Kobina Ebo Abruquah v. State of Maryland, No. 10, September Term, 2022. Opinion by Fader, C.J.

EVIDENCE – EXPERT EVIDENCE

Firearms identification examiner testifying as an expert witness should not have been permitted to offer an unqualified opinion that crime scene bullets and a bullet fragment were fired from the petitioner's gun. The reports, studies, and testimony presented to the circuit court demonstrate that the firearms identification methodology employed by the examiner in this case can support reliable conclusions that patterns and markings on bullets are consistent or inconsistent with those on bullets fired from a particular known firearm. Those reports, studies, and testimony do not, however, demonstrate that the methodology used can reliably support an unqualified conclusion that such bullets were fired from a particular firearm.

Circuit Court for Prince George's County
Case No. CT121375X
Argued: October 4, 2022

IN THE SUPREME COURT

OF MARYLAND*

No. 10

September Term, 2022

KOBINA EBO ABRUQUAH

v.

STATE OF MARYLAND

Fader, C.J.,
Watts,
Hotten,
Booth,
Biran,
Gould,
Eaves,

JJ.

Opinion by Fader, C.J.
Hotten, Gould, and Eaves, JJ., dissent.

Filed: June 20, 2023

* At the November 8, 2022 general election, the voters of Maryland ratified a constitutional amendment changing the name of the Court of Appeals of Maryland to the Supreme Court of Maryland. The name change took effect on December 14, 2022.

Firearms identification, a subset of toolmark identification, is "the practice of investigating whether a bullet, cartridge case or other ammunition component or fragment can be traced to a particular suspect weapon." *Fleming v. State*, 194 Md. App. 76, 100-01 (2010). The basic idea is that (1) features unique to the interior of any particular firearm leave unique, microscopic patterns and marks on bullets and cartridge cases that are fired from that firearm, and so (2) by comparing patterns and marks left on bullets and cartridge cases found at a crime scene ("unknown samples") to marks left on bullets and cartridge cases fired from a known firearm ("known samples"), firearms examiners can determine whether the unknown samples were or were not fired from the known firearm.

At the trial of the petitioner, Kobina Ebo Abruquah, the Circuit Court for Prince George's County permitted a firearms examiner to testify, without qualification, that bullets left at a murder scene were fired from a gun that Mr. Abruquah had acknowledged was his. Based on reports, studies, and testimony calling into question the reliability of firearms identification analysis, Mr. Abruquah contends that the circuit court abused its discretion in permitting the firearms examiner's testimony. The State, relying on different studies and testimony, contends that the examiner's opinion was properly admitted.

Applying the analysis required by *Rochkind v. Stevenson*, 471 Md. 1 (2020), we conclude that the examiner should not have been permitted to offer an unqualified opinion that the crime scene bullets were fired from Mr. Abruquah's gun. The reports, studies, and testimony presented to the circuit court demonstrate that the firearms identification methodology employed in this case can support reliable conclusions that patterns and markings on bullets are consistent or inconsistent with those on bullets fired from a

particular firearm. Those reports, studies, and testimony do not, however, demonstrate that that methodology can reliably support an unqualified conclusion that such bullets were fired from a particular firearm.

The State also contends that any error in the circuit court's admission of the examiner's testimony was harmless. Because we are not convinced "beyond a reasonable doubt, that the error in no way influenced the verdict," *Dionas v. State*, 436 Md. 97, 108 (2013) (quoting *Dorsey v. State*, 276 Md. 638, 659 (1976)), we must reverse and remand for a new trial.

BACKGROUND

Factual Background

On August 3, 2012, police responded to three separate calls complaining of disturbances at the house that Mr. Abruquah shared with his roommate, Ivan Aguirre-Herrera. On the third of these occasions, just before midnight, two officers arrived at the house. According to the officers, Mr. Abruquah appeared "agitated," "very aggressive," and uncooperative. One of the officers testified that Mr. Aguirre-Herrera appeared to be terrified of Mr. Abruquah. Before leaving around 12:15 a.m., the officers told the men to stay away from each other.

A neighbor of Messrs. Abruquah and Aguirre-Herrera testified that he heard multiple gunshots sometime between 11:30 p.m. on August 3 and 12:30 a.m. on August 4.

Four days later, officers discovered Mr. Aguirre-Herrera's body decomposing in his bedroom. An autopsy revealed that he had been shot five times, including once in the back

of the head. The police recovered four bullets and two bullet fragments from the crime scene.

During questioning, Mr. Abruquah told the police that he owned two firearms, both hidden in the ceiling of the basement of the residence he shared with Mr. Aguirre-Herrera. The police recovered both firearms, a Glock pistol and a Taurus .38 Special revolver.

A jailhouse informant testified that Mr. Abruquah had said that he had engaged in "a heated argument" with Mr. Aguirre-Herrera, "snapped," and shot him with "a 38" that he kept in the ceiling of his basement.[1]

Procedural Background

Mr. Abruquah was convicted by a jury of first-degree murder and related handgun offenses in December 2013. *Abruquah v. State*, No. 246, Sept. Term 2014, 2016 WL 7496174, at *1 & n.1 (Md. App. Dec. 20, 2016). In an unreported opinion, the Appellate Court of Maryland (then named the Court of Special Appeals)[2] reversed the judgment and remanded the case for a new trial on grounds that are not relevant to the current appeal. *Id.* at *9.

On remand, Mr. Abruquah filed a motion in limine to exclude firearms identification evidence the State intended to offer through its expert witness, Scott McVeigh, a senior firearms examiner with the Firearms Examination Unit of the Prince George's County

[1] The jailhouse informant testified at Mr. Abruquah's first trial in 2013. At his second trial, in 2018, the State read into the record a transcript of that prior testimony.

[2] At the November 8, 2022 general election, the voters of Maryland ratified a constitutional amendment changing the name of the Court of Special Appeals of Maryland to the Appellate Court of Maryland. The name change took effect on December 14, 2022.

Police Department, Forensic Science Division. The circuit court held a four-day *Frye-Reed* hearing[3] during which both parties introduced evidence and elicited testimony that we summarize below.

Following the hearing, the circuit court largely denied, but partially granted, the motion. The court concluded that "firearm and toolmark identification is still generally accepted and sufficiently reliable under the *Frye-Reed* standard" and therefore should not be "excluded in its entirety." Nonetheless, the court agreed with Mr. Abruquah that the subjective nature of the matching analysis made it inappropriate for an expert to "testify to any level of practical certainty/impossibility, ballistic certainty, or scientific certainty that a suspect weapon matches certain bullet or casing striations." The court thus restricted the expert to opining whether the bullets and bullet fragment "recovered from the murder scene fall into any of" a particular set of five classifications, one of which is "[i]dentification" of the unknown bullet as a match to a known bullet.

At trial, Mr. McVeigh testified about the process by which he eliminated the Glock pistol as a source of the unknown crime scene samples, created known samples from the Taurus revolver, and compared the microscopic patterns and markings on the two sets of samples. Over defense objection, Mr. McVeigh opined that four bullets and one bullet

[3] Prior to our decision in *Rochkind v. Stevenson*, 471 Md. 1 (2020), courts in Maryland determined the admissibility of expert testimony using the *Frye-Reed* evidentiary standard, which "turned on the 'general acceptance' of such evidence 'in the particular field in which it belongs.'" *Rochkind*, 471 Md. at 4 (discussing *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923) and *Reed v. State*, 283 Md. 374 (1978)).

fragment recovered from the crime scene "at some point had been fired from [the Taurus revolver]."[4]

Mr. Abruquah was again convicted of first-degree murder and use of a handgun in the commission of a crime. His first appeal from that conviction resulted in a remand to the circuit court to consider whether it "would reach a different conclusion concerning the admission of firearm and toolmark identification testimony" applying our then-new decision in *Rochkind v. Stevenson*, 471 Md. 1, 27 (2020). In that decision, which was issued after Mr. Abruquah's second conviction while his appeal was pending, we abandoned the *Frye-Reed* standard for admissibility of expert testimony in favor of the standard set forth in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993), and its progeny. *Abruquah v. State*, 471 Md. 249, 250 (2020).

On remand, the circuit court held a hearing in which it once again received evidence from both sides, which is discussed further below. The court ultimately issued an opinion in which it reviewed each of the ten factors this Court set forth in *Rochkind* and concluded that the testimony remained admissible. The court noted that although Mr. Abruquah "ha[d] made a Herculean effort to demonstrate why the evidence should be heavily scrutinized, questioned and potentially impeached, the State has met the burden for admissibility of this evidence." The court therefore sustained Mr. Abruquah's prior conviction.

[4] Four bullets and two bullet fragments were recovered from the crime scene but Mr. McVeigh found that one of the fragments was not suitable for comparison. As a result, his testimony was limited to the bullets and one of the fragments.

Mr. Abruquah filed another timely appeal to the intermediate appellate court and, while that appeal was pending, he filed a petition for writ of certiorari in this Court. We granted that petition to address whether the firearms identification methodology employed by Mr. McVeigh is sufficiently reliable to allow a firearms examiner, without any qualification, to identify a specific firearm as the source of a questioned bullet or cartridge case found at a crime scene. *See Abruquah v. State*, 479 Md. 63 (2022).

DISCUSSION

We review a circuit court's decision to admit expert testimony for an abuse of discretion. *Rochkind*, 471 Md. at 10. Under that standard, we will "not reverse simply because . . . [we] would not have made the same ruling." *State v. Matthews*, 479 Md. 278, 305 (2022) (quoting *Devincentz v. State*, 460 Md. 518, 550 (2018)). In connection with the admission of expert testimony, where circuit courts are to act as gatekeepers in applying the factors set out by this Court in *Rochkind*, a circuit court abuses its discretion by, for example, admitting expert evidence where there is an analytical gap between the type of evidence the methodology can reliably support and the evidence offered.[5] *See Rochkind*, 471 Md. at 26-27.

[5] This Court has frequently described an abuse of discretion as occurring when "no reasonable person would take the view adopted by the circuit court" or when a decision is "well removed from any center mark imagined by the reviewing court and beyond the fringe of what the court deems minimally acceptable." *Matthews*, 479 Md. at 305 (first quoting *Williams v. State*, 457 Md. 551, 563 (2018), and next quoting *Devincentz v. State*, 460 Md. 518, 550 (2018)). In our view, the application of those descriptions to a trial court's application of a newly adopted standard, such as that adopted by this Court in *Rochkind* as applicable to the admissibility of expert testimony, is somewhat unfair. In this case, in the absence of additional caselaw from this Court implementing the newly adopted standard, the circuit court acted deliberately and thoughtfully in approaching, analyzing,

Part I of our discussion sets forth the standard for the admissibility of expert testimony in Maryland following this Court's decision in *Rochkind v. Stevenson*, 471 Md. 1 (2020). In Part II, we discuss general background on the firearms identification methodology employed by the State's expert witness, criticisms of that methodology, studies of the methodology, the testimony presented to the circuit court, and caselaw from other jurisdictions. In Part III, we apply the factors set forth in *Rochkind* to the evidence before the circuit court.

I. THE ADMISSIBILITY OF EXPERT TESTIMONY

The admissibility of expert testimony is governed by Rule 5-702, which provides:

> Expert testimony may be admitted, in the form of an opinion or otherwise, if the court determines that the testimony will assist the trier of fact to understand the evidence or to determine a fact in issue. In making that determination, the court shall determine
>
> (1) whether the witness is qualified as an expert by knowledge, skill, experience, training, or education,
>
> (2) the appropriateness of the expert testimony on the particular subject, and
>
> (3) whether a sufficient factual basis exists to support the expert testimony.

Trial courts analyzing the admissibility of evidence under Rule 5-702 are to consider the following non-exhaustive list of "factors in determining whether the proffered expert testimony is sufficiently reliable to be provided to the trier of facts," *Matthews*, 479 Md. at 310:

---

and resolving the question before it. This Court's majority has come to a different conclusion concerning the outer bounds of what is acceptable expert evidence in this area.

> (1) whether a theory or technique can be (and has been) tested;
>
> (2) whether a theory or technique has been subjected to peer review and publication;
>
> (3) whether a particular scientific technique has a known or potential rate of error;
>
> (4) the existence and maintenance of standards and controls; . . .
>
> (5) whether a theory or technique is generally accepted[;]
>
> . . .
>
> (6) whether experts are proposing to testify about matters growing naturally and directly out of research they have conducted independent of the litigation, or whether they have developed their opinions expressly for purposes of testifying;
>
> (7) whether the expert has unjustifiably extrapolated from an accepted premise to an unfounded conclusion;
>
> (8) whether the expert has adequately accounted for obvious alternative explanations;
>
> (9) whether the expert is being as careful as [the expert] would be in [the expert's] regular professional work outside [the expert's] paid litigation consulting; and
>
> (10) whether the field of expertise claimed by the expert is known to reach reliable results for the type of opinion the expert would give.

Rochkind, 471 Md. at 35-36 (first quoting *Daubert*, 509 U.S. at 593-94 (for factors 1-5) and next quoting Fed. R. Evid. 702 Advisory Committee Note (cleaned up) (for factors 6-10)).

In applying these "*Daubert-Rochkind* factors," we have observed that the guidance provided by the United States Supreme Court in *Daubert* and its progeny, especially *General Electric Co. v. Joiner*, 522 U.S. 136 (1997), and *Kumho Tire Co. v. Carmichael*,

526 U.S. 137 (1999), "is critical to a trial court's reliability analysis." *Rochkind*, 471 Md. at 36. In *Matthews*, we summarized that guidance in five principles:

- "[T]he reliability inquiry is 'a flexible one.'" *Matthews*, 479 Md. at 311 (quoting *Rochkind*, 471 Md. at 36).
- "[T]he trial court must focus solely on principles and methodology, not on the conclusions that they generate. However, conclusions and methodology are not entirely distinct from one another. Thus, [a] trial court . . . must consider the relationship between the methodology applied and conclusion reached." *Id.* (internal citations and quotation marks omitted).
- "[A] trial court need not admit opinion evidence that is connected to existing data only by the *ipse dixit* of the expert; rather, [a] court may conclude that there is simply too great an analytical gap between the data and the opinion proffered." *Id.* (internal quotation marks omitted).
- "[A]ll of the *Daubert* factors are relevant to determining the reliability of expert testimony, yet no single factor is dispositive in the analysis. A trial court may apply some, all, or none of the factors depending on the particular expert testimony at issue." *Id.* at 37.
- "*Rochkind* did not upend [the] trial court's gatekeeping function. Vigorous cross-examination, presentation of contrary evidence, and careful instruction on the burden of proof are the traditional and appropriate means of attacking shaky but admissible evidence." *Id.* at 38 (internal quotation marks omitted).

The overarching criterion for the admission of relevant expert testimony under *Rochkind*, and the goal to which each of the ten *Daubert-Rochkind* factors and the five principles summarized in *Matthews* are all addressed, is reliability. The question for a trial court is not whether proposed expert testimony is right or wrong, but whether it meets a minimum threshold of reliability so that it may be presented to a jury, where it may then be questioned, tested, and attacked through means such as cross-examination or the submission of opposing expert testimony.

Because we evaluate a trial court's decision to admit or exclude expert testimony under an abuse of discretion standard, our review is necessarily limited to the information

that was before the trial court at the time it made the decision. A trial court can hardly abuse its discretion in failing to consider evidence that was not before it.[6]

II. FIREARMS IDENTIFICATION EVIDENCE

Through multiple submissions by the parties and two evidentiary hearings over the course of five days, the circuit court ultimately received the testimony of five witnesses (one twice); 18 reports or articles discussing firearms identification, describing studies testing firearms identification, or criticizing the theory or the results of the studies testing it; and a chart identifying dozens of additional or planned studies or reports. In section A of this Part II, we discuss firearms identification evidence generally. In sections B and C, we review criticisms and studies of the methodology, respectively. In section D, we summarize the testimony presented to the circuit court. Finally, in section E, we discuss how some other courts have resolved challenges to the admissibility of firearms identification evidence.

[6] On appeal, the State cited articles presenting the results of studies that were not presented to the circuit court and, in some cases, that were not even in existence at the time the circuit court ruled. *See, e.g.*, Maddisen Neuman et al., *Blind Testing in Firearms: Preliminary Results from a Blind Quality Control Program*, 67 J. Forensic Scis. 964 (2022); Eric F. Law & Keith B. Morris, *Evaluating Firearm Examiner Conclusion Variability Using Cartridge Case Reproductions*, 66:5 J. Forensic Scis. 1704 (2021). We have not considered those studies in reaching our decision. If any of those studies materially alters the analysis applicable to the reliability of the Association of Firearm and Tool Mark Examiners theory of firearms identification, they will need to be presented in another case.

A. Firearms Identification

1. The Theory Underlying Firearms Identification Generally

Firearms identification is a subset of toolmark identification. A toolmark—literally, a mark left by a particular tool—is "generated when a hard object (tool) comes into contact with a relatively softer object," such as the marks that result "when the internal parts of a firearm make contact with the brass and lead that comprise ammunition." *United States v. Willock*, 696 F. Supp. 2d 536, 555 (D. Md. 2010) (quoting Nat'l Rsch. Council, Nat'l Acad. of Scis., *Strengthening Forensic Science in the United States: A Path Forward* 150 (2009)), *aff'd sub nom. United States v. Mouzone*, 687 F.3d 207 (4th Cir. 2012). The marks are then viewable using a "comparison microscope," which a firearms examiner uses "to compare ammunition test-fired from a recovered gun with spent ammunition from a crime scene[.]" *United States v. Monteiro*, 407 F. Supp. 2d 351, 359 (D. Mass. 2006).

As a forensic technique to identify a particular firearm as the source of a particular ammunition component, firearms identification is based on the premise that no two firearms will make identical marks on a bullet or cartridge case. *United States v. Natson*, 469 F. Supp. 2d 1253, 1260 (M.D. Ga. 2007). That, the theory goes, is because the method of manufacturing firearms results in the interior of each firearm being unique and, therefore, making unique imprints on ammunition components fired from it. *Id.*

As the United States District Court for the District of Massachusetts explained:

> When a firearm is manufactured, the "process of cutting, drilling, grinding, hand-filing, and, very occasionally, hand-polishing . . . will leave individual characteristics" on the components of the firearm. *See* Brian J. Heard, *Handbook of Firearms and Ballistics* 127 (1997). Although modern manufacturing methods have reduced the amount of handiwork performed

> on an individual gun, the final step in production of most firearm parts requires some degree of hand-filing which imparts individual characteristics to the firearm part. *See id.* at 128. This process results in "randomly produced patterns of individual stria," or thin grooves or markings, being left on firearm parts. *Id.* These parts are assembled to compose the final firearm.
>
> When a round (a single "shot") of ammunition is fired from a particular firearm, the various components of the ammunition come into contact with the firearm at very high pressures. As a result, the individual markings on the firearm parts are transferred to the ammunition. *Id.* The ammunition is composed primarily of the bullet and the cartridge case. The bullet is the missile-like component of the ammunition that is actually projected from the firearm, through the barrel, toward the target. . . . The cartridge case is the part of the ammunition situated behind the bullet containing the primer and propellant, the explosive mixture of chemicals that causes the bullet to be projected through the barrel. *Id.* at 42.

Monteiro, 407 F. Supp. 2d at 359-60.

The patterns and marks left on bullets and cartridge cases are classified into three categories. First, "class characteristics" are common to all bullets and cartridge cases fired from "weapons of the make and model that fired the ammunition." *Willock*, 696 F. Supp. 2d at 557-58. "Examples of class characteristics include the bullet's weight and caliber; number and width of the lands and grooves in the gun's barrel; and the 'twist' (direction of turn, i.e., clockwise or counterclockwise, of the rifling in the barrel)."[7] *Id.* at 558.

Second, "subclass characteristics" are common to "a group of guns within a certain make or model, such as those manufactured at a particular time and place." *Monteiro*, 407

[7] "Rifling" refers to "a pattern of channels that run the length of a firearm barrel, manufactured with a helical pattern, or twist," which has raised areas called "lands," and lowered areas called "grooves." Ass'n of Firearms & Tool Mark Exam'rs, *What Is Firearm and Tool Mark Identification?*, *available at* https://afte.org/about-us/what-is-afte/what-is-firearm-and-tool-mark-identification (last accessed June 14, 2023), archived at https://perma.cc/UYA4-99CS. "The number and width of lands and grooves is determined by the manufacturer and will be the same for a large group of firearms." *Id.*

F. Supp. 2d at 360. "An example would include imperfections 'on a rifling tool that imparts similar toolmarks on a number of barrels before being modified either through use or refinishing.'" *Willock*, 696 F. Supp. 2d at 558 (quoting Ronald G. Nichols, *Defending the Scientific Foundations of the Firearms and Tool Mark Identification Discipline: Responding to Recent Challenges*, 52 J. Forensic Scis. 586, 587 (2007)).

Third, "individual characteristics" are those unique to an individual firearm that therefore "distinguish [the firearm] from all others." *Willock*, 696 F. Supp. 2d at 558 (quoting *Monteiro*, 407 F. Supp. 2d at 360). Individual characteristics include "[r]andom imperfections produced during manufacture or caused by accidental damage." *Id.* Notably, not all individual characteristics are unique, *Willock*, 696 F. Supp. 2d at 558, and individual characteristics can change over the life of a firearm as a result of, for example, wear, polishing, or damage. As will be discussed further below, one dispute between proponents of firearms identification and its detractors is the degree to which firearms examiners can reliably identify the difference between subclass and individual characteristics when performing casework.

2. The Association of Firearm and Tool Mark Examiners Methodology

The leading methodology used by firearms examiners, and the methodology employed in this case by Mr. McVeigh, is the Association of Firearm and Tool Mark Examiners ("AFTE") "Theory of Identification" (the "AFTE Theory").[8] *See* Committee

[8] According to its website, the AFTE "is the international professional organization for practitioners of Firearm and/or Toolmark Identification and has been dedicated to the exchange of information, methods and best practices, and the furtherance of research since

for the Advancement of the Science of Firearm & Toolmark Identification, *Theory of Identification as it Relates to Toolmarks: Revised*, 43 AFTE J. 287 (2011). Examiners employing the AFTE Theory follow a two-step process. At step one, the examiner evaluates class characteristics of the unknown and known samples. *See* AFTE, *Summary of the Examination Method*, *available at* https://afte.org/resources/swggun-ark/summary-of-the-examination-method (last accessed June 14, 2023), archived at https://perma.cc/4D8W-UDW9. If the class characteristics do not match—i.e., if the samples have different numbers of lands and grooves or a different twist direction—the firearm that produced the known sample is excluded as the source of the unknown sample. *Id.* If the class characteristics match, the second step involves "a comparative examination . . . utilizing a comparison microscope." *Id.* At that step, the examiner engages in "pattern matching" "to determine: 1) if any marks present are subclass characteristics and/or individual characteristics, and 2) the level of correspondence of any individual characteristics."[9] *Id.*

---

its creation in 1969." AFTE, *What is AFTE?*, *available at* https://afte.org/about-us/what-is-afte (last accessed June 14, 2023), archived at https://perma.cc/4VKT-EZW7. According to AFTE's bylaws, individuals are eligible to become members if they are, among other things, "a practicing firearm and/or toolmark examiner," which is defined to mean a person who "derives a substantial portion of their livelihood from the examination, identification, and evaluation of firearms and related materials and/or toolmarks; or an individual whose present livelihood is a direct result of the knowledge and experience gained from the examination, identification, and evaluation of firearms and related materials and/or toolmarks." AFTE, *AFTE Bylaws*, Art. III, § 1, *available at* https://afte.org/about-us/bylaws (last accessed June 14, 2023), archived at https://perma.cc/Y2PF-XWUF.

