufficient to serve as a basis for assessing or demonstrating conformance to a method or labeling a result as "appropriate." Conformance must be assessed and empirically demonstrated based on adherence to procedures that define the method. Once conformance has been demonstrated, performance data for that method can be used to evaluate the weight of an "appropriate" result. Fig. 1 uses a simplified flow diagram to illustrate the process for evaluating examination results and the distinctions between results labeled as "appropriate" vs. "inappropriate," "justifiable" vs. "not justifiable," and "correct" vs. "incorrect."

3. Conclusion

Different treatments of inconclusive decisions and calculations of error rates in forensic feature comparison disciplines have led to different representations and interpretations of the reliability of forensic science results. In this paper, we explored these issues in further detail from a metrology perspective and distinguished between the concepts of

[14] Different approaches for analyzing quality, quantity, similarity, or rarity of comparison features (e.g., subjective versus algorithmic) or decision criteria or thresholds different from those specified by the method can impact performance and therefore reflect deviations from established procedures that define a particular method.

method conformance and *method performance*. We also considered the broader implications of these concepts when determining reliability of analysts' examination results.

The issues discussed in this paper have several practical implications to researchers and forensic service providers alike. They impact studies and activities relating to method validation and performance monitoring, as well as how results are characterized and communicated—all of which are prescribed by ISO/IEC 17025:2017 [23], the prevailing international standard to which many forensic laboratories conform—and the extent to which performance data are useful for determinations of reliability in casework.[15] Major implications of these issues and key takeaways from this paper are as follows:

First, determinations of the reliability of analysts' examination results require consideration of those results in the context of both method conformance *and* method performance—a result alone is not sufficient for one to assess its reliability.

Second, error rates alone do not adequately characterize method performance for non-binary scales. Instead, the entirety of possible outcomes should be provided as it relates to measures of discriminability (i.e., 2×3 table) and reproducibility (i.e., 3×3 table) constructed from relevant validation testing.

Third, inconclusive decisions are neither "correct" nor "incorrect" (in terms of method performance) but can be either "appropriate" or "inappropriate" (in terms of method conformance).

Fourth, studies that purport to characterize the performance of a *particular* method (i.e., validation studies) are only relevant if conformance to that method can be demonstrated. Therefore, forensic service providers that do not have well documented and detailed step-by-step procedures that define their method, including conditions for method application and decision criteria for results for which performance data can be associated are unlikely to be able to meaningfully support a claim that the outcome of their examination is the product of a reliable method.

Fifth, studies that characterize aggregate measures of performance across a discipline (e.g., black-box studies or interlaboratory comparisons) but do not specify the methods used can provide information about the performance characteristics that can be expected for the practice overall. While these studies are helpful to users of the information, they cannot necessarily serve as a validation or provide generalizable performance characteristics of a *particular* method relevant to a specific case unless it can be shown that the same method was used by all participants. The development and use of standard methods by multiple laboratories is an important step toward reducing variability and ensuring that aggregate measures of performance can be represented as generalized measures of performance for those methods. This standardization, in turn, strengthens the evidence-base[16] supporting the validation of those methods and reduces the resource burdens that would otherwise be placed on individual laboratories to accomplish these studies independently.

CRediT authorship contribution statement

H. Swofford: Writing – review & editing, Writing – original draft, Conceptualization. **S. Lund:** Writing – review & editing, Writing – original draft, Conceptualization. **H. Iyer:** Writing – review & editing, Writing – original draft, Conceptualization. **J. Butler:** Writing – review & editing, Writing – original draft, Conceptualization. **J. Soons:** Writing – review & editing, Writing – original draft, Conceptualization. **R. Thompson:** Writing – review & editing, Writing – original draft, Conceptualization. **V. Desiderio:** Writing – review & editing, Writing – original draft, Conceptualization. **J.P. Jones:** Writing – review & editing, Writing – original draft, Conceptualization. **R. Ramotowski:** Writing – review & editing, Writing – original draft, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix I

Explanation for the inadequacy of error rate summaries for factfinder decision making

Ultimately, an expert's opinion is information provided to a factfinder, who is tasked with assessing what weight to give that opinion as part of their decision-making process. Understanding what outcomes have been produced in known ground-truth scenarios (i.e., validation testing) can help factfinders assess the weight of an expert's opinion. Oftentimes, attention centers around the error rates of a given method. However, two pillars of statistical inference dealing with optimal decision making—Bayesian decision theory [20] and the Neyman-Pearson Lemma [22]—show that likelihood ratios, rather than error rates, are the quantities of interest from a 2×3 table for factfinders. Computing likelihood ratios requires assessing additional probabilities beyond those that represent error rates. Providing only error rates suppresses information relevant to assessing these additional probabilities. We elaborate on these concepts below.

