



950 Pennsylvania Ave, N.W.
Washington, D.C. 20530

November 6, 2020

Hon. Patrick Schiltz
United States District Judge
United States Courthouse
300 South Fourth Street, Room 14E
Minneapolis, MN 55415

Re: Possible Amendment to Federal Rule of Evidence 702

Dear Judge Schiltz:

We write respectfully, in advance of our upcoming November 13 meeting, to supplement the agenda materials with some additional reference materials and thoughts. Since the virtual nature of our meeting may make free-flowing discussion more difficult, we hope that having our views in advance will help further the conversation.

Uniform Language for Testimony and Reports

As the Committee will recall, the Department has proposed that the Committee table any amendment to Rule 702 in order to gauge the effectiveness of Department's initiatives with respect to Uniform Language for Testimony and Reports ("ULTRs"). The Department's Forensic Science webpage currently contains 16 ULTRs, many updated this past summer to further address important qualifications and limitations of expert testimony in various forensic disciplines.

In the forensic geology discipline, for example, an examiner may testify to a (1) Fracture fit; (2) Inclusion (i.e., included); (3) Exclusion (i.e., excluded); or (4) Inconclusive. When explaining his or her conclusion, "[a]n examiner shall not assert that two or more geologically-derived materials were once part of the same object unless the materials physically fit together." In addition, when offering a conclusion, an examiner shall not assert that a fracture fit is based on the "uniqueness" of an item of evidence; use the term "individualize" or "individualization;" or claim that the geologically-derived materials originated from the same object "to the exclusion of all other objects." Nor may an examiner assert absolute or 100% certainty or claim that forensic geology examinations are infallible or have a zero-error rate. Moreover, the ULTRs make clear that an examiner's source identification opinion is not based on a statistically derived or verified measurement or comparison to all other potential sources of a questioned sample. See <https://www.justice.gov/olp/page/file/1284776/download>.

Beginning in 2018, and continuing to the present, there are ample examples of federal, state, and D.C. courts that have limited or excluded testimony regarding the source of a spent bullet or shell casing. These cases, some of which are included in the case law digest, include:

United States v. Jovon Medley, No. PWG 17-242 (S.D. Md. April 24, 2018)
Williams v. United States, 210 A.3d 734 (D.C. Ct. App. June 27, 2019)
United States v. Tibbs, 2019 D.C. Super. LEXIS 9 (D.C. Sup. Ct. September 5, 2019)
United States v. Davis, 2019 U.S. Dist. LEXIS 155037 (W.D. Va. September 11, 2019)
United States v. Shipp, 2019 U.S. Dist. LEXIS 205397 (E.D.N.Y. November 26, 2019)
United States v. Adams, 2020 U.S. Dist. LEXIS 45125 (D. Oregon March 16, 2020)
People v. A.M., 2020 N.Y. Misc. LEXIS 2961 (Sup Ct. Bronx June 30, 2020)

In each of these cases—whether or not one agrees with the analysis and ultimate decision—the court used the existing rules of evidence to preclude the examiner from offering identification testimony. In contrast, the meeting memo (“Memo”) discusses *U.S. v. Simmons*, 2018 U.S. Dist. LEXIS 18606 (E.D. Va), decided January 12, 2018, as an example of a case that failed to heed the Department’s directives. *Simmons*, however, predated the publication of the ULTR documents. In addition, *Simmons* was a case in which the government—not the witness—offered alternative formulations of the expert’s conclusion for the court’s consideration during pretrial proceedings.

Although the Memo correctly notes that the ULTRs are not binding on state laboratories or state courts, neither are the Federal Rules of Evidence. Nevertheless, the ULTRs may well have an important impact on the states. The Organization of Scientific Area Committees¹ (“OSAC”), whose primary mission is to develop uniform national standards across forensic disciplines, and whose membership includes experts from federal, state, county, and local government, academia, and the private sector, has drawn from language provided in the ULTRs to draft national forensic standards. By allowing this industry-wide standards-building process to continue and develop, the guidance articulated in ULTRs may take hold faster and more effectively than any federal rule change. Indeed, in two recently published opinions, one from the D.C. District Court and another from the Western District of Oklahoma, the court utilized the Department’s ULTRs to properly limit the scope of firearms-toolmarks testimony.²

The Conceptual and Practical Differences Between “Match” and “Source Identification”

The conceptual formulation of a “match” and a “source identification” opinion is not the same. The traditional “match” paradigm in the forensic pattern comparison disciplines employed an essentially deductive reasoning process in which a sufficient combination of corresponding features was considered to be “unique” in the natural world. It followed that if a questioned sample exhibited a sufficient combination of features that corresponded to those observed in the known item, then the questioned sample (pattern) was considered “unique.” As such, an examiner “individualized” the questioned sample “to the exclusion of all other” such items (e.g. fingerprints, shell casings).

¹ <https://www.nist.gov/topics/organization-scientific-area-committees-forensic-science>

² *U.S. v. Harris*, 2020 U.S. Dist. LEXIS 205810 (D.D.C.) (Nov. 4, 2020); *see also U.S. v. Hunt*, 2020 U.S. Dist. LEXIS 95471 (W.D. Okla.) (June 1, 2020).

In contrast to the “match” paradigm, a “source identification”³ conclusion is the result of an inductive reasoning process that makes no universal claims of deductive certainty. During an examination, a known item and a questioned sample are examined for a sufficient combination of corresponding features.⁴ If an examiner determines that there is sufficient correspondence such that she (based on her knowledge, training, experience, and skill) would not expect to find the same combination of features repeated in another source, and there is insufficient disagreement to conclude that the combination of features came from a different source, then the correspondence provides extremely strong support for the proposition that the questioned sample came from the known item. Similarly, it provides extremely weak or no support for the proposition that the questioned sample came from a different source. The examiner then inductively infers (from the observed data) that the questioned sample originated from the known item.⁵ The resulting classification as a “source identification,” “source exclusion,” “inconclusive,” is ultimately an examiner’s skill and experience-based opinion.

Importantly, at the conclusion of this process, an examiner makes no claim that the observed combination of corresponding features in the questioned sample (class and individual

³ “Identification is the decision process of establishing with sufficient confidence (not absolute certainty), that some identity-related information describes a specific entity in a given context, at a certain time.” Casey Eoghan & David-Oliver, *Do Identities Matter?* 13 *Policing: A Journal of Policy & Practice* 21, 21 (March 2019).

⁴ “The question for the scientist is not ‘are this mark and print identical’ but, ‘given the detail that has been revealed and the comparison that has been made, what inference might be drawn in relation to the propositions that I have set out to consider.’” Christophe Champod & Ian Evett, *A Probabilistic Approach to Fingerprint Evidence*, *Journal of Forensic Identification*, 101-22, 103 (2001).

⁵ See David Kaye, *Probability, Individualization, and Uniqueness in Forensic Science Evidence: Listening to the Academies*, 75 *Brooklyn L. Rev.* 1163, 1176 (2010) (“In appropriate cases . . . it is ethical and scientifically sound for an expert witness to offer an opinion as to the source of the trace evidence. Of course, it would be more precise to present the random-match probability instead of the qualitative statement, but scientists speak of many propositions that are merely highly likely as if they have been proved. They are practicing rather than evading science when they round off in this fashion.”).

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

Paul Roberts & Colin Aitken, *Communicating and Interpreting Statistical Evidence in the Administration of Criminal Justice*, 3. *The Logic of Forensic Proof—Inferential Reasoning in Criminal Evidence and Forensic Science, Guidance for Judges, Lawyers, Forensic Scientists, and Expert Witnesses*, Royal Statistical Society 43 (2014) <https://www.maths.ed.ac.uk/~cgga/Guide-3-WEB.pdf>.

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

Colin Aitken & Franco Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists* (Wiley 2nd Ed. 2004).

characteristics) is “unique”⁶ in the natural world, or that the examiner can universally “individualize”⁷ the item or person from which the questioned sample originated. Moreover, given the limitations of inductive reasoning, an examiner cannot logically “exclude all other” potential sources of the questioned sample with certainty.⁸ Accordingly, ULTR documents that authorize a “source identification”⁹ conclusion also prohibit an examiner from asserting that a questioned sample originated from a known source “to the exclusion of all other sources.” They also disallow claims of absolute or 100% certainty, infallibility, or a zero-error rate.¹⁰

From a legal perspective, a “source identification” conclusion is properly characterized as technical or specialized knowledge under Rule 702,¹¹ as it is based on an examiner’s training, skill, and experience—not statistical methods or measurements. As such, the PCAST Report erred when it claimed that all forensic pattern comparison disciplines are “metrology” (measurement science).¹² Although many of these disciplines are grounded in scientific principles, source identification conclusions provided by forensic examiners are “skill and experience-based”

⁶ “Every entity is unique; no two entities can be ‘Identical’ to each other because an entity may only be identical to itself. Thus, to say ‘this mark and this print are identical to each other’ invokes a profound misconception: they might be indistinguishable but they cannot be identical.” Champod, *supra* note 4, at 103.

⁷ “[I]ndividualization—the conclusion that ‘this trace came from this individual or this object’—is not the same as, and need not depend on, the belief in universal uniqueness. Consequently, there are circumstances in which an analyst reasonably can testify to having determined the source of an object, whether or not uniqueness is demonstrable.” Kaye, *supra* note 5, at 1166. The Department uses the term “identification” rather than “individualization.”

⁸ “We cannot consider the entire population of suspects - the best we can do is to take a *sample*... We use our observations on the sample, whether formal or in formal, to draw inferences about the *population*. No matter how large our sample, it is not possible for us to say that we have eliminated every person in the population with certainty. . . . This is the classic scientific problem of *induction* that has been considered in the greatest depth by philosophers.” Champod, *supra* note 4, at 104-105.

⁹ See also Kaye, *supra* note 5, at 1185 (“Radical skepticism of all possible assertions of uniqueness is not justified. Absolute certainty (in the sense of zero probability of a future contradicting observation) is unattainable in any science. But this fact does not make otherwise well-founded opinions unscientific or inadmissible. Furthermore, whether or not global uniqueness is demonstrable, there are circumstances in which an analyst can testify to scientific knowledge of the likely source of an object or impression.”).

¹⁰ <https://www.justice.gov/olp/uniform-language-testimony-and-reports>.

¹¹ See, e.g. *U.S. v. Herrera*, 704 F.3d 480 (7th Cir. 2013) (“[E]xpert evidence is not limited to ‘scientific’ evidence, however such evidence might be defined. . . . It includes any evidence created or validated by expert methods and presented by an expert witness that is shown to be reliable.” (Latent print decision); *Restivo v. Hessemann*, 846 F.3d 547, 576 (2d Cir. 2017) (“Rule 702 ‘makes no relevant distinction between ‘scientific’ knowledge and ‘technical’ or ‘other specialized’ knowledge,’ and ‘makes clear that any such knowledge might become the subject of expert testimony.’ *Kumho Tire Co.*, 526 U.S. at 147”); see also *U.S. v. Harris*, 2020 U.S. Dist. LEXIS 205810 (D.C. November 4, 2020) (characterizing firearms-toolmarks testimony as technical/specialized knowledge); *Accord U.S. v. Hunt*, 2020 U.S. Dist. LEXIS 95471 (W.D. Okla.); *U.S. v. Johnson*, 2019 U.S. Dist. LEXIS 39590 (S.D.N.Y. 2019); *U.S. v. Otero*, 849 F. Supp. 2d 425 (D.N.J. 2012); *U.S. v. Mouzone*, 696 F. Supp. 2d 536 (D. Md. 2009); *U.S. v. Monteiro*, 407 F. Supp. 2d 351 (D. Mass. 2006).

¹² *President’s Council of Advisors on Sci. & Tech., Executive Office of the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature Comparison Methods* 23, 44, 143 (2016) (original emphasis) at 23, 44 n.93, 143.

opinions, similar to those offered by an electrical engineer, and discussed in the meeting Memo (pp. 132-33). It is also important to note that the PCAST Report chose to use the term “proposed identification” as the appropriate way for a forensic pattern examiner to articulate his or her conclusion. By adding the word “proposed,” PCAST meant to convey the possibility that the opinion might be incorrect¹³ As such, a “proposed identification” is essentially equivalent to a “source identification” conclusion. Both formulations recognize that an examiner’s opinion is potentially fallible.

Cross-Examination as a Solution to Perceived “Overstatement”

The meeting Memo suggests that empirical studies have shown that cross-examination is an ineffective means by which to challenge the credibility of expert witnesses—citing a 2008 study by McQuiston-Surrett & Saks. That study, however, is inconsistent with more recent research, including a 2013 paper authored by Professor Brandon Garrett. That study found that

[p]articipants exposed to the examiner who testified on direct that his method was reliable and then acknowledged on cross a possible misidentification rated the general reliability of fingerprint identifications the lowest. Thus, our results suggest that an examiner who claims infallibility on direct will be viewed skeptically after a cross that elicits error-risk concessions, but an examiner who on direct describes her method in reasonable terms, including acknowledging some risk of error, may be able to limit the negative impact of an effective cross-examination or contrary fingerprint evidence presented by the defense.¹⁴

In another study published in 2015, Joseph Eastwood and Jiana Caldwell found that educating jurors about the limitations of forensic procedures by presenting opposing expert witnesses can be effective in raising legitimate doubts about the forensic conclusions.¹⁵

A 2019 study—authored by PCAST contributor William Thompson—reported that participants found an expert less credible and were less likely to convict when the expert admitted that his interpretation rested on subjective judgment and when he admitted to having been exposed to potentially biasing task-irrelevant contextual information.¹⁶ Thompson found that,

¹³ *Id.* at 46. (“We suggest the term “*proposed* identification” to appropriately convey the examiner’s conclusion, along with the possibility that it might be wrong. We will use this term throughout this report.”) (original emphasis).