[9] An alternative to the AFTE method is the "consecutive matching striae method of toolmark analysis" ("CMS"). *Fleming*, 194 Md. App. at 105. "The CMS method . . . calls

Based on that "pattern matching," the examiner makes a determination in accordance with the "AFTE Range of Conclusions," which presents the following options:

> 1. "Identification" occurs when there is "[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."
>
> 2. There are three categories of "Inconclusive," all of which require full agreement of "all discernible class characteristics":
>
> (a) when there is "[s]ome agreement of individual characteristics . . . but insufficient for an identification";
>
> (b) when there is neither "agreement [n]or disagreement of individual characteristics"; and
>
> (c) when there is "disagreement of individual characteristics, but insufficient for an elimination."
>
> 3. "Elimination" occurs when there is "[s]ignificant disagreement of discernible class characteristics and/or individual characteristics."

AFTE, *Range of Conclusions*, *available at* https://afte.org/about-us/what-is-afte/afte-range-of-conclusions (last accessed June 14, 2023), archived at https://perma.cc/WKF5-M6HD.

According to the AFTE, a positive "Identification" can be made when there is "sufficient agreement" between "two or more sets of surface contour patterns" on samples. AFTE, *AFTE Theory of Identification as It Relates to Toolmarks*, *available at*

---

for the examiner to consider the number of consecutive matching striae, or 'scratches' appearing on a projectile fragment. The theory provides that a positive 'match' determination can be made only when a certain, statistically established number of striae match." *Id.* Proponents of the CMS method argue that it has a "greater degree of objective certainty" than other methods. *Id.* The CMS method was not used in this case.

https://afte.org/about-us/what-is-afte/afte-theory-of-identification (last accessed June 14, 2023), archived at https://perma.cc/E397-U8KM. "[S]ufficient agreement," in turn: (1) occurs when the level of agreement "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"; and (2) 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." *Id.*

The AFTE acknowledges that "[c]urrently the interpretation of individualization/identification is subjective in nature[.]" *Id.* The AFTE Theory provides no objective criteria to determine what constitutes the "best agreement demonstrated" between toolmarks produced by different tools or what rises to the level of "quantity and quality" of agreement demonstrating a "practical impossibility" of a different tool having made the same mark. There are also no established standards for classifying a particular pattern or mark as a subclass versus an individual characteristic.

B. Critiques of Firearms Identification

Firearms identification has existed as a field for more than a century.[10] Throughout most of that time, it has been accepted by law enforcement organizations and courts without

[10] The first prominent use of firearms identification in the United States is attributed to examinations made in the aftermath of the 1906 race-related incident in Brownsville, Texas, known as the "Brownsville Affair." There, Army personnel matched 39 out of 45 cartridge cases to two types of rifles "through the use of only magnified photographs of firing pin impressions[.]" Kathryn E. Carso, *Amending the Illinois Postconviction Statute to Include Ballistics Testing*, 56 DePaul L. Rev. 695, 700 n.43 (2007).

significant challenge. However, the advent of *Daubert*, work exposing the unreliability of other previously accepted forensic techniques,[11] and recent reports questioning the foundations underlying firearms identification have led to greater skepticism.

Reports issued since 2008 by two blue-ribbon groups of experts outside of the firearms and toolmark identification field have been critical of the AFTE Theory. In 2008, the National Research Council of the National Academies of Science (the "NRC") published a report concerning the feasibility of developing a national database of ballistic images to aid in criminal investigations. National Research Council, National Academy of Sciences, Committee to Assess the Feasibility, Accuracy, and Technical Capability of a National Ballistics Database, *Ballistic Imaging* 1-2 (2008), *available at* https://nap.nationalacademies.org/read/12162/chapter/1 (last accessed June 14, 2023), archived at https://perma.cc/X6NG-BNVN. In the report, the committee identified challenges that complicate firearms identifications, and ultimately determined that the creation of a national ballistic image database was not advisable at the time. *Id.* at 4-5.

[11] For example, comparative bullet lead analysis was initially widely accepted within the scientific and legal community, and admitted successfully in criminal prosecutions nationwide, yet its validity was subsequently undermined and such evidence is now inadmissible. *See Chesson v. Montgomery Mut. Ins. Co*., 434 Md. 346, 358-59 (2013) (stating that, despite the expert's "use of th[e] technique for thirty years," comparative bullet lead analysis evidence was inadmissible because its "general and underlying assumption . . . was no longer generally accepted by the relevant scientific community"); *Clemons v. State*, 392 Md. 339, 364-72 (2006) (comprehensively discussing comparative bullet lead analysis and holding that it does not satisfy *Frye-Reed*); *Sissoko v. State*, 236 Md. App. 676, 721-27 (2018) (discussing that the "methodology underlying [comparative bullet lead analysis], which was developed in the 1960s and became a widely accepted forensic tool by the 1980s[,] . . . [was] undermined by many in the relevant scientific community" and was "no longer . . . 'valid and reliable'" (quoting *Clemons v. State*, 392 Md. 339, 359 (2006))).

Then, in 2009, the NRC published a report in which it addressed "pressing issues" within several forensic science disciplines, including firearms identification. National Research Council, National Academy of Sciences, *Strengthening Forensic Science in the United States: A Path Forward* 2-5 (2009) (the "2009 NRC Report"), *available at* https://www.ojp.gov/pdffiles1/nij/grants/228091.pdf (last accessed June 14, 2023), archived at https://perma.cc/RLT6-49C3.[12] The NRC observed that advances in DNA evidence had revealed flaws in other forensic science disciplines that "may have contributed to wrongful convictions of innocent people," *id.* at 4, and pointed especially to the relative "dearth of peer-reviewed, published studies establishing the scientific bases and validity of many forensic methods," *id.* at 8.

With respect to firearms identification specifically, the NRC criticized the AFTE Theory as lacking specificity in its protocols; producing results that are not shown to be accurate, repeatable, and reproducible; lacking databases and imaging that could improve the method; having deficiencies in proficiency training; and requiring examiners to offer opinions based on their own experiences without articulated standards. *Id.* at 6, 63-64, 155. In particular, the lack of knowledge "about the variabilities among individual tools and guns" means that there is an inability of examiners "to specify how many points of similarity are necessary for a given level of confidence in the result." *Id.* at 154. Indeed, the NRC noted, the AFTE's guidance, which is the "best . . . available for the field of

[12] The lead NRC "Committee" behind the report was the "Committee on Identifying the Needs of the Forensic Science Community." The committee was co-chaired by Judge Harry T. Edwards of the United States Court of Appeals for the District of Columbia Circuit and included members from a variety of distinguished academic and scientific programs.

toolmark identification, does not even consider, let alone address, questions regarding variability, reliability, repeatability, or the number of correlations needed to achieve a given degree of confidence." *Id.* at 155. The NRC concluded that "[t]he validity of the fundamental assumptions of uniqueness and reproducibility of firearms-related toolmarks has not yet been fully demonstrated." *Id.* at 70, 80-81, 154-55 (citation omitted).

In 2016, the President's Council of Advisors on Science and Technology ("PCAST")[13] issued a report identifying additional concerns about the scientific validity of, among other forensic techniques, firearms identification. *See* Executive Office of the President, 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* (2016) (the "PCAST Report"), *available at* https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_fo

[13] The PCAST Report provides the following description of PCAST's role:

> The President's Council of Advisors on Science and Technology (PCAST) is an advisory group of the Nation's leading scientists and engineers, appointed by the President to augment the science and technology advice available to him from inside the White House and from cabinet departments and other Federal agencies. PCAST is consulted about, and often makes policy recommendations concerning, the full range of issues where understandings from the domains of science, technology, and innovation bear potentially on the policy choices before the President.

PCAST Report at iv. Members of PCAST included scholars and senior executives at institutions and firms including Harvard University; the University of Texas at Austin, Honeywell; Princeton University; the University of Maryland; the University of Michigan; the University of California, Berkeley; United Technologies Corporation; Washington University of St. Louis; Alphabet, Inc.; Northwestern University; and the University of California, San Diego. *Id.* at v-vi. PCAST also consulted with "Senior Advisors" including eight federal appellate and trial court judges, as well as law school and university professors. *Id.* at viii-ix.

rensic_science_report_final.pdf (last accessed June 14, 2023), archived at https://perma.cc/3QWJ-2DGR. With respect to all six forensic disciplines addressed in the report, including firearms identification, PCAST focused on whether there had been a demonstration of both "foundational validity" and "validity as applied." *Id.* at 4-5. Foundational validity, according to PCAST, requires that the method "be shown, based on empirical studies, to be *repeatable*, *reproducible*, and *accurate*, at levels that have been measured and are appropriate to the intended application." *Id.* Validity as applied requires "that the method has been reliably applied *in practice*." *Id.* at 5.

With respect to firearms identification specifically, PCAST described the AFTE Theory as a "circular" method that lacks "foundational validity" because appropriate studies had not confirmed its accuracy, repeatability, and reproducibility. *Id.* at 60, 104-05. PCAST concluded that the studies performed to that date, with one exception, were not properly designed, had severely underestimated the false positive and false negative error rates, or otherwise "differ[ed] in important ways from the problems faced in casework." *Id.* at 106. Among other things, PCAST noted design flaws in existing studies, including: (1) many were not "black-box" studies,[14] *id.* at 49; and (2) many were closed-set studies,

[14] "A black box study assesses the accuracy of examiners' conclusions without considering how the conclusions were reached. The examiner is treated as a 'black-box' and the researcher measures how the output of the 'black-box' (examiner's conclusion) varies depending on the input (the test specimens presented for analysis). To test examiner accuracy, the 'ground truth' regarding the type or source of the test specimens must be known with certainty." Organization of Scientific Area Committees for Forensic Science, *OSAC Draft Guidance on Testing the Performance of Forensic Examiners* (2018), *available at* https://www.nist.gov/document/drafthfcguidancedocument-may8pdf (last accessed June 14, 2023), archived at https://perma.cc/3LH5-KURT.

in which comparisons are dependent upon each other and there is always a "correct" answer within the set, *id.* at 106.

The sole exception to PCAST's negative critique of study designs was a study performed by the United States Department of Energy's Ames Laboratory (the "Ames I Study"), which PCAST called "the first appropriately designed black-box study of firearms [identification]." *Id.* at 11. Nonetheless, PCAST observed that that study, which we discuss below, was not published in a scientific journal, had not been subjected to peer review, and stood alone. *Id.* PCAST therefore concluded that "firearms analysis currently falls short of the criteria for foundational validity" and called for additional testing. *Id.* at 111-14.

C. Recent Studies of the AFTE Theory

Numerous studies of the AFTE Theory have been performed over the course of several decades. The State contends that many of those studies are scientifically valid, reflect extremely low false positive error rates, and therefore support the reliability of the methodology. Mr. Abruquah argues that the studies on which the State relies are flawed and were properly discounted by the NRC and PCAST, that even the best studies present artificially low error rates by treating inconclusive findings as correct, and that the most recent and authoritative study reveals "shockingly" low rates of repeatability and reproducibility.

The State is correct that numerous studies have purported to validate the AFTE Theory, including by identifying relatively low false positive error rates. One of the State's expert witnesses, Dr. James E. Hamby, is the lead author on one such study, in which 697

examiners inspected "over 240 test sets consisting of bullets fired through 10 consecutively rifled RUGER P-85 pistol barrels." James E. Hamby et al., *A Worldwide Study of Bullets Fired from 10 Consecutively Rifled 9MM Ruger Pistol Barrels—Analysis of Examiner Error Rate*, 64:2 J. Forensic Scis. 551, 551 (Mar. 2019) (the "Hamby Study"). In that closed-set study, of 10,455 unknown bullets examined, 10,447 "were correctly identified by participants to the provided 'known' bullets," examiners could not reach a definitive conclusion on eight bullets, and none were misidentified.[15] *Id.* at 556. The error rate, excluding inconclusive results, was thus 0.0%. *See id.*

Examples of other studies on which the State relies, all of which identify relatively low error rates based on the study method employed, include: (1) Jamie A. Smith, *Beretta barrel fired bullet validation study*, 66 J. Forensic Scis. 547 (2021) (comparison testing of 30 consecutively manufactured pistol barrels, producing a 0.55% error rate); and (2) Tasha P. Smith et al., *A Validation Study of Bullet and Cartridge Case Comparisons Using Samples Representative of Actual Casework*, 61 J. Forensic Scis. 939 (2016) (within-set study of 31 examiners matching bullets and cartridge cases, yielding a 0.0% false-positive rate for bullet comparisons and a 0.14% false-positive error rate for cartridge cases).

The NRC and PCAST both are critical of closed-set studies like the Hamby Study and others that provide examiners with multiple "unknown" bullets or cartridge cases and a corresponding number of "known" bullets or cartridge cases that the examiners are asked

[15] Of the eight, the authors point out that three examiners "reported insufficient individual characteristics for two of the test bullets and two trainees could not associate five of the test bullets to their known counterpart bullets." Hamby Study, at 556.

to match. The NRC and PCAST criticize such studies as not being representative of casework because, among other reasons: (1) examiners are aware they are being tested; (2) a correct match exists within the set for every sample, which the examiners also know; and (3) the use of consecutively manufactured firearms (or barrels) in a closed-set study has the effect of eliminating any confusion concerning whether particular patterns or marks constitute subclass or individual characteristics. PCAST Report, at 32-33, 52-59, 107-09; 2009 NRC Report, at 154-55.

The Ames I Study, which PCAST had identified as the only one that had been "appropriately designed" to that point, PCAST Report, at 111, was a 2014 open-set, black-box study designed to measure error rates in the comparison of "known" and "unknown" cartridge cases (the Ames I Study did not involve bullets). *See* David P. Baldwin et al., *A Study of False-Positive and False-Negative Error Rate in Cartridge Case Comparisons*, Defense Biometrics & Forensics Office, U.S. Dep't of Energy (Apr. 2014). In the Ames I Study, 15 sets of four cartridge cases fired from 25 new, same-model handguns using the same type of ammunition were sent to 218 examiners. Ames I Study, at 3. Each set included one unknown sample and three known samples fired from the same known gun, which might or might not have been the source of the unknown sample. *Id.* at 4. Even though there was a known correct answer of either an identification or an elimination for every set, examiners were permitted to make "inconclusive" responses, which were "not counted as an error or as a non-answer[.]" *Id.* at 6. Of the 1,090 comparisons where the "known" and "unknown" cartridge cases were fired from the same source firearm, the examiners incorrectly excluded only four cartridge cases, yielding a false-negative rate of

0.367%. *Id.* at 15. Of the 2,180 comparisons where the "known" and "unknown" cartridge cases were fired from different firearms, the examiners incorrectly matched 22 cartridge cases, yielding a false-positive rate of 1.01%.[16] *Id.* at 16. However, of the non-matching comparison sets, 735, or 33.7%, were classified as inconclusive, *id.*, a significantly higher percentage than in any closed-set study.

The Ames Laboratory later conducted a second open-set, black-box study that was completed in 2020, in between the *Frye-Reed* and *Daubert-Rochkind* hearings in this case. *See* Stanley J. Bajic et al., *Report: Validation Study of the Accuracy, Repeatability, and Reproducibility of Firearm Comparisons*, U.S. Dep't of Energy 1-2 (2020) (the "Ames II Study"). The Ames II Study, which was undertaken in direct response to PCAST's call for further studies to demonstrate the foundational validity of firearms identification, *id.* at 12, enrolled 173 examiners for a three-phase study to test for all three elements PCAST had identified as necessary to support foundational validity: accuracy (in Phase I), repeatability (in Phase II), and reproducibility (in Phase III). In each of three phases, each participating examiner received 15 comparison sets of known and unknown cartridge cases and 15 comparison sets of known and unknown bullets. *Id.* at 23. The firearms used for the bullet comparisons were either Beretta or Ruger handguns and the firearms used for the cartridge case comparisons were either Beretta or Jimenez handguns. *Id.* Only the researchers knew the "ground truth" for each packet; that is, which "unknown" cartridges and bullets matched or did not match the included "known" cartridges and bullets. *Id.* As with the

[16] The authors stressed that a significant majority of the false positive responses—20 out of 22—came from just five of the 165 examiners. Ames I Study, at 16.

Ames I Study, although there was a "ground truth" correct answer for each sample set, examiners were permitted to pick from among the full array of the AFTE Range of Conclusions—identification, elimination, or one of the three levels of "inconclusive." *Id.* at 12-13.

The first phase of testing was designed to assess accuracy of identification, "defined as the ability of an examiner to correctly identify a known match or eliminate a known nonmatch." *Id.* at 33. In the second phase, each examiner was given the same test set examined in phase one, without being told it was the same, to test repeatability, "defined as the ability of an examiner, when confronted with the exact same comparison once again, to reach the same conclusion as when first examined." *Id.* In the third phase, each examiner was given a test set that had previously been examined by one of the other examiners, to test reproducibility, "defined as the ability of a second examiner to evaluate a comparison set previously viewed by a different examiner and reach the same conclusion." *Id.*

In the first phase, the results, shown in percentages, were:

Bullet Evaluations						
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Total Sets
Matching	76.6%	9.04%	8.90%	2.56%	**2.92%**	1405
Nonmatching	**0.70%**	9.43%	29.8%	26.2%	33.8%	2842
Cartridge case Evaluations						
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Total Sets
Matching	74.4%	12.5%	9.86%	1.55%	**1.76%**	1420
Nonmatching	**0.92%**	6.24%	22.5%	21.9%	48.5%	2835

Id. at 35. Treating inconclusive results as appropriate answers, the authors identified a false negative rate for bullets and cartridge cases of 2.92% and 1.76%, respectively, and a false positive rate for each of 0.7% and 0.92%, respectively. *Id.* Examiners selected one of the three categories of inconclusive for 20.5% of matching bullet sets and 65.3% of non-matching bullet sets. *Id.* As reflected in the following table, the results overall varied based on the type of handgun that produced the bullet/cartridge, with examiners' results reflecting much greater certainty and correctness in classifying bullets and cartridge cases fired from the Beretta handguns than from the Ruger (for bullets) and Jimenez (for cartridge cases) handguns:[17]

Bullet Set Evaluations						
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Total Sets
Matching						
Beretta	89.7%	4.13%	2.59%	1.30%	**2.24%**	848
Ruger	56.6%	16.5%	18.5%	4.49%	**3.95%**	557
Nonmatching						
Beretta	**0.54%**	9.59%	22.9%	28.2%	38.7%	2022
Ruger	**1.10%**	9.02%	47.0%	21.2%	21.7%	820
Cartridge case Set Evaluations						
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Total Sets
Matching						
Beretta	80.7%	9.57%	8.40%	0.350%	**0.93%**	857
Jimenez	64.7%	16.9%	12.1%	3.37%	**3.02%**	563
Nonmatching						
Beretta	**0.86%**	6.65%	24.0%	22.7%	45.7%	1971
Jimenez	**1.04%**	5.32%	18.9%	19.9%	54.9%	864

[17] "Of the 27 Beretta handguns used in the study, 23 were from a single recent manufacturing run, and four were guns produced in separate earlier manufacturing runs." Ames II Study, at 56. The Ames II Study does not identify similar information for the Ruger or Jimenez handguns.

Id. at 53.

Comparing the results from the second phase of testing against the results from the first phase, intended to test repeatability, the outcomes, shown in percentages, were:

Paired Bullet Classifications		
	Proportion of paired agreements	Proportion of paired disagreements
Matching Sets	79.0%	21.0%
Nonmatching Sets	64.7%	35.3%
Paired Cartridge case Classifications		
	Proportion of paired agreements	Proportion of paired disagreements
Matching Sets	75.6%	24.4%
Nonmatching Sets	62.2%	37.8%

Id. at 39. Thus, an examiner classifying the same matching bullet or cartridge case set a second time classified it in the same AFTE category 79% and 75.6% of the time, respectively, and an examiner classifying the same non-matching bullet or cartridge case set a second time did so 64.7% and 62.2% of the time, respectively. *Id.* The authors viewed these percentages favorably, concluding that this level of "observed agreement" exceeded the level of their "expected agreement."[18] *Id.* at 39-41. They did so, however, based on an expected level of agreement reflecting the overall pattern of results from the first phase of

[18] The study authors also produced alternate calculations in which they merged either (1) all inconclusive results together or (2) positive identifications with "Inconclusive A" results and eliminations with "Inconclusive B" results. Ames II Study, at 40. As expected, those results produced greater agreement, although still ranging only from 71.3% agreement to 85.5% agreement. *Id.* at 42.

testing. *Id.* at 39-40. In other words, the metric against which the authors gauged repeatability was, in essence, random chance.

Comparing the results from the third phase of testing against the results of the first phase, intended to test reproducibility, the outcomes, shown in percentages, were:

Bullet Classifications		
	Proportion of paired agreements	Proportion of paired disagreements
Matching Sets	67.8%	32.2%
Nonmatching Sets	30.9%	69.1%
Cartridge case Classifications		
	Proportion of paired agreements	Proportion of paired disagreements
Matching Sets	63.6%	36.4%
Nonmatching Sets	40.3%	59.7%

Id. at 47. Thus, an examiner classifying a matching bullet or cartridge case set previously classified by a different examiner classified it in the same AFTE category 67.8% and 63.6% of the time, respectively, and an examiner classifying a nonmatching bullet or cartridge case set previously classified by a different examiner classified it in the same AFTE category 30.9% and 40.3% of the time, respectively. *Id.* The authors again viewed these percentages largely favorably. *Id.* at 47-49. Again, however, that conclusion was based on a level of expected agreement that was essentially random based on the overall results from the first phase of testing. *Id.* at 48-49.

The State claims support from the Ames I and Ames II Studies based on what it calls their relatively low overall false positive rates. The State contends that those results confirm the low false positive rates produced in every other study of firearms identification,

which are worthy of consideration even if they were not as robust in design as the Ames studies. By contrast, Mr. Abruquah claims that the high rates of inconclusive responses in both studies and the low rates of repeatability and reproducibility in the Ames II Study further support the concerns raised by NRC and PCAST about the lack of demonstrated foundational validity of firearms identification.

D. Witness Testimony

1. The *Frye-Reed* Hearing

Five witnesses testified at the two hearings conducted by the circuit court. In the *Frye-Reed* hearing, Mr. Abruquah called William Tobin, a 27-year veteran of the Federal Bureau of Investigation with 24 years' experience at the FBI Laboratory and an expert in forensic metallurgy. Mr. Tobin's testimony was broadly critical of firearms identification generally and the AFTE Theory specifically. Citing support from multiple sources, he opined that: (1) firearms identification is "not a science," does not follow the scientific method, and is circular; (2) the AFTE Theory is wholly subjective and lacks any guidance for examiners to determine the number of similarities needed to achieve an identification; (3) in the absence of standards, examiners ignore or "rationalize away" dissimilarities in samples; (4) examiners are incapable of distinguishing between subclass characteristics and individual characteristics—a phenomenon referred to as "subclass carryover"—thus undermining a fundamental premise of the AFTE Theory; (5) the studies on which the State had relied are flawed, do not reflect actual casework, and underestimate error rates; and (6) the AFTE Theory had not been subject to any "valid hypothesis testing" because the studies cited as support for it "lack any indicia of scientific reliability." Mr. Tobin opined

that, in the absence of a pool of samples from all other possible firearms that might have fired the bullets at issue, the most a firearms examiner could accurately testify to in reliance on the AFTE Theory is whether it was possible that the recovered bullets were fired from Mr. Abruquah's revolver.

The State presented three witnesses. It first presented Dr. James Hamby, an AFTE firearms examiner with a Ph.D. in forensic science who had been Chief of the Firearms Division for the United States Army Lab, authored dozens of articles and studies in the firearms examination field, trained firearms examiners domestically and internationally, and who, over the course of nearly 50 years in the field, managed his own forensic laboratory and two others. Dr. Hamby testified generally about the AFTE Theory, which he asserted had been accepted by the relevant scientific community and by courts, and proven by numerous studies, for more than a century. Dr. Hamby agreed with PCAST that to have foundational validity, a methodology dependent on subjective analysis must be subjected to empirical testing by multiple groups, be repeatable and reproducible, and provide valid estimates of the method's accuracy. He opined that studies of firearms identification proved that the AFTE Theory meets all those criteria and has consistently low error rates. Dr. Hamby acknowledged that false positives can result when similarities in subclass characteristics are mistaken for individual characteristics, but testified that trained examiners would not make that mistake.