Consider a factfinder evaluating the prosecution hypothesis H_p that the two impressions share the same source, relative to the defense hypothesis H_d that they do not. For simplicity, we assume that the factfinder has only two actions available—find the defendant "guilty" or find the defendant "not guilty." If the factfinder finds the defendant guilty when H_d is true it will lead to a "wrongful conviction." If the factfinder finds the defendant not guilty when H_p is true, the result will be a "false acquittal." It is desirable to avoid both situations. Bayesian decision theory provides a principled approach for arriving at an optimal decision strategy and the reader is referred to Ref. [21] for a detailed discussion of how this theory can guide a decision maker in the criminal justice system. In general terms, Bayesian decision theory suggests that, among all available decision strategies, one should choose a decision strategy that minimizes the "expected cost" of the decision.

Assessing the expected cost of a given decision requires the probabilities for various scenarios of interest and the cost the factfinder would associate with errant decisions under each of those scenarios. Suppose the costs the factfinder associates with a wrongful conviction or false acquittal are given by C_{wc} or C_{fa}, respectively. Suppose the factfinder has a prior probability p for H_p (and 1- p for H_d).[17] This prior probability reflects the factfinder's state of uncertainty before hearing from the expert and will be updated after learning the result from the forensic analysis.

[15] When considering these issues, it is important to keep in mind that ISO/IEC 17025:2017 specifies that the term "method" is "considered synonymous with the term 'measurement procedure' as defined in ISO/IEC Guide 99" and is referred to as being defined by a *specific* and *detailed* step-by-step procedure, referred to as a standard operating procedure [23,26].

[16] The term "evidence-base" refers to empirical data reflecting the performance of the method under varying conditions.

[17] The prior probability is independently determined by the factfinder based on their prior belief that H_p is true or H_d is true.

The process of updating uncertainty in response to new information can be conducted using Bayes' equation, which requires a likelihood of the new information under each of the scenarios of interest.

Table AI-1 provides the probabilities for the different outcomes an analyst might reach in H_p-true and H_d-true scenarios, respectively.[18] In table AI-1, the value of P1 represents the probability that an expert would provide an "ID" after evaluating a pair of impressions for which H_p is true, and the value of Q1 represents the probability that an expert would provide an "ID" after evaluating a pair of impressions for which H_d is true.

Table AI.1
A 2 × 3 table of the probabilities for different conclusions an analyst might reach in H_p-true and H_d-true scenarios.

Scenario	Identification (ID)	Inconclusive	Exclusion	Total
H_p-true	P1	P2	P3	100 %
H_d-true	Q1	Q2	Q3	100 %

Let us focus on the situation where the analyst result is "ID". In this case, the factfinder would like to update their prior probability estimate p in light of the expert's decision. Using Bayes rule, we get:

$$P(H_p|Expert\ says\ ID) = \frac{p \bullet P1}{p \bullet P1 + (1-p) \bullet Q1} = \frac{\frac{p}{1-p} \bullet \frac{P1}{Q1}}{1 + \frac{p}{1-p} \bullet \frac{P1}{Q1}} \quad \text{Equation (1)}$$

We extended equation (1) to illustrate that the evaluation requires the values of P1 and Q1, and includes the ratio of P1/Q1. A factfinder can use their estimated posterior probability to assess their expected cost associated with a decision to convict or to acquit. The expected cost is used to assess whether one decision is better than another—a decision with a lower expected cost is preferred. In this setup, the expected costs for the factfinder's available decisions are:

$$\text{Expected cost of acquittal} = C_{fa} \bullet P(H_p|Expert\ says\ \text{ID})$$

$$\text{Expected cost of conviction} = C_{wc} \bullet P(H_d|Expert\ says\ \text{ID}),$$

where C_{fa} is the cost of a false acquittal and C_{wc} is the cost of a wrongful conviction. Ultimately, it is only the ratio $C = C_{wc}/C_{fa}$ that matters when comparing expected costs of different decisions. The quantity C represents how many false acquittals the factfinder would exchange to avoid one false conviction. For simplicity, and without loss of generality, it is common to consider relative costs by taking $C_{fa} = 1$ and $C_{wc} = C$. Thus, we get:

$$\text{Expected cost of acquittal} = P(H_p|Expert\ says\ \text{ID})$$

$$\text{Expected cost of conviction} = \text{C} \bullet P(H_d|Expert\ says\ \text{ID}).$$

To apply the Bayesian decision-making paradigm, which is generally accepted as normative [21], the factfinder simply picks whichever choice has the lower expected cost. Note that equation (1) makes clear that this process depends on the value of $P1/Q1$. Thus, P1/Q1 is an important component of Bayesian reasoning.