¹⁴ Brandon Garrett & Gregory Mitchell, *How Jurors Evaluate Fingerprint Evidence: The Relative Importance of Match Language, Method Information, and Error Acknowledgment*, J. of Empirical Legal Stud., 484, 505-06 (2013); see also Brandon Garrett & Gregory Mitchell, *How Jurors Evaluate Fingerprint Evidence: The Relative Importance of Match Language, Method Information, and Error Acknowledgment*, Journal of Empirical Legal Studies, 484, 507 (“[W]hen the fingerprint examiner admitted that his method is not foolproof and that his conclusion in this case could be in error, that disclosure had a significant negative impact on the evidence.”).

¹⁵ Joseph Eastwood & Jiana Caldwell, *Educating Jurors About Forensic Evidence: Using an Expert Witness and Judicial Instructions to Mitigate the Impact of Invalid Forensic Science Testimony*, 60 J. Forensic Sci. 1523, 1528.

¹⁶ William Thompson & Nicholas Scurich, *How Cross-Examination on Subjectivity and Bias Affect Jurors’ Evaluations of Forensic Evidence*, 64 J. Forensic Sci. 1379-88 (2019).

[o]verall, the results indicate that jurors were skeptical of the expert's claim that he had ignored the task-irrelevant information, and this skepticism increased when the expert also admitted that his interpretation of the findings required subjective judgment in the absence of objective standards for interpretation.¹⁷

* * *

From a legal perspective, the finding suggests that lawyers can successfully challenge the credibility of a non-blind forensic expert in two ways: either by revealing the subjectivity of the expert's methods or by revealing the expert's exposure to task irrelevant information.¹⁸

Accordingly, recent research supports the position that conceding the fallibility of forensic findings on direct examination, during cross-examination, or through contrary evidence by an opposing expert, *does* affect the persuasiveness of a forensic examiner's opinion. Moreover, cross-examination is enhanced by the timely production of information underlying the expert's opinion. This was the reason that the Criminal Rules Committee—with the Department's support—has worked on a proposed amendment to Rule 16. The proposed timeliness requirement in Rule 16 is also being supplemented with additional DOJ training to ensure that prosecutors understand and adhere to their disclosure obligations.

The Department recognizes that a forensic examiner's past performance on relevant, skill-based testing is an important measure for evaluating her performance in a given case. As such, FBI proficiency test results are routinely provided to defense counsel upon request. The FBI Laboratory will soon begin disclosing proficiency test results without a specific defense request as part of their general discovery and disclosure procedures. In addition, Department laboratory quality assurance manuals, standard operating procedures, testing methodologies, and other laboratory policies are currently available online to defense attorneys and the general public.¹⁹ Moreover, the Department's ULTRs, which set forth the qualifications and limitations for sixteen forensic disciplines, are available to defense counsel in each case and are available on-line.²⁰

In a recent study, Professor Garrett examined the impact of proficiency test results and laboratory error rates on jury-eligible adults. His study found that,

[w]hen jurors receive information about flaws or weaknesses in a forensic method or receive general information about a field's error rates, the juror cannot be sure how that information applies to the particular analyst in the case at hand. But when jurors receive information about the testifying expert's own performance on a proficiency test that simulates the task involved in the case at hand, the relevance of this information is easy to comprehend and hard to ignore.²¹

¹⁷ *Id.* at 1386.

¹⁸ *Id.*

¹⁹ <https://www.justice.gov/olp/forensic-science#posting>.

²⁰ See <https://www.justice.gov/olp/uniform-language-testimony-and-reports>. “This document is intended to describe and explain terminology that may be provided by Department examiners. *It shall be attached to, or incorporated by reference in, laboratory reports or included in the case file.*” (Emphasis added).

²¹ Gregory Mitchell & Brandon Garrett, *The Impact of Proficiency Testing Information and Error Aversions on the Weight Given to Fingerprint Evidence*, 37 *Behav. Sci. Law*, 1, 14 (2019).

Regarding the impact of proficiency test information in particular, the Garrett study found that,

[t]he fingerprint examiner's level of performance on a proficiency test (high, medium, low, or very low), but not the type of error committed on the test (false positive identifications, false negative identifications, or a mix of both types of error), affected the weight that jury-eligible adults gave to an examiner's opinion that latent fingerprints recovered from a crime scene matched the defendant's fingerprints, which in turn affected judgments about the defendant's guilt.²²

Collectively, these recent studies undermine the position that cross-examination is an ineffective means of challenging the credibility of a forensic examiner. Instead, the findings clearly support the position that conceding the potential fallibility of forensic results on direct examination or during cross-examination, or challenging forensic evidence by use of an opposing expert, impacts the credibility of a forensic examiner's opinion.

Strength of Evidentiary Support versus Opinion Testimony

The meeting Memo appears to favor “strength of evidence” testimony over an expert’s opinion about the source of a questioned item. Memo at 110. Recent research, however, has shown that jurors do not correctly discern differences between subtle gradations of evidentiary strength, such as those endorsed by the American Statistical Association and described in the Memo.

Specifically, Eleanor Arscott found that study participants performed poorly when attempting to distinguish between strength of evidence expressions at the strong end of the scale (“strong,” “very strong,” and “extremely strong”).²³ As a result, she concluded that it was possible “to question the effectiveness of the scale of verbal expressions in communicating the intended evidential strength at the higher end of the scale.”²⁴ Arscott also noted the same can be argued for distinctions between “weak” and “moderate” strength, and between “moderate” and “moderately strong” evidence.²⁵ She concluded that “[t]hese results suggest we may not be able to assume that decision makers will be able to discern between these expressions.”²⁶

Separate research by Kristy Martire²⁷ on verbally described gradations in evidentiary strength revealed what she described as “the weak evidence effect.” That is, study participants presented with evidence that weakly supported guilt tended to invert that finding and wrongly think that “weak” evidence in support of the prosecution’s case actually meant that the evidence favored the accused.²⁸ Participants presented with weakly exculpatory evidence, however, were

²² *Id.* at 1.

²³ Eleanor Arscott et al., *Understanding Forensic Expert Evaluative Evidence: A Study of the Perception of Verbal Expressions of the Strength of Evidence*, 57 *Sci. and Just.* 222, 224, n.13 (2017).

²⁴ *Id.* at 224.

²⁵ *Id.*

²⁶ *Id.* at 227.

²⁷ Kristy Martire et al., *The Expression and Interpretation of Uncertain Forensic Science Evidence: Verbal Equivalence, Evidence Strength, and the Weak Evidence Effect*, 37 *Law and Hum.* 197, 205-06 (2013).

²⁸ *Id.* at 205-06.

not affected in the same way.²⁹ These studies demonstrate that testimony based on gradations of evidentiary support may actually confuse rather than clarify the intended meaning of an examiner's conclusion. This is surely not the intended result of a proposed rule change to FRE 702.

Assumptions Underlying the Proposed Rule Change and Note

1. Studies on the Baseline Valuation of Forensic Evidence by Potential Jurors: The So-Called "CSI Effect"

The draft Committee Note that accompanies the proposed amendments to FRE 702 suggests that jurors may overvalue scientific evidence and either unquestionably accept it or fail to understand expert testimony. *See, e.g.*, Memo at p. 143. ("Just as jurors are unable to evaluate meaningfully the reliability of scientific and other methods underlying expert opinion, jurors lack a basis for assessing critically the conclusions of an expert that go beyond what the expert's methodology may reliably support.").

Recent research, however, contradicts the notion that jurors overvalue forensic evidence. To the contrary, the findings show that jurors approach forensic evidence with a critical eye and tend to underweight its probative value. For example, a 2020 study by Jacob Kaplan and colleagues³⁰ reached the following conclusion:

We find that individuals in the United States hold a pessimistic view of the forensic science investigation process, believing that an error can occur about half of the time at each stage of the process. We find that respondents believe that forensics are far from perfect, with accuracy rates ranging from a low of 55% for voice analysis to a high of 83% for DNA analysis, with most techniques being considered between 65% and 75% accurate.³¹

The results differed from the researchers' expectations:

While we expected respondents to have a high level of confidence in the forensic science investigation process and for the accuracy of each forensic science technique (Hypothesis 1), our results suggest that members of the US public hold significant doubts about the accuracy of forensic techniques and believe that each technique contains high levels of human judgement. The technique perceived to be most accurate was DNA evidence at 83% accuracy, while voice analysis at 55% and footwear analysis at 57% were perceived to be least reliable. Most forensic techniques were considered to be in the range of 65–75% accurate. Our results align with prior work indicating that DNA is often perceived to be among the most accurate forensic techniques, though our study yields lower perceptions of accuracy for DNA than reported elsewhere. Additionally, respondents indicated that they

²⁹ *Id.* at 205.

³⁰ Jacob Kaplan et al., *Public Beliefs About the Accuracy and Importance of Forensic Evidence in the United States*, 60 *Sci. & Just.* 263-72 (2020).

³¹ *Id.* at 263.

believed there was a substantial risk of error at each stage of the forensic science process, and that each stage involves a large amount of human judgement.³²

In short, the authors found that,

US respondents believe that there is a high degree of human judgement involved and high risk of an error occurring at each stage of the forensic science process. When considering forensic science techniques specifically, those in the US hold a skeptical view of the vast majority of techniques, viewing some of them as little more accurate than a coin flip, and no technique more than 84% accurate.³³

Kaplan's results corroborate the findings of a similar study from Australia. In that work, Gianni Ribeiro and colleagues³⁴ found, contrary to their expectations, that study participants believed that the forensic process involved considerable human judgment and was relatively prone to error. Specifically, the researchers found:

[P]articipants had wide-ranging beliefs about the accuracy of various forensic techniques, ranging from 65.18% (document analysis) up to 89.95% (DNA). For some forensic techniques, estimates were lower than that found in experimental proficiency studies, suggesting that our participants are more skeptical of certain forensic evidence than they need to be.³⁵

Ribeiro concluded that, “[i]n this study, we have demonstrated that participants do not just blindly believe that all forensic techniques are highly accurate, which has previously been assumed in the CSI effect literature. Instead, our participants believe that the forensic science process is error prone and involves a considerable amount of human judgment at each and every stage.”³⁶

As surprising as these findings may be, they are not anomalous. Indeed, they are consistent with other research finding that study participants consistently undervalue the significance of forensic evidence. For example, Dale Nance, in a study that involved people called for jury service in Illinois, concluded that, “[l]ooking at the forest rather than the trees, the dominant problem the empirical research reveals is that jurors as a group tend to undervalue the scientific evidence.”³⁷

In a separate large-scale empirical study—again using members of an Illinois jury pool—Nance confirmed the findings of his earlier research that jurors tend to minimize forensic

³² *Id.* at 270.

³³ *Id.* at 271.

³⁴ Gianni Ribeiro et al., *Beliefs About Error Rates and Human Judgment in Forensic Science*, 297 *Forensic Sci. Int'l.* 138-47 (2019).

³⁵ *Id.* at 138.

³⁶ *Id.* at 146.

³⁷ Dale Nance & Scott Morris, *An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Large and Quantifiable Random Match Probability*, 42 *Jurimetrics J.* 403 (2002).

evidence.³⁸ Specifically, he found that “for the most part jurors’ innate skepticism and need to be convinced create a dominating undervaluation of the evidence.”³⁹

In a later study, Jason Schklar⁴⁰ found that “[a]lthough no published study has reported jurors’ naive expectancies of how likely it was that a DNA match report could have resulted from either random chance or a laboratory error, some evidence indicates that people think human errors in the DNA lab are *more likely* than proficiency test results have revealed.”⁴¹ In addition, Schklar concluded, “[t]he results of this study also suggest that jurors may not infer that DNA test results are error-free when they do not receive an LE [error rate] estimate.”⁴²

Most recently, William Thompson and Edward Newman⁴³ found that study participants undervalued forensic footwear evidence.⁴⁴ Their findings “indicate that perceptions of forensic science evidence are shaped by prior beliefs and expectations as well as expert testimony and consequently that the best way to characterize and explain forensic evidence may vary across forensic disciplines.”⁴⁵ The authors concluded, “The complexity of our findings suggests that the problem of how “best” to present forensic evidence to lay audiences may not have a single, simple solution.”⁴⁶

2. Error Rates

Professor Brandon Garrett, in a letter to the Committee, claimed that “[n]o conclusion can be reached about a method without qualification or discussion of error rates, because there is no type of expertise that does not have some error rate.” Memo, p. 121. The draft Committee Note reflects this view. See Memo, p. 143 (“Accurate testimony will ordinarily include a fair assessment of the rate of error of the methodology employed, based where appropriate on empirical studies of how often the method produces correct results, as well as other relevant limitations inherent in the methodology.”). But it is scientifically incorrect to assume that a single error rate can be attributed to a particular method or generally applied to all forensic examiners who practice that method.⁴⁷

³⁸ Dale Nance & Scott Morris, *Juror Understanding of DNA Evidence: An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Small Random-Match Probability*, 34 J. Legal Stud. 395 (2005).

³⁹ *Id.* at 436.

⁴⁰ Jason Schklar & Shari Diamond, *Juror Reactions to DNA Evidence: Errors and Expectancies*, 23 Law & Hum. Behav. 159 (1999).

⁴¹ *Id.* at 165 (emphasis added).

⁴² *Id.* at 178.

⁴³ See William Thompson & Eryn Newman, *Lay Understanding of Forensic Statistics: Evaluation of Random Match Probabilities, Likelihood Ratios, and Verbal Equivalents*, 39 Law & Hum. Behav. 332 (2015).