Dr. Hamby also discussed the controls and standards governing the work of firearms identification examiners, including internal laboratory procedures, the AFTE training manual, and periodic proficiency training required of every examiner. He testified that one

way forensic labs guard against the possibility of false positive results is by having a second examiner review all matches to ensure the correctness of the first examiner's decision. In his decades of experience, Dr. Hamby was not personally aware of a second examiner ever having reached a different conclusion than the first in actual casework, which he seemed to view as a positive reflection on the reliability of the methodology.

The State's second witness was Torin Suber, a forensic scientist manager with the Maryland State Police. Like Dr. Hamby, Mr. Suber testified about the low false-positive error rates identified in the Ames I and other studies. Mr. Suber agreed that some examiners could potentially mistake subclass characteristics for individual characteristics, but testified that such errors would be limited to novice examiners who "don't actually have that eye or knack for identification yet."

The final witness presented at the *Frye-Reed* hearing was the State's testifying expert, Mr. McVeigh, whom the court accepted as an expert in firearms and toolmark examinations generally, as well as "the specifics of the examination conducted in this matter." Mr. McVeigh testified that 100% of his work is in firearms examinations and that firearms identification is generally accepted as reliable in the relevant scientific community. Mr. McVeigh acknowledged the subjective standards and procedures used in the AFTE methodology but claimed that it is "a forensic discipline with a fairly strict methodology and a lot of rules and accreditation standards to follow." He also relied heavily on what he described as low error rates revealed by the Ames I Study and a separate

study out of Miami-Dade County.[19] Although acknowledging the concern that examiners might mistake subclass characteristics for individual characteristics, Mr. McVeigh testified that possibility is "the number one thing[] that firearm examiners guard against." He said that the "current thinking in the field" is that a trained examiner can overcome that concern.

With respect to the examination he conducted in Mr. Abruquah's case, Mr. McVeigh testified that he received for analysis two firearms, a Glock pistol and a Taurus revolver, along with "six fired bullet items," one of which was unsuitable for comparison. Based on class characteristics, he first eliminated the Glock pistol. He then fired two rounds from the Taurus revolver and compared markings on those bullets against the crime scene bullets using the comparative microscope. In doing so, he focused on the "land impressions," rather than the "groove impressions[, which] are the most likely place where the subclass [characteristics] would occur[.]" Mr. McVeigh opined, without qualification, that, based on his analysis, "at some point each one of those five projectiles had been fired from the Taurus revolver." He testified that his conclusion had been confirmed by another examiner in his lab.

[19] Mr. McVeigh referred to the Miami-Dade Study as an open-set study. Although neither party introduced a report of the Miami-Dade Study, PCAST described it as a "partly open" study. PCAST Report, at 109. According to PCAST, examiners were provided 15 questioned samples, 13 of which matched samples that were provided and two of which did not. *Id.* Of the 330 non-matching samples that were provided, the examiners eliminated 188 of them, reached an inconclusive determination for 138 more, and made four false classifications. *Id.* The inconclusive rate for the non-matching samples was thus 41.8% with a false positive rate of 2.1%. *Id.* PCAST observed that even in that "partly open" study, the inconclusive rate was "200-fold higher" and the false positive rate was "100-fold higher" than in closed set studies. *Id.*

On cross-examination, Mr. McVeigh admitted that he did not know how Taurus manufactured its .38 Special revolver, how many such revolvers had been consecutively manufactured and shipped to the Prince George's County area, or how many in the area might show similar subclass characteristics. He also admitted that the proficiency testing he had undergone during his career is not blind testing and is "straight forward." Indeed, to his knowledge, no one in his lab had ever failed a proficiency test. Mr. McVeigh asserted that bias is not a concern in firearms examinations because the examiners are not provided any details from the police investigation before conducting an examination.

2. The *Daubert-Rochkind* Hearing

At the *Daubert-Rochkind* hearing, each party presented only one witness to supplement the record that had been created at the *Frye-Reed* hearing. The State began with Dr. Hamby. In addition to reviewing many of the same points from his original testimony, Dr. Hamby testified that the AFTE Theory had been tested since 1907 and peer reviewed hundreds of times. He highlighted the low error rates produced in studies, including those in which examiners matched bullets fired from consecutively manufactured barrels. He was also asked about the more recent Ames II Study, but seemed to have limited familiarity with it.

Mr. Abruquah presented testimony and an extensive affidavit from David Faigman, Dean of the University of California Hastings College of Law, whom the court accepted as an expert in statistical and methodological bases for scientific evidence, including research design, scientific research, and methodology. Dean Faigman discussed several concerns with the validity of the AFTE Theory, which were principally premised on the subjective

nature of the methodology, including: (1) the difference in error rates between closed- and open-set tests; (2) potential biases in testing that might skew the results in studies, including (a) the "Hawthorne effect," which theorizes that participants in a test who know they are being observed will try harder; and (b) a bias toward selecting "inconclusive" responses in testing when examiners know it will not be counted against them, but that an incorrect "ground truth" response will; (3) an absence of pre-testing and control groups; (4) the "prior probability problem," in which examiners expect a certain result and so are more likely to find it; and (5) the lack of repeatability and reproducibility effects.

Dean Faigman agreed with PCAST that the Ames I Study "generally . . . was the right approach to studying the subject." He observed, however, that if inconclusives were counted as errors, the error rate from that study would "balloon[]" to over 30%. In discussing the Ames II Study, he similarly opined that inconclusive responses should be counted as errors. By not doing so, he contended, the researchers had artificially reduced their error rates and allowed test participants to boost their scores. By his calculation, when accounting for inconclusive answers, the overall error rate of the Ames II Study was 53% for bullet comparisons and 44% for cartridge case comparisons—essentially the same as "flipping a coin." Regarding the other two phases of the Ames II Study, Dean Faigman found the rates of repeatability and reproducibility "shockingly low."

E. The Evolving Caselaw

Until the 2008 NRC Report, most courts seem to have accepted expert testimony on firearms identification without incident. *See* David H. Kaye, *Firearm-Mark Evidence: Looking Back and Looking Ahead*, 68 Case Western Reserve L. Rev. 723, 723-26 (2018);

see also, e.g., *United States v. Davis*, 103 F.3d 660, 672 (8th Cir. 1996); *United States v. Natson*, 469 F. Supp. 2d 1253, 1261 (M.D. Ga. 2007) (permitting an expert to testify "to a 100% degree of certainty"); *United States v. Foster*, 300 F. Supp. 2d 375, 376 n.1, 377 (D. Md. 2004) (stating that "numerous cases have confirmed the reliability" of firearms and toolmark identification); *United States v. Santiago*, 199 F. Supp. 2d 101, 111 (S.D.N.Y. 2002); *State v. Mack*, 653 N.E.2d 329, 337 (Ohio 1995); *Commonwealth v. Moore*, 340 A.2d 447, 451 (Pa. 1975).

However, "[a]fter the NRC Report issued, some jurisdictions began to limit the scope of a ballistics expert's testimony." *Gardner v. United States*, 140 A.3d 1172, 1183 (D.C. 2016); *see also Commonwealth v. Pytou Heang*, 942 N.E.2d 927, 938 (Mass. 2011) ("Concerns about both the lack of a firm scientific basis for evaluating the reliability of forensic ballistics evidence and the subjective nature of forensic ballistics comparisons have prompted many courts to reexamine the admissibility of such evidence."). Initially, those limitations consisted primarily of precluding experts from testifying that their opinions were offered with something approaching absolute certainty. In *United States v. Willock*, for example, Judge William D. Quarles, Jr. of the United States District Court for the District of Maryland, in adopting a report and recommendation by then-Chief Magistrate Judge, later Judge, Paul W. Grimm of that court, permitted an examiner to testify as to a "match" between a crime scene cartridge case and a particular firearm, but "without any characterization as to degree of certainty." 696 F. Supp. 2d at 572, 574; *see also United States v. Ashburn*, 88 F. Supp. 3d 239, 250 (E.D.N.Y. 2015) (limiting an expert's conclusions to those within a "reasonable degree of certainty in the ballistics field"

or a "reasonable degree of ballistics certainty"); *Monteiro*, 407 F. Supp. 2d at 372 (stating that the proper standard is a "reasonable degree of ballistic certainty"); *United States v. Taylor*, 663 F. Supp. 2d 1170, 1180 (D.N.M. 2009) ("[The expert] will be permitted to give . . . his expert opinion that there is a match . . . . [He] will not be permitted to testify that his methodology allows him to reach this conclusion as a matter of scientific certainty."); *United States v. Glynn*, 578 F. Supp. 2d 567, 574-75 (S.D.N.Y. 2008) (allowing expert testimony that it was "more likely than not" that certain bullets or casings came from the same gun, "but nothing more").

Following issuance of the PCAST Report, some courts have imposed yet more stringent limitations on testimony. One example of that evolution—notable because it involved the same judicial officer as *Willock*, Judge Grimm, as well as the same examiner as here, Mr. McVeigh—is in *United States v. Medley*, No. PWG-17-242 (D. Md. Apr. 24, 2018), ECF No. 111. In *Medley*, Judge Grimm thoroughly reviewed the state of knowledge at that time concerning firearms identification, including developments since his report and recommendation in *Willock*. Judge Grimm restricted Mr. McVeigh to testifying only "that the marks that were produced by the . . . cartridges are consistent with the marks that were found on the" recovered firearm, and precluded him from offering any opinion that the cartridges "were fired by the same gun" or expressing "any confidence level" in his opinion. *Id.* at 119.

Some other courts, although still a minority overall, have recently imposed similar or even more restrictive limitations. *See United States v. Shipp*, 422 F. Supp. 3d 762, 783 (E.D.N.Y. 2019) (limiting expert's testimony to opining that "the recovered firearm cannot

be excluded as the source of the recovered bullet fragment and shell casing"); *Williams v. United States*, 210 A.3d 734, 744 (D.C. 2019) ("[I]t is plainly error to allow a firearms and toolmark examiner to unqualifiedly opine, based on pattern matching, that a specific bullet was fired by a specific gun."); *United States v. Adams*, 444 F. Supp. 3d 1248, 1256, 1261, 1267 (D. Or. 2020) (precluding expert from offering testimony of a match but permitting testimony about "limited observational evidence").[20]

III. ANALYSIS

In granting in part Mr. Abruquah's motion in limine to exclude firearms identification evidence, the circuit court ruled that Mr. McVeigh could not testify "to any level of practical certainty/impossibility, ballistic certainty, or scientific certainty that a suspect weapon matches certain bullet or casing striations." However, the court ruled that Mr. McVeigh could opine the bullets and fragment "recovered from the murder scene fall into any of the AFTE Range of Conclusions[,]" i.e., identification, any of the three levels of inconclusive, or elimination. Accordingly, at trial, after explaining how he analyzed the samples and compared their features, Mr. McVeigh testified, over objection and separately with respect to each of the four bullets and the bullet fragment, that each "at some point" "had been fired" from or through "the Taurus revolver." He testified neither that his

[20] In *United States v. Davis*, citing Judge Grimm's reasoning in *Medley* with approval, a federal district court judge in West Virginia also precluded Mr. McVeigh and other examiners from testifying that marks on a cartridge case indicated a "match" with a particular firearm, while permitting the examiners to testify that marks on the cartridges were "similar and consistent with each other." 2019 WL 4306971, at *7, Case No. 4:18-cr-00011 (W.D. Va. 2019).

opinion was offered to any particular level of certainty nor that it was subject to any qualifications or caveats.

In his appeal, Mr. Abruquah does not challenge all of Mr. McVeigh's testimony or that firearms identification is sufficiently reliable to be admitted for some purposes. Instead, he contends that the methodology is insufficiently reliable to support testimony "identify[ing] a specific firearm as the source of a questioned bullet," and argues that an examiner should be limited to opining, "at most, that a firearm cannot be excluded as the source of the questioned projectile[.]" In response, the State argues that firearms identification evidence has been accepted by courts applying the *Daubert* standard as reliable, has repeatedly been proven reliable in studies demonstrating very low false-positive rates, and that, "[a]t best, [Mr. Abruquah] has demonstrated that there are ongoing debates regarding *how* to assess the AFTE methodology[,]" not whether it is admissible.

In light of the scope of Mr. Abruquah's challenge, our task is to assess, based on the information presented to the circuit court, whether the AFTE Theory can reliably support an unqualified opinion that a particular firearm is the source of one or more particular bullets. Our analysis of the *Daubert-Rochkind* factors is thus tailored specifically to that issue, not to the reliability of the methodology more generally.

Before turning to the specific *Daubert-Rochkind* factors, we offer two preliminary observations. First, our analysis is not dependent on whether firearms identification is a "science." "*Daubert*'s general holding," adopted by this Court in *Rochkind*, "applies not only to testimony based on 'scientific' knowledge, but also to testimony based on 'technical' and 'other specialized' knowledge." *Rochkind*, 471 Md. at 36 (quoting *Kumho*

Tire Co., 526 U.S. at 141). Second, it is also not dispositive that firearms identification is a subjective endeavor. *See, e.g.*, *United States v. Romero-Lobato*, 379 F. Supp. 3d 1111, 1120 (D. Nev. 2019) ("The mere fact that an expert's opinion is derived from subjective methodology does not render it unreliable."); *Ashburn*, 88 F. Supp. 3d at 246-47 (stating that "the subjectivity of a methodology is not fatal under [Federal] Rule 702 and *Daubert*"). The absence of objective criteria is a factor that we consider in our analysis of reliability, but it is not dispositive.

We now turn to consider each of the ten *Daubert-Rochkind* factors. Of course, those factors "are neither exhaustive nor mandatory," *Matthews*, 479 Md. at 314, but they provide a helpful framework for our analysis in this case.

A. Testability

Although significant dispute surrounds many of the studies conducted on firearms identification to date, and especially their applicability to actual casework, it is undisputed that firearms identification can be tested. Indeed, the bottom-line recommendation of the most significant critics of firearms identification to date, the authors of the 2009 NRC and PCAST Reports, was to call for more and better testing, not to question whether such testing is possible.

B. Peer Review and Publication

The second *Daubert-Rochkind* factor considers whether a methodology has been submitted "to the scrutiny of the scientific community," under the belief that doing so "increases the likelihood that substantive flaws in methodology will be detected." *Daubert*, 509 U.S. at 593. The circuit court concluded that the State satisfied its burden to show that

the firearms and toolmark identification methodology has been peer reviewed and published. We think the evidence is more mixed.

The two most robust studies of firearms identification—Ames I and II—have not been peer reviewed or published in a journal. The record does not disclose why. Some of the articles on which the State and its witnesses rely have been published in the AFTE Journal, a publication of the primary trade group dedicated to advancing firearms identification. The required steps in the AFTE Journal's peer review process involve a review by "a member of [AFTE's] Editorial Review Panel" for "grammatical and technical correctness" and review by an AFTE "Assistant Editor[]" for "grammar and technical content." *See* AFTE, *Peer Review Process*, available at https://afte.org/afte-journal/afte-journal-peer-review-process (last accessed June 14, 2023), archived at https://perma.cc/822Y-C7G8. That process appears designed primarily to review articles and studies to determine their adherence to the AFTE Theory, not to test the methodology.

Although a handful of other firearms identification studies have been published in other forensic journals, the record is devoid of any information about the extent or quality of peer review as concerns the validity of the methodology. Nonetheless, NRC's and PCAST's critiques of some of those same studies, and of the AFTE Theory more generally, have served many of the same purposes that might have been served by a robust peer review process. *See Shipp*, 422 F. Supp. 3d at 777 (concluding that the AFTE Theory had been adequately subjected to peer review and publication due in large part to "the scrutiny of PCAST and the flaws it perceived in the AFTE Theory").

C. Known or Potential Rate of Error

The circuit court found that the parties did not dispute "that a known or potential rate of error has been attributed to firearms identification evidence," and treated that as favoring admission of Mr. McVeigh's testimony. (Emphasis removed). Neither party disputes that there is a potential rate of error for firearms identification or that a number of studies have purported to identify such an error rate. However, they do dispute whether the studies to date have identified a reliable error rate. On that issue, we glean several relevant points from the record.

First, the reported rates of "ground truth" errors—i.e., "identification" of a non-matching sample or "elimination" of a matching sample—from studies in the record are relatively low.[21] Error rates in most closed-set studies hover close to zero and the overall error rates calculated in the Ames I and II Studies were in the low single digits.[22] It thus

[21] Most of the parties' attention in this case is naturally focused on the "false positive" rate. Although false positives create the greatest risk of leading directly to an erroneous guilty verdict, an examiner's erroneous failure to eliminate the possibility of a match could also contribute to an erroneous guilty verdict if the correct answer—elimination—would have led to an acquittal. To that extent, it is notable that in the first round of testing in the Ames II Study, examiners correctly eliminated only 33.8% of non-matching bullets and 48.5% of non-matching cartridge cases. *See* Ames II Study, at 35.

[22] The Ames I Study identified a false negative rate of 0.367%, with a 95% confidence interval of up to 0.94%, a false-negative-plus-inconclusive rate of 1.376%, with a 95% confidence interval of up to 2.26%, and a false positive rate of 0.939%, with a 95% confidence interval of up to 2.26%. Ames I Study, at 17. The Ames II Study reports its results for bullets as having a false positive error probability of 0.656%, with a 95% confidence interval of up to 1.42%, and a false negative error probability of 2.87%, with a 95% confidence interval of up to 4.26%. The Ames II Study results for cartridge cases showed a false positive error probability of 0.933%, with a 95% confidence interval of up to 1.57% and a false negative error probability of 1.87%, with a 95% confidence interval of up to 2.99%. Ames II Study, at 77.

appears that, at least in studies conducted thus far, it is relatively rare for an examiner in a study environment to identify a match between a firearm and a non-matching bullet.

Second, the low error rates from closed-set, matching studies utilizing bullets or cartridges fired from consecutively manufactured firearms or barrels, offer strong support for the propositions that: (1) firearms produce some unique collections of individual patterns and markings on bullets and cartridges they fire; and (2) such collections of individual patterns and markings can be reliably identified when subclass characteristics are removed from the equation.[23]

Third, the rate of "inconclusive" responses in closed-set studies is negligible to non-existent, *see, e.g.*, Hamby Study, at 555-56 (finding that examiners classified eight out of 10,445 responses as inconclusive); but the rate of such responses in open-set studies is significant, *see, e.g.*, Ames I Study, at 16 (finding that examiners classified 33.7% of "true different-source comparisons" as inconclusive); Ames II Study, at 35 (finding that examiners classified more than 20% of matching bullet sets and more than 65% of non-matching bullet sets as inconclusive), suggesting that examiners choose "inconclusive" even when it is not a "correct" response. The State, its witnesses, and the studies on which they rely suggest that responses of "inconclusive" are properly treated as appropriate

[23] The use of bullets and cartridges from consecutively manufactured firearms or barrels, although more difficult in the sense that the markings in total can be expected to be more similar than those fired from non-consecutively manufactured firearms or barrels, also makes it easier to eliminate any confusion concerning whether marks or patterns are subclass or individual characteristics. *See* Tasha P. Smith et al., *A Validation Study of Bullet and Cartridge Case Comparisons Using Samples Representative of Actual Casework*, 61 J. Forensic Scis. 939 (2016) (noting that toolmarks on consecutively manufactured firearms may be identified "when subclass influence is excused").

responses because, as stated in the Ames I Study, if "the examiner is unable to locate sufficient corresponding individual characteristics to either include or exclude an exhibit as having been fired in a particular firearm," then "inconclusive" is the only appropriate response. Ames I Study, at 6. That answer would be more convincing if rates of inconclusive findings were consistent as between closed-set and open-set studies or if the Ames II Study had produced higher levels of consistency in the repeatability or reproducibility portions of the study. Instead, whether an examiner chooses "inconclusive" in a study seems to depend on something other than just the "corresponding individual characteristics" themselves.

Fourth, if at least some inconclusives should be treated as incorrect responses, then the rates of error in open-set studies performed to date are unreliable. Notably, if just the "Inconclusive-A" responses—those for which the examiner thought there was almost enough agreement to identify a match—for non-matching bullets in the Ames II Study were counted as incorrect matches, the "false positive" rate would balloon from 0.7% to 10.13%. That is particularly noteworthy because in all the studies conducted to date, the participating examiners knew that (1) they were being studied and (2) an inconclusive response would not be counted as incorrect. There is no evidence in the record that examiners in a casework environment—when processing presumably less pristine samples than those included in studies and that were provided to them by law enforcement officers in the context of an investigation—select inconclusive at the same rate they do in an open-set testing environment.

Fifth, it is notable that the accuracy rate in the Ames II Study varied significantly between the two different types of firearms tested. Examiners correctly classified 89.7% of matching bullet sets fired from Beretta handguns but only 56.6% of those fired from Ruger handguns. Ames II Study, at 53. They also correctly eliminated 38.7% of non-matching bullet sets fired from Beretta handguns and only 21.7% of those fired from Ruger handguns. *Id.* Given that variability, it is significant that the record provides scant information about where Taurus revolvers might fall on the error rate spectrum.[24]

Finally, we observe that even if the studies reflecting potential error rates of up to 2.6% reflected error rates in actual casework—a proposition for which this record provides no support—that rate must be assessed in the context of the evidence at issue. Not all expert witness testimony is created the same. Unlike testimony that results in a determination that the perpetrator of a crime was of a certain height range, *see Matthews*, 479 Md. at 285, a conclusion that a bullet found in a victim's body was fired from the defendant's gun is likely to lead much more directly to a conviction. That effect is compounded by the fact that a defendant is almost certain to lack access to the best evidence that could potentially contradict (or, of course, confirm) such testimony, which would be bullets fired from other firearms from the same production run.

[24] During the *Frye-Reed* hearing, Dr. Hamby testified, using Glock as an example, that high-quality firearms would produce bullets and cartridge cases with very consistent patterns and markings, even across 10,000 cartridges, because the process of firing has little effect on the firearm. He also testified that, by contrast, an examiner might not be able to tell the difference between cartridge cases from rounds fired even consecutively from a low-quality firearm, because each bullet "just eats up the barrel." Asked where a Taurus .38 revolver falls on the spectrum between a "cheap gun versus the most expensive," Dr. Hamby offered that "it's mid-level."

The relatively low rate of "false positive" responses in studies conducted to date is by far the most persuasive piece of evidence in favor of admissibility of firearms identification evidence. On balance, however, the record does not demonstrate that that rate is reliable, especially when it comes to actual casework.

D. Existence and Maintenance of Standards and Controls

The circuit court found the evidence with respect to the existence and maintenance of standards and controls to be "muddled" and so to weigh against admission. We mostly agree. On the one hand, to the extent that this factor encompasses operating procedures designed to ensure a consistency in process, *see, e.g.*, *Adams*, 444 F. Supp. 3d at 1266 (discussing annual proficiency testing, second reviewer verification, technical review, and training as relevant to the analysis of standards and quality control), the State presented evidence of such standards and controls. That evidence includes the AFTE training manual, laboratory standard operating procedures, and laboratory accreditation standards. Together, those sources provide standards and controls applicable to: (1) the training and certification of firearms examiners; (2) proficiency testing of firearms examiners; and (3) the mechanics of how examiners treat evidence and conduct examinations. *Accord Willock*, 696 F. Supp. 2d at 571-72 (finding the existence of "standards governing the methodology of firearms-related toolmark examination").

Notably, however, the record also contains evidence that severely undermines the value of some of those same standards and controls. For example, one control touted by advocates of firearms identification is a requirement that a second reviewer confirm every identification classification. *See Taylor*, 663 F. Supp. 2d at 1176 (noting an expert's

testimony that "industry standards require confirmation by at least one other examiner when the first examiner reaches an identification"). Indeed, Dr. Hamby testified that he believes error rates identified in firearms identification studies are overstated because those studies do not permit confirmatory review by a second examiner. However, Dr. Hamby also testified that the confirmatory review process is not blind, meaning that the second reviewer knows the conclusion reached by the first. Even more significantly, Dr. Hamby testified that in his decades of experience in firearms identification in multiple laboratories in multiple states, he was not aware of a single occasion in which a second reviewer had reached a different conclusion than the first. In light of the findings in the reproducibility phase of the Ames II Study concerning how frequently examiners in the study environment come to different conclusions, Dr. Hamby's testimony strongly suggests that study results do not, in fact, reliably represent what occurs in actual casework.