We continue this explanation to provide another theoretical motivation for the importance of $P1/Q1$. Under the above setup, a factfinder's expected cost of conviction would be lower than their expected cost of acquittal if and only if:

$$C < \frac{P(H_p|Expert\ says\ ID)}{P(H_d|Expert\ says\ ID)} \quad \text{Equation (2)}$$

The right-hand side of this expression is the posterior odds. In the case where exactly two propositions are considered, Bayes rule shows this is equal to:

$$\frac{P(H_p|Expert\ says\ ID)}{P(H_d|Expert\ says\ ID)} = \frac{P1}{Q1} \bullet \frac{p}{1-p} \quad \text{Equation (3)}$$

where $P1/Q1$ represents the likelihood ratio (LR) that links the prior odds to the posterior odds. (A more general form of Bayes rule, in which the LR is replaced with a Bayes factor, applies to situations when more than two propositions are considered.)

With some algebra, this means the factfinder's expected cost of conviction would be lower than their expected cost of acquittal if and only if:

$$\frac{P1}{Q1} > C \bullet \frac{1-p}{p} = \frac{C}{Prior\ odds\ of\ H_p} = \tau \quad \text{Equation (4)}$$

This provides a decision rule in the form of: "find the defendant guilty if and only if LR $P1/Q1$ is bigger than the threshold τ," where τ indicates the factfinder's threshold for how probative the expert's opinion must be in order for them to decide the defendant is guilty.

According to the Neyman-Pearson Lemma [22], decision rules based on whether or not a LR is greater than a given threshold are optimal in the sense that no other type of decision rule can produce a higher true positive rate for any given false positive rate (i.e., no other rule could produce more just convictions while maintaining a given rate of false convictions).[19] Implementing this optimal decision rule required the value $P1/\ Q1$.

We have shown, under two hallmarks of statistical reasoning, that the ratio $P1/Q1$ is directly relevant to the factfinder when the expert says "ID."

[18] A 2 × 3 table is used in this discussion; however, this generalizes to a 2xk table, where k is the total number of possible outcomes that can be produced by the method, such as those feature comparison disciplines that might employ a 5-level scale, 7-level scale, 9-level scale, or another similar type of scale.

[19] The optimality only applies with respect to expected performance according to the provided probabilities. In theoretical exercises where the probabilities represent long-run relative frequencies, the optimality is in terms of long-run observed performance.

Similar reasoning shows that the ratio $P2/Q2$ is important to the factfinder when the expert says "inconclusive," and the ratio $P3/Q3$ is important when the expert says "exclusion." Thus, it is critical that factfinders have access to information that would assist their assessments of these ratios. Summarizing performance using error rates alone (or any other summary from which the 2×3 table cannot be reconstructed) deprives the factfinder of information relevant for updating their beliefs.

Appendix II

Limitation of Empirical Cross Entropy

Empirical cross entropy (ECE) produces identical curves for tables AII-1 and AII-2 below. See equation 6.4 in Ref. [19]. However, the implied likelihood ratios for an "ID" in tables AII-1 and AII-2 are 59 %/1 % = 59 and 40 %/10 % = 4, respectively. This illustrates that ECE curves do not convey all the relevant information from a 2×3 table.

Table AII.1
A 2×3 table representing performance metrics relating to hypothetical Method A.

Method A	Identification (ID)	Inconclusive	Exclusion
Mated Comparisons	59 %	40 %	1 %
Non-Mated Comparisons	1 %	10 %	89 %

Table AII.2
A 2×3 table representing performance metrics relating to hypothetical Method B.

Method B	Identification (ID)	Inconclusive	Exclusion
Mated Comparisons	40 %	59 %	1 %
Non-Mated Comparisons	10 %	1 %	89 %

References

[1] National Research Council Committee on Identifying the Needs of the Forensic Sciences Community, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington, D.C. USA, 2009, https://doi.org/10.17226/12589.

[2] President's Council of Advisors on Science and Technology, Report to the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, U.S. Executive Office of the President, Washington, D.C., USA, 2016.

[3] v Daubert, Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 1993.

[4] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Accuracy and reliability of forensic latent fingerprint decisions, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 7733–7738.

[5] I. Pacheco, B. Cerchiai, S. Stoiloff, Miami-dade Research Study for the Reliability of the ACE-V Process: Accuracy & Precision in Latent Fingerprint Examinations (Final Report for NIJ Award 2010-DN-BX-K268), National Institute of Justice, 2014. http://www.ncjrs.gov/pdffiles1/nij/grants/248534.pdf.

[6] D.P. Baldwin, S.J. Bajic, M. Morris, D. Zamzow, A study of false-positive and false-negative error rates in cartridge case comparisons. https://www.ojp.gov/pdffiles1/nij/249874.pdf, 2014.