⁴⁴ Consistent with these results, other research has also found that study participants underutilize forensic evidence. See William Thompson & Edward Schumann, *Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor’s Fallacy and the Defense Attorney’s Fallacy*, 11 Law & Hum. Behav. 167 (1987); David Faigman & A.J. Baglioni, *Bayes’ Theorem in the Trial Process: Instructing Jurors on the Value of Statistical Evidence*, 12 Law & Hum. Behav. 1 (1988); Jane Goodman, *Jurors’ Comprehension and Assessment of Probabilistic Evidence*, 16 Am. J. Trial Advoc. 361 (1992).

⁴⁵ *Id.* at 332.

⁴⁶ *Id.* at 348.

⁴⁷ See, e.g., William Thompson et al., American Academy for the Advancement of Science *Forensic Science Assessments: A Quality and Gap Analysis* (2017) (“[I]t is unreasonable to think that the “error rate” of latent fingerprint examination can meaningfully be reduced to a single number or even a single set of numbers. At best, it might be

First, many experts, including skill and experience-based experts, will be unable to testify to a specific error rate. Consider the brain surgeon testifying in a medical malpractice suit. Based on the surgeon's experience performing and observing a procedure thousands of times, she opines that the failure to correctly clamp a particular artery led to the plaintiff's excess bleeding and subsequent paralyzing stroke. The surgeon's opinion, and her confidence in that opinion, may be tested on cross-examination and through rebuttal experts. But there is no error rate that accompanies the methodology used to reach that opinion. Similarly, the structural engineer who studies the collapse of a bridge and testifies that, in his opinion, the bridge had a specific design flaw need not provide an error rate in order to offer his skill and experience-based opinion.

Second, even error rate advocates concede that it is exceedingly difficult to accurately establish scientifically valid and generally applicable figures. PCAST contributor and Boston College Symposium participant Itiel Dror addressed this point in a recent paper in which he discussed the complexities and practical difficulties of establishing a valid error rate.⁴⁸ These include knowing ground truth facts, establishing appropriate databases, determining what counts as an error, deciding on an acceptable metric, and problems with the external or ecological validity⁴⁹ of generalizing a given rate to different situations and circumstances.⁵⁰ Dror observed that, "[p]roviding 'an error rate' for a forensic domain may be misleading because it is a function of numerous parameters and depends on a variety of factors."⁵¹ He then posed the following rhetorical question:

The need to properly establish error rates in forensic science is clear. But, given the time and effort it requires, as well as the inherent limitations of the very notion of error rates, is it worth it? And, how does it compare (or complement) other measures of performance (e.g., effective proficiency testing, quality assurance checks such as dip sampling and blind verification, accreditation, and ongoing training and development).⁵²

Given these limitations, perhaps the best one can do is to examine the compendium of relevant studies and view them as a composite measure of the potential range of error rates across a discipline⁵³—but one that is not necessarily applicable to any particular case or examiner (due

possible to describe, in broad terms, the rates of false identifications and false exclusions likely to arise for comparisons of a given level of difficulty.”).

⁴⁸ Itiel Dror, *The Error in Error Rate: Why Error Rates Are So Needed, Yet So Elusive* 65 *J. Forensic Sci.*, 1034 (2020).

⁴⁹ Ecological validity refers to “a kind of external validity referring to the generalizability of findings from one group to another group.” W. Paul Vogt, *Dictionary of Statistics and Methodology* 78 (Sage Publications 1993).

⁵⁰ Dror, *supra* note 48, at 1034.

⁵¹ *Id.* at 1037.

⁵² *Id.* at 1038.

⁵³ *Daubert* discussed the known or *potential* rate of error. See also The American Association for the Advancement of Science (AAAS) recently published a study on latent fingerprints (William Thompson et al., *Forensic Science Assessments: A Quality and Gap Analysis* (2017)) that discussed the concept of “convergent validity,” an approach that draws conclusions about method validity from the body of relevant literature *as a whole*, recognizing that various study designs have different strengths and weaknesses. It also recognized that some studies can reinforce others and collectively support conclusions not otherwise warranted. Thompson, at 44. See also NAT'L RESEARCH COUNCIL, NAT'L ACADS., *THE EVALUATION OF FORENSIC DNA EVIDENCE* 85, 87 (1996) (“The question to be

to the scientific limitations imposed by external/ecological validity). For example, the composite false positive error rate derived from extant firearms-toolmarks studies is at or below 1%—a rate consistent with that detected by the largest latent fingerprint study to date.⁵⁴

The Department provided the Committee with the results of an ongoing firearms-toolmarks experiment by Mark Keisler and Stacey Hartman, *Isolated Pairs Research Study*, 50 AFTE Journal 56 (Winter 2018). The false positive error rate for that study is currently zero. This finding is consistent with the low false positive error rates recorded by numerous research studies in the firearms-toolmarks discipline of various experimental design.

A new firearms-toolmarks open-set black box study conducted by Jamie A. Smith was recently accepted for publication in the peer-reviewed *Journal of Forensic Sciences*.⁵⁵ The study was undertaken in response to the PCAST Report’s criticism of closed set experimental designs used in some past firearms-toolmarks studies. Smith’s study involved 72 qualified firearms examiners who compared bullets fired from 30 consecutively manufactured barrels (which makes comparisons much more difficult than those typically encountered during casework). The study’s false positive error rate was calculated to be 0.08% with only 1 false association recorded in 1,250 comparisons.⁵⁶

Finally, consider that the PCAST Report said the following about forensic error rates: “To be considered reliable, the FPR [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.”⁵⁷ The extant studies (including black box and other designs) for firearms-toolmarks and latent fingerprints consistently record false positive error rates at or less than 1%—well below PCAST’s recommended 5% upper threshold.

A table of firearms-toolmarks studies that have measured false positive error rates for examiner-participants who conducted forensic comparisons of spent bullets and/or shell casings is appended to this letter as Attachment A.

decided is not the general error rate for a laboratory or laboratories over time but rather whether the laboratory doing DNA testing in this particular case made a critical error.”) and (“The risk of error is properly considered case by case, taking into account the record of the laboratory performing the tests, the extent of redundancy, and the overall quality of the results”).

⁵⁴ For latent prints, in the largest-scale study to date, involving 169 examiners and 17,121 total decisions, the false positive error rate was 0.1%. Bradford Ulery et al., *Accuracy of Forensic Latent Fingerprint Decisions*, 108 Proceedings of the National Academy of Sciences 7733-38 (2011).

⁵⁵ Jamie A. Smith, Beretta Barrel Fired Bullet Validation Study, *Journal of Forensic Sciences* (accepted for publication October 2, 2020).

⁵⁶ It is important to note that experimental study error rates do not translate to laboratory error rates, as comparisons performed during studies do not have the benefit of verification performed by a second examiner or a laboratory’s quality assurance measures. In this regard, see BALDWIN ET AL., A STUDY OF FALSE POSITIVE AND FALSE NEGATIVE ERROR RATES IN CARTRIDGE CASE COMPARISON 18 (2014), <https://www.ncjrs.gov/pdffiles1/nij/249874.pdf>: (“This finding [a 1.0% false positive error rate] does not mean that 1% of the time each examiner will make a false-positive error. Nor does it mean that 1% of the time laboratories or agencies would report false positives, since this study did not include standard or existing quality assurance procedures, such as peer review or blind reanalysis.”).

⁵⁷ PCAST Report, *supra* note 12, at 152.

The Transactional Cost of a Rule Change to FRE 702

During the October 2017 roundtable that the Committee hosted in Boston, there seemed to be a consensus among participants that Rules 702 and 104(a) already provide the correct standard by which courts should assess the admissibility of expert testimony. The discussion was more focused on whether there was value in tweaking the rules to emphasize that courts should follow the existing rules, and in so doing, use the rule change to more broadly discuss the topic in a committee note. On the issue of admissibility versus weight, Judge James O. Browning—a participant in the Boston roundtable—subsequently wrote the following in a published opinion:

Rule 702’s most prominent hurdle is the sufficiency of basis. Yet the judiciary’s uncomfortableness with analyzing an opinion’s basis can be seen in the conflict in the cases. The current conflict is whether the questions of sufficiency of basis, and of application of principles and methods, are matters of weight or admissibility. *** There should not be a conflict. Rule 702 states that these are questions of admissibility. Yet many courts treat them as questions of weight. *** The Court is concerned that the federal courts will overact to the wayward opinions that have created a split whether sufficiency of basis and application of methods is for the court or goes to the evidence’s weight. The Court is concerned that the federal courts are going in the direction of new rules. *** The development of new rules burdens the federal judiciary and the bar -- all of which are overworked -- with mandatory changes each year, often constituting little more than stylistic changes. Everyone has to get new rule books every year. The burden of new rules often does not justify the meager benefits of the changes.

Walker v. Spina, et al, Civil Action No. 17-0991 JB\SCY (D.N.M. Jan. 9, 2019) (Doc. 111), p. 32, n. 11 (internal citations omitted).

Judge Browning’s observation is especially apt here, where proposed textual changes are not strictly necessary, but open the door to sweeping commentary in the note. Here, the proposed note is already obsolete, and would only become further outdated by the time an amendment takes effect. Forensic science is a quickly evolving discipline where new studies constantly add to a growing body of knowledge. *See, e.g., Harris, supra* at *2 (“recent advancements in the field in the four years since the PCAST Report address many of Mr. Harris’s concerns). Studies conducted in the last few years already undermine the lead premise of the proposed note, *i.e.*, that jurors overvalue forensic testimony. Given the swift pace of forensic and social science research, the slow pace of rulemaking, and the permanence of Committee notes, we propose restraint. Other methods exist to educate courts on the correct application of Rule 702. The language of the Federal Rules already provide courts the tools necessary to regulate expert testimony, and many courts are actively doing so.

Respectfully,

/s/ Elizabeth J. Shapiro
Elizabeth J. Shapiro, Deputy Director
Ted R. Hunt, Senior Advisor on Forensic Science
U.S. Department of Justice

Appendix A

Significant Firearms-Toolmarks False Positive Error Rate Studies

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

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

^a The error rate reported by the author appears to be (1-True Positive Rate). There were three false positive identifications made but the number of true negative comparisons is not reported. 259

correct positive identifications were made. The False Discovery Rate (FDR) for the study is $3/(3+259) = 1.1\%$.

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

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

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

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

^f The empirically observed false positive rate is 0%. Using Bayesian estimation methods, the authors' most conservative (worst case) estimate of the average examiner false positive error rate for the study is .053% with a 95% credible interval of $(1.1 \times 10^{-5}\%, 0.16\%)$.

List of References

1. Brundage, D. (Summer 1998). The Identification of Consecutively Rifled Gun Barrels, *AFTE Journal*, 30(3), 438-44 (Bullets).
2. Bunch, S.G., & Murphy, D.P. (Spring 2003). A Comprehensive Validity Study for the Forensic Examination of Cartridge Cases, *AFTE Journal*, 35(2), 201-03 (Cartridge Cases).
3. DeFrance, C.S. & Van Arsdale, M.D. (Winter 2003). Validation Study of Electrochemical Rifling, *AFTE Journal*, 35(1), 35-37 (Bullets).
4. Smith, E.D. (Fall 2004). Cartridge Case and Bullet Comparison Validation Study with Firearms Submitted in Casework, *AFTE Journal*, 36(4), 130-35 (Bullets and Cartridge Cases).
5. Hamby, J.E., Brundage, D.J., & Thorpe, J.W. (Spring 2009). The Identification of Bullets Fired from 10 Consecutively Rifled 9mm Ruger Pistol Barrels: A Research Project Involving 507 Participants from 20 Countries, *AFTE Journal*, 41(2), 99-110 (Bullets).
6. Lyons, D.J. (Summer 2009). The Identification of Consecutively Manufactured Extractors, *AFTE Journal*, 41(3), 246-56 (Cartridge Cases).
7. Mayland, B. & Tucker, C. (Spring 2012). Validation of Obturation Marks in Consecutively Reamed Chambers, *AFTE Journal*, 44(2), 167-69 (Cartridge Cases).
8. Cazes, M. & Goudeau, J. (Spring 2013). Validation Study Results from Hi-Point Consecutively Manufactured Slides, *AFTE Journal*, 45(2), 175-77 (Cartridge Cases).

9. Fadul Jr., T.G., Hernandez, G.A., Wilson, E., Stoiloff, S., & Gulati, S. (Fall 2013). An Empirical Study to Improve the Scientific Foundation of Forensic Firearm and Tool Mark Identification Utilizing 10 Consecutively Manufactured Slides, *AFTE Journal*, 45(4), 376-93 (Cartridge Cases).
10. Fadul Jr., T.G., Hernandez, G.A., Wilson, E., Stoiloff, S., & Gulati, S. (December 2013). An Empirical Study to Improve the Foundation of Firearm and Tool Mark Identification Utilizing Consecutively Manufactured Glock EBIS Barrels with the Same EBIS Pattern. <https://www.ncjrs.gov/pdffiles1/nij/grants/244232.pdf> (Bullets).
11. Stroman, A. (Spring 2014), Empirically Determined Frequency of Error in Cartridge Case Examinations Using a Declared Double Blind Format, *AFTE Journal*, 46(2), 157-75 (Cartridge Cases).
12. Baldwin, D.P., Bajic, S.J., Morris, M., & Zamzow, D. (April 7, 2014). A Study of False-Positive and False-Negative Error Rates in Cartridge Case Comparisons. <https://apps.dtic.mil/dtic/tr/fulltext/u2/a611807.pdf> (Cartridge Cases).
13. Kerkhoff, W. et al. (2015). Design and Results of an Exploratory Double Blind Testing Program in Firearms Examination, *Science & Justice*, 55, 514-19 (Bullets and Cartridge Cases).
14. Smith, T.P., Smith, A.G., & Snipes, J.B. (July 2016). A Validation Study of Bullet and Cartridge Case Comparisons Using Samples Representative of Actual Casework, *Journal of Forensic Sciences*, 61(4), 939-45 (Cartridge Cases).
15. Duez, P. et al. (July 2018). Development and Validation of a Virtual Examination Tool for Firearm Forensics, *Journal of Forensic Sciences*, Vol. 63(4), 1069-1084 (Cartridge Cases).
16. Keisler, M. et al. (Winter 2018). Isolated Pairs Research Study, *AFTE Journal*, 50(1), 56-58 (Cartridge Cases).
17. Hamby, J. et al. (March 2019). A Worldwide Study of Bullets Fired From 10 Consecutively Rifled 9MM Ruger Pistol Barrels—Analysis of Examiner Error Rates, *Journal of Forensic Sciences*, 64(2), 551-57 (Bullets).
18. Smith, J. (2020). Beretta Barrel Fired Bullet Validation Study, *Journal of Forensic Sciences* (accepted for publication October 2, 2020) (Bullets).