As a second example, although advocates of firearms identification tout periodic proficiency testing by Collaborative Testing Services Inc. ("CTS") as a method of ensuring the quality of firearms identification, the record contains no evidence supporting efficacy of that testing. To the contrary, the evidence suggests that examiners rarely, if ever, fail CTS proficiency tests. Dr. Hamby confirmed that the industry's mandate to CTS with respect to proficiency tests "was to try to make them [as] inexpensive as possible."

To the extent that "standards and controls" encompasses standards applicable to the analysis itself, *see, e.g.*, *Shipp*, 422 F. Supp. 3d at 779-81 (discussing the "circular and subjective" nature of the sufficient agreement standard and the inability of examiners "to protect against false positives" as an absence of "standards controlling the technique's

operation" (quoting *Daubert*, 509 U.S. at 594)), firearms identification faces an even greater challenge. As noted, "sufficient agreement," the threshold for reaching an "identification" classification, lacks any guiding standard other than the examiner's own subjective judgment. The AFTE Theory states that:

> "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.

The theory then observes that:

> [a]greement 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.

AFTE Theory (emphasis removed). The theory offers no guidance as to the quality or quantity of shared individual characteristics—even assuming it is possible to reliably differentiate these from subclass characteristics—that should cause an examiner to determine that two bullets were fired from the same firearm or the quality or quantity of different individual characteristics that should cause an examiner to reach the opposite conclusion.[25] *See* William A. Tobin & Peter J. Blau, *Hypothesis Testing of the Critical Underlying Premise of Discernible Uniqueness in Firearms-Toolmarks Forensic Practice*, 53 Jurimetrics J. 121, 125 (2013); 2009 NRC Report, at 153-54; *see also* Itiel E. Dror, Commentary, *The Error in "Error Rate": Why Error Rates Are So Needed, Yet So Elusive*,

[25] On cross-examination, Mr. McVeigh answered that he could not identify the "least number of matching individual characteristics" that he had "ever used to make an identification[,]" declining to say even whether it may have been as low as two shared characteristics.

65 J. Forensic Scis. 1034, 1037 (2020) (stating that "forensic laboratories vary widely in what decisions are verified").

As explained in the findings of the authors of the Ames II Study, in defending the decision not to treat inconclusive results as errors:

> When confronted with a myriad of markings to be compared, a decision has to be made about whether the variations noted rise above a threshold level the examiner has unconsciously assigned for each examination.

Ames II Study, at 75; *see also id.* ("[A]ll examiners must establish for themselves a threshold value for evaluation[.]"). A "standard" for evaluation that is dependent on each individual examiner "unconsciously assign[ing]" a threshold level "for each examination" may not undermine the reliability of the methodology to support generalized testimony about the consistency of patterns and marks on ammunition fired from a particular firearm and crime scene bullets. It does not, however, support the reliability of the methodology to identify, without qualification, a particular crime scene bullet as having been fired from a particular firearm.

On this issue, we find the results of phases two and three of the Ames II Study particularly enlightening. The PCAST Report identified accuracy, repeatability, and reproducibility as the key components of the foundational validity of any forensic technique. PCAST Report, at 5. Dr. Hamby testified at the *Frye-Reed* hearing that he agreed with that as a general proposition. The Ames II Study, which was not available at the time of the *Frye-Reed* hearing, was designed specifically to test the repeatability and reproducibility of the AFTE Theory methodology. For purposes of reviewing the

reliability of firearms identification to support the admissibility of expert testimony of a "match," the level of inconsistency identified through that study is troublesome.

Notably, at the *Frye-Reed* hearing, Mr. McVeigh rejected the notion that a firearms examiner looking at a bullet multiple times might come to different conclusions, stating that he believed that firearms identification's "repeatability is not in question." By the time of the *Daubert* hearing, however, the Ames II Study had been released, with data revealing that an examiner reviewing the *same* bullet set a second time classified it in the same AFTE category only 79% of time for matching sets and 65% of the time for non-matching sets. Ames II Study, at 39. In light of the black-box nature of the study, there is no explanation of this lack of consistency or of the lack of reproducibility shown in the same study.[26] Nonetheless, it highlights both (1) the absence of any standards or controls to guide the analysis of examiners and (2) the importance of testing unverified (though undoubtedly genuinely held) claims about reliability.

The lack of standards and controls is perhaps most acute in discerning whether a particular characteristic is a subclass or an individual characteristic. As noted, subclass characteristics are those shared by a group of firearms made using the same tools, such as those made in the same production run at a facility. Individual characteristics are those

[26] As noted above, the Ames II Study also found that an examiner reviewing a bullet set previously classified by a different examiner classified it in the same AFTE category 68% of the time for matching sets and 31% of the time for non-matching sets. Ames II Study, at 47. Even when the authors of the Ames II Study paired Identifications with Inconclusive-A responses and Eliminations with Inconclusive-C responses, second examiners still reached the same results as the first examiners looking at the same set of matching bullet sets only 77.4% of the time, and did so when looking at the same set of non-matching bullet sets only 49.0% of the time. Ames II Study, at 49.

specific to a particular firearm. Both can result from aspects of the manufacturing process; individual characteristics can also result from later events, such as ordinary wear and cleaning and polishing. Currently, there are no published standards or controls to guide examiners in identifying whether any particular pattern or mark is a subclass or an individual characteristic. Mr. McVeigh testified that examiners attempt to "guard against" this "subclass carryover," and that it is possible for a "trained examiner" to do so.[27] However, neither he nor any other witness identified any industry standards or controls addressing that topic.

On balance, consideration of the existence and maintenance of standards and controls weighs against admission of testimony of a "match" between a particular firearm and a particular crime scene bullet. *Accord Shipp*, 422 F. Supp. 3d at 782 ("[T]he court finds that the subjective and circular nature of AFTE Theory weighs against finding that a firearms examiner can reliably identify when two bullets or shell casings were fired from the same gun.").

[27] Mr. Abruquah relies on a 2007 study published in the AFTE Journal that was designed to test the possibility that cartridge cases fired from two pistols that had been shipped to the same retailer on the same date would show similarities in subclass characteristics. *See* Gene C. Rivera, *Subclass Characteristics in Smith & Wesson SW40VE Sigma Pistols*, 39 AFTE J. 247 (2007) (the "Rivera Study"). The Rivera Study found "alarming similarities" among the marks from the two different pistols, which, the author concluded, "should raise further concern for the firearm and tool mark examiner who may rely only on one particular type of mark for identification purposes." *Id.* at 250. The Rivera Study suggested that the AFTE Theory's "currently accepted standard for an identification" may need to be reconsidered as a result of very "significant" agreement between the two different pistols. *Id.* The AFTE seems to have responded by clarifying in the statement of its theory that an examiner's decision should be based on individual characteristics, but it has not provided standards for distinguishing those from subclass characteristics.

E. General Acceptance

Whether the AFTE Theory of firearms identification is generally accepted by the relevant community is largely dependent on what the relevant community is. Based on materials included in the record, as well as caselaw, the community of firearms identification examiners appears to be overwhelmingly accepting of the AFTE Theory. *See, e.g.*, *Romero-Lobato*, 379 F. Supp. 3d at 1122 (stating that "[t]he AFTE method certainly satisfies th[e general acceptance] element"); *United States v. Otero*, 849 F. Supp. 2d 425, 435 (D.N.J. 2012) (stating that the AFTE Theory is "widely accepted in the forensic community and, specifically, in the community of firearm and toolmark examiners"); *Willock*, 696 F. Supp. 2d at 571 ("[D]espite its inherent subjectivity, the AFTE theory . . . has been generally accepted within the field of toolmark examiners[.]"); *Monteiro*, 407 F. Supp. 2d at 372 ("[T]he community of toolmark examiners seems virtually united in their acceptance of the current technique.").

On the other hand, groups of eminent scientists and other academics have been critical of the absence of studies demonstrating the validity of firearms identification generally and the AFTE Theory specifically. *See, e.g.*, 2009 NRC Report, at 155; PCAST Report, at 111. Indeed, the record does not divulge evidence of general acceptance of the methodology by any group outside of firearms identification examiners and law enforcement.

We conclude that the relevant community for the purpose of determining general acceptance consists of both firearms examiners and the broader scientific community that has weighed in on the reliability of the methodology. The widespread acceptance of the

methodology among those who have vast experience with it, study it, and devote their careers to it is of great significance. However, we would be remiss were we to rely exclusively on a community that, by definition, is dependent for its livelihood on the continued viability of a methodology to sustain it, while ignoring the relevant and persuasive input of a different, well-qualified, and disinterested segment of professionals.[28] We consider this factor to be neutral.

F. Whether Opinions Emerged Independently or Were Developed for Litigation

The circuit court found that Mr. McVeigh's testimony grew naturally out of research independent of the litigation because "the ultimate purpose of th[e firearms and toolmark] evidence is investigation [into the victim's death], not litigation." We disagree. "Historically, forensic science has been used primarily in two phases of the criminal-justice process: (1) *investigation*, which seeks to identify the likely perpetrator of a crime, and (2) *prosecution*, which seeks to prove the guilt of a defendant beyond a reasonable doubt."[29] *See* PCAST Report, at 4. The use of firearms identification in a criminal prosecution is not independent of its investigative use. Nonetheless, the purpose of this factor is to determine whether there is reason for skepticism that the opinion reached might

[28] In his dissent, Justice Gould takes Mr. Abruquah to task for not retaining his own firearms examiner to provide a different analysis of the bullets at issue. Dissenting Op. of Gould, J. at 42. In doing so, Justice Gould assumes that there are firearms examiners whose services were readily available to Mr. Abruquah, i.e., who are willing and able to take on work for criminal defendants in such cases. The record contains no support for that proposition.

[29] Here, for example, it appears that Mr. Abruquah was already identified as the likely perpetrator of the murder before Mr. McVeigh began his analysis of the Taurus revolver and the crime scene bullets.

be tailored to the preferred result for the litigation, rather than the expert's considered, independent conclusion. Here, the circuit court lauded Mr. McVeigh's integrity and forthrightness, and we have no reason to second-guess that view.[30] Crediting the court's findings about Mr. McVeigh's testimony, we are confident that the court would not weigh this factor against admissibility and so we will not either.

G. Unjustified Extrapolation from Accepted Premise

Citing Mr. Abruquah's "voluminous data indicating that firearms identification evidence is unjustifiably extrapolated from the toolmarks" and the State's "credible and persuasive evidence that all extrapolations are justifiably calculated and well-reasoned[,]" the circuit court found "this factor to be in equipoise" and so to weigh against admission. In *Rochkind*, we explained that this factor invokes the concept of an analytical gap, as "[t]rained experts commonly extrapolate from existing data[,]" but a circuit court is not required "to admit opinion evidence that is connected to existing data only by the *ipse dixit* of the expert." 471 Md. at 36 (quoting *Joiner*, 522 U.S. at 146). "An 'analytical gap' typically occurs as a result of 'the failure by the expert witness to bridge the gap between [the expert's] opinion and the empirical foundation on which the opinion was derived.'" *Matthews*, 479 Md. at 317 (quoting *Savage v. State*, 455 Md. 138, 163 (2017)).

[30] We observe that another seasoned trial judge, even while limiting Mr. McVeigh's testimony more than we do here, was equally profuse in his laudatory comments about Mr. McVeigh's integrity. *See United States v. Medley*, No. PWG-17-242 (D. Md. April 24, 2019), ECF No. 111, at 14 ("Mr. McVeigh, who was, for an expert witness, . . . remarkably forthcoming in his testimony and credible."); *id.* at 53-54 ("I've seldom seen an expert who is as sincere and straightforward and no baloney and genuine in what he did as Mr. McVeigh."). Nothing about our opinion or our conclusion in this case should be understood as contradicting that sentiment.

Although we do not preclude the possibility that the gap may be closed in the future, for the reasons already discussed, this case presents just such an analytical gap. That gap should have foreclosed Mr. McVeigh's unqualified testimony that the crime scene bullets and bullet fragment were fired from Mr. Abruquah's Taurus revolver. Although the court precluded Mr. McVeigh from testifying to his opinions to a "certainty," an unqualified statement that the bullets were fired from Mr. Abruquah's revolver is still more definitive than can be supported by the record. To be sure, the AFTE Theory is intended to allow firearms examiners to reach conclusions linking particular firearms to particular unknown bullets. Mr. McVeigh's testimony was thus not an unjustified departure from the methodology employed by those practicing in his field. We conclude, however, for reasons discussed above, that although the studies and other information in the record support the use of the AFTE Theory to reliably identify whether patterns and lines on bullets of unknown origin are consistent with those known to have been fired from a particular firearm, they do not support the use of that methodology to reliably opine without qualification that the bullets of unknown origin were fired from the particular firearm.

H. Accounting for Obvious Alternative Explanations

The court found this factor "definitively weighs in favor of admission" because Mr. McVeigh and Dr. Hamby "clearly and concisely addressed how alternative interpretations of toolmarks are generally accounted for in the field of firearms identification," and Mr. Abruquah's "counters in this area were ineffective." We disagree. For reasons already addressed, without the ability to examine other bullets fired from other firearms in the same production run as the firearm under examination, the record simply

does not support that firearms identification can reliably eliminate all alternative sources so as to permit unqualified testimony of a match between a particular firearm and a particular crime scene bullet.

I. Level of Care

Mr. McVeigh's testimony here was given as part of his regular professional work, rendering this factor technically inapplicable. Nonetheless, to the extent this factor can be re-cast as a general inquiry into the level of care he exhibited, we have no qualms about accepting the circuit court's determination that Mr. McVeigh is a "consummate professional in his field" and demonstrated a "level of care in this case" that was not "assailed in any convincing manner."

J. Relationship Between Reliability of Methodology and Opinion to Be Offered

Based on the State's evidence concerning the reliability of firearms examinations and "a dearth of real-life examples of erroneous examinations," the circuit court concluded that "firearm and toolmark evidence is known to reach reliable results" and, therefore, that this final factor favors admission of the evidence. We do not question that firearms identification is generally reliable, and can be helpful to a jury, in identifying whether patterns and markings on "unknown" bullets or cartridges are consistent or inconsistent with those on bullets or cartridges known to have been fired from a particular firearm. For that reason, to the extent Mr. Abruquah suggests that testimony about the consistency of

such patterns and markings should be excluded, we disagree.[31] It is also possible that experts who are asked the right questions or have the benefit of additional studies and data may be able to offer opinions that drill down further on the *level* of consistency exhibited by samples or the likelihood that two bullets or cartridges fired from different firearms might exhibit such consistency. However, based on the record here, and particularly the lack of evidence that study results are reflective of actual casework, firearms identification has not been shown to reach reliable results linking a particular unknown bullet to a particular known firearm.

For those reasons, we conclude that the methodology of firearms identification presented to the circuit court did not provide a reliable basis for Mr. McVeigh's unqualified opinion that four bullets and one bullet fragment found at the crime scene in this case were fired from Mr. Abruquah's Taurus revolver. In effect, there was an analytical gap between the type of opinion firearms identification can reliably support and the opinion Mr. McVeigh offered.[32] Accordingly, the circuit court abused its discretion in permitting Mr. McVeigh to offer that opinion.

---

[31] As noted, Mr. Abruquah argues that the testimony of a firearms identification examiner should be limited to opining, "at most, that a firearm cannot be excluded as the source of the questioned projectile[.]" It is not entirely clear to us whether Mr. Abruquah believes that testimony about the consistency of patterns and markings on bullets would be permissible—and, indeed, necessary to establish the basis for an opinion that a firearm cannot be excluded—or whether he believes that testimony about the consistency of such patterns and markings goes too far and should be excluded. If the latter, we disagree for the reasons identified.

[32] Both dissenting opinions contend that we have been insufficiently deferential to the trial court's determination. Although they observe, quite correctly, that we do not ask trial judges to play the role of "amateur scientists," Dissenting Op. of Hotten, J. at 4

IV. HARMLESS ERROR

The State argues in the alternative that any error in admitting Mr. McVeigh's testimony was harmless. We disagree.

"The harmless error doctrine is grounded in the notion that a defendant has the right to a fair trial, but not a perfect one." *State v. Jordan*, 480 Md. 490, 505 (2022). The doctrine is strictly limited only to "error[s] in the trial process itself" that may warrant reversal. *Id.* at 506 (quoting *Weaver v. Massachusetts*, 137 S. Ct. 1899, 1907 (2017)). For an appellate court to conclude that the admission of expert testimony was harmless, the State must show "beyond a reasonable doubt, that the error in no way influenced the verdict." *Dionas*, 436 Md. at 108 (quoting *Dorsey*, 276 Md. at 659).

Upon our review of the record, we are not convinced beyond a reasonable doubt that the expert testimony in no way contributed to the guilty verdict. The firearm and toolmark identification evidence was the only direct evidence before the jury linking Mr. Abruquah's gun to the crime. Absent that evidence, the guilty verdict rested upon circumstantial evidence of a dispute between the men, a witness who heard gunfire around the time of the dispute, a firearm recovered from the residence, and testimony of a jailhouse

---

(quoting *Rochkind*, 471 Md. at 33-34); Dissenting Op. of Gould, J. at 1, 50, we also do not provide increased deference simply because the subject matter of the expert testimony is scientific. The forensic technique under review was, until relatively recently, accepted almost entirely without critical analysis. *See* discussion above at 16-17. *Daubert* and *Rochkind* demand more than adherence to an orthodoxy simply because it has long been accepted or because of the number of impressive-sounding statistics generated by studies that do not establish the reliability of the specific testimony offered. They require that the party proffering such evidence, whatever type of evidence it is, establish that it meets a minimum threshold of reliability.

informant. To be sure, that evidence is strong. But the burden of showing that an error was harmless is high and we cannot say, beyond a reasonable doubt, that the admission of the particular expert testimony at issue did not influence or contribute to the jury's decision to convict Mr. Abruquah. *See Clemons v. State*, 392 Md. 339, 372 (2006) (stating that "[l]ay jurors tend to give considerable weight to 'scientific' evidence when presented by 'experts' with impressive credentials" (quoting *Reed v. State*, 283 Md. 374, 386 (1978))).

CONCLUSION

Based on the evidence presented at the hearings, we hold that the circuit court did not abuse its discretion in ruling that Mr. McVeigh could testify about firearms identification generally, his examination of the bullets and bullet fragments found at the crime scene, his comparison of that evidence to bullets known to have been fired from Mr. Abruquah's Taurus revolver, and whether the patterns and markings on the crime scene bullets are consistent or inconsistent with the patterns and markings on the known bullets. However, the circuit court should not have permitted the State's expert witness to opine without qualification that the crime scene bullets were fired from Mr. Abruquah's firearm. Because the court's error was not harmless beyond a reasonable doubt, we will therefore

reverse the circuit court's ruling on Mr. Abruquah's motion in limine, vacate Mr. Abruquah's convictions, and remand for a new trial.

RULING ON MOTION IN LIMINE CONCERNING EXPERT TESTIMONY REVERSED; JUDGMENT OF THE CIRCUIT COURT FOR PRINCE GEORGE'S COUNTY VACATED; CASE REMANDED FOR A NEW TRIAL. COSTS TO BE PAID BY PRINCE GEORGE'S COUNTY.

Circuit Court for Prince George's County
Case No. CT121375X
Argued: October 4, 2022

IN THE SUPREME COURT

OF MARYLAND*

No. 10

September Term, 2022

KOBINA EBO ABRUQUAH

v.

STATE OF MARYLAND

Fader, C.J.,
Watts,
Hotten,
Booth,
Biran,
Gould,
Eaves,

JJ.

Dissenting Opinion by Hotten, J., which Eaves, J., joins.

Filed: June 20, 2023

*During the November 8, 2022 general election, the voters of Maryland ratified a constitutional amendment changing the name of the Court of Appeals to the Supreme Court of Maryland. The name change took effect on December 14, 2022.

Respectfully, I dissent. I would hold that the Circuit Court for Prince George's County did not abuse its discretion in admitting the State's expert firearm and toolmark identification testimony and evidence, following its analysis and consideration of the factors outlined in *Rochkind v. Stevenson*, 471 Md. 1, 236 A.3d 630 (2020). "When the basis of an expert's opinion is challenged pursuant to Maryland Rule 5-702, the review is abuse of discretion." *Id.* at 10, 236 A.3d at 636 (citation omitted); *State v. Matthews*, 479 Md. 278, 305, 277 A.3d 991, 1007 (2022) (citation omitted). We have declared it "the rare case in which a Maryland trial court's exercise of discretion to admit or deny expert testimony will be overturned." *Matthews*, 479 Md. at 286, 306, 277 A.3d at 996, 1008. This should not be one of those instances.

The Circuit Court Did Not Abuse Its Discretion in Admitting the State's Firearm Toolmark Identification Testimony Under *Rochkind*.

In *Rochkind*, this Court abandoned the *Frye-Reed* standard in favor of the more "flexible" analysis set forth in *Daubert v. Merrell Dow Pharmaceuticals, Inc*., 509 U.S. 579, 113 S. Ct. 2786 (1993), concerning the admissibility of expert testimony. *Rochkind*, 471 Md. at 29, 34, 236 A.3d at 646, 650. *Rochkind* prescribes ten factors for trial judges to consider when applying Maryland Rule 5-702.[1] *See id.* at 35, 236 A.3d at 650 (emphasis added). First, the trial court must consider the original five *Daubert* factors:

---

[1] Rule 5-702 pertains to the admissibility of expert testimony and provides, in full:

> Expert testimony may be admitted, in the form of an opinion or otherwise, if the court determines that the testimony will assist the trier of fact to understand the evidence or to determine a fact in issue. In making that determination, the court shall determine[:]

(continued . . .)

(1) whether a theory or technique can be (and has been) tested;

(2) whether a theory or technique has been subjected to peer review and publication;

(3) whether a particular scientific technique has a known or potential rate of error;

(4) the existence and maintenance of standards and controls; and

(5) whether a theory or technique is generally accepted.

Id., 236 A.3d at 650 (quoting *Daubert*, 509 U.S. at 593–94, 113 S. Ct. 2786). Next, "courts have developed additional factors for determining whether expert testimony is sufficiently reliable[,]" including:

(6) whether experts are proposing to testify about matters growing naturally and directly out of research they have conducted independent of the litigation, or whether they have developed their opinions expressly for purposes of testifying;

(7) whether the expert has unjustifiably extrapolated from an accepted premise to an unfounded conclusion;

(8) whether the expert has adequately accounted for obvious alternative explanations;

(. . . continued)

(1) whether the witness is qualified as an expert by knowledge, skill, experience, training, or education,

(2) the appropriateness of the expert testimony on the particular subject, and

(3) whether a sufficient factual basis exists to support the expert testimony.

"[S]ufficient factual basis" includes two subfactors: "(1) an adequate supply of data; and (2) a reliable methodology." *Rochkind*, 471 Md. at 22, 236 A.3d at 642 (citation omitted). Without either, the expert's testimony is considered to be mere "speculation or conjecture." *Id.*, 236 A.3d at 642 (internal quotations and citation omitted).

> (9) whether the expert is being as careful as he [or she] would be in his [or her] regular professional work outside his [or her] paid litigation consulting; and
>
> (10) whether the field of expertise claimed by the expert is known to reach reliable results for the type of opinion the expert would give.

Id. at 35–36, 236 A.3d at 650 (quoting Fed. R. Evid. 702 Advisory Committee Note).