[7] B.T. Ulery, R.A. Hicklin, J. Buscaglia, M.A. Roberts, Repeatability and reproducibility of decisions by latent fingerprint examiners, PLoS One 7 (2012) e32800.

[8] B.T. Ulery, R.A. Hicklin, G.I. Kiebuzinski, M.A. Roberts, J. Buscaglia, Understanding the sufficiency of information for latent fingerprint value determinations, Forensic Sci. Int. 230 (2013) 99–106, https://doi.org/10.1016/j.forsciint.2013.01.012.

[9] B.T. Ulery, R.A. Hicklin, M.A. Roberts, J. Buscaglia, Measuring what latent fingerprint examiners consider sufficient information for individualization determinations, PLoS One 9 (2014) e110179, https://doi.org/10.1371/journal.pone.0110179.

[10] B.T. Ulery, R.A. Hicklin, M.A. Roberts, J. Buscaglia, Changes in latent fingerprint examiners' markup between analysis and comparison, Forensic Sci. Int. 247 (2015) 54–61, https://doi.org/10.1016/j.forsciint.2014.11.021.

[11] I.E. Dror, G. Langenburg, "Cannot decide": the fine line between appropriate inconclusive determinations versus unjustifiably deciding not to decide, J. Forensic Sci. 64 (2019) 10–15.

[12] I.E. Dror, N. Scurich, (Mis)use of scientific measurements in forensic science, Forensic Sci. Int.: Synergy 2 (2020) 333–338, https://doi.org/10.1016/j.fsisyn.2020.08.006.

[13] T.J. Weller, M.D. Morris, Commentary on: I. Dror, N Scurich "(Mis)use of scientific measurements in forensic science", Forensic Sci. Int.: Synergy (2020) https://doi.org/10.1016/j.fsisyn.2020.08.006. ForensicSci.Int.: Synergy 2 (2020) 701–702.

[14] H. Hofmann, A. Carriquiry, S. Vanderplas, Treatment of inconclusives in the AFTE range of conclusions, Law Probab. Risk 19 (2020) 317–364.

[15] A. Biedermann, K.N. Kotsoglou, Forensic science and the principle of excluded middle: "inconclusive" decisions and the structure of error rate studies, Forensic Sci. Int.: Synergy 3 (2021) 1–11, https://doi.org/10.1016/j.fsisyn.2021.100147.

[16] H.R. Arkes, J.J. Koehler, Inconclusives and error rates in forensic science: a signal detection theory approach, Law Probab. Risk 20 (2021) 153–168.

[17] A.H. Dorfman, R. Valliant, Inconclusives, errors, and error rates in forensic firearms analysis: three statistical perspectives, Forensic Sci. Int.: Synergy 5 (2022) 1–8, https://doi.org/10.1016/j.fsisyn.2022.100273.

[18] M. Guyll, S. Madon, Y. Yang, K.A. Burd, G. Wells, Validity of forensic cartridge-case comparisons, Proc. Natl. Acad. Sci. U. S. A. 120 (2023) e2210428120.

[19] G. Zadora, A. Martyna, D. Ramos, C. Aitken, Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data, John Wiley & Sons, 2013.

[20] J.O. Berger, Statistical Decision Theory and Bayesian Analysis, Springer Science & Business Media, 2013.

[21] A. Biedermann, S. Bozza, F. Taroni, Analysing and exemplifying forensic conclusion criteria in terms of bayesian decision theory, Sci. Justice 58 (2018) 159–165.

[22] J. Neyman, E.S. Pearson, IX. On the Problem of the most efficient Tests of statistical Hypotheses, Philos. Trans. R. Soc. Lond. - Ser. A Contain. Pap. a Math. or Phys. Character 231 (1933) 289–337. https://royalsocietypublishing.org/doi/10.1098/rsta.1933.0009.

[23] International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC), 17025:2017 General Requirements for the Competence of Testing and Calibration Laboratories, (n.d.). https://www.iso.org/standard/66912.html...

[24] M. Cuellar, S. Vanderplas, A. Luby, M. Rosenblum, Methodological problems in every black-box study of forensic firearm comparisons, ArXiv Preprint ArXiv: 2403.17248 (2024) 1–51, https://doi.org/10.48550/arXiv.2403.17248.

[25] U.S. Food and Drug Administration, Guidance for Industry and FDA Staff: Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests, U.S. Department of Health and Human Services, 2007. https://www.fda.gov/media/71147/download.

[26] International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC), ISO/IEC Guide 99:2007 International Vocabulary of Metrology—Basic and General Concepts and Associated Terms (VIM), (n.d.). https://www.iso.org/standard/45324.html...