PAPER

Criminalistics

Beretta barrel fired bullet validation study

Jaimie A. Smith MS

Prince George's County Police
Department, Forensic Science Division,
Firearms Examination Unit, Landover, MD,
USA

Correspondence

Jaimie A. Smith MS, Prince George's
County Police Department, Forensic
Science Division, Firearms Examination
Unit, 7600 Barlowe Road, Landover, MD
20785, USA.

Email: Jasmith1@co.pg.md.us

Funding information

Funding provided by Collaborative Testing
Service (CTS).

Abstract

A report published in 2016 by the President's Council of Advisors on Science and Technology (PCAST) criticized studies that have been published regarding the discipline of firearm identification. This study was designed to answer some of these criticisms and involved 30 consecutively manufactured Beretta brand 9 mm Luger caliber barrels. This study had an "open set" design to help the discipline of firearm identification establish "Foundational Validity" which is outlined in the PCAST report. Seventy-two qualified firearm examiners completed and submitted answers for this study that included 15 knowns and 20 unknowns. There were an additional 5 firearms with similar characteristics as the Beretta barrels that were also included as unknowns which provided "known non-match" comparisons. Test sets were created using the random function in Microsoft Excel. Collaborative Testing Services (CTS) funded, facilitated, distributed the tests, and collected the answers from qualified firearm examiners throughout the United States and the world. Firearm examiners were able to complete the test of fired bullets with a low error rate. The error rate for the corrected data was 0.08% (1 in 1250) with the lower confidence interval as low as 0.01% (1 in 10,000) and the upper confidence interval being as high as 0.4% (1 in 250).

KEYWORDS

barrels, Beretta, comparison, consecutively manufactured, error rate, firearm identification, fired bullets, foundational validity, microscopic examination, PCAST, validation study

2 | INTRODUCTION

Firearm and toolmark identification is a discipline within forensic science whose primary objective is to determine if a fired bullet or fired cartridge case was fired in a specific firearm or the same firearm by comparison to each other if a suspected firearm is not submitted. A firearm examiner can determine if a fired bullet from a victim or from a crime scene was fired from a specific firearm that was recovered at a scene or from a suspect. If no firearm is recovered, a firearm examiner can determine how many firearms were discharged at the scene. A firearm examiner microscopically evaluates fired evidence using an optical comparison microscope and observes the stria on the bearing surface of a fired bullet. These striae are marked on the bullet as it travels down the barrel of the firearm. They are accidental in nature and occur because of random imperfections within the barrel of the firearm. The patterns of these striations are considered by firearm examiners to be

unique. Many studies have been published supporting the idea that the striations on a bullet are unique (1–11). The striations are considered unique because the rifling tools during barrel manufacturing wear during their use and change microscopically. The greatest similarities between two barrels would be expected to occur in two barrels that were manufactured by the same rifling tool consecutively. There have also been many studies of a firearm examiners ability to differentiate evidence involving consecutively manufactured tools (2,3,5,11–38). Even though there is strong evidence supporting the discipline of firearm identification, there have been some expected criticisms considering the subjective nature of the analysis.

In 2009, the National Academy of Science Report (NAS) questioned the scientific validity of firearm and toolmark identification (39). Additional studies have been published after the NAS report that help support the scientific validity of firearm and toolmark comparisons (11,27–38,40–47). However, in September of 2016, the

Executive Office of the President President's Council of Advisors on Science and Technology (PCAST) published a Report to the President titled: Forensic Science in the Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods (48). It criticized several different forensic disciplines as well as the scientific validity of firearm and toolmark identification. In this report, PCAST outlines reasons they believed firearm/toolmark examinations did not meet the scientific criteria for "foundational validity". PCAST coined and defined the term "Foundational Validity". According to PCAST, since firearm identification is a feature-comparison method, its foundational validity can only be established through multiple independent black box studies ([48, p. 68]). In order to meet the scientific criteria for foundational validity, PCAST states that the following criteria must be met:

1. Studies must involve sufficiently a large number of examiners and be based on sufficiently large collections of known and representative samples from relevant populations to reflect the range of features or combination of features that will occur in the application.
2. Empirical studies should be conducted so that neither the examiner nor those with whom the examiner interacts have any information about the correct answer.
3. Study design and analysis framework should be specified in advance.
4. The empirical studies should be conducted or overseen by individuals or an organization that do not have a stake in the outcome of the studies.
5. Data, software, and results of the validation studies should be available to allow other scientists to review the conclusions.
6. To ensure that conclusions are reproducible and robust, there should be multiple studies by separate groups reaching similar conclusions. ([48, pp.52-53])

PCAST reviewed several studies that have been conducted in the field of firearm/toolmark identification in the past 15 years. They stated that many of the studies were not appropriate for assessing scientific validity and estimating the reliability because they employed artificial designs that differ in important ways from the problems faced in casework ([48, p.106]). These studies employed a "closed set" design where the source firearm is always present. They stated that the closed-set design is problematic in principle and underestimates the false positive rate in practice ([48, p.106]). Therefore, PCAST concluded that this design is not appropriate for assessing scientific validity and measuring reliability ([48, p.109]).

In order to address this criticism, more "open set" studies need to be conducted to have a black-box study that meets the scientific criteria for "foundational validity" set forth by PCAST as much as possible for firearm and toolmark identification.

With this goal in mind, the author's laboratory obtained 30 consecutively manufactured Beretta 9 mm Luger caliber barrels. These Beretta barrels were obtained by the laboratory in 1996 from Beretta

Highlights

- PCAST criticized firearm identification because of the few studies to support "Foundational Validity".
- A study of 30 consecutively manufactured Beretta barrels was created to address the concerns of PCAST.
- This test uses an "open set" design which was deemed appropriate by PCAST.
- CTS was used as a third party so that the participant did not communicate with the test designer.
- A low error rate was observed for firearm examiners when comparing fired bullets for this study.

U.S.A. Corp. of Accokeek, Maryland with the intent of performing a consecutively manufactured study. Given that the barrels were obtained in 1996, no one from the laboratory was present during the collection of the barrels and there is no formal documentation other than a packing list. The barrels are stamped numerically from 1 to 30 indicating the order of production. This experiment will provide participants in this study with a selection of known test standards from the 30 consecutively manufactured barrels and also provide them with 20 unknowns (a sample where the participant needs to determine if the bullet was fired from one of the barrels provided or some other barrel).

This experiment will be set up similarly to the Ten Consecutive Manufactured Ruger Barrel Study by James Hamby (49); however, instead of a "closed set", it will be an "open set". In an "open set", the participant should have no expectation that all questioned bullets should match one or more of the unknowns. Only firearm examiners who were qualified to do work by their laboratory were selected to participate in this experiment. There was an administrative section with several questions that each participant filled out, such as, years of experience in the field, type of lighting, type of scope, laboratory accreditation, certification, etc.

Two hundred tests were created for this study. Within the 200 tests, there were 20 different answer keys of 10 sets each. The 30 consecutive Beretta manufactured barrels and 5 "known non-matching" (in this study, "known non-match" refers to a bullet fired from a barrel that is not present in the provided knowns) 9 mm Luger caliber firearms with similar rifling characteristics as the 30 consecutive barrels from the laboratory's reference collection were included in the test sets. Each set of 10 was determined using the random number function present in Microsoft Excel. The random number function was generated and then repeated for the next 19 unknowns for each test set. Using this process for the 20 unknowns, it was possible to have multiple bullets from the same barrel. It was also possible for the unknown bullets to have been fired in a barrel which did not correspond to any of the knowns.

3 | MATERIALS AND METHODS

Thirty consecutively manufactured barrels were obtained from Beretta U.S.A. Corp in January of 1996 by a local laboratory. These barrels have been test fired many times, so there was no concern a “break-in” period would significantly affect the test samples. A “break-in” period is a short period after the barrel has been manufactured where several bullets have to be fired in the barrel before the striations mark in a reproducible manner (16,18). There were five additional pistols used in the test structure to provide “known non-match” fired bullets. All of the pistols have similar general rifling characteristics (GRC) to the known Beretta barrels that were provided. The general rifling characteristics (GRC) were six lands and grooves with a right hand twist where the land impression widths ranged from 0.072 to 0.076 inches and the groove impression widths ranged from 0.100 to 0.106 inches. The following pistols were used: Beretta model 92F, Ruger model P85 MKII, FEG model PJK-9HP, Fabrique Nationale model Hi-Power, and CZ model 75.

For this study, over 14,000 9 mm Luger caliber Federal FMJ cartridges with Lot# AE9AP were obtained and test fired through the barrels.

Figure 1 is a simplified flow chart to help visualize the procedure of how the test sets were created in this study. Each barrel/pistol was lubricated and cleaned prior to test firing the test set (there were approximately 400 bullets fired through each known barrel). Ten percent of the fired bullets were verified, by an AFTE certified firearm examiner, to display sufficient microscopic individual characteristics for identification. Prior to the firing process, every 10th bullet (1, 11, 21, 31, 41, etc) was marked with a sharpie for microscopic comparison to other fired bullets in that set of 100. The ten bullets from each set of 100 were intracompared. A bullet from each set of 100 was then microscopically intercompared to a bullet from each of the 4 sets of 100. Therefore, all of the bullets from 1 to the total number of bullets fired for that barrel should be identifiable; however, not all fired bullets were microscopically compared. A dry patch was run down the barrel after each set of 100 test fires.

After all 30 Beretta barrels were fired, the “known non-matching” pistols received from the laboratory's Firearms Reference Collection were fired using the same process outlined above; however, only about 100 bullets were fired through these pistols because the known exemplars did not need to be fired and therefore, lessened the number of test fires needed.

The Beretta barrels used in this study were manufactured using a broaching tool (50). Since the potential for subclass characteristics may be present, the procedure Ronald Nichols outlined in his journal article (51) was utilized. A cast was made from the muzzle to the chamber of the 30 Beretta barrels using Forensic Sil casting material. The cast was then cut in half and the muzzle end of the cast was compared to the chamber end of the cast. This comparison was conducted by an AFTE certified firearm examiner and no subclass characteristics were observed. Due to the exorbitant cost of making the cast, it was not possible to ship casts of the barrel to each examiner. If any participant asked about the potential for subclass

characteristics, they were told this method had been utilized to verify, there were no subclass characteristics.

Each test consisted of a set of three fired bullets each fired from 15 known standards (numbered 1 through 15) and 20 unknowns (labeled A through T). The random number generator feature on Microsoft Excel was used to determine the test sets. The function used to create the random number was `RANDBETWEEN(x,y)` where x is the lowest number and y is the highest number. Excel could select any number between x and y . This means that there could be multiple unknowns from the same barrel whether it is from a known barrel or an unknown non-matching barrel.

There were two sets of tests: the first set included barrels from 1 to 15, barrels 16 and 17 (not provided in this test as a known), the Beretta model 92F pistol, the Ruger model P85 MKII pistol, and the FEG model PJK-9HP pistol. The second set included barrels from 16 to 30, barrel 14 and 15 (not provided in this test as a known), the Beretta model 92F pistol, the FN model Hi-Power pistol, and the CZ model 75 pistol.

Once all of the test firing was completed, the bullets were scribed according to the Excel spreadsheet and packaged to be sent to Collaborative Testing Services (CTS). For each known of a particular test set, each bullet was scribed with the barrel number, and the set of standards were packaged into a coin envelope labeled with the barrel number. These knowns were placed in a large zip top plastic bag with the test set range (#1-#10, #11-#20, etc.) and the barrel number written on the bag. After all of the knowns were scribed for a particular barrel, the unknowns for that barrel were scribed with the appropriate letter, packaged in a coin envelope with that letter written on it, and put in a small zip top bag labeled with the test range and the appropriate letter. This procedure was performed for all 30 barrels.

For the fired bullets from barrels where a corresponding known was not present, the bullet was scribed with the appropriate letter and packaged in a coin envelope with the letter written on it and put in a small zip top bag with the test set range and the appropriate letter. For each test set, a large zip top plastic bag was labeled with the test set range and that it contained unknowns without a known present, incorporating “known non-match” in the test design.

Therefore, there were 15 large zip top plastic bags for each test set which contained the fifteen knowns (labeled 1-15 or 16-30) and unknowns (labeled A-T). In addition, there was one large zip top plastic bag labeled with the test set range and “unknowns without a known present” written on it.

These test sets were then sent to CTS for packaging and shipment. CTS assigned each test set a unique webcode. If more than one test set was ordered by a specific laboratory, different test sets were sent. This meant that no examiner in the same laboratory would have the same test. CTS managed communication with all of the participants in the study. At no time did the developer of the test know which particular tests were received by the participants.