We adopted *Rochkind* to "refine" and "streamline the evaluation of scientific expert testimony under [Md.] Rule 5-702." *Id.* at 30, 35, 236 A.3d at 647, 650. As a threshold matter, scientific testimony must be relevant and reliable. *Id.* at 14, 236 A.3d at 638 (citation omitted). *Rochkind* provided more flexibility for the gatekeeping mechanism of ascertaining whether the expert evidence should be admitted in its analytical shift to a "reliability" standard (*Daubert*), as opposed to "general acceptance" (*Frye-Reed*). The *Rochkind* elements "provide guidance on *how* to determine if scientific reasoning is, indeed, sound, or a scientific theory adequately justifies an expert's conclusion." *Id.* at 33, 236 A.3d at 649. "[A]ll of the *Daubert* factors are relevant to determining the reliability of expert testimony, *yet no single factor is dispositive in the analysis*. A trial court may apply some, all, or none of the factors depending on the particular expert testimony at issue." *Id.* at 37, 236 A.3d at 651 (emphasis added) (citation omitted). As the U.S. Supreme Court recognized, "*Daubert's* list of specific factors *neither necessarily nor exclusively* applies to all experts or in every case. Rather, the law grants a [trial] court the same broad latitude when it decides *how* to determine reliability as it enjoys in respect to its ultimate reliability determination." *Kumho Tire Co. v. Carmichael*, 526 U.S. 137, 141–42, 119 S. Ct. 1167, 1171 (1999) (emphasis added) (citation omitted); *Matthews*, 479 Md.

at 314, 277 A.3d at 1012 (quoting *Kumho Tire*, 526 U.S. at 141–42, 119 S. Ct. at 1171); *Rochkind*, 471 Md. at 37, 236 A.3d at 651 (quoting *Kumho Tire*, 526 U.S. at 141–42, 119 S. Ct. at 1171); *Savage v. State*, 455 Md. 138, 178, 166 A.3d 183, 206 (2017) ("[A] trial court is not required to consider any or all of the *Daubert* factors in making its reliability determination—they were 'meant to be helpful, not determinative.'" (Adkins, J., concurring) (quoting *Kumho Tire*, 526 U.S. at 151, 119 S. Ct. at 1175)). Trial judges, therefore, assume the critical role as "gatekeepers" against unreliable scientific evidence. *Rochkind*, 471 Md. at 38, 236 A.3d at 652; *Matthews*, 479 Md. at 322, 277 A.3d at 1017 (citation omitted); *Daubert*, 509 U.S. at 597, 113 S. Ct. at 2798; *Gen. Elec. Co. v. Joiner*, 522 U.S. 136, 142, 118 S. Ct. 512, 517 (1997); *Kumho Tire*, 526 U.S. at 147, 119 S. Ct. at 1174 (noting that *Daubert's* gatekeeping function also applies to expert testimony); Victor E. Schwartz, *Expert Testimony Needs Judges to Act As "Gatekeepers": The Maryland Court of Appeals Teaches Why*, 13 J. Tort L. 229, 231 (2020).

Trial judges were provided these factors to assist in the evidence-based management of their judicial gatekeeping function in criminal, civil, and equitable causes. The gatekeeping function is significant, particularly for the ability of the finders of fact to evaluate the scientific evidence and testimony and determine whether it should be accepted or rejected in their ultimate determination. This Court has rejected the argument that judges are "amateur scientists[:]"

> [T]rial judges are *not* required to make a determination of the ultimate scientific validity of any scientific propositions. Instead, they need only make a much more limited inquiry: *whether sufficient indicia of legitimacy exist to support the conclusion that evidence derived from the principle may be profitably considered by a fact finder at trial*. We are confident that trial

> judges are duly capable of undertaking the reliability analysis absent scientific training.

Rochkind, 471 Md. at 34, 236 A.3d at 649 (emphasis added) (internal quotations and citations omitted). "Applying these standards, we determine that the [circuit court's] decision in this case . . . was within its discretion and therefore lawful." *Kumho Tire*, 526 U.S. at 142, 119 S. Ct. at 1171.

In light of the newly adopted *Rochkind* standard, the circuit court reconsidered the admissibility of the firearm and toolmark identification expert evidence, known as the AFTE methodology.[2] The court examined the *Rochkind* factors "by way of the pleadings, testimony and evidence presented during the 5-day hearing conducted prior to the second jury trial, coupled with the supplemental hearing and pleadings conducted after the most recent remand." The court found that the State's evidence for factors one, two, three, five, six, eight, nine, and ten weighed in favor of admission. The evidence for factors four and seven weighed against admission. Based on the *Rochkind* factors and the "totality of the evidence and arguments presented," the circuit court admitted the State's firearms and toolmark examination evidence.

As the majority notes, Mr. Abruquah argues on appeal that the State's expert, Mr. McVeigh, should have been "limited to opining, 'at most, that a firearm cannot be excluded as the source of the questioned projectile'" because the methodology is "insufficiently reliable" to support Mr. McVeigh's testimony. Maj. Op. at 38. To determine this "tailored"

[2] The State's expert, Mr. Scott McVeigh, uses the "AFTE method" to "compar[e] microscopic markings on a bullet or cartridge case to make an 'identification,' *i.e.*, to opine that a specific firearm is the source of a fired ammunition component."

issue of "whether the AFTE Theory can reliably support an unqualified opinion that a particular firearm is the source of one or more particular bullets[,]" the majority conducts its own *Rochkind* analysis. *Id.* at 38–39.

The majority holds that factors one, six, and nine weighs in favor of admission. *Id.* at 39, 52–53, 55. Factors two, three, four, seven, eight, and ten, the majority concludes, weighs against admission. *Id.* at 40, 45, 50, 54–56. The majority notes that factor five is "neutral." *Id.* at 52. Upon consideration of the factors, the majority determines that the record, "on balance," does not support Mr. McVeigh's "unqualified testimony that the crime scene bullets and bullet fragments were fired from Mr. Abruquah's Taurus revolver." *Id.* at 45, 54. According to the majority, "the studies and other information in the record . . . do not support the use of [the AFTE Theory] to reliably opine without qualification that the bullets of unknown origin were fired from the particular firearm." *Id.* at 54. Specifically, the "firearms identification has not been shown to reach reliable results linking a particular unknown bullet to a particular known firearm." *Id.* at 56. The majority, therefore, holds that "there was an analytical gap between the type of opinion firearms identification can reliably support and the opinion Mr. McVeigh offered." *Id.* (footnote omitted). To the majority, this "gap should have foreclosed" Mr. McVeigh's unqualified testimony. *Id.* at 54.

I disagree, finding no error with the circuit court's analysis. The concept of the "analytical gap" originated in *Joiner*, 522 U.S. 136, 118 S. Ct. 512. *Rochkind*, 471 Md. at 43, 236 A.3d at 654 (Watts, J., dissenting). Over the years, it has become a "critical" component in Maryland's evidentiary analysis. *Id.* at 14, 236 A.3d at 638. As we've

explained, the role of the expert is to "connect[] the dots" or "provide[] a causal link" between the data and/or science used by the expert and the expert's ultimate conclusions. *See id.* at 14–19, 236 A.3d at 638–40 (internal quotation marks and citations omitted). In essence, an "analytical gap" results when the expert fails to "'bridge'" the gap between the expert's opinion and "'the empirical foundation on which the opinion was derived.'" *Matthews*, 479 Md. at 317, 277 A.3d at 1014 (quoting *Savage*, 455 Md. at 163, 166 A.3d at 198). In determining reliability, the trial judge "must also consider the relationship between the methodology applied and conclusion reached." *Rochkind*, 471 Md. at 36, 236 A.3d at 651. However, neither *Daubert* nor the Federal Rules of Evidence require trial judges "'to admit opinion evidence that is connected to existing data only by the *ipse dixit* of the expert.'" *Id.*, 236 A.3d at 651 (quoting *Joiner*, 522 U.S. at 144, 118 S. Ct. at 519). "[T]he question then becomes: is this *specific* causation case . . . , where the analytical gap was too vast, or a [] case where the analytical gap was sufficiently bridged?" *Id.* at 25, 236 A.3d at 644.

The circuit court thoroughly followed the *Rochkind* factors as prescribed by the U.S. Supreme Court and this Court. Under the previous *Frye-Reed* standard, trial courts determined whether an expert's methodology was "generally accepted in the scientific community." *Id.* at 12–13, 236 A.3d at 637; *Matthews*, 479 Md. at 307, 277 A.3d at 1008 ("[P]rior to . . . *Daubert* [], the *Frye* 'general acceptance' test was the dominant standard that courts used to determine the admissibility of novel scientific evidence."). "Under *Daubert*, judges are charged with gauging only the threshold *reliability—not the ultimate validity*—of a particular methodology or theory." *Rochkind*, 471 Md. at 33, 236 A.3d at

649 (emphasis added). In conducting the "reliability assessment," courts are to consider the non-exhaustive list of factors provided in *Daubert*. *Matthews*, 479 Md. at 307, 277 A.3d at 1008 (citing *Daubert*, 509 U.S. at 593–94, 113 S. Ct. at 2797). Trial courts are granted "broad latitude to determine[]" "reliability in a particular case[.]" *Kumho Tire*, 526 U.S. at 153, 119 S. Ct. at 1176; *Rochkind*, 471 Md. at 37, 236 A.3d at 651 (quoting *Kumho Tire*, 526 U.S. at 141–42, 119 S. Ct. at 1171). Thus, if "a trial court is satisfied that an expert has applied a reliable methodology to an adequate supply of data, the court should not exclude the expert's testimony merely because the court is concerned that the expert's particular conclusions may be inaccurate." *Matthews*, 479 Md. at 316, 277 A.3d at 1013.

While *Rochkind* requires trial judges to conduct an analysis under the *Rochkind-Daubert* factors, this Court does not require trial judges to be "scientists" and arrive at a conclusion with some measure of mathematical certainty. *See Rochkind*, 471 Md. at 33, 236 A.3d at 649; *Daubert*, 509 U.S. at 597, 113 S. Ct. at 2799 ("[T]he Rules of Evidence . . . do assign to the trial judge the task of ensuring that an expert's testimony both rests on a reliable foundation and is relevant to the task at hand. Pertinent evidence based on scientifically valid principles will satisfy those demands."). As the U.S. Supreme Court observed, "there are no certainties in science." *Daubert*, 509 U.S. at 590, 113 S. Ct. at 2795 (citation omitted). Accordingly, a trial judge's gatekeeping role isn't to determine whether the expert is "right" or "wrong;" rather, the judge's role is to determine whether the expert's testimony is "adequately grounded in reliable and sound science, and that there is not 'too great an analytical gap' between the expert's methodology and conclusions." Schwartz, *supra*, at 233 (quoting *Rochkind*, 471 Md. at 36, 236 A.3d at 651); Maj. Op. at

9. This is exactly what the circuit court did when it recognized Mr. Abruquah's "Herculean effort" in demonstrating "why the evidence should be heavily scrutinized, questioned and potentially impeached[.]" While the court did not, however, expressly address an "analytical gap," it observed that "the crux" of Mr. Abruquah's arguments "address impeachment rather than admissibility."

In *Rochkind*, we rejected the argument that the *Rochkind-Daubert* standard "enable[d] judges to . . . 'usurp[] the role of juries.'" *Rochkind*, 471 Md. at 33, 236 A.3d at 649. The power to weigh the validity of the evidence still sits with the jury or fact finder. *Id.*, 236 A.3d at 649. While I recognize the importance of "published standards," it is for *the jury or the factfinder* to determine the validity of the methodology of the firearm identification testimony presented—*not the circuit court or this Court*. Maj. Op. at 50; *Rochkind*, 471 Md. at 33, 236 A.3d at 649. We reaffirmed this principle a year later in *State v. Matthews*, acknowledging that "[t]he unknown degree of uncertainty concerning the accuracy of [the expert testimony] went to the weight the jury should give to the expert testimony, not to its admissibility." 479 Md. at 313, 277 A.3d at 1012 (footnote omitted). Here, the circuit court continuously reaffirmed this notion, stating that Mr. Abruquah's critiques of the firearm identification evidence "are more suited to the *weight* such evidence should be given *at trial*." (Emphasis added). In doing so, it fulfilled its obligation under *Rochkind*.

The majority notes that Mr. McVeigh's testimony "was the only direct evidence before the jury linking Mr. Abruquah's gun to the crime." Maj. Op. at 57. "Absent that evidence," the majority observes that Mr. Abruquah's "guilty verdict rested upon

circumstantial evidence[.]" *Id*. at 57. Accordingly, the majority concludes that the admission of such testimony was not harmless because it "cannot say, beyond a reasonable doubt, that the admission of the particular expert testimony at issue did not influence or contribute to the jury's decision to convict Mr. Abruquah." *Id*. at 58. This is especially so, the majority recognizes, because "[l]ay jurors tend to give considerable weight to 'scientific evidence' when presented by 'experts' with impressive credentials[.]" *Id*. (quoting *Clemons v. State*, 392 Md. 339, 372, 896 A.2d 1059, 1078 (2006) (internal quotations and citation omitted)).

Assuming, *arguendo*, that the circuit court erred in admitting Mr. McVeigh's testimony, such error *was* harmless considering the overwhelming circumstantial and direct evidence of guilt tying Mr. Abruquah to the shooting. As the majority notes, the responding officers left Mr. Abruquah and Mr. Aguirre-Herrera after their third response to the men's shared residence around 12:15 a.m. According to the officers, Mr. Aguirre-Herrera appeared to be terrified of Mr. Abruquah. A nearby witness testified that he heard gunshots between 11:30 p.m. and 12:30 a.m. During questioning, Mr. Abruquah told police where to find his firearms in the men's shared residence, including Mr. Abruquah's Taurus .38 Special revolver. The State also introduced into evidence the transcript of the testimony of Cecil Muhammed, Mr. Abruquah's jail cellmate. Mr. Muhammed testified that, while they were incarcerated together, Mr. Abruquah confessed to shooting Mr. Aguirre-Herrera with his Taurus .38 on the night in question. According to Mr. Muhammed, Mr. Abruquah and Mr. Aguirre-Herrera were in a relationship, but Mr. Aguirre-Herrera engaged in prostitution through Craigslist and conducted such business in

the men's shared residence. Mr. Muhammed testified that this allegedly enraged Mr. Abruquah and made him jealous. We recognize this evidence is circumstantial; yet, as the majority itself observes, is "strong." Maj. Op. at 58.

Trial judges provide jury instructions "to aid the jury in clearly understanding the case, to provide guidance for the jury's deliberations, and to help the jury arrive at a correct verdict." *Stabb v. State*, 423 Md. 454, 464, 31 A.3d 922, 928 (2011) (internal quotation marks and citation omitted). As such, trial judges instruct juries that "[a] conviction may rest on *circumstantial evidence alone*, on direct evidence alone, or on a combination of circumstantial and direct evidence." *Taylor v. State*, 473 Md. 205, 218 n.8, 249 A.3d 810, 818 n.8 (2021) (emphasis added) (internal quotation marks and citation omitted). "The law makes no distinction between the weight to be given to either direct or circumstantial evidence." Maryland Criminal Pattern Jury Instructions ("MPJI-Cr") 3:01 (Maryland State Bar Association 2d ed. 2022). "'Circumstantial evidence may support a conviction if the circumstances, taken together, do not require the trier of fact to resort to speculation or conjecture . . . . It must afford the basis for an inference of guilt beyond a reasonable doubt.'" *Beckwitt v. State*, 477 Md. 398, 429, 270 A.3d 307, 325 (2022) (quoting *Smith v. State*, 415 Md. 174, 185, 999 A.2d 986, 992 (2010)). Accordingly, even in a case relying solely on circumstantial evidence, "the finder of fact has the 'ability to choose among differing inferences that might possibly be made from a factual situation[.]'" *Smith*, 415 Md. at 183, 999 A.2d at 991 (quoting *State v. Smith*, 374 Md. 527, 534, 823 A.2d 664, 668 (2003)).

"[A] fundamental principle underlying trial by jury is that the credibility of a witness and the weight to be accorded the witness' testimony are solely within the province of the jury." *Fallin v. State*, 460 Md. 130, 154, 188 A.3d 988, 1002 (2018) (internal quotation marks and citation omitted); *see also* MPJI-Cr 3:10. MPJI-Cr 3:14 provides, in part, that jurors should:

> [C]onsider an expert's testimony *together with all the other evidence*. . . . You should give expert testimony the weight and value *you believe it should have*. You are *not required* to accept an expert's testimony, even if it is uncontradicted. As with any other witness, *you may believe all, part, or none of the testimony of any expert*.

(Emphasis added). In adopting *Daubert*, we reiterated this notion, affirming that "juries will continue to weigh competing, but still reliable, testimony." *Rochkind*, 471 Md. at 33, 236 A.3d at 649. We, therefore, cannot hold that the firearm evidence did or "did not influence or contribute to the jury's decision[.]" Maj. Op. at 58; *Stokes v. State*, 379 Md. 618, 638, 843 A.2d 64, 75 (2004) ("Jury deliberations are private and are to be conducted in secret." (citation omitted)). "We need not decide whether the jury could have drawn other inferences from the evidence, refused to draw inferences, or whether we would have drawn different inferences from the evidence." *Smith*, 415 Md. at 184, 999 A.2d at 991 (citation omitted). The jury could have based its verdict upon a weighing of *all* the evidence, including the scientific and circumstantial evidence. *See* MPJI-Cr 3:14; *see also Howling v. State*, 478 Md. 472, 507, 274 A.3d 1124, 1144 (2022) ("[O]ur concern is only whether the verdict was supported by sufficient evidence, direct or circumstantial, which could fairly convince a trier of fact of the defendant's guilt of the offenses charged beyond a reasonable doubt." (internal quotation marks and citations omitted)); *State v. Manion*,

442 Md. 419, 437, 112 A.3d 506, 517 (2015) (recognizing that "a rational trier of fact could conclude, beyond a reasonable doubt, that" the defendant intended to commit a crime based circumstantial evidence). Concluding otherwise, as the majority does here, minimizes the importance of both the role of the jury and jury instructions if we expect juries to believe that direct evidence, especially if it's scientific evidence, provides any more persuasive value than circumstantial evidence. *See* MPJI-Cr 3:01; *see also Taylor*, 473 Md. at 218 n.8, 249 A.3d at 818 n.8 ("No greater degree of certainty is required when the evidence is circumstantial than when it is direct." (internal quotation marks and citation omitted)).

The circuit court followed *Rochkind* within the letter of the law as prescribed. *See Matthews*, 479 Md. at 305, 277 A.3d at 1007 (citing *Jenkins v. State*, 375 Md. 284, 296, 825 A.2d 1008, 1015 (2003)). Based upon the extensive hearings, pleadings, testimony, and evidence presented, the circuit court was satisfied that the State met its burden to admit the firearm identification expert evidence for consideration by the jury.[3] Its decision was neither "well removed from any center mark" nor "beyond the fringe of what [this] [C]ourt

[3] The majority's analysis is largely predicated on a consideration of the *Daubert* factors. Maj. Op. at 39-56. While the majority recognizes that the factors "are neither exhaustive nor mandatory," its rationale seems to suggest otherwise. *Id.* at 39 (internal quotation marks and citation omitted). As we expressed, the list of *Daubert* factors is not exhaustive or mandatory. *Kumho Tire*, 526 U.S. at 141, 119 S. Ct. at 1171; *Rochkind*, 471 Md. at 36–37, 236 A.3d at 651 (internal quotation marks and citation omitted). The circuit court had "broad latitude" to consider how to determine the reliability of the State's firearm and toolmark identification expert evidence. *Kumho Tire*, 526 U.S. at 141, 119 S. Ct. at 1171. In addition to its *Daubert* analysis, the circuit court considered other aspects of the case expressed herein in making its ultimate reliability determination. It was within the court's discretion and capacity to do so as the gatekeeper. *Id.*, 119 S. Ct. at 1171; *Rochkind*, 471 Md. at 36–37, 236 A.3d at 651 (internal quotation marks and citation omitted). "For these [] reasons *taken together*, it concluded that [the] testimony was" admissible. *Id.* at 156, 119 S. Ct. at 1178.

deems minimally acceptable." *Id.*, 277 A.3d at 1007 (internal quotation marks and citation omitted). The majority considers this "standard" to be "somewhat unfair." Maj. Op. at 6 n.5. While it observes that "the circuit court acted deliberately and thoughtfully in approaching, analyzing, and resolving the question before it[,]" the majority nonetheless "c[a]me to a different conclusion concerning the outer bounds of what is acceptable expert evidence in this area[]" and provides no guidance for the trial court in terms of what standard applies. *Id*. at 6–7 n.5. This Court does not "reverse simply because [*we*] *would not* have made the same ruling." *Devincentz v. State*, 460 Md. 518, 550, 191 A.3d 373, 391 (2018) (emphasis added) (internal quotations and citation omitted). A reasonable person *would* take the view adopted by the circuit court here. *Williams v. State*, 457 Md. 551, 563, 179 A.3d 1006, 1013 (2018). It was, therefore, within the realm of the jury, as the triers of fact, to resolve the firearm toolmark analysis and opinion, along with the other evidence presented, in rendering its verdict. *See* MPJI-Cr 3:14.

The majority's holding blurs the role of the trial judge, allowing judges to "exclude . . . legitimate opinions of experts[] that [] are for a jury to weigh credibility." *Rochkind*, 471 Md. at 33, 236 A.3d at 649. The majority appears to conflate the role of the trial judge as gatekeepers, with the evaluation of the science or the expert opinion that is presented for consideration of its admissibility by the judge. That is not what *Rochkind* required. At the time of *Rochkind*, we did not "foresee th[is] gloomy outlook." *Id.*, 236 A.3d at 649. However, the majority's decision does exactly that.

CONCLUSION

For these reasons, I respectfully dissent and would affirm the judgment of the Circuit Court for Prince George's County. Justice Eaves has authorized me to state that she joins in this opinion.

Circuit Court for Prince George's County
Case No. CT121375X
Argued: October 4, 2022

IN THE SUPREME COURT

OF MARYLAND*

No. 10

September Term, 2022

KOBINA EBO ABRUQUAH

v.

STATE OF MARYLAND

Fader, C.J.,
Watts,
Hotten,
Booth,
Biran,
Gould,
Eaves,

JJ.

Dissenting Opinion by Gould, J.

Filed: June 20, 2023

*During the November 8, 2022 general election, the voters of Maryland ratified a constitutional amendment changing the name of the Court of Appeals to the Supreme Court of Maryland. The name change took effect on December 14, 2022.

In *Rochkind v. Stevenson*, 471 Md. 1 (2020), this Court adopted the *Daubert*[1] framework for the admission of expert testimony and embraced certain important principles. Justice Hotten highlights these principles in her dissent, some of which I reiterate here for context. Dissenting Op. of Justice Hotten 1-4.

First, the *Daubert-Rochkind* factors ("*Daubert* factors") provide trial courts with a flexible guide—not a mandatory scoresheet—for serving as gatekeepers with respect to scientific or technical evidence. *See, e.g.*, *Rochkind*, 471 Md. at 36-37; *Daubert*, 509 U.S. at 589, 596.

Second, trial courts are not tasked with determining "the ultimate scientific validity of any scientific propositions." *Rochkind*, 471 Md. at 34 (quoting *State v. Porter*, 698 A.2d 739, 757 (Conn. 1997)). Instead, the trial court's duty is far more modest—to determine only "whether sufficient indicia of legitimacy exist to support the conclusion that evidence derived from the principle may be profitably considered by a fact finder at trial." *Id.* As Justice Hotten reminds us, this Court emphasized that "[w]e are confident that trial judges are duly capable of undertaking the reliability analysis absent scientific training." Dissenting Op. of Justice Hotten 4-5 (quoting *Rochkind*, 471 Md. at 34). We do not ask judges to be "amateur scientists." *Id.* at 4 (quoting *Rochkind*, 471 Md. at 33).