The procedure outline below was the procedure that CTS used to package the test:

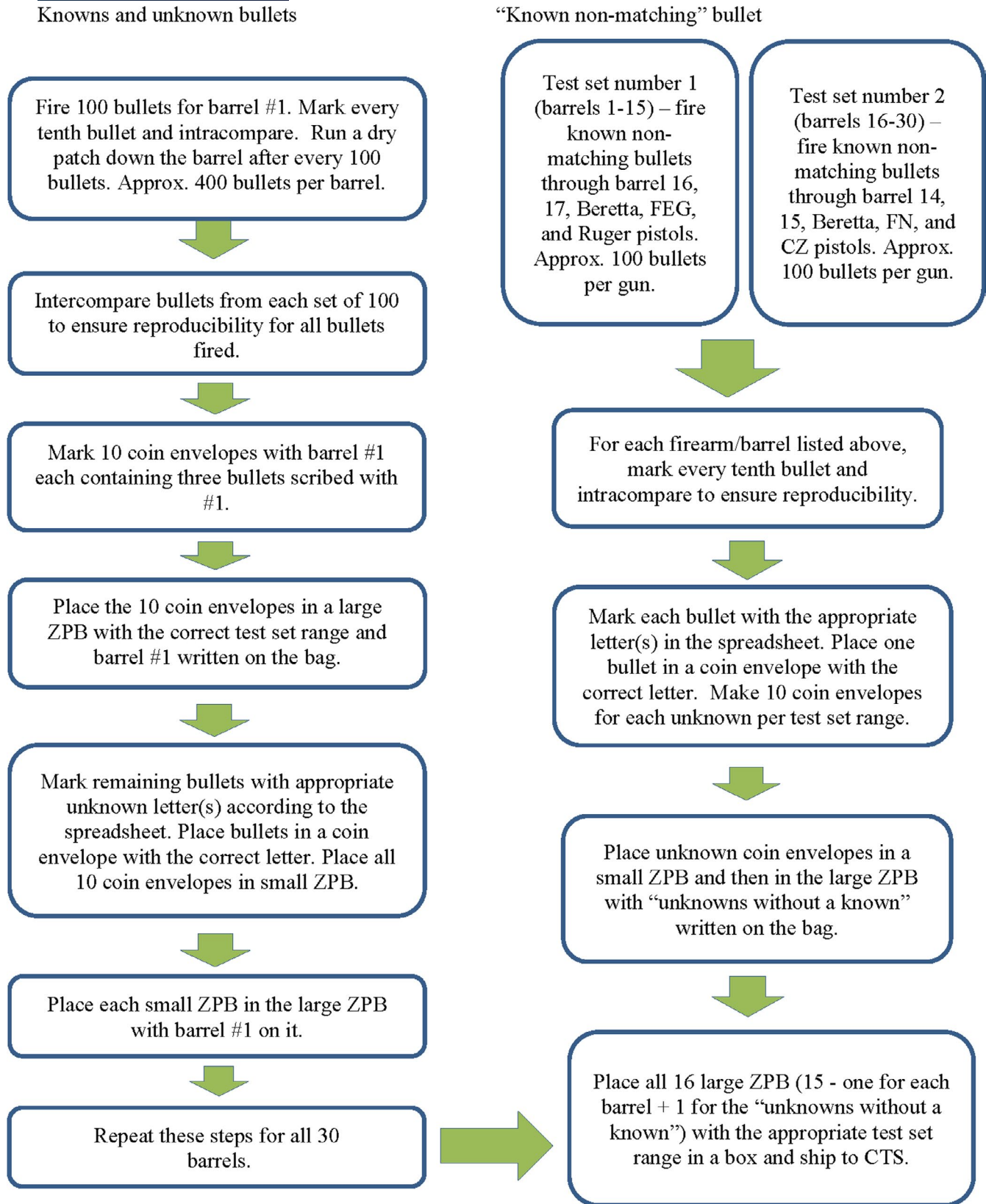


FIGURE 1 Simplified flow chart for procedure to create the test sets

1. With approximately 120 participants over a generated 200 Kits, the participants were spread out as evenly as possible, by utilizing up to 6 kits from each set of 10. Participants were

assigned a random alpha-numeric 6 digit code (WebCode). This was sorted alphabetically and the kits were assigned numerically to this sorted list.

2. CTS received boxes of the materials for the Kits in 10 kit ranges.
3. CTS unpacked the bags of Known and Questioned envelopes and laid them out on tables for the required number of kits to be used per range. As stated above this was approximately 6 per range.
4. The attached picture illustrates one of the multiple stations that were set up to lay out the envelopes as they were unpacked from the provided bags. The known bullets were numerical, so no assistance in laying them out was used. However, to assist with the Questioned Bullets, paper with the alphabetical range was laid down so that no letter was missed during unpacking.
5. Once all the envelopes were laid out from the provided bags, it was verified that all items were present on the table for all of the necessary kits.
6. Then the full range of envelopes were picked up and packaged into the appropriately labeled zip top bag.
7. The kit ranges and their assigned webcodes were checked prior to laying out the samples, after they were packaged into the zip top bags, and again when the zip top bags were placed inside of a sample pack box.

Each participant would receive a box from CTS with a label containing the participant number and the appropriate webcode. Within that box, there would be 15 coin envelopes containing three bullets from each of the test standards (either labeled #1 through #15 or labeled #16 through #30) and 20 coin envelopes containing one bullet from an unknown (questioned) sample labeled letter A through T.

For test set number 1 (barrels #1-#15) and test set number 2 (barrels #16-#30), an average of 22% of the unknowns provided did not have a corresponding known provided. The first test set ranged from having three unknowns (15%) not provided to having seven unknowns (35%) not provided. While the second test set ranged from having three unknowns (15%) not provided to having six unknowns (30%) not provided. The number of duplicates for test set number 1 and number 2 range from two to five. The number of triplicates for test set number 1 and number 2 range from zero to two. Because of the importance of the consecutive nature of this study, the number of unknowns provided from consecutively produced barrels within each 15 barrel grouping was reviewed. For test set number 1, the number of unknowns from consecutively produced barrels ranged from 7 to 10 barrels and for test set number 2, it ranged from 6 to 13 consecutive barrels; however, the set with six (6) unknowns from consecutive barrels also had another set of 5 unknowns from another subgroup of consecutively produced barrels.

4 | RESULTS

After soliciting qualified examiners from the firearm examination community, there were a total of 110 participants who volunteered to receive the test and participate. All of the data was collected by CTS via their website; there were 74 participants (67.3%) who submitted results.

From the tests distributed, there were 1149 possible identifications to a known barrel, 151 possible identifications to another bullet present in the unknowns that are not present in the knowns, and 180 true eliminations (bullet where a known or another unknown is not present in the test). Therefore, there was a total of 1300 possible identifications and 6120 true eliminations ($180 * 34 [15 \text{ knowns} + 19 \text{ unknowns}] = 6120$).

Upon initial submission of the test results, there were 7 false identifications, 18 false eliminations, 23 missed identifications when the known was present and 22 missed identifications when only the unknown was present. See Table 1 for the data associated with results. In Table 1, the percentage of false identifications was calculated by dividing the number of false identifications by the number of correct identifications.

After looking at several of the false identification responses, it was realized that two examiners appeared to have incorrectly transferred the information into the wrong cell on the CTS website. One examiner made four false identifications because they had transposed the letters. On the answer sheet, the examiner had identified one of the unknowns to a specific barrel and then included other unknowns that had been identified to a different barrel. Another examiner made one false identification which was off by one letter; this would indicate that they read the wrong letter when filling in the answer sheet. A generic letter was sent by CTS to the participants who had incorrect responses stating that it was believed that they had made a typographical error and had ended up identifying one bullet to two different barrels. Below is the text of the email that was sent:

"It was noticed that there is an entry that appears to be a transcription error because there was an entry with more than one identification and your answers reference two different barrels. Any clarification that you could provide would be appreciated."

A response to the email was received from both examiners and their email response identified where the error was and what the correct answer should have been.

Another false identification was a typographical error. In the answer sheet, an unknown was identified as having been fired from barrel #1; however, barrel #1 was not one of the barrels provided for

TABLE 1 Error calculation based on original data submission

Type of error	Number	Total ^a	Error rate (%)
False identification	7	1251	0.56
False elimination	18	10935	0.16
Total (false identification and false elimination)	25	12186	0.21%
Missed identification (known present)	23	1251	1.84
Missed identification (unknown present)	22	1251	1.76

^aThe information for identifications was always filled in; however, for the false elimination data, some examiners left the area blank.

that test set. The barrels provided for that test set were barrels #16 through #30. Therefore, this had to be a typographical error. Two emails were sent to try and clarify what the correct response should have been; however, no response was received.

From the text described above, it is reasonable to determine that the five transferring errors and the one typographical error are administrative in nature and therefore, should not be counted as false positives. Since these tests were not technically or administratively reviewed, which is part of the normal process in most forensic laboratories, these errors would likely have been discovered during the administrative review process. For the results submitted, there was one false identification. Therefore, corrected responses from this test are in Table 2. In Table 2, the percentage of false identifications was calculated by dividing the number of false identifications by the number of correct identifications.

There were 18 false eliminations present in the study. These false eliminations were made by six examiners. Four examiners were responsible for 16 of the false eliminations (8, 3, 3, and 2), and two examiners made one false elimination each. The false elimination response in Tables 1 and 2 were calculated based upon the total number of eliminations present because not all examiners filled in the area designated for eliminations. This area was left blank by many examiners because most firearm examiners do not feel it is necessary to eliminate all other firearms if they have made an identification to a specific firearm.

After calculating the overall error rates of the examiners, the sensitivity and specificity were also calculated. Sensitivity is the number of identifications reported divided by the number of identifications present in the test. The number of identifications submitted in this test was 1251 and the identifications present in this test was 1300. Therefore, the sensitivity of this test is 96.2%. The specificity is the number of eliminations reported divided by the number of eliminations present in the test. The number of eliminations reported in this test was 10,935 and the number of eliminations present in this test was 47,876. Therefore, the specificity of this test is 22.8%. While the specificity of this test is on the low side, possible reasons are explained in the discussion.

TABLE 2 Error calculation based on corrected data from participants

Type of error	Number	Total ^a	Error rate (%)
False identification	1	1257	0.080
False elimination	18	10935	0.16
Total (false identification and false elimination)	19	12192	0.16
Missed identification (known present)	23	1257	1.83
Missed identification (unknown present)	22	1257	1.75

^aThe information for identifications was always filled in; however, for the false elimination data, some examiners left the area blank.

5 | DISCUSSION

The overall goal of a consecutive manufactured barrel study is to support the firearm identification community with scientific studies that show qualified firearm examiners can identify a fired bullet or fired cartridge case to a specific firearm within a small degree of error. The consecutively manufactured study is a “worst case scenario” where multiple barrels are manufactured consecutively (one after the other) with the same tool at the factory. In this and other consecutively manufactured studies, a firearm examiner can identify an unknown bullet to the correct barrel with a very low error rate. PCAST and other critics have found fault with many of the previous studies.

The first criterion that PCAST outlined: in order to establish foundational validity was the studies need to include a sufficiently large number of examiners and have large collections of representative samples that are typically found in casework. This is the largest consecutively manufactured barrel study known to date. Prior to this study, 10 consecutively manufactured barrels was the largest study that had been completed [3, 5, 6, 9, 11, 16, 20, 49]. Seventy-four examiners of the 110 that signed up completed the test (67.3%). This result is similar to other studies, such as the Ames Study where 218 out of 284 (76.8%) examiners participated (40) and the Smith study where 31 out of 47 (65.9%) examiners participated [41].

Since there are approximately 1200 firearm examiners (AFTE membership: Provisional [304], Regular [685] and Distinguished [174]) throughout the world, the number of participants in this study would have incorporated 6.3% of the firearm examiner in the world. This is obviously lower than desired; however, to be expected given the study had a large number of knowns and unknowns, it required a significant amount of time to complete the task. Since many firearm laboratories throughout the country and world have large backlogs and minimum manpower, it is reasonable to conclude participation could put an undue strain on their laboratories and participation would not be permitted by the employer in most cases. Also, examiners who would eagerly volunteer must manage time effectively and choose which studies to participate in because casework is still the priority.

In this study, Beretta barrels and pistols present in the Firearms Reference Collection were used. Many people purchase firearms chambered for the 9 mm Luger cartridge including the military, police departments, and civilian consumers for home defense. Since 1999, more than 44,000 firearms have been submitted to the firearm identification section of a local laboratory in a variety of different types of cases. Of those 44,000 firearms, more than 12% of those firearms have been chambered in 9 mm Luger caliber. Beretta is a popular manufacturer and they manufacture many different firearms chambered for the 9 mm Luger cartridge. For many years, the local police department used the 9 mm Luger Beretta model 92FS as their duty weapon. Beretta manufactured firearms are also commonly found in casework. Of the 5365 firearms chambered in 9 mm Luger submitted to the local police department since 1999, 515 of them were manufactured by Beretta. Therefore, Beretta accounted for approximately 9.6% of the 9 mm Luger submitted firearms. All of the pistols

selected for the unknowns were from the local laboratory's firearms reference collection. The local laboratory's firearms reference collection is a collection of firearms that have been seized during police investigations that occurred within the county. Therefore, all of the firearms used in this study are often seen in casework.

The second criterion for PCAST was: Empirical studies should be conducted so that neither the examiner nor those with whom the examiner interacts have any information about the correct answer. In this study, this criterion was met by a company called Collaborative Testing Services, Inc. (CTS). CTS is a company widely known throughout the forensic community as a proficiency test provider. All qualified firearm examiners filled out an application and submitted the application to CTS which served as the main point of contact for all of the participants in this study. CTS determined which tests were going to be shipped to the participant. In the event that a technical question needed to be answered, the test developer was contacted through CTS. In that event, the test developer did not know which specific test the participant was given because the webcode did not correlate to any information the test developer had.

The third criterion for PCAST was: Study design and analysis framework should be specified in advance. The study design and analysis framework were specified in advance. The local laboratory in collaboration with CTS specified in advance the design and analysis framework of the study. This was necessary so both parties knew and understood their responsibilities.

The fourth criterion for PCAST was: The empirical studies should be conducted or overseen by individuals or an organization that do not have a stake in the outcome of the studies. The study was conducted and overseen by CTS. In its capacity in this study, CTS served as the administrator of the test. CTS had no stake in the outcome of results of this study. CTS collected all of the answers submitted via their website and then forwarded the responses to the developer of the test.

The fifth criterion for PCAST was: Data, software, and results of the validation studies should be available to allow other scientists to

review the conclusions. The test materials and results of this validation study are available upon request.

The sixth criterion for PCAST was: To ensure that conclusions are reproducible and robust, there should be multiple studies by separate groups reaching similar conclusions. This study, along with many other studies that are currently being distributed, will help ensure that the conclusions are robust and reproducible. This study reaches similar conclusions previous studies have demonstrated which is that within a low error rate, firearm examiners are able to identify an unknown bullet to a specific firearm.

Along with the criterion described above, PCAST also found fault with previous studies because they did not incorporate an "open set". As described in the study design, this study incorporated an "open-set" concept. Known non-matching samples were included.

It was suggested in the PCAST report, that a 95% confidence interval be calculated for these studies using the Clopper-Pearson/Exact Binomial method, the Wilson Score interval, the Agresti-Coull (adjusted Wald) interval, and the Jeffreys interval. These calculations were done using the following website <https://epitools.ausvet.com.au/ciproportion>. The data is included in Table 3.

The 95% confidence interval for this study at the upper limit for the corrected results was an error between 0.24% and 0.5%. The 95% confidence interval at the upper limit for the reported results was a range of 0.97%–1.17%. According to sources (52,53), for a study this size, the best confidence interval method calculations would be either the Wilson Score, Agresti-Coull (adjusted Wald), or Jeffreys Interval.

In the PCAST report, it was stated that closed-set studies have inconclusive and false-positive rates that are dramatically lower than those for an open designed study (p. 109). If one includes inconclusive results with false positive answers, the error rate will increase; however, it is inappropriate to include inconclusive results with false positive errors. An inconclusive result is reserved for an examiner when the class characteristics are the same and there are insufficient individual characteristics to reach a conclusion. If the firearm examiner believes that there is not enough

TABLE 3 Calculation of binomial confidence intervals for false identifications for both the original submission and the corrected data

Sample size	Positive number	Confidence	Proportion	Lower 95%	Upper 95%
1258	7	0.95			
Normal			0.0056	0.0015	0.0097
Clopper-Pearson			0.0056	0.0022	0.0114
Wilson			0.0056	0.0027	0.0114
Jeffreys			0.0056	0.0025	0.0109
Agresti-Coull			0.0056	0.0024	0.0117
1258	1	0.95			
Normal			0.0008	0.0008	0.0024
Clopper-Pearson			0.0008	0.0000	0.0044
Wilson			0.0008	0.0001	0.0045
Jeffreys			0.0008	0.0001	0.0037
Agresti-Coull			0.0008	0.0001	0.0050



information on the sample to come to a conclusion, then an inconclusive result is appropriate. Firearm examiners approach these tests as if they are casework; therefore, it would be inappropriate for an examiner to be forced to come to an identification or elimination if sufficient information is not observed on the items in question. A laboratory would not want to have a policy that forces a scientist to render an opinion if there is not enough information to make a determination. The same approach should be used for firearm examiners in this study. Also, inconclusive is neither a correct answer nor an incorrect answer. From the perspective of the defense attorney, this conclusion could be a benefit because it would allow for "reasonable doubt".

As stated above, it is not accurate to include inconclusive answers in the error rate because an inconclusive result is neither positive nor negative. These confidence interval calculations are based upon the theory that the result is either positive or negative, and an inconclusive result is not possible. However, in order to compare information that was published in the PCAST report, below the inconclusive result has been included in the error rate. For the submitted results, if one included false positive and inconclusive results, the results would be 52 out of the 1303 (4.0%) for the submitted result and 46 out of 1303 (3.5%) for the corrected result. When comparing the error rates of the submitted results, the false positive error was 0.56% and when the inconclusive results are included, the false positive and inconclusive error is 4.0% (7-fold increase). When comparing the errors rates of the corrected results, the false positive error was 0.08% and when the inconclusive results are included the false positive and inconclusive error is 3.5% (44-fold increase). This is by far much lower than the 100-fold error reported in the PCAST report ([48, p.11).

Some of the inconclusive results can be explained due to laboratory policies. In the additional questions that were provided with the answer sheet, one of the questions was whether there was a laboratory policy that did not allow examiners to eliminate two items based on differences in individual characteristics. There were 3 examiners who reported that their laboratory prohibited eliminating based on differences in individual characteristics because of a laboratory policy. Two examiners reported that they could only eliminate based on individual characteristics if it was verified by another qualified examiner. Since all of the fired bullets in this study have similar rifling characteristics, an examiner would have to eliminate based upon individual characteristics. For those two examiners who needed verification from another examiner to eliminate an item based on individual characteristics, it is unknown as to whether that examiner requested this procedure for the purposes of this test.

The number of inconclusive results for this study may be higher than other studies. This was a large test with many known samples. There were 15 knowns which typically represents far more knowns than an examiner would evaluate in routine casework. For a comparison of one unknown to the fifteen knowns, the examiner is comparing potentially conducting ninety ($90 = 15 \text{ knowns} * 6 \text{ per bullet}$) land impression comparisons. Therefore, there would be 1800 (90 land

impressions * 20 unknown bullets). In addition, with an average of more than 4 unknowns present per test, there would be potentially 24 comparisons (4 comparisons * 6 land impressions) per unknown for a total of about 1824 comparisons per test.

A sensitivity of 96.2% and specificity of 22.8% were calculated for this test. While the sensitivity is very good, the specificity was evaluated further. Of the 74 examiners who submitted results, many examiners either left the elimination area blank, put "N/A", or did not have a response. If an examiner left the elimination answer blank or put an "N/A", this meant that there were as many as 34 eliminations for one bullet that were missing (depending upon the test set). If the examiner left it blank for all of the bullets in a single test, this would mean that up to 680 ($34 * 20$) eliminations were potentially missing. There were several examiners who would eliminate the knowns, but did not eliminate the unknowns. Therefore, the number of eliminations went from 34 eliminations to 15 eliminations. This could be because the examiners did not realize that they were supposed to eliminate each unknown bullet from all of the unknowns. The normal process in most laboratories in casework is to compare all of the evidence to each other and to the tests, the directions for the study could have been more explicit. As discussed earlier, many firearms examiners did not fill in this area because they do not think it is necessary to eliminate all other firearms if they have made an identification to a specific firearm. Given this information, this perception has skewed the data for specificity for this study.

There were 16 examiners who had an inconclusive result for all of the eliminations in the test. The examiners in this study were asked to follow the AFTE Range of Conclusions and designate which inconclusive result that they were reporting. Below is the definition of inconclusive from the AFTE Range of Conclusions (54):

2. Inconclusive

- a. Some agreement of individual characteristics and all discernible class characteristics, but insufficient for an identification.
- b. Agreement of all discernible class characteristics without agreement or disagreement of individual characteristics due to an absence, insufficiency, or lack of reproducibility.
- c. Agreement of all discernible class characteristics and disagreement of individual characteristics, but insufficient for an elimination.

Of the 16 examiners who gave an inconclusive result, 9 examiners have a result of inconclusive (c), four have a result of inconclusive (b), and two have a result of inconclusive (a).

Therefore, the majority of the examiners who gave an inconclusive result, thought that it was inconclusive (c) and that there was disagreement of individual characteristics; however, just not enough disagreement of individual characteristics to come to a conclusion of an elimination.

There can be several reasons why an examiner would choose an inconclusive result over elimination. As discussed earlier, it may be a laboratory policy not to eliminate based on individual characteristics.

Therefore, five of sixteen examiners who have this result was due to a laboratory policy. Another reason that an examiner may give the result of inconclusive in a test like this is because they feel it was not feasible to determine the reproducibility of the marks. If there are two representations of the bullet fired from a specific barrel, then an examiner can determine what striations are reproducing and what striations are not. Often times in casework, an examiner will compare the tests to each other and the evidence to each other. If the evidence marks consistently and the tests mark consistently and the evidence and tests mark differently, then the examiner can come to the result that the evidence and the tests are from different firearms. However, if there is only one representative of the evidence, this decision becomes more complicated if some of the marks are similar. If this is the case, the conservative approach is for the result to be inconclusive.

For eliminations, there were 18 false eliminations and 10,935 correct eliminations for a false-elimination error rate of 0.16%. Of the 18 false eliminations, eight false eliminations occurred with one examiner (almost half of the errors). In recent journal publications [28, 40], false identifications and false eliminations are calculated separately. As a scientific discipline, it is important for the firearm examiners to pay attention to both false identifications and false eliminations. However, a false elimination is less problematic than a false identification because the subject of an investigation is not going to be imprisoned for a false elimination. After calculating both the false identifications and the false eliminations, total error rate was calculated for this study. The total error was calculated to be 0.21% for the original submission and 0.16% for the corrected results (Tables 1 and 2).

Besides what was discussed earlier, there were other additional questions that asked about the examiner such as, the years of experience, whether the examiner's laboratory was accredited, whether the examiner was certified, and the method the examiner used for the examination (pattern matching, QCMS, or both). All but two of the participants responded to these questions, so this information was based on 72 responses. From this information, the examiners had a range of experience that went from 1 year of experience to 50+ years of experience. The average years of experience was 12.3 years. 91.7% (66) of examiners were from accredited laboratories. 33.3% (24) of the examiners were certified firearm examiners. 92.9% (65) reported that they used pattern matching as the method for their comparison while 7.1% (5) reported that they used both pattern matching and QCMS (Quantifiable Consecutive Matching Stria) (2 of the responses were incomplete). While none of this information appeared to have an effect on the results of the test, it does represent the information pertaining to the background of the examiners in this test.

6 | CONCLUSIONS

This study was designed to respond to many of the criticisms presented in the PCAST report. It was modeled after the requirements outlined in the PCAST report to enable forensic disciplines which analyze impression evidence to establish Foundational Validity.

From the results of this study, trained and qualified firearm examiners throughout the United States and world are able to identify unknown samples to a known barrel in an "open set" format with a very low error rate. This test incorporated 30 consecutively manufactured Beretta barrels. It was divided into two different test sets, but combined results indicate, examiners are able to identify unknown bullets to the correct barrel from 30 consecutively manufactured barrels within a low error rate. Consecutively manufactured barrels are a firearm examiner's "worst case scenario" because a barrel manufactured by the same tool one after the next will have striations that are the most similar and it is more likely that an examiner could make an error. From the data submitted, the false identification error rate of the 74 examiners was 0.55% (1 in 182) with the result for the lower confidence interval as low as 0.2% (1 in 500) and with the upper confidence interval as high as 1.1% (1 in 91). The false identification error rate for the corrected data (data where the typographical errors were corrected) was 0.08% (1 in 1250) with the lower confidence interval being as low as 0.01% (1 in 10,000) and as high as 0.4% (1 in 250) for the upper confidence interval. These error rates are similar to previous studies (which may or may not have followed the model outlined in the PCAST Report) that have been published in the firearms examination discipline indicating that the specific requirements set up by PCAST have little effect on the overall error rates of firearm examiners.

ACKNOWLEDGEMENTS

The author would like to thank Collaborative Testing Service (CTS) for assisting in designing, administering, packaging, distributing, and compiling the answers for this study which includes Cathy Brown and Richard Hockensmith. I would like to acknowledge my fellow laboratory employees to include Director Kristen Lease, Firearms Manager Joseph Young, and all of the examiners in my laboratory for their help and assistance with this project. I would specifically like to thank Corporal Tara Mattingly for helping fire the 14,000 cartridges needed to complete this study. I would also like to thank all of the examiners and their agencies who took time out of their busy day to complete the many comparisons that were needed for this study.

REFERENCES

1. Biasotti AA. A statistical study of the individual characteristics of fired bullets. *J Forensic Sci.* 1959;4(1):34-50.
2. Goddard C. The Valentine's Day massacre: a study in ammunition tracing. *AFTE J.* 1980;12(1):44-59.
3. Lutz M. Consecutive revolver barrels. *AFTE Newslett.* 1970;9:24-48.
4. Murdock J. The effects of crowning on gun barrel individuality. *AFTE Newslett.* 1970;7:12-3.
5. Hall E. Bullet markings from consecutively rifled Shilen DGA barrels. *AFTE J.* 1983;15(1):35-53.
6. Miller J. An examination of two consecutively rifled barrels and a review of the literature. *AFTE J.* 2000;32(3):259-70.
7. Miller J. An examination of the conservative criteria for identification of striated toolmarks using bullet fired from ten consecutively rifled barrels. *AFTE J.* 2001;33(2):125-32.
8. Smith E. Cartridge case and bullet comparison validation study with firearms submitted in casework. *AFTE J.* 2005;37(4):130-5.