Third, we apply an abuse of discretion standard of review to the trial court's admission or exclusion of expert testimony. *Rochkind*, 471 Md. at 37. This deferential posture is inextricably linked to our expectation that trial judges need only serve as

[1] *Daubert v. Merrell Dow Pharm., Inc.*, 509 U.S. 579 (1993).

gatekeepers: that is, we expect trial judges to make reasonable decisions based on the evidence presented, not to become experts in their own right. If a trial court's decision is supported by competent evidence and untainted by a mistake of law, we defer to its determination even if we would have reached a different conclusion. *See State v. Matthews*, 479 Md. 278, 305 (2022) (quoting *Devincentz v. State*, 460 Md. 518, 550 (2018) (we do "not reverse simply because . . . [we] would not have made the same ruling")); *Id.* at 306 ("[I]t is still the rare case in which a Maryland trial court's exercise of discretion to admit or deny expert testimony will be overturned.").

This Court has articulated the abuse of discretion standard in several ways. We have held that an abuse of discretion occurs "where no reasonable person would take the view adopted by the circuit court," *Williams v. State*, 457 Md. 551, 563 (2018); "when a trial judge exercises discretion in an arbitrary or capricious manner or when he or she acts beyond the letter or reason of the law[,]" *Jenkins v. State*, 375 Md. 284, 295-96 (2003); and when "the trial court's decision [is] 'well removed from any center mark imagined by the reviewing court and beyond the fringe of what that court deems minimally acceptable[,]'" *Devincentz*, 460 Md. at 550 (quotation omitted). We reiterated these standards in *Matthews*. 479 Md. at 305-06.

Although the Majority acknowledges the abuse of discretion standard, it suggests that its application here is unfair "in the absence of additional caselaw from this Court implementing the newly adopted standard[.]" Maj. Op. 6 n.5. The Majority thus sidesteps the deferential standard of review by recasting its decision as establishing the "outer bounds of what is acceptable expert evidence in this area." *See* Maj. Op. 7 n.5.

This misses the mark. First, the Majority does not in practice establish *any* boundaries for the admission of forensic firearms evidence. Second, though this Court has not yet evaluated this type of evidence under *Daubert*, the Majority's disagreement with the trial court does not arise from a lack of judicial guidance. To the contrary, there is no shortage of federal cases applying the *Daubert* factors to determine the reliability of the Association of Firearm and Toolmark Examiners ("AFTE") Theory of Identification (the "AFTE Theory"),[2] *see* Maj. Op. 14 n.8, some of which expressly considered the 2016 report issued by the President's Council of Advisors on Science and Technology on the scientific validity of forensic techniques, including firearms identification.[3] At bottom, the Majority simply disagrees with the trial court's application of the *Daubert* factors and its interpretation of the evidence—a classic de novo review.

That the Majority thinks the abuse of standard is unfair in this context does not justify setting it aside and applying what is, in practice if not in name, a de novo standard

[2] *See, e.g.*, *United States v. Brown*, 973 F.3d 667, 702-04 (7th Cir. 2020) (affirming admission of expert testimony that cartridge cases found in different locations matched and acknowledging PCAST findings); *United States v. Johnson*, 875 F.3d 1265, 1280-81 (9th Cir. 2017) (affirming admission of expert testimony that matched a bullet recovered from crime scene to defendant's pistol and acknowledging 2009 NAS report's criticisms of the AFTE Theory).

[3] *See* EXECUTIVE OFFICE OF THE PRESIDENT, 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 (Sept. 2016) ("PCAST" or the "PCAST Report"), available at https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf (last accessed June 12, 2023) archived at https://perma.cc/3QWJ-2DGR.

of review. If the abuse of discretion standard is not appropriate here, then we should reconsider whether that standard is appropriate for reviewing *Daubert* decisions. But we do not serve well the parties and trial judges who apply our decisions if we inconsistently apply the standards of review to a trial court's discretionary ruling.

Usually, when we hold that a trial court abuses its discretion, we identify what it did wrong and explain how to do it properly going forward. *See, e.g.*, *State v. Robertson*, 463 Md. 342, 365 (2019) (explaining the trial court's error underlying abuse of discretion holding and correcting the mistake for future cases); *State v. Heath*, 464 Md. 445, 462-65 (2019) (holding that the trial court abused its discretion, and explaining how the abuse occurred and how it could be avoided in future cases). Not so today. Though the Majority cabins its analysis to the record here and acknowledges that trial courts may consider other studies in future cases, Maj. Op. 9-10 & 10 n.6, the Majority fails to instruct trial courts *how* to determine the levels at which the accuracy, repeatability, reproducibility, and inconclusive determination rates of firearm identification would be sufficiently reliable for the evidence to be "profitably considered by a fact finder at trial," *Rochkind*, 471 Md. at 34. From the Majority's opinion today, trial courts can only glean that these metrics, based on the studies discussed by the Majority, fail to establish reliability. The Majority's opinion leaves trial courts rudderless at sea in evaluating this type of evidence henceforth.

As discussed below, the focus of our inquiry should not be the reliability of the AFTE Theory in general, but rather the reliability of conclusive determinations produced when the AFTE Theory is applied. Of course, an examiner applying the AFTE Theory might be unable to declare a match ("identification") or a non-match ("elimination"),

resulting in an inconclusive determination. But that's not our concern. Rather, our concern is this: when the examiner *does* declare an identification or elimination, we want to know how reliable *that* determination is. The record shows that conclusive determinations of either kind (identification or elimination) are highly reliable. So, given the record before it, the trial court here made a ruling well within the bounds of its considerable discretion.

I join Justice Hotten's dissent but write separately to explain how the evidence at the center of the Majority's analysis was sufficient to support the trial court's admission of Scott McVeigh's unqualified opinion that bullets recovered from the murder scene were fired from Mr. Abruquah's Taurus revolver.[4] In so doing, I assume familiarity with the defined terms and discussion of the various studies (Ames I and II,[5] in particular) in the Majority's opinion.

I.

PCAST

Before delving into the results of Ames I and Ames II, the two studies that garnered the lion's share of the Majority's attention, we should recognize that the trial court was

[4] I join Justice Hotten's dissent on the issue of whether the trial court abused its discretion in admitting Mr. McVeigh's unqualified opinion, but not as to harmless error. In that regard, I agree with the Majority that, to the extent that Mr. McVeigh's opinion was inadmissible, the error would not be harmless.

[5] *See generally* David P. Baldwin, et al., *A Study of False-Positive and False Negative Error Rates in Cartridge Case Comparisons*, U.S. DEP'T OF ENERGY (2014) ("Ames I"); Stanley J. Bajic, et al., *Report: Validation Study of the Accuracy, Repeatability, and Reproducibility of Firearm Comparisons*, U.S. DEP'T OF ENERGY 1-2 (2020) ("Ames II").

presented with at least three other studies also supporting the conclusion that the AFTE Theory could reliably link bullets to specific guns.[6]

The Majority, however, finds limited value in all studies but Ames I and II. In discounting these other studies, the Majority relies heavily on criticisms made by the PCAST Report.[7] PCAST concluded that the foundational validity, and thus reliability, of subjective forensic feature-comparison methods such as the AFTE Theory "can *only* be established through multiple independent black box studies[.]" PCAST Report at 106. At that time, according to PCAST, the only appropriately designed black box study of firearms examination was Ames I.[8] *Id.* at 111. PCAST concluded that, though Ames I supported

[6] Tasha P. Smith, et al., *A Validation Study of Bullet and Cartridge Case Comparisons Using Samples Representative of Actual Casework*, 61 J. FORENSIC SCIS. 939 (May 2016) ("Validation Study"); James E. Hamby, et al., *A Worldwide Study of Bullets Fired From 10 Consecutively Rifled 9MM RUGER Pistol Barrels – Analysis of Examiner Error Rate*, 64 J. FORENSIC SCIS. 551 (Mar. 2019) ("Worldwide Study"); Jamie A. Smith, *Beretta barrel fired bullet Validation Study*, 66 J. FORENSIC SCIS. 547 (Oct. 2020) ("Bullet Validation Study").

[7] These "set-based" studies can be described variously as "within-set," "closed-set," or "set-to-set comparison" studies. The distinguishing characteristic of such studies is that determinations are not independent of each other; an examiner's determination for a sample changes the likelihood of a correct response for a subsequent sample. For this reason, PCAST discounts these studies. PCAST Report at 106-109.

[8] Black box studies "measure the accuracy outcomes absent information on how they are reached." Lucas Zarwell and Gregory Dutton, *The History and Legacy of the Latent Fingerprint Black Box Study*, NAT'L INST. OF JUST. 1 (Dec. 2022), available at https://nij.ojp.gov/topics/articles/history-and-legacy-latent-fingerprint-black-box-study (last accessed June 12, 2023) archived at https://perma.cc/MMS5-3S4P. Accordingly, black box studies are often used to assess the accuracy of subjective methods. Here, a black box study can measure the accuracy of the AFTE Theory without investigating *how* examiners arrive to conclusions, instead measuring only *whether* the method produces accurate outcomes.

the reliability of the AFTE Theory, the available evidence at the time "[fell] short of the scientific criteria for foundational validity." *Id.* at 111. According to PCAST, more studies were needed. *Id.*

"Foundational validity," as defined by PCAST, however, is not the legal standard by which we evaluate expert testimony under *Daubert*. PCAST itself acknowledges this distinction. PCAST Report at 4 ("Judges' decisions about the admissibility of scientific evidence rest solely on *legal* standards; they are exclusively the province of the courts and PCAST does not opine on them."). Moreover, PCAST apparently created the term "foundational validity" as used in this context; the term began to appear in court opinions only after PCAST was published.[9] And the requirements for foundational validity were developed by PCAST.

The trial judge was not required to credit PCAST's notion of foundational validity at all, let alone apply it strictly. What's more, the trial judge was presented with evidence expressly challenging positions asserted by PCAST. Specifically, the record included a statement by the United States Department of Justice ("DOJ") sharply disagreeing with PCAST in important respects.[10] United States Department of Justice Statement on the PCAST Report: *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* (Jan. 13, 2021), available at

[9] A search of Westlaw for "Daubert" and "foundational validity" returns no cases from before October 2016.

[10] The record also included a response to the DOJ Statement from the Innocence Project.

https://www.justice.gov/olp/page/file/1352496/download ("DOJ Statement") (last accessed June 12, 2023).

Among other things, the DOJ Statement forcefully disagreed with PCAST's conclusion that only "appropriately designed" black box studies could be used to validate a scientific method. *Id.* at 10-12. Although the DOJ did not object to the individual criteria that PCAST deemed necessary for such a study, the DOJ disagreed with the rigidity of PCAST's approach to establishing scientific validity. *Id.* at 11. The DOJ observed that "PCAST failed to cite a single authority that supports its sweeping claim that the collective and *non-severable* application of *all* of these experimental design requirements in multiple black box studies is the *sine qua non* for establishing the scientific validity of forensic 'feature comparison' methods." *Id.* The DOJ also observed that the authorities relied upon by PCAST instead supported the proposition "that no single experimental design is either essential or required." *Id.*; *see also* 1 DAVID L. FAIGMAN ET AL., MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY 66 (2018-2019 ed. 2018) ("There is no single way to conduct research to answer a particular question and research programs rarely answer factual questions definitively. Thus, there are no 'perfect' studies or 'final' answers in science.").

In treating PCAST as near gospel, the Majority applies non-legal and overly demanding requirements to what should be, at its core, a "screening" exercise. *See Daubert*, 509 U.S. at 589, 596 (describing a "screening role" for the trial judge). In doing so, the Majority discounts to the point of irrelevance a substantial body of useful, if

imperfect or incomplete, information from which the trial court reasonably concluded that the method used by Mr. McVeigh was reliable.

II.

A.

A Hypothetical

To set the stage for showing that the Ames I and Ames II studies can be understood to support the trial court's ruling, consider the following thought experiment. Suppose you know nothing about firearms. You know nothing about the basic principles of forensic pattern comparison methods in general or the AFTE Theory in particular. You have never touched a gun or bullet, let alone examined one. You have never been in a crime lab. Now, you are tested on your ability to determine whether a particular bullet was fired from a particular gun.

The test administrator fires two bullets from each of 10 consecutively manufactured handguns. The administrator then gives you two sets of 10 bullets each. One set consists of 10 "unknown" bullets—where the source of the bullet is unknown to the examiner—and the other set consists of 10 "known" bullets—where the source of the bullet is known. You are given unfettered access to a sophisticated crime lab, with the tools, supplies, and equipment necessary to conduct a forensic examination. And, like the vocabulary tests from grade school requiring you to match words with pictures, you must match each of the 10 unknown bullets to the 10 known bullets.

Even though you know that each of the unknowns can be matched with exactly one of the knowns, you probably wouldn't know where to begin. If you had to resort to

guessing, your odds of correctly matching the 10 unknown bullets to the 10 knowns would be one out of 3,628,800.[11] Even if you correctly matched five unknown bullets to five known bullets and guessed on the remaining five unknowns, your odds of matching the remaining unknowns correctly would be one out of 120.[12] Not very promising.

The closed-set and semi-closed-set studies before the trial court—the studies which PCAST discounted—show that if you were to properly apply the AFTE Theory, you would be very likely to match correctly each of the 10 unknowns to the corresponding knowns. *See* Validation Study; Worldwide Study; Bullet Validation Study.

Your odds would thus improve from virtually zero (one in 3,628,800) to 100 percent. Yet according to PCAST, those studies provide *no* support for the scientific validity of the AFTE Theory. PCAST reasons that, in set-based studies, examiners can rely on the process of elimination, aided by deductive reasoning. Thus, by affording examiners a decisional crutch, PCAST reasons, such studies likely underestimate error rates in actual casework.

Now let's assume you take a different type of test, one designed in the image of Ames I and Ames II. This time, the administrator gives you 30 sets of bullets, with three bullets in each set. Within each set of three bullets, two are identified as having been fired

---

[11] With 10 unknown bullets and 10 known bullets, the odds of guessing the first pair correctly are one out of 10. And if you get the first right, the odds of getting the second right are one out of nine. If you get the first two right, the odds of getting the third right are one out of eight, and so on. Thus, the odds of matching each unknown bullet to the correct known is represented by the following calculation: (1/10) x (1/9) x (1/8) x (1/7) x (1/6) x (1/5) x (1/4) x (1/3) x (1/2) x (1/1).

[12] (1/5) x (1/4) x (1/3) x (1/2) x (1/1).

from the same gun. Your task is to determine whether the third bullet, the unknown bullet, was also fired from the same gun. The administrator, of course, knows the correct answer (the "ground truth"). In contrast to the set-based studies, however, you know nothing about the source(s) of bullets and the relationship between knowns and unknowns. Thus, your answers for each set are independent of each other.

Assume again that you know nothing about guns or the AFTE Theory. You might as well guess or flip a coin to determine if there is a match between the unknown and two knowns, which means that you can expect to answer, on average, 15 out of 30 sets correctly.

But now assume that you are properly trained in the AFTE Theory. You examine each of the 30 sets. Suppose you determine that 10 sets lack sufficient information to make a conclusive determination of identification or elimination, so you mark 10 sets as inconclusive,[13] but you reach *conclusive* determinations for the remaining 20 sets. The results of Ames I and II indicate a high likelihood that all 20 of those determinations would be correct.

B.

The Treatment of Inconclusive Determinations

But let's suppose you made one error—a false positive identification—and correctly determined the remaining 19 sets. The question then becomes how your error rate should

[13] "[A] finding of inconclusive is an appropriate answer" if "the examiner does not find sufficient matching detail to uniquely identify a common source for the known and questioned samples, and there are no class characteristics such as caliber that would preclude the cases as having been fired from the same-source firearm[.]" Ames I at 6.

be calculated, which turns on how your 10 inconclusive determinations are treated. This issue was heavily debated in the trial court and looms large in the Majority's analysis.

The parties have focused on two ways to account for inconclusive determinations. The State argues that an inconclusive should be counted neither as a correct response, because the examiner failed to obtain the ground truth, nor as an error, because the examiner did not make an incorrect conclusive determination. Accordingly, the State advocates calculating error rates according to the method used in Ames I and Ames II: to include inconclusive determinations ("inconclusives") in the denominator[14] but exclude them from the numerator. Applying this method to the example above, the error rate would be 1/30, or 3.33 percent.

Mr. Abruquah's expert, Professor David Faigman, did not mince words about this method, declaring that "in the annals of scientific research or of proficiency testing, it would be difficult to find a more risible manner of measuring error." To Mr. Faigman, the issue was simple: in Ames I and II, the ground truth was known, thus "there are really only two answers to the test, like a true or false exam[ple]." Mr. Faigman explained that "the common sense of it is if you know the answer is either A or B and the person says I don't know, in any testing that I've ever seen that's a wrong answer." He argued, therefore, that inconclusives should be counted as errors. In the above example, under that approach, the error rate would be 11 out of 30, or 36.7 percent.

[14] Under this method, the denominator thus represents the total number of responses.

The Majority doesn't expressly choose between the competing views, but its analysis favors Mr. Faigman's approach. As the Majority sees it, an inconclusive should be deemed an error if there was sufficient information to make a conclusive determination.[15] Maj. Op. 44 ("[I]f at least some inconclusives should be treated as incorrect responses, then the rates of error in open-set studies performed to date are unreliable.").

The Majority is skeptical that the inconclusive rates in the studies mirror the rates of inconclusive determinations in real field work. Maj. Op. 43. The Majority points to the disparity observed in the inconclusive rates between closed-set and open-set studies. Maj. Op. 42-43. This concern echoes that of Mr. Faigman, who testified that the disparity suggests that "something crazy is going on here" and that inconclusives should thus be deemed errors unless "you can demonstrate that they are somehow right."

But there is no mystery here. The examiner in closed-set studies knows that the unknown bullets match one of the known bullets. Thus, the examiner trained in the AFTE Theory can use the process of elimination, aided by deductive reasoning, to make a conclusive determination from what would otherwise be an inconclusive finding. In other words, by its nature, the closed-set design reduces the rate of inconclusives. In contrast, open-set studies do not permit the examiner to use the process of elimination to make a conclusive determination, resulting in higher rates of inconclusives. The Majority's

[15] In making this argument, the Majority implicitly acknowledges the validity of the foundational assumption of the AFTE Theory—that, at least sometimes, bullets and cartridges display markings sufficient to match them to a specific source gun.

concern about the disparity in inconclusive rates between closed-set and open-set tests is thus explained away by this fundamental difference in test design.

The Majority, however, infers from the disparity and what it considers unimpressive repeatability and reproducibility results that "whether an examiner chooses 'inconclusive' in a study seems to depend on something other than just the 'corresponding individual characteristics' themselves." Maj. Op. 43. The Majority implies that, because the ground truth is known in a test environment, the examiner, who makes a living performing these examinations, changes his behavior, consciously or otherwise, to minimize error rates by over-reporting inconclusives. *See id.* (concluding that the rates of inconclusives reported in Ames II "suggest[] that examiners choose 'inconclusive' even when it is not a 'correct' response"). This view mirrors Mr. Faigman's testimony that, in the face of any ambiguity, the examiners who participated in the studies "default[ed] to inconclusive because [they] know that [they're] in the business that a false positive has the worst optics associated with it."

Based on the premise that at least some inconclusives should be treated as errors, the Majority declares that the resulting "rates of error in open-set studies performed to date are unreliable." Maj. Op. 43. As an example, the Majority observes that if Inconclusive-A responses for non-matching bullets in Ames II were counted as false positives, then the false positive rate "would balloon from 0.7% to 10.13%." *Id.* An "Inconclusive-A" determination, you might recall, is the inconclusive level closest to a positive identification. So, the Majority's reasoning goes, if we know the ground truth is a non-match and we treat

the "almost a match" determination as a match (a false positive), the false positive rate increases substantially.

The logic behind that view escapes me. If an examiner makes an Inconclusive-A determination, that means the examiner affirmatively chose *not* to make a positive identification. The examiner's determination of Inconclusive-A does not necessarily mean that he *almost* made a false positive, as the Majority's exercise presumes. Rather, Ames II instructed examiners to determine a result to be Inconclusive-A when they observed "[s]ome agreement of individual characteristics and all discernible class characteristics, but insufficient for an identification." Ames II Report at 100. What's more, the record lacks any evidence that trained and ethical examiners, which undisputedly describes Mr. McVeigh, are less concerned about making a false positive in actual field work than in a study setting.

The Majority supposes that "in all the studies conducted to date," examiners deemed some samples inconclusive that they would have instead deemed matches in actual casework, reasoning that "the participating examiners knew that (1) they were being studied and (2) an inconclusive response would not be counted as incorrect." Maj. Op. 43. The record, however, does not establish either assertion.

First, the examiners themselves were not the primary object of these studies. Rather, the AFTE Theory was. In Ames II, the examiners were told that the study would "assess[] the accuracy, repeatability, and reproducibility of decisions involving forensic comparisons," that "[r]eported results and findings [would] be completely anonymized," and that "[i]ndividual results [would] not be disclosed to the subjects or their employers."

Thus, in contrast to proficiency testing, which evaluates the performance of individual examiners, examiners took no personal reputational or professional risk by participating and providing responses that faithfully reflected their determinations. So even though examiners knew they were participating in a study, they did not think "they" were the object of study.[16]

Second, the written instructions provided to examiners in Ames II did not indicate how inconclusives would be accounted for in the study results. As a result, there's no basis to conclude, as the Majority does, that the examiners knew how the authors of Ames II would account for inconclusives in the final analysis.

C.

How to Calculate Error Rates

This brings us to a different way of looking at error rates, one that received no consideration by the Majority but should mitigate much of its concerns about the reliability of the error rates in Ames I and Ames II. I am referring to calculating error by excluding inconclusives from *both* the numerator and the denominator. This measure calculates the rate of error of false positives and/or false negatives against only conclusive determinations. Under this measure, in our example above, the error rate would be one out

[16] Relatedly, the examiners would not likely have been concerned about the viability of their profession when they participated in Ames II. Even if the results of the study did not support the admissibility of unqualified conclusive opinions at trial, there is no basis to believe that participating examiners would have perceived a threat to their livelihood of assisting criminal investigations, particularly in light of the numerous studies supporting, at the very least, the viability of the AFTE Theory as an investigative tool.

of 20, or five percent. Thus, by excluding inconclusive determinations altogether, the error rate in our example *increases* from 3.33 percent to five percent.

Before explaining the merits of this calculation, I point out that Mr. Faigman, as he so often did when disagreeing with an opposing view, chastised this way of calculating error rates, saying, "that's completely crazy from any testing regime that I have ever heard of and from just common sense." But, contrary to Mr. Faigman's unsupported criticism, excluding inconclusives from the numerator and denominator accords with both common sense and accepted statistical methodologies. It is also supported by competent evidence in the record.

As a matter of common sense, the measure used to calculate error rates should align with the specific purpose of our inquiry. The Majority notes that "Mr. Abruquah does not challenge all of Mr. McVeigh's testimony or that firearms identification is sufficiently reliable to be admitted for some purposes." Maj. Op. 38. The Majority correctly defines the issue in narrow terms: our "task is to assess, based on the information presented to the circuit court, whether the AFTE Theory can reliably support an unqualified opinion that a particular firearm is the source of one or more particular bullets." *Id.* at 38. Put another way, here, we are not concerned with the likelihood that application of the AFTE Theory will, in actual field work, conclusively tell us whether or not a specific bullet was fired from a specific gun. Rather, we are concerned with the likelihood that *when* application of the AFTE Theory yields a conclusive determination—here, a positive identification—the

result is correct.[17] PCAST framed the issue similarly, albeit in statistical language, explaining that "[f]orensic feature-comparison methods typically aim to determine how likely it is that two samples came from the same source," PCAST Report at 151, and that false positives are "especially important because [they] can lead directly to wrongful convictions," *Id.* at 50 (footnote omitted).

PCAST identified two accepted measures of accuracy: sensitivity and false positive rates. *Id.* at 50. PCAST defined sensitivity as "the probability that the method declares a proposed identification between samples that actually come from the *same* source" and the false positive rate as the "probability that the method declares a proposed identification between samples that actually come from *different* sources." *Id.*

Of critical importance here, PCAST explained that the sensitivity and false positive rates can be calculated "based on the *conclusive* examinations or on *all* examinations." *Id.* at 153. PCAST even went a step further and contended that even though both measures "are of interest," false positive rates should be based only on conclusive examinations "because evidence used against a defendant will typically be based on *conclusive*, rather

[17] That's not to say that false eliminations should not concern us. A false elimination could also lead to the conviction of an innocent person. *See* PCAST Report at 44 n.94. That could happen if, for example, an individual commits a homicide with a firearm and is not charged due to a false elimination, and instead another person is wrongly convicted. The implications of this observation, in my view, support the admissibility of conclusive determinations of both kinds, i.e., identifications and eliminations. As explained below, the rates of false eliminations are low, and repeatability and reproducibility data show that when a false elimination is made, the error is typically corrected on a second examination. So, in this hypothetical, the innocent defendant could hire his own firearms examiner, who may determine that the other suspect's gun was the murder weapon. Under the Majority's ruling, that examiner would not be allowed to testify that the other suspect's gun was the murder weapon.

than inconclusive, determinations." *Id.* ("The quantity of most interest in a criminal trial is . . . the probability that the samples are from the same source given that a match has been declared.") (cleaned up).[18]

So, far from being "crazy," as Mr. Faigman argued, excluding inconclusives from error rate calculations when assessing the reliability of a positive identification is not only an acceptable approach, but the preferred one, at least according to PCAST. Moreover, from a mathematical standpoint, excluding inconclusives from the denominator actually *penalizes* the examiner because errors accounted for in the numerator are measured against a smaller denominator, i.e., a smaller sample size. That's why the error rate in the above example *increased* from 3.33 percent to five percent.

This brings us back to Mr. Faigman's and the Majority's speculation that the examiners in Ames II were biased toward inconclusives. Ames II was conducted in response to PCAST, Ames II at 12, and the firearms examiner community was acutely aware of PCAST when it was published. Thus, although Mr. Faigman and the Majority assume that the participating examiners "knew" that inconclusives would not be counted against them and consequently over-relied on them, it is just as likely that examiners assumed that inconclusives would be accounted for in the manner advocated by PCAST,

[18] The authors of Ames II write that "[a]lthough some might propose an inconclusive decision as an unsuccessful outcome, or failure ('error') to identify a [known match], such a decision rightly represents judgment that the comparison presents insufficient information (quality and/or quantity of individual characteristics) for a definitive statement (minimization of false positive being paramount[).]" Keith L. Monson, et al., *Planning, design and logistics of a decision analysis study: The FBI/Ames study involving forensic firearms examiners*, 4 FORENSIC SCI. INT'L.: SYNERGY 1, 5 (Feb. 19, 2022) (footnotes omitted). The authors of Ames I agree. Ames I at 6.

with every inconclusive driving up the error rate. Perhaps that's why Mr. McVeigh rejected the premise that examiners were not penalized for making an inconclusive determination. Because Mr. Abruquah and the Majority rely heavily on PCAST, we should at least consider how PCAST's preferred measurement of error rate affects the results of Ames I and Ames II. I take up that task next.

III.

ACCURACY

Before turning to the specific error rates reported in Ames I and Ames II, let's first address where to draw the line between an acceptable and unacceptable error rate to establish a minimum threshold of reliability. I turn again to PCAST, which posits that "[t]o be considered reliable, the [false positive rate] should certainly be less than 5 percent and it may be appropriate that it be considerably lower, depending on the intended application." PCAST Report at 152.

PCAST is not definitive on any topic, let alone the maximum false positive rate for a reliability determination. But given the Majority's reliance on PCAST, PCAST's standard provides a helpful benchmark when assessing whether the trial court appropriately exercised its discretion. At bottom, however, trial courts should be left to their own discretion to make such a judgment call.

A.

Ames I

Let's start with Ames I. With respect to matching bullet sets, the number of inconclusives was so low that whether inconclusives are included in the denominator

makes little difference to error rates. Of the 1,090 matching sets, only 11, or 1.01 percent, were inconclusives. Of the conclusive determinations, 1,075 were correctly identified as a match ("identifications") and four were incorrectly eliminated ("eliminations"). The four false eliminations were committed by three examiners; 215 of the 218 examiners did not report *any* false eliminations. Measured against the total number of matching sets (1,090), the false elimination rate was 0.36 percent. Against only the conclusive determinations (1,079), the false elimination rate was 0.37 percent.

The error rates for non-matching bullets vary more significantly if inconclusive determinations are excluded from the denominator. Of 2,178 non-matching sets, examiners reported 735 inconclusives for an inconclusive rate of 33.7 percent, 1,421 sets as correct eliminations, and 22 sets as incorrect identifications (false positives). The false positives were concentrated among a few examiners: 20 of the 22 false positives were made by the same five examiners. As a percentage of the total 2,178 non-matching sets, the false positive rate was 1.01 percent. As a percentage of the 1,443 conclusive determinations, however, the false positive rate was 1.52 percent. Either way, the results show that the risk of a false positive is very low, particularly when measured against the five percent benchmark recommended by PCAST.

Combining the results of the matching and non-matching sets is also instructive. Of the total number of sets (3,268), 746 were inconclusives, for an inconclusive rate of 22.83 percent, and 26 were either erroneous eliminations or identifications. Measured against the total number of sets, the overall error rate was 0.79 percent. Measured against only conclusive determinations, the error rate was 1.03 percent.

In sum, the results of Ames I show that, with inconclusives either included or excluded in the denominator in the error calculation, identifications and eliminations boast extremely low error rates.

B.

Ames II

Although PCAST found Ames I to be an appropriately designed black box study, PCAST concluded that one such study was not enough to establish the scientific validity of the AFTE Theory. PCAST Report at 113. Eric Lander, a co-chair of PCAST and President of the Broad Institute of MIT and Harvard when the PCAST Report was published, wrote: "With only a single well-designed study estimating accuracy, PCAST judged that firearms analysis fell just short of the criteria for scientific validity, which requires reproducibility. A second study would solve this problem." Eric S. Lander, *Fixing Rule 702: The PCAST Report and Steps to Ensure the Reliability of Forensic Feature-Comparison Methods in the Criminal Courts,* 86 FORDHAM L. REV. 1661, 1672 (2018). Ames II was that second study.

Matching Bullet Sets

Because Mr. McVeigh's testimony linked bullets, not cartridges, to Mr. Abruquah's gun, I will focus on the results of the bullet examinations in Ames II. There were 1,405 recorded results for matching sets of bullets. Of those, 288 were placed in any one of the three inconclusive categories, Inconclusive-A, Inconclusive-B, and Inconclusive-C, for an inconclusive rate of 20.50 percent. Of the 1,117 conclusive determinations, 1,076 were correct identifications. Measured against the total number of recorded results (1,405), the

identification rate (sensitivity) was 76.6 percent, which the Majority perceives as low. But, when measured against the total conclusive determinations (1,117), the identification rate jumps to 96.3 percent, indicating far greater reliability of identifications.

There were 41 false eliminations. As a percentage of the 1,405 recorded results, the false elimination rate was 2.9 percent. As a percentage of only the conclusive results, the false elimination rate increased to 3.7 percent—still below PCAST's recommended five percent threshold.

Non-Matching Bullet Sets

There were 2,842 recorded results for non-matching sets, 1,861 were inconclusives, for an inconclusive rate of 65.48 percent, and 961 were correct eliminations. Measured against the total number of recorded results (2,842), the correct elimination rate was only 33.8 percent. But measured against only the total number of conclusive determinations (981), the correct elimination rate jumps to 97.9 percent—another indication that conclusive determinations under the AFTE Theory are reliable.

Of course, we are most concerned about the risk of false positives—that is, incorrect identifications. There were 20 false positives. Measured against the total number of recorded results (2,842), the false positive rate was 0.7 percent. Measured against only the *conclusive* determinations, however, the false positive rate increases to 2.04 percent. Under either measure, the false positive rate was well below PCAST's recommended threshold of five percent.

In sum, using PCAST's preferred method of calculating error rates and its five percent threshold for an acceptable error rate, the error rates observed in Ames II show that the trial court's determination of reliability was reasonable.[19]

IV.

REPEATABILITY AND REPRODUCIBILITY

The Majority focuses on what it perceives to be unimpressive showings of repeatability and reproducibility in Ames II. Maj. Op. 43, 48-49. To the Majority, the "inconsistent" results in these respects are "troublesome" and undermines the reliability of the AFTE Theory. Maj. Op. 45, 48-49.

Before proceeding, I offer an observation about repeatability and reproducibility: consistent results from separate examinations of the same sample, by either the same or a different examiner, are not necessarily desirable. Certainly, consistency is good if the initial determination is correct. But consistency is undesirable if the initial determination is wrong, in which case we would prefer disagreement. That is, we would prefer that the same examiner or another examiner get it right the second time rather than repeat the mistake. Disagreement with an incorrect determination would increase our confidence that a peer review process would catch and correct mistakes, particularly false positives, and that the traditional tools for challenging "shaky" evidence—cross-examination, opposing

[19] I recognize that PCAST acknowledged that an appropriate error rate threshold could be lower than five percent, depending on the purpose for which the evidence would be used. PCAST Report at 152. But how much lower or higher should be a matter for the trial judge to determine.

expert testimony, and presentation of contrary evidence—would expose errors. *Matthews*, 479 Md. at 312 (quoting *Daubert*, 509 U.S. at 596).

So, as to repeatability and reproducibility rates: (1) the higher the better for initial correct identifications and correct eliminations and (2) the lower the better for initial false eliminations and false positive identifications. And, because our primary concern is the reliability of an identification, we are less concerned whether the initial and subsequent examination, by the same or different examiner, yielded the same particular level of inconclusives. Thus, the repeatability and reproducibility figures relied upon by the Majority, which include all categories (identification, elimination, and three levels of inconclusive), do not align well with the specific nature of our inquiry.

A.

Repeatability

Repeatability is the likelihood that the same examiner will make the same determination for a particular sample on a subsequent examination. Ames II refers to an examiner's initial examination as "Round One" of the study and that examiner's subsequent examination as "Round Two."

Matching Bullet Sets

As noted by the Majority, the overall repeatability rate was 79.0 percent for matching bullets and 64.7 percent for non-matching bullets. The Majority is not impressed by these results but doesn't tell us what levels would, in its view, support reliability. In my view, reasonable minds can differ. As for the matching sets, given the wide range of responses, identification, elimination, Inconclusive-A, Inconclusive-B, and Inconclusive-

C, one might reasonably be impressed that, on independent examinations of the same sets months apart, examiners reached the same result nearly 80 percent of the time.

But there is more to glean from the results. The following table reproduces the data from Table IX of the Ames II report, with percentages, to show how the Round One results for the matching bullet sets were distributed in Round Two. Ames II Report at 38.

Table IX: Matching Sets (Bullets)[20]							
Classification on First Evaluation	**Classification on Second Evaluation**						
	ID	**Inc. A**	**Inc. B**	**Inc. C**	**Elimination**	**Unsuitable[21]**	**Total**
ID	665	27	26	14	8	2	**742**
	89.62%	3.64%	3.50%	1.89%	1.08%	0.27%	
Inc. A	31	28	12	6	2	0	**79**
	39.24%	35.44%	15.19%	7.59%	2.53%	0.00%	
Inc. B	13	14	45	5	2	2	**81**
	16.05%	17.28%	55.56%	6.17%	2.47%	2.47%	
Inc. C	2	3	3	5	3	0	**16**
	12.50%	18.75%	18.75%	31.25%	18.75%	0.00%	
Elimination	8	7	3	2	13	0	**33**
	24.24%	21.21%	9.09%	6.06%	39.39%	0.00%	
Unsuitable	1	3	3	0	0	2	**9**
	11.11%	33.33%	33.33%	0.00%	0.00%	22.22%	
Total	**720**	**82**	**92**	**32**	**28**	**6**	**960**

This table shows that the repeatability rate of a correct identification, which is a focus of our inquiry, was 89.62 percent (665/742). Given the subjective nature of the AFTE

[20] For ease of reference, the numbers and titles of the tables in this dissent correspond to the corresponding numbers and titles of the tables in the Ames II Report.

[21] A determination of "unsuitable" is appropriate when "a comparison can not be made due to [the] quality of the provided samples." Ames II Report at 11.

Theory, this repeatability rate for correct identifications could reasonably be viewed as an indicator of reliability.

The table also shows what happened to the 77 correct identifications from Round One (742 - 665 = 77) that were not repeated in Round Two. Ames II refers to different determinations for the same sample as a "paired disagreement." Sixty-seven of those 77 (87.0 percent) paired disagreements were placed in an inconclusive category: 27 in Inconclusive-A, 26 in Inconclusive-B, and 14 in Inconclusive-C. So, while the change from a correct identification reduces the repeatability rate, the different determination suggests that examiners exercised caution in making an identification, the determination typically most inculpatory to a defendant. Examiners changed only eight correct identifications from Round One to false eliminations in Round Two, which weighs against consistency but again suggests the examiners' tendency to err on the side of caution.

The table also sheds light on the inconclusives. Of the 960 matching bullet sets examined in Round Two, 176 were in one of the inconclusive levels in Round One, 121 of which were again in an inconclusive level in Round Two. Individual repeatability rates of Inconclusive-A, Inconclusive-B, and Inconclusive-C were 35.44 percent, 55.56 percent, and 31.25 percent, respectively. Those rates are a drag on the overall repeatability rate.

But, if we return to the primary focus of our inquiry—the reliability of a conclusive determination—we should be less concerned with movement within the inconclusive categories. The version of Table IX below presents the same repeatability results as the previous table, but with the inconclusive determinations pooled, that is, we eliminate the distinctions between the three inconclusive categories.

Table IX: Matching Sets (Bullets)					
Classification on First Evaluation	**Classification on Second Evaluation**				
	ID	**Inconclusive (pooled)**	**Elimination**	**Unsuitable**	**Total**
ID	665	67	8	2	**742**
	89.62%	9.03%	1.08%	0.27%	
Inconclusive (pooled)	46	121	7	2	**176**
	26.14%	68.75%	3.98%	1.14%	
Elimination	8	12	13	0	**33**
	24.24%	36.36%	39.39%	0.00%	
Unsuitable	1	6	0	2	**9**
	11.11%	66.67%	0.00%	22.22%	
Total	**720**	**206**	**28**	**6**	**960**

When the inconclusives are pooled, the overall repeatability rate increases from 79.0 percent to 83.4 percent.[22] That is because the repeatability rate for pooled inconclusives, 68.75 percent (121/176), is higher than the individual repeatability rates for the three inconclusive categories.[23]

An examination of what happened to inconclusives that were *not* repeated in the second round supports reliability. Of the 176 inconclusives in Round One, examiners placed 55 (176 - 121 = 55) into a different category in Round Two. Of those 55, 46 were correct identifications. So, though the movement from the inconclusive category reduced the overall repeatability rate, the vast majority (46/55, or 83.6 percent) of that movement resulted in a determination of the ground truth. Conversely, a comparatively low

[22] Overall repeatability here is calculated as: (665 paired agreement identifications + 121 paired agreement inconclusives + 13 paired agreement eliminations + 2 paired agreement unsuitables)/960 = 83.4 percent.

[23] The results when inconclusives are pooled were available to the trial court, as the Ames II Report presented results under various pooling scenarios. Ames II Report *passim*.

proportion (7 out of 55) moved into the elimination column. That strong trend toward accuracy—the movement from inconclusive to ground truth on an examiner's second attempt—supports the reliability of the AFTE Theory.

Finally, let's look at the repeatability of a false elimination. Of the 33 false eliminations from the first round, 13 were likewise eliminations in Round Two, a repeatability rate of 39.39 percent. Though this reduces the overall repeatability rate, we can take solace that examiners did not repeat most of their mistakes, a trend that reflects well on the methodology. Drilling down even further, of the 20 false eliminations which were not repeated, eight became correct identifications in Round Two, which also speaks well of the methodology. And 12 of the false eliminations were judged inconclusive in the second round, another shift in the direction of ground truth.

Non-Matching Bullet Sets

Without pooling the inconclusive results, the overall repeatability rate for non-matching bullets was 64.7 percent. The Majority highlights the disparity between this rate and the repeatability rate of 79.0 percent for non-matching sets. Maj. Op. 49.

The following table reproduces the data from Table IX of the Ames II report, with percentages, to show how the Round One results for the matching bullet sets were distributed in Round Two. Ames II Report at 38.

Table IX: Non-Matching Sets (Bullets)							
Classification on First Evaluation	**Classification on Second Evaluation**						
	ID	**Inc. A**	**Inc. B**	**Inc. C**	**Elimination**	**Unsuitable**	**Total**
ID	2	3	6	2	6	0	**19**
	10.53%	15.79%	31.58%	10.53%	31.58%	0.00%	
Inc. A	0	52	37	42	27	0	**158**
	0.00%	32.91%	23.42%	26.58%	17.09%	0.00%	
Inc. B	5	31	341	98	45	7	**527**
	0.95%	5.88%	64.71%	18.60%	8.54%	1.33%	
Inc. C	1	32	109	284	53	1	**480**
	0.21%	6.67%	22.71%	59.17%	11.04%	0.21%	
Elimination	1	20	35	66	514	4	**640**
	0.16%	3.13%	5.47%	10.31%	80.31%	0.63%	
Unsuitable	0	0	13	6	4	8	**31**
	0.00%	0.00%	41.94%	19.35%	12.90%	25.81%	
Total	**9**	**138**	**541**	**498**	**649**	**20**	**1855**

With a focus on the primary inquiry here—the reliability of conclusive determinations—we can make several observations. The repeatability rate of a correct elimination was 80.31 percent (514/640), significantly higher than the overall repeatability rate of 64.7 percent and a stronger indicator of reliability.

Of the 126 correct eliminations from Round One (640 - 514 = 126) that were *not* repeated in Round Two, 121 of those 126 (96.0 percent) were placed in an inconclusive category. This movement shows a caution in making conclusive eliminations that does not undermine the reliability of a correct identification. Only one set went from a correct elimination to a false positive, showing that the risk of such a flip-flop is low: one out of 126, or 0.79 percent.

Now let's look at repeatability rates for Round One inconclusives. Repeatability rates of Inconclusive-A, Inconclusive-B, and Inconclusive-C were 32.91 percent, 64.71

percent, and 59.17 percent, respectively. Those low repeatability rates drag down the overall repeatability rate. But, again, should we really be concerned with the repeatability of a particular level of inconclusive, given that the heart of the inquiry here is the reliability of a positive identification?

Let's see what happens when the three levels of inconclusive determinations are pooled, again using the data from Table IX of the Ames II report, with percentage calculations added:

Table IX: Non-Matching Sets (Bullets)					
Classification on First Evaluation	**Classification on Second Evaluation**				
	ID	**Inconclusive (pooled)**	**Elimination**	**Unsuitable**	**Total**
ID	2	11	6	0	**19**
	10.53%	57.89%	31.58%	0.00%	
Inconclusive (pooled)	6	1026	125	8	**1165**
	0.52%	88.07%	10.73%	0.69%	
Elimination	1	121	514	4	**640**
	0.16%	18.91%	80.31%	0.63%	
Unsuitable	0	19	4	8	**31**
	0.00%	61.29%	12.90%	25.81%	
Total	**9**	**1177**	**649**	**20**	**1855**

For starters, the repeatability rate for inconclusives for non-matching bullets improves to 88.07 percent (1026/1165). More importantly, by pooling the inconclusives, the overall repeatability rate of all determinations, both inconclusive and conclusive, increases from 64.74 percent to 83.56 percent.[24] Recall that the Majority noted the disparity between the overall repeatability rates of matching bullet sets (79.0 percent) and

[24] Calculated as: (2 paired agreement identifications + 1026 paired agreement inconclusives + 514 paired agreement eliminations + 8 paired agreement unsuitables)/1855 = 83.56 percent.

non-matching bullet sets (64.7 percent). When inconclusive results are pooled, however, the disparity all but disappears—the repeatability rate for matching sets and non-matching sets converge at 83.4 percent and 83.6 percent, respectively. Put differently, the Majority's concern about the disparity between the repeatability rates of matching and non-matching bullets can be entirely explained by changes within the three levels of inconclusive determinations, which do not compromise the reliability of a conclusive determination.

Now let's examine what happened to inconclusives from Round One that were not judged inconclusive in Round Two. Of those 139 sets, 125 were correctly determined to be an elimination. So, although the movement out of inconclusive reduced the repeatability rate, nearly all of that movement (125/139 = 89.9 percent) was to the ground truth. Only six of the 139 sets turned into false positives. These shifts indicate reliability.

Finally, let's look at the repeatability rate of false identifications or false positives. Of the 19 false identifications from Round One, only two remained in that category in Round Two (10.5 percent). Thus, examiners were highly unlikely to repeat the most prejudicial type of mistake. Of the 17 false positives from Round One that were not repeated in Round Two, six were judged correct eliminations and 11 inconclusive.

B.

Reproducibility

Reproducibility is the likelihood that, for a particular sample, a different examiner will make the same determination as the initial examiner. Ames II refers to the second examiner's evaluation as "Round Three" of the study.

As the Majority notes, the overall reproducibility rate was 68.0 percent for matching bullets and 31.0 percent for non-matching bullets. Maj. Op. 49 n.26. The Majority is again unimpressed by these results. Maj. Op. 49. But, again, if we focus on the reliability of *conclusive* determinations, the data tell a different story, one more supportive of the reliability of the AFTE Theory.

Matching Bullet Sets

Let's start with the matching bullet sets. The following table reproduces the data from Table XIV of the Ames II report, with percentages, to show how the Round One results for the matching bullet sets were distributed in Round Three when examined by different examiners. Ames II Report at 46.

Table XIV: Matching Sets (Bullets)							
Classification by First Round Examiner	**Classification by Third Round Examiner**						
	ID	**Inc. A**	**Inc. B**	**Inc. C**	**Elimination**	**Unsuitable**	**Total**
ID	601	38	39	14	12	5	**709**
	84.77%	5.36%	5.50%	1.97%	1.69%	0.71%	
Inc. A	42	18	7	6	6	0	**79**
	53.16%	22.78%	8.86%	7.59%	7.59%	0.00%	
Inc. B	34	15	22	4	6	0	**81**
	41.98%	18.52%	27.16%	4.94%	7.41%	0.00%	
Inc. C	9	7	5	2	6	0	**29**
	31.03%	24.14%	17.24%	6.90%	20.69%	0.00%	
Elimination	13	5	14	6	3	0	**41**
	31.71%	12.20%	34.15%	14.63%	7.32%	0.00%	
Unsuitable	3	2	8	1	1	2	**17**
	17.65%	11.76%	47.06%	5.88%	5.88%	11.76%	
Total	**702**	**85**	**95**	**33**	**34**	**7**	**956**

According to this table, the reproducibility rate of correct identifications—the primary focus of our inquiry—was 84.77 percent (601/709). Given the subjectivity of the

AFTE Theory, that result can be reasonably viewed as an indicator of reliability. At a minimum, it renders far less concerning the 68 percent overall reproducibility rate on which the Majority focuses. Moreover, there were 108 correct identifications from Round One (709 - 601 = 108) that were judged differently in Round Three, 91 (84.3 percent) of which went into an inconclusive category. Meanwhile, only 12 of the 108 became false eliminations. Thus, although the movement from a correct identification reduces the overall reproducibility rate, the difference indicates the examiners exercised caution, even at the expense of making a correct identification. That is another reason we can have confidence in conclusive determinations resulting from application of the AFTE Theory.

Of the 956 bullet sets examined in Round Three, 189 were inconclusive in both Round One and Round Three. Individual reproducibility rates of Inconclusive-A, Inconclusive-B, and Inconclusive-C were 22.78 percent, 27.16 percent, and 6.90 percent, respectively. Those low rates reduced the overall reproducibility rate for matching bullets.