9. Brundage D. The identification of consecutively rifled gun barrels. *AFTE J.* 1998;30(3):438-44.
10. Hamby J. Identification of projectiles. *AFTE J.* 1974;6(5-6):22.
11. Monkres J, Luckie C, Petraco NDK, Milam A. Comparison and statistical analysis of land impressions from consecutively rifled barrels. *AFTE J.* 2013;45(1):3-20.
12. Flynn E. Toolmark identification. *J Forensic Sci.* 1957;2(1):95-106.
13. Butcher S, Pugh D. A study of marks made by bolt cutters. *J Forensic Sci Soc.* 1975;15(2):115-26. [https://doi.org/10.1016/S0015-7368\(75\)70965-9](https://doi.org/10.1016/S0015-7368(75)70965-9).
14. Watson D. The identification of toolmarks produced from consecutively manufactured knife blades in soft plastics. *AFTE J.* 1978;10(3):43-45.
15. Cassidy F. Examination of toolmarks from sequentially manufactured tongue and groove pliers. *J Forensic Sci.* 1980;25(4):796-809. <https://doi.org/10.1520/JFS11294J>.
16. Murdock J. A general discussion of gun barrel individuality and an empirical assessment of individuality of consecutively rifled 22 caliber rifles. *AFTE J.* 1981;13(3):84-111.
17. Taira Y. Tire stabbing with consecutively manufactured knives. *AFTE J.* 1982;14(1):50-2.
18. Raven Matty W. 25 automatic pistol breech face tool marks. *AFTE J.* 1984;16(3):57-60.
19. Matty W. A comparison of three individual barrels produced from one button-rifled barrel blank. *AFTE J.* 1985;17(3):64-9.
20. Brown C, Bryant W. Consecutively rifled gun barrels present in most crime labs. *AFTE J.* 1995;27(3):254-8.
21. Lopez L, Grew S. Consecutively machined Ruger bolt faces. *AFTE J.* 2000;32(1):19-24.
22. Eckerman S. A study of consecutively manufactured chisels. *AFTE J.* 2002;34(4):379-90.
23. Lee S. Examination of consecutively manufactured slotted screwdrivers. *AFTE J.* 2003;35(1):66-70.
24. Thompson E, Wyant R. Knife identification project (KIP). *AFTE J.* 2003;35(4):366-70.
25. Bunch S, Murphy D. A comprehensive validity study for the forensic examination of cartridge cases. *AFTE J.* 2003;35(2):201-3.
26. Clow C. Cartilage stabbing with consecutively manufactured knives: A response to Ramirez v. State of Florida. *AFTE J.* 2005;37(2):86-116.
27. Weller T, Zheng A, Thompson R, Tullerners F. Confocal microscopy analysis of breech face marks of fired cartridge cases from 10 consecutively manufactured pistol slides. *J Forensic Sci.* 2012;57(4):912-7. <https://doi.org/10.1111/j.1556-4029.2012.02072.x>.
28. Giroux B. Empirical and validation study: consecutively manufactured screwdrivers. *AFTE J.* 2009;41(2):153-5.
29. Lancon D. Toolmarks in bone: continuing research with consecutively made knife blades. *AFTE J.* 2009;41(2):130-7.
30. Lyons D. The identification of consecutively manufactured extractors. *AFTE J.* 2009;41(3):246-56.
31. Mayland B, Tucker C. Validation of obturation marks in consecutively reamed chambers. *AFTE J.* 2012;44(2):167-9.
32. Cazes M, Goudeau J. Validation study results from Hi-Point consecutively manufactured slides. *AFTE J.* 2013;45(2):175-7.
33. Chu W, Tong M, Song J. Validation tests for the congruent matching cells (CMC) method using cartridge cases fired with consecutively manufactured pistol slides. *AFTE J.* 2013;45(4):361-6.
34. Zheng XA, Soones J, Thompson R, Villanova J, Kakal T. 2D and 3D topography comparisons of toolmarks produced from consecutively manufactured chisels and punches. *AFTE J.* 2014;46(2):143-7.
35. King E. Validation study of computer numerical control (CNC), consecutively manufactured screwdrivers. *AFTE J.* 2015;47(3):171-6.
36. Owens S. An examination of five consecutively rifled Hi-Point 9mm pistol barrels with three lands and grooves left twist rifling to assess identifiability and the presence of subclass characteristics. *AFTE J.* 2017;49(4):208-15.
37. Laporte D. An empirical and validation study of breechface marks on .380 ACP caliber cartridge cases fired from ten consecutively finished Hi-Point model C9 pistols. *AFTE J.* 2011;43(4):303-9.
38. Fadul T, Hernandez G, Stoiloff S, Gulati S. An empirical study to improve the scientific foundation of forensic firearm and tool mark identification utilizing 10 consecutively manufactured slides. *AFTE J.* 2013;45(4):376-93.
39. National Research Council, Committee on Identifying the Needs of the Forensic Science Community. Strengthening forensic science in the United States: A path forward. Washington, DC: National Academies Press; 2009. <https://www.ncjrs.gov/pdffiles1/nij/grant/s/228091.pdf>. Accessed 2 Oct 2020.
40. Baldwin D, Bajic S, Morris M, Zamzow D. A study of false-positive and false-negative error rates in cartridge case comparisons; 2016. <https://www.ncjrs.gov/pdffiles1/nij/249874.pdf>. Accessed 12 Apr 2020.
41. Smith T, Smith G, Snipes J. A validation study of bullet and cartridge case comparisons using samples representative of actual casework. *J Forensic Sci.* 2016;61(4):939-46. <https://doi.org/10.1111/1556-4029.13093>.
42. Hamby J, Norric S, Petraco N. Evaluation of Glock 9 mm firing pin aperture shear mark individuality based on 1,632 different pistols by traditional pattern matching and IBIS pattern recognition. *J Forensic Sci.* 2016;61(1):170-6. <https://doi.org/10.1111/1556-4029.12940>.
43. Hamby J, Thorpe J. The examination, evaluation and identification of 9mm cartridge cases fired from 617 different Glock Model 17 & 19 semiautomatic pistols. *AFTE J.* 2009;41(4):310-24.
44. Stroman A. Empirically determined frequency of error in cartridge case examinations using a declared double-blind format. *AFTE J.* 2014;46(2):157-75.
45. McClarin D. Adding an objective component to routine casework: use of confocal microscopy for the analysis of 9 mm caliber bullets. *AFTE J.* 2015;47(3):161-70.
46. Song J. Proposed, "congruent matching cells (CMC)" method for ballistic identification and error rate estimation. *AFTE J.* 2015;47(3):177-85.
47. Keisler M, Hartman S, Kilmon A, Oberg M, Templeton M. Isolated pairs research study. *AFTE J.* 2018;50(1):56-8.
48. President's Council of Advisors on Science and Technology (PCAST). Report to the President - Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods. Washington, DC: President's Council of Advisors on Science and Technology; 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf. Accessed 2 Oct 2020.
49. Hamby J, Brundage D, Thorpe J. The identification of bullet fired from 10 consecutively rifled 9mm Luger Ruger pistol barrels: a research project involving 507 participants from 20 countries. *AFTE J.* 2009;41(2):99-110.
50. Smith J. Method of rifling by manufacturer. *AFTE J.* 2011;43(1):45-50.
51. Nichols R. Subclass characteristics: from origin to evaluation. *AFTE J.* 2018;50(2):68-88.
52. Calculate confidence limits for a sample proportion. <https://epito.ols.ausvet.com.au/ciproportion>. Accessed 9 Mar 2020.
53. Brown L, Cai T, Dasgupta A. Interval estimation for binomial proportion. *Stat Sci.* 2001;16(2):101-33.
54. AFTE Range of Conclusions. <https://afte.org/about-us/what-is-afte/afte-range-of-conclusions>. Accessed 22 May 2020.

How to cite this article: Smith JA. Beretta barrel fired bullet validation study. *J Forensic Sci.* 2020;00:1-10. <https://doi.org/10.1111/1556-4029.14604>



Public beliefs about the accuracy and importance of forensic evidence in the United States

Jacob Kaplan¹, Shichun Ling^{1,*}, Maria Cuellar

Department of Criminology, University of Pennsylvania, United States

ARTICLE INFO

Keywords:

forensic science
forensic evidence
CSI effect
public perceptions

ABSTRACT

Recent advances in forensic science, especially the use of DNA technology, have revealed that faulty forensic analyses may have contributed to miscarriages of justice. In this study we build on recent research on the general public's perceptions of the accuracy of 10 forensic science techniques and of each stage in the investigation process. We find that individuals in the United States hold a pessimistic view of the forensic science investigation process, believing that an error can occur about half of the time at each stage of the process. We find that respondents believe that forensics are far from perfect, with accuracy rates ranging from a low of 55% for voice analysis to a high of 83% for DNA analysis, with most techniques being considered between 65% and 75% accurate. Nevertheless, respondents still believe that forensic evidence is a key part of a criminal case, with nearly 30% of respondents believing that the absence of forensic evidence is sufficient for a prosecutor to drop the case and nearly 40% believing that the presence of forensic evidence – even if other forms of evidence suggest that the defendant is not guilty – is enough to convict the defendant.

1. Introduction

The collection and use of forensic evidence have increasingly become vital to criminal investigations and prosecutions [22]. Forensic evidence has been valuable in establishing key elements of a crime, identifying people who were at the crime scene, exonerating innocent defendants, and corroborating victim testimonies [10]. However, recent advances in forensic science, especially the use of DNA technology, have revealed that faulty forensic analyses have contributed to miscarriages of justice. This has led to calls to strengthen scientific foundations of the analysis and presentation of forensic evidence by identifying the types of errors that could occur, describing key concepts that clarify the sources of error, and developing strategies for how to reduce error in forensic analyses [34,35]. Given the importance of recognizing the limitations of forensic science, and the potential devastating consequences that the misuse of forensic science can yield, research on perceptions of forensic science is an important endeavor.

In the United States (US) criminal justice system, jurors are expected to determine guilt based upon relevant facts of a case. While there are attempts to minimize biases in juries, there remains concern that jurors may still hold preconceptions that influence their decisions. In recent years, one such concern relates to juror perceptions of forensic science. Dubbed the “CSI effect”, this term refers to how television crime shows

may affect juror expectations and perceptions, including creating unreasonable expectations among jurors; elevating forensic evidence over other forms of evidence; and perceiving forensic evidence as infallible, objective and free from human judgement or error [2,25,29]. While there have been multiple studies examining the influence of television crime shows on perceptions of forensic evidence or testimony, to the authors' knowledge, only one study to date [29] has directly examined public beliefs about how accurate various forensic techniques are and the role that human judgements plays in the forensic science investigation process. Ribeiro et al. [29] surveyed 101 members of the public in Australia to measure general perceptions of human judgement and error involved in forensic techniques and did not find support for a CSI effect. In fact, their findings suggest that participants believed forensic science was relatively error-prone, involved an appreciable amount of human judgement, and that different forensic techniques yielded different levels of accuracy.

While Ribeiro et al.'s [29] study provides important insights into perceptions of human judgement and error in the context of forensic science, the study was based upon an Australian sample, so it may not immediately translate to the American context. The Australian legal system is similar to that of the US in many ways (e.g., presumption of innocence, requirements to ensure voluntariness of confessions), but there are also crucial differences. These differences include whether

* Correspondence author: Department of Criminology, 3718 Locust Walk, Suite 216, University of Pennsylvania, United States.

E-mail address: lings@sas.upenn.edu (S. Ling).

¹ Joint first authorship.

illegally obtained evidence is excluded from trial, who has the power to determine charges (prosecutors in the United States but police officers and other criminal investigative units in Australia) as well as plea bargaining and sentencing practices [21,37]. Differences between the US and Australian criminal justice system more broadly necessitate an investigation into US perceptions of forensic science. The US serious crime rate, as well as its high rate of incarceration, give the criminal justice system a much broader role in public life in the United States than in Australia because it affects a far greater percent of the population. Moreover, while there have been acknowledgements of national reports outlining forensic science reliability concerns and errors among legal practitioners in the United States, other countries, such as Australia, have been slower to conduct independent inquiries into the validity and reliability of claims made in forensic science [9]. While there is some evidence that this situation is changing [20], there are differences between the two countries in the knowledge of legal practitioners regarding the fallibility of forensic science, and it is unknown whether such differences also exist among in the general public. Differences of opinion between the two populations could also be attributed to cultural differences distinct from institutional differences between the criminal justice systems of each nation. A sociological comparison of attitudes towards forensic science between Australia and the United States would be an interesting contribution to this discussion. However, this article will focus on documenting the differences in opinion rather than on attempting to explain their cause. As such, it is important to understand the extent to which Ribeiro et al.'s [29] findings are generalizable.