But the following table, also drawn from Table XIV of the Ames II report, shows what happens to the reproducibility results presented above when the three levels of inconclusive are pooled:

Table XIV: Matching Sets (Bullets)					
Classification by First Round Examiner	**Classification by Third Round Examiner**				
	ID	**Inconclusive (pooled)**	**Elimination**	**Unsuitable**	**Total**
ID	601	91	12	5	**709**
	84.8%	12.8%	1.7%	0.7%	
Inconclusive (pooled)	85	86	18	0	**189**
	45.0%	45.5%	9.5%	0.0%	
Elimination	13	25	3	0	**41**
	31.7%	61.0%	7.3%	0.0%	
Unsuitable	3	11	1	2	**17**
	17.6%	64.7%	5.9%	11.8%	
Total	**702**	**213**	**34**	**7**	**956**

When inconclusive results are pooled, the reproducibility rate of inconclusives improves to 45.50 percent (86/189). And more importantly, the reproducibility rate of all determinations, both inconclusive and conclusive, increases from 67.8 percent to 72.4 percent.[25]

Let's examine what happened to the inconclusives from Round One that were not inconclusive in Round Three. Of the 189 inconclusives from Round One, subsequent examiners placed 103 into a different category in Round Three. Of those 103, 85 became a correct identification. So, although that change dragged down the reproducibility rate, most of that movement (85/103 = 82.5 percent) produced the ground truth. Conversely, 18 of the 103 were incorrectly judged eliminations.

This table also shows that subsequent examiners reproduced only three of the 41, or 7.32 percent, of the false eliminations in Round One. Here, the failure to reproduce a result

[25] Calculated as: (601 paired agreement identifications + 86 paired agreement inconclusives + 3 paired agreement eliminations + 2 paired agreement unsuitables)/956 = 72.4 percent.

is welcome; that subsequent examiners were unlikely to reproduce the mistake of the first examiner should be viewed favorably. Moreover, it shows that most of the time, a false elimination is discernible, which means that a rigorous peer review process and the traditional tools for challenging expert testimony, cross-examination and opposing experts, are likely to be effective.

Non-Matching Bullet Sets

The following table reproduces the data from Table XIV of the Ames II report, with percentages, to show how the Round One results for the non-matching bullet sets were distributed by different examiners in Round Three. Ames II Report at 46.

Table XIV: Non-Matching Sets (Bullets)							
Classification by First Round Examiner	**Classification by Third Round Examiner**						
	ID	**Inc. A**	**Inc. B**	**Inc. C**	**Elimination**	**Unsuitable**	**Total**
ID	0	5	8	5	1	0	**19**
	0.00%	26.32%	42.11%	26.32%	5.26%	0.00%	
Inc. A	1	15	58	33	60	0	**167**
	0.60%	8.98%	34.73%	19.76%	35.93%	0.00%	
Inc. B	5	61	180	125	159	10	**540**
	0.93%	11.30%	33.33%	23.15%	29.44%	1.85%	
Inc. C	2	35	134	114	142	4	**431**
	0.46%	8.12%	31.09%	26.45%	32.95%	0.93%	
Elimination	1	71	162	193	274	0	**701**
	0.14%	10.13%	23.11%	27.53%	39.09%	0.00%	
Unsuitable	0	1	13	5	9	0	**28**
	0.00%	3.57%	46.43%	17.86%	32.14%	0.00%	
Total	**9**	**188**	**555**	**475**	**645**	**14**	**1886**

Of the 2,842 recorded results for non-matching bullets, 1,886 sets were examined by a different examiner in Round Three, including 19 of the 20 false identifications from Round One. Of these 701 correct eliminations from Round One, different examiners again

correctly eliminated 274 in Round Three, for a reproducibility rate of 39.09 percent (274/701), while placing the remaining 427 into another category.

Let's examine the 427 correct eliminations that were not again eliminated by the second examiner. Of that total, 426 (99.8 percent) were judged inconclusive by the subsequent examiner, which again indicates that examiners were generally cautious about making conclusive determinations.

Only one set moved from a correct elimination to a false identification, indicating that it is very unlikely that different examiners, when independently examining non-matching sets, would reach opposite conclusive determinations. This finding supports the notion that through cross-examination and opposing experts, a rare false positive by the State's expert can be neutralized.

Now let's look at the reproducibility rate of inconclusives. Of the 1,886 sets examined in Round Three, 1,138 were placed in one of the three inconclusive levels in Round One. Of those, 755 were again judged inconclusive in Round Three by a different examiner. Individual reproducibility rates of Inconclusive-A, Inconclusive-B, and Inconclusive-C were 8.89 percent, 33.33 percent, and 26.45 percent, respectively. Those rates drag down the overall reproducibility rate for non-matching bullets.

The following table illustrates what happens to the results if we pool the three inconclusive levels:

Table XIV: Non-Matching Sets (Bullets)					
Classification by First Round Examiner	**Classification by Third Round Examiner**				
	ID	**Inconclusive (pooled)**	**Elimination**	**Unsuitable**	**Total**
ID	0	18	1	0	**19**
	0.00%	94.74%	5.26%	0.00%	
Inconclusive (pooled)	8	755	361	14	**1138**
	0.70%	66.34%	31.72%	1.23%	
Elimination	1	426	274	0	**701**
	0.14%	60.77%	39.09%	0.00%	
Unsuitable	0	19	9	0	**28**
	0.00%	67.86%	32.14%	0.00%	
Total	**9**	**1218**	**645**	**14**	**1886**

The overall reproducibility rate of the pooled inconclusive determinations for non-matching bullets is 66.34 percent (755/1138), a dramatic increase from the reproducibility rates of the individual levels of inconclusive. And the reproducibility rate of all determinations, both inconclusive and conclusive, increases from 30.9 percent to 54.6 percent.[26]

The results evidence greater reliability if we examine the 383 inconclusives from Round One that were *not* deemed inconclusive in Round Three. Of the 383 inconclusives, 361 moved from the inconclusive column in Round One to elimination in Round Three. So, though the migration out of inconclusive reduced the reproducibility rate for the non-matching bullet sets, nearly all (361/383 = 94.2 percent) moved in favor of the ground truth of elimination—the best possible directional change for non-matching bullets. Conversely, only eight sets of the 383 were incorrectly moved to the identification column. These shifts

[26] Calculated as: (0 paired agreement identifications + 755 paired agreement inconclusives + 274 paired agreement eliminations + 0 paired agreement unsuitables)/1886 = 54.6 percent.

between rounds reduced the overall reproducibility rate but increases confidence that the traditional tools for contesting expert testimony would be effective.

Finally, and most importantly for this case, let's look at the reproducibility rate of a false identification. There were 19 false identifications from Round One that were reviewed by a subsequent examiner, *and not a single examiner reproduced the initial examiner's mistake*. Of those 19 sets, one was correctly placed in the elimination column and the other 18 were deemed inconclusive.

In sum, the accuracy rates from Ames I and Ames II show that the risk of a false positive is both low and concentrated among a small number of examiners. The reproducibility results indicate that a subsequent examiner will very likely catch the rare false positive. Of the 1,886 sets of non-matching bullets that were reviewed by two different examiners, none were twice judged false positives.

C.

Recap

Let's recap the foregoing analysis. By focusing on the repeatability and reproducibility rates of the primary issue before us, the reliability of conclusive determinations for bullets, and by pooling the three levels of inconclusive results, we can make the following observations:

1. Repeatability rates of *correct* conclusive determinations were substantially higher than overall repeatability rates. While the overall repeatability rate for matching bullets was 79.0 percent and 64.7 percent for non-matching bullets, it was 89.6 percent for correct identifications (true positives) and 80.3 percent for correct eliminations (true negatives).

2. Repeatability rates of *incorrect* conclusive determinations were much lower: 10.5 percent for false identifications (false positives) and 39.4 percent for false eliminations (false negatives). Low rates of repeatability of incorrect conclusive results are, of course, preferred, because they indicate that mistakes are likely to be caught upon review.

3. When inconclusives are pooled, the overall repeatability rate for matching bullets improves from 79.0 percent to 83.4 percent. For non-matching bullets, the rate improves from 64.7 percent to 83.6 percent. Thus, by pooling inconclusives, the disparity noted by the Majority in repeatability between matching and non-matching bullets disappears.

4. Even when examiners were inconsistent with themselves between rounds, their responses were not far apart. Of the correct identifications from Round One that were *not* again judged to be identifications, 87.0 percent were judged inconclusive in Round Two. Of the correct eliminations from Round One that were *not* again judged to be eliminations, 96.0 percent were judged inconclusive in Round Two. This indicates that examiners exercised caution in making conclusive determinations.

5. Where examiners made an inconclusive determination in Round One but a conclusive determination in Round Two, they trended strongly toward ground truth, an indicator of reliability:

    i. Of the *matching* bullets that examiners initially judged inconclusive in Round One but judged differently in Round Two, examiners made a correct identification 83.6 percent of the time in Round Two.

    ii. Of the *non-matching* bullets that examiners initially judged inconclusive in Round One but judged differently in Round Two, examiners made a correct elimination 89.9 percent of the time in Round Two.

6. Reproducibility rates of *correct* conclusive determinations were substantially higher than overall reproducibility rates. While the overall reproducibility rate for matching bullets was 67.8 percent and 30.9 percent for non-matching bullets, it was 84.8 percent for correct identifications (true positives) and 39.1 percent for correct eliminations (true negatives).

7. Reproducibility rates of *incorrect* conclusive determinations were much lower: 5.3 percent for false Identifications (false positives) and 7.3 percent for false eliminations (false negatives). Low rates of reproducibility are preferred here because they indicate that mistakes are likely to be caught upon review.

8. When inconclusives are pooled, the overall reproducibility rate for matching bullets improves from 67.8 percent to 72.4 percent. For non-matching bullets, the rate improves from 30.9 percent to 54.6 percent. Thus, by pooling inconclusives, the disparity in reproducibility rates between matching and non-matching bullets decreases substantially.
9. Even when examiners were inconsistent with each other, their responses were not far apart. Of the correct identifications from Round One that were *not* again judged identifications, 84.3 percent were judged inconclusive in Round Three, another indication that the examiners exercised caution in making conclusive determinations. Of the correct eliminations from Round One that were *not* again judged eliminations, 99.8 percent were judged inconclusive in Round Two, again indicating caution.
10. When an examiner made an inconclusive determination in Round One but a subsequent examiner made a conclusive determination in Round Three, the subsequent examination trended strongly toward ground truth, an indicator of reliability:
    i. Of the *matching* bullets that examiners initially judged inconclusive in Round One but judged differently in Round Two, examiners made a correct Identification 82.5 percent of the time in Round Two.
    ii. Of the *non*-matching bullets that examiners initially judged inconclusive in Round One but judged *differently* in Round Two, examiners made a correct elimination 94.2 percent of the time in Round Two.

These findings and conclusions support a few takeaways: (1) examiners are not perfectly consistent, either with themselves or others, which is neither surprising nor disqualifying for a subjective pattern-matching discipline; (2) inconsistencies typically appear at the margins between two adjacent categories, showing that even where there is some "madness," there is also "method"; (3) the vast majority of inconsistencies do not prejudice the defendant but instead reflect caution by examiners in making conclusive determinations; (4) subsequent review by the same examiner, and especially by a different

examiner, is likely to catch errors and steer toward ground truth; and (5) independent examinations by two examiners almost never both produce false positives.

As we said in *State v. Matthews*, "[v]igorous cross-examination, presentation of contrary evidence, and careful instruction on the burden of proof are the traditional and appropriate means of attacking shaky but admissible evidence." 479 Md. at 312 (quotations omitted). Here, Mr. Abruquah could have called a competing firearms examiner to challenge Mr. McVeigh's opinions. Instead, he called two experts to opine solely on the reliability of the methodology, not Mr. McVeigh's analysis and conclusions. Similarly, when cross-examining Mr. McVeigh at trial, defense counsel did challenge Mr. McVeigh's analysis of the specimens recovered from the crime scene but focused instead on the reliability of the methodology generally. If Ames II tells us anything, it's that if a false positive is made, another trained examiner will be able to, at a minimum, poke substantial holes in the initial examiner's analysis. That Mr. Abruquah made no such effort at trial is, in my view, telling.[27]

[27] This is not, as the Majority asserts, Maj. Op. 52 n.28, a criticism of Mr. Abruquah, but rather an observation. The Majority's analysis hinges on the proposition that when making a positive identification, examiners show greater caution in studies than in field work. If this were a case in which the examiner was less cautious than was warranted by the facts and made a positive identification based on ambiguous or insufficient markings, one could reasonably expect that Mr. Abruquah would have attempted to expose such weaknesses in the examiner's analysis through cross-examination or his own expert. I am not criticizing Mr. Abruquah for not doing so, but rather inviting the reader to draw reasonable inferences from the fact that he did not. Further, the Majority states that the record "contains no support for th[e] proposition" that "that there are firearms examiners whose services were readily available to Mr. Abruquah, i.e., who are willing and able to take on work for criminal defendants in such cases." *Id.* However, the website for The Association of Firearm and Tool Mark Examiners—the same website to which the

D.

Additional Evidence

Not only does the Majority, in my view, fail to appreciate that Ames II has shown the AFTE Theory to be generally reliable, but the Majority also discounts specific standards and controls employed here by the Firearms Examination Unit of the Prince George's County Police Department Forensic Science Division ("FEU"). Those standards and controls, which were presented at trial, reduce the risk of error. Two central elements of those protocols are examiner independence and peer review, which I discuss briefly here.

At the *Frye-Reed* hearing, Mr. McVeigh identified protocols that the FEU follows to ensure examiners are independent and unbiased: examiners do not participate in investigations, read narratives of crime reports, or discuss cases with detectives. Mr. McVeigh affirmed that those protocols were followed in this case. He received unknown bullets and the Taurus revolver, knowing they were collected as part of the same case, and was asked to determine whether the firearm fired the bullets. He also received a report that included two paragraphs stating only the date, time, and location of the incident, and that officers "located the victim unresponsive in the residence suffering from an apparent gunshot wound."

Majority cites several times—has an "Expert Referral List" for "individuals requesting the assistance of a firearms/tool marks examiner in private casework." *Expert Referral*, AFTE, https://afte.org/resources/expert-referral (last visited June 13, 2023).

The FEU also requires peer review, which includes technical and administrative review of all cases. Technical review consists of a second examiner reviewing "all forensic conclusions," including "all bench notes, data, and other information that the examiner employs to form an opinion[.]" To be sure, technical review is not a blind second opinion, but it is nonetheless a form of peer review. In administrative review, the FEU manager or designee reviews all of the examiner's forensic conclusions.

Here, Mr. McVeigh's identification work was reviewed by another examiner, who approved of his conclusions. Is that a perfect check against the danger of a false positive? No. But it is a check, and the efficacy of that safeguard is not a function of the reliability underlying methodology—the focus of our inquiry—but rather of the competence and skill of individual examiners. The Majority dismisses these procedural safeguards.[28]

VI.

EXAMINERS' ABILITY TO DISTINGUISH INDIVIDUAL CHARACTERISTICS

The Majority acknowledges, without fully embracing, the underlying premise of the AFTE Theory—that firearms produce distinctive markings on bullets and cartridges ("individual characteristics"), and that examiners can identify those markings. Maj. Op. 42 (finding "strong support for the propositions that: (1) firearms produce some unique collections of individual patterns and markings on bullets and cartridges they fire; and

[28] Citing testimony from Dr. James E. Hamby, the Majority implies that the peer review process is a pro forma rubber stamp of the initial determination. Maj. Op. 46. The Majority may draw its own conclusions from Dr. Hamby's testimony, but having reviewed that same testimony carefully, in my view, a trial court could have reasonably drawn different conclusions.

(2) such collections of individual patterns and markings can be reliably identified [under certain conditions]") (footnote omitted).

The Majority, however, raises the prospect that examiners can reliably identify individual characteristics only "when subclass characteristics are removed from the equation."[29] *Id.* The Majority expresses particular concern with the apparent absence of published standards or controls guiding examiners on how to distinguish individual from subclass characteristics. Maj. Op. 49 ("The lack of standards and controls is perhaps most acute in discerning whether a particular characteristic is a subclass or an individual characteristic.").

The Majority, however, discounts studies showing that examiners can indeed make reliable determinations despite the risk of subclass carryover. Though published standards and controls would certainly be helpful, we should not ignore evidence that examiners make correct determinations in practice. Indeed, the concept of a black box study is premised on the assumption that when a process is not itself testable, we should study the accuracy of outcomes.

Before discussing these studies, I must clarify a point relating to study design. A first category of studies has controlled for subclass characteristics, i.e., eliminated the risk of subclass carryover, to determine only (1) whether firearms produce individual characteristics and (2) whether examiners can reliably identify those individual

[29] Subclass characteristics "are those shared by a group of firearms made using the same tools, such as those made in the same production run at a facility." Maj. Op. 49.

characteristics. *See, e.g.* Bullet Validation Study at 3.[30] As the Majority recognizes, these studies show that when the risk of subclass carryover is controlled for or eliminated, examiners can reliably identify individual characteristics. Maj. Op. 42 & 42 n.23.

A second category of studies assumes that examiners can identify individual characteristics and instead assesses whether examiners can reliably do so when there *is* a risk of subclass carryover. In these studies, similar subclass characteristics are likely present, but examiners do not know anything about the weapons used. Accordingly, examiners cannot assume that certain shared characteristics are subclass and thereby disregard them for purposes of individual determinations.

---

[30] In the Bullet Validation Study, Jamie Smith of the Firearms Examination Unit of Prince George's County Police Department Forensic Science Division sought to replicate Dr. Hamby's consecutive gun study while also introducing elements of open-set design. Here, examiners received 15 known samples fired from consecutively manufactured Beretta pistol barrels and 20 unknown samples. All samples used the same type of ammunition. The test administrators verified through inspection that the weapons produced no subclass characteristics, of which participants were made aware.

Of the unknown samples, some were also fired from the known Beretta barrels, while others were fired from other pistols of similar characteristics. This intentional mismatch between the firearms used for known and unknown samples introduced the possibility that unknown samples would not match any of the known samples and that, consequently, examiners could not count on using the process of elimination. And, because these tests were designed to possibly include multiple unknowns from the same source, the study abandoned the one-to-one relationship between known and unknown samples, which was characteristic of many closed studies.

Though this study was not fully open-set and had other design limitations, it is hard to ignore that only seven false identifications were reported, six of which the test administrators reported as resulting from typos. If those alleged typos were indeed typos, then the false positive rate was just 0.07 percent. Even if those alleged typos were treated as false positives, the false positive rate was just 0.47 percent.

At least two studies of this variety were introduced at trial. Both found that subclass characteristics did not undermine examiners' ability to reliably identify individual characteristics. First, Dr. Hamby tested examiners' ability to identify bullets fired from ten consecutively manufactured pistol barrels, which were expected to share subclass characteristics. *See generally* Worldwide Study.[31] Examiners were not provided any information about the barrels. *Id.* Dr. Hamby observed that "[e]rrors due to subclass characteristics, which one could conjecture would be a significant issue when consecutively rifled barrels are involved, have not been a problem for the examiners," concluding that "there are identifiable features on the surface of bullets that may link them to the barrel that fired them." *Id.* at 556.

[31] Dr. Hamby worked with others to develop an ongoing study that tested the examiners' ability to identify bullets fired from ten consecutively manufactured Ruger P-85 9mm pistol barrels. A total of 697 examiners from 32 countries participated. Each test set included a set of 20 known bullets, two fired from each of the ten barrels, and 15 unknown bullets, comprised of one to three bullets from each of the barrels.

To be sure, the modified closed-set design of this study limits its value, despite Dr. Hamby's introduction of greater uncertainty by abandoning the one-to-one known-to-unknown relationship of past studies (the so-called "Sudoku style" test). Regardless, the results cannot be ignored: examiners correctly matched all but eight of the 10,455 unknown bullets to the known match. Examiners reported inconclusive determinations on the remaining eight and made no misidentifications.

The authors concluded that "there are identifiable features on the surface of bullets that may link them to the barrel that fired them" and that shared subclass characteristics did not confound an examiners' ability to draw accurate conclusions. The Majority might reasonably disagree with these conclusions, but, by the same token, a trial judge would not be unreasonable to place credence in them when determining threshold reliability under *Daubert.*

The second study, Ames II, compared examiners' performance on samples with similar subclass characteristics against their performance on samples that likely had distinct subclass characteristics. Each sample was the same type of ammunition and fired from a weapon of the same make and model. Examiners were not made aware of the characteristics of the weapons used in the study.

Researchers assessed, among other things, performance with respect to two variables: manufacturing run and sequential groups within a single manufacturing run. Guns manufactured in the same run, which were produced by the same tool, would presumably produce greater shared subclass characteristics than guns manufactured by different tools in different runs. Similarly, guns manufactured in the same group within a single manufacturing run would presumably produce greater shared subclass characteristics than those manufactured in different groups within the same manufacturing run.[32]

Examiners performed somewhat better overall, with lower rates of false positives, for guns from different manufacturing runs. Ames II at 56-67. The same was observed for guns from different groups within a single manufacturing run. *Id.* These observations suggest that samples featuring different subclass characteristics might be "easier" to correctly determine than those with shared subclass characteristics.

The authors nonetheless concluded that examiners' responses for bullets did not, as a whole, differ in a statistically significant way between same-run and different-run

[32] Researchers performed this analysis for only the Beretta group of guns.

samples. *Id.* For cartridges, though, examiners' responses did meaningfully differ, attributable mostly to differences in elimination determinations. *Id.* Responses did not meaningfully differ between same-group and different-group samples *within* the same manufacturing run for bullets or cartridges. *Id.* [italicized the word within]

Notably, the false positive rates within any of the categories—cartridges included—ranged from 0.38 percent to 1.14 percent.[33] *Id.* Thus, even though examiners may have reported false positives more frequently for certain categories of guns, that the highest false positive rate was just 1.14 percent paints a picture of a reliable discipline.

From these studies, the trial court could have reasonably concluded that, despite the risk of subclass carryover, the AFTE Theory is sufficiently reliable to admit Mr. McVeigh's proffered testimony. Additionally, the trial court could have credited testimony by the State's experts acknowledging the risk of subclass carryover but emphasizing the caution examiners exercise to protect against it.

CONCLUSION

The Majority, "misunderstand[ing] *Daubert* to demand unassailable expert testimony," *United States v. Mooney*, 315 F.3d 54, 63 (1st Cir. 2002), misses the forest for the trees. The trial court's task is not to ensure that an expert's conclusion is "correct," but only "that the expert's conclusion has been arrived at in a scientifically sound and methodologically reliable fashion." *Id.* (quoting *Ruiz–Troche v. Pepsi Cola of P.R. Bottling Co.*, 161 F.3d 77, 85 (1st Cir. 1998)).

[33] This data is taken from Table XXIII and Table XXIV in Ames II.

Although I recognize the substantial deference that jurors may accord to experts, particularly on forensic, technical, and scientific matters, we are dealing with the admissibility of expert *opinion*, and opinion, by definition, carries a risk of error or inaccuracy. This Court's job is not to inoculate the jury from *all* risk of inaccurate expert opinion; to do so would be "overly pessimistic about the capabilities of the jury and of the adversary system generally." *Daubert*, 509 U.S. at 596. On this basis, courts have admitted expert opinions on imperfect subjective methods, such as handwriting analysis and coin-grading. *See Mooney*, 315 F.3d at 61-63 (allowing handwriting expert to opine that defendant authored specific letters, despite evidence that handwriting examiners had a potential rate of error of 6.5 percent); *United States v. Romano*, 794 F.3d 317, 330-33 (2d Cir. 2015) (allowing testimony on grades of coins).

Indeed, the sort of extensive statistical investigation that the Majority and I engage in here is precisely what *Daubert* and *Rochkind* told courts *not* to do. Contrary to the admonitions of those Courts, this Court unwisely assumes the role of "amateur scientist," *see Rochkind*, 471 Md. at 33, in our "exhaustive search for cosmic understanding[,]" *Daubert*, 509 U.S. at 597.

None of the foregoing is to suggest that the Majority's reasoning is irrational or unreasonable, or that admitting the testimony was the only correct decision. Rather, I contend that the trial court made a *reasonable* decision supported by the evidence, and the fact that others may disagree merely signifies to me that on this difficult topic, reasonable minds can and do differ. That, in my view, is not the stuff of abuse of discretion.

Respectfully, I therefore dissent.