1.1. Miscarriages of justice

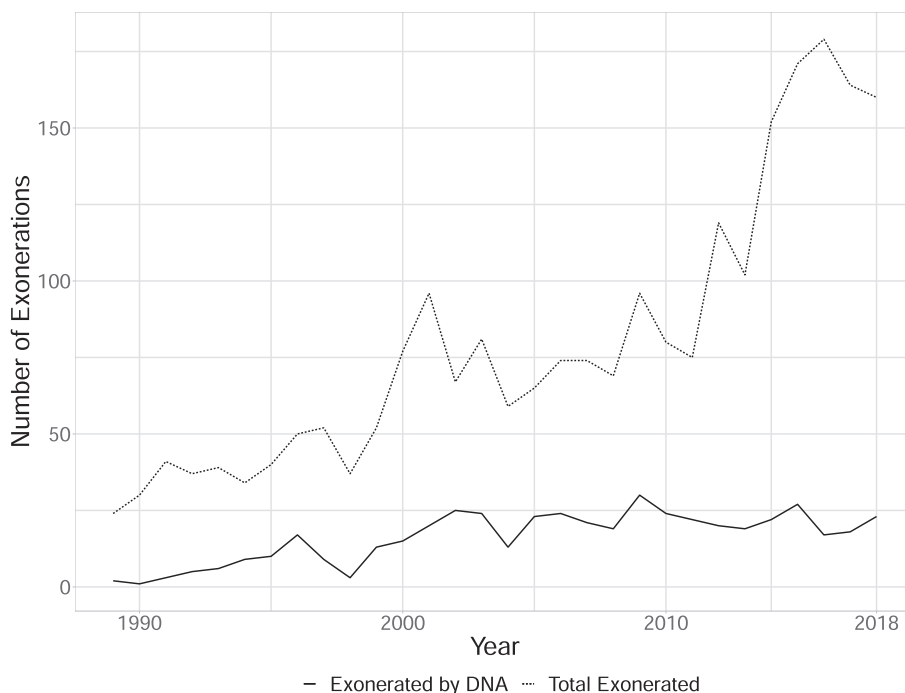
1.1.1. Exonerations

With the increased use and application of forensic science over the years come increasing concern over the misuse of forensic evidence. The inappropriate use or application of forensic science has been estimated to contribute to almost a quarter of all wrongful convictions nation-wide [27]. In a study by Garrett and Neufeld [12], 60% of cases involved unsubstantiated or misleading forensic testimonies. There is

an increasing trend in the annual number of exonerations in the United States (Fig. 1) and the number of exonerations due, at least in part, to inaccurate or misleading forensic evidence (Fig. 2) over the last two decades. These concerns are especially troubling when considering potential racial disparities in exoneration rates, with evidence that Blacks are exonerated at higher rates than Whites [31]. In an effort to review, rectify, and prevent cases of wrongful convictions, a growing number of prosecutorial offices are establishing conviction integrity units (CIUs). One tool that CIUs use to review cases involves the re-examination of forensic evidence. In 2018, CIUs have been responsible for 58 exonerations, some of which involved official misconduct such as falsifying forensic results [23]. Ultimately, flawed interpretations or misrepresentation by forensic analysts may negatively impact jury perceptions. This has augmented concerns about how forensic science may contribute to miscarriages of justice, and how pre-existing and contextual biases may play a role in how forensic evidence is perceived [16].

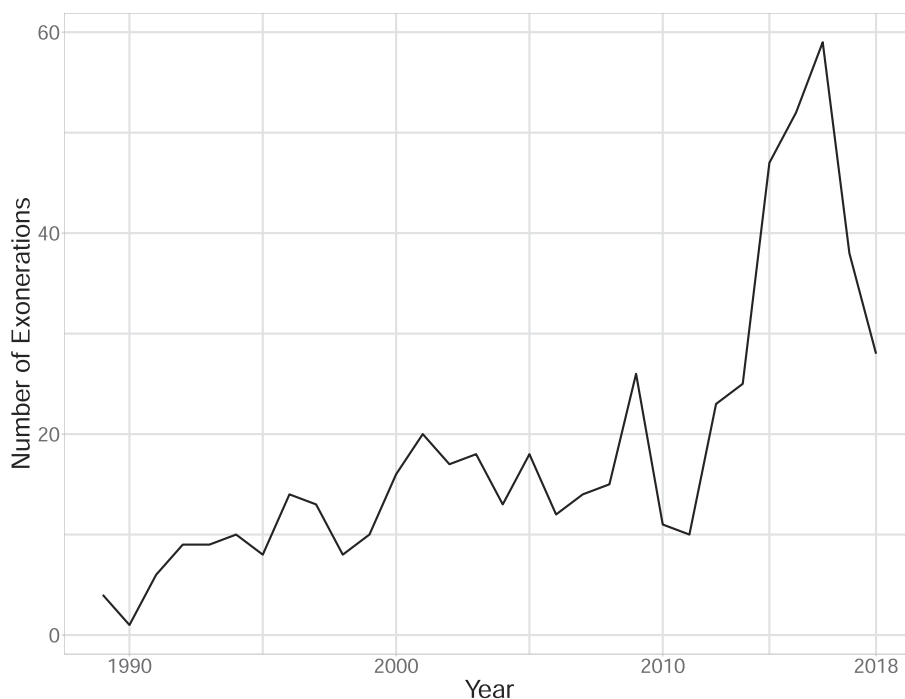
1.1.2. Community relations

The consequences of erroneous use or interpretation of forensic techniques may disproportionately affect racial and ethnic minorities in the US, who have disproportionate contact throughout the criminal justice system. In recent years, there has been a spotlight on compounding racial tensions between criminal justice system and minority community members in particular. This has manifested in several ways, including the establishment and growth of the Black Lives Matters movement as well as the elections of progressive prosecutors. These efforts are part of a growing movement seeking to redress perceived wrongs that certain groups disproportionately experience within the criminal justice system. Indeed, perceptions of injustice or unfair treatment by the criminal justice system can undermine the perception of legitimacy of the system as a whole. This could foster distrust of certain types of evidence during trials, such as police or eyewitness testimony, if they are perceived as biased or subjective. If forensic evidence is seen as more objective than other types of evidence, there may be more reliance on these measures to avoid the flaws of other evidence types. However, there remain ethical concerns over various



Source: National Registry of Exonerations, <http://www.law.umich.edu/special/exoneration/Pages/browse.aspx>

Fig. 1. Annual Number of People Exonerated in the United States.



Note: Source: National Registry of Exonerations, <http://www.law.umich.edu/special/exoneration/Pages/browse.aspx>.

Fig. 2. Annual Number of People Exonerated in the United States Whose Conviction Included Inaccurate or Misleading Forensic Evidence.

aspects of forensic evidence. The existence of DNA databases, for example, may be helpful in identifying DNA recovered from a crime scene if the perpetrator has a record in the DNA database already. However, Amankwaa [1] and Machado and Silva [19] identify key risks that may occur with the improper use of these databases, including exacerbating existing stigmas and stereotypes due to the over-representation of certain social and racial groups in criminal DNA databases, as well as mistaken identification resulting from erroneous interpretations of the information provided by DNA profiles that can lead to wrongful convictions.

1.2. How frequently is forensic evidence used?

A study analyzing forensic science collection practices by law enforcement in Denver and San Diego found that in nearly all homicide cases, at least one type of forensic evidence – primarily DNA, fingerprints, evidence from the weapon used, or hair – was collected [22]. For the crime of sexual assault, over half of cases in Denver and two-thirds of cases in San Diego collected forensic evidence, with the vast majority being DNA or hair. Forensic evidence collection is far less common in other crimes with under one-third of burglaries in San Diego and < 16% of burglaries in Denver having a single type of forensic evidence collected. The cases which do collect evidence primarily collect fingerprints. While forensic evidence is primarily collected in cases of violent crime, there is growing interest in collecting forensic evidence – in particular DNA evidence – at property crime scenes, vastly expanding the scope of cases in which forensic evidence may play a role [30]. Recent advances in technology have reduced the cost of DNA collection and dramatically increased the speed at which DNA collected at a crime scene can be compared against a DNA registry [14]. This had led to even small police agencies collecting forensic evidence for violent as well as property crimes. As forensic evidence becomes increasingly common in criminal cases, research on how the general public – specifically, jury-eligible members of the public – respond to this evidence is crucial to understanding how they will behave when presented with forensic evidence in a criminal trial.

1.3. Levels of accuracy from literature reports

While differences in public opinion about the validity and reliability of forensic methods are of intrinsic interest to policy makers and other researchers, it is also important to compare public opinion to the findings of scientific experts about the validity and reliability of these methods. At the time of this writing, the authors are not aware of a single standard by which the claims of forensic science can be evaluated. However, a number of studies have been conducted in the US to determine the validity and reliability of forensic methods. In this study, we will compare our survey findings to the expert opinions articulated in one prominent report from the United States, the President's Council of Advisors on Science and Technology (PCAST) report [35]. We use this report because it is a recent, careful analysis by independent scientists of the validity and reliability of a number of forensic methods.

There is no simple score from zero to 100 for the levels of accuracy of forensic methods. However, there are available reviews about whether these methods are valid, meaning accurate and consistent. In the United States, Rule 702 (Fed. R. Evid. 702), from the Federal Rules of Evidence sets the standards of admissibility of scientific evidence in court.² Among other sections, it states that the expert may testify if the testimony is “the product of reliable principles and methods” and “the expert has reliably applied the principles and methods to the facts of the case.” PCAST called these two standards *foundational validity* and *validity as applied*, respectively. The report reviewed the research about seven forensic disciplines (DNA single-source and simple mixture, DNA complex mixture, bitemarks, fingerprint, firearms, footwear, and hair). The reviewed research consisted of studies of error rates of the methods, and consistency if an analyst performs the analysis at different times and if different analysts perform the same analysis with the same

² While Rule 702 establishes federal standards for the admissibility of evidence, the standards within states are somewhat more heterogeneous. States typically adopt the Frye (Frye v. United States, 293F. 1013 (D.C. Cir. 1923)) or Daubert (Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993) at 592) standards, which are based on precedents from case law.

materials. While PCAST is not the only review that could be used for comparison (for instance, The National Research Council [34] could be used as well), we chose it because it is a recent, careful analysis by independent scientists that provides a clear and supported categorization of the validity and reliability of a number of forensic methods. It is left as future work to use other reviews for comparison with our survey responses.

PCAST [35] determined that, out of the seven disciplines reviewed, only DNA analysis of single-source simple mixture (two sources where one source is known) samples and latent fingerprint analysis were foundationally valid. DNA analysis of complex-mixture samples with probabilistic genotyping and firearms analysis were not foundationally valid, but had the potential to be so with current and future research. DNA analysis of complex-mixture samples with combined-probability-of-inclusion (CPI) methods, bitemark analysis, footwear analysis, and microscopic hair comparison were not foundationally valid and/or were missing serious research.

Regarding the techniques from our survey not included in the PCAST report, there is no single review that gives a definitive answer about their foundational validity. The National Research Council [34] concluded that for bloodstain analysis, “some experts extrapolate far beyond what can be supported” and “the uncertainties associated with bloodstain pattern analysis are enormous.” For gunshot residue, there are no studies of which the authors are aware that estimate the accuracy or evaluate the validity of the technique, and thus they have not been demonstrated to be foundationally valid. For voice analysis, there is a recent review of the scientific validity of various methods by the Scientific Literature Working Group [36]. The review does not make a final conclusion about the scientific validity, but it does show promising research on the accuracy of various methods. For this study we leave voice analysis unranked in terms of actual accuracy. Toxicology is multidisciplinary since it uses analytical chemistry, pharmacology, and clinical chemistry to aid medical or legal investigation of death, poisoning, and drug use. There are studies of the accuracy of many of the methods used, so it should be considered foundationally valid. However, neither the National Research Council nor the PCAST present a careful review of its methodologies. Finally, while the current study includes brain imaging as a technique, it is not a traditional forensic discipline or a component of crime scene investigation. However, it has been offered as a potential method of gaining insight into individuals’ psychological states after a suspect is in custody, and has been used as evidence in multiple phases of criminal trials by prosecutors and defense attorneys [6,7,13].

1.4. Current study

The current study aims to bridge the gap between the increasing importance of forensic evidence in criminal cases and the dearth of knowledge of the US public’s view of that evidence. We do so by surveying members of the US public to assess their beliefs on the accuracy of forensic evidence and the process of collecting, analyzing, and reporting of such evidence. We approach this study with four hypotheses:

1. Respondents will have a high level of confidence in the forensic science investigation process as well as for the accuracy of each forensic science technique. Given the relatively high confidence found in Ribeiro et al.’s [29] Australian sample, we expect that our US sample will have a similar high degree of confidence in forensic science.
2. Respondents will overestimate the accuracy of forensic evidence. While determining the objective accuracy of forensic evidence is a difficult and ongoing process, we expect that respondents will perceive the evidence to be of a higher quality than supported by research.
3. Respondents will support the *CSI* effect by believing that what they see on fictional TV shows about forensic science reflects actual forensic science techniques and outcomes.

4. Forensic evidence will be given great weight in criminal trials and be considered a decisive factor in whether a defendant is considered guilty or not guilty. We expect that respondents will prioritize forensic evidence in criminal trials over other types of evidence, and consider its presence to be strong evidence that the defendant is guilty.

2. Method

2.1. Participants

This study utilized Amazon’s Mechanical Turk, an online survey platform, to collect information about the general public’s perceptions of various forensic science techniques. The survey consisted of 49 questions and took approximately 24 min to complete. Only Mechanical Turk users in the United States were eligible to take the survey. All surveys were collected between June 26th and 27th, 2019. Participants were financially compensated up to \$1 for their participation. All study procedures were approved by the University of Pennsylvania’s institutional review board. Users who agreed to take the survey were directed to a link on the Mechanical Turk website to the survey which was administered through the Qualtrics survey software.

In total, 180 people completed the survey. Two attention-check questions were used to determine whether responses were reliable. Following the introductory page explaining the purpose and topic of the survey, respondents were asked a multiple-choice question (the first attention-check question) on what the survey was about. Fifteen respondents chose an option other than “Forensic evidence.” The second attention-check asked if the respondent had “ever been a victim of murder?” An additional 10 respondents said that they had. In total, 25 respondents failed the attention check and were dropped from the study analyses. Responses from the remaining 155 participants were used for the analyses.

Respondents varied in age from 19 to 70 with most respondents being in their 30s (Mean = 35.6, SD = 10.6). The majority of respondents identified as male (59%), 39% identified as female, and 2% identified as neither male nor female. Over two-thirds (70%) identified as White-only, 10% identified as Black-only, 6.5% identified as Asian or Pacific Islander, and 9% identified as Hispanic. The remaining respondents identified as mixed-race or as American Indians. This is similar to the United States population as a whole where 60.4% of residents are White-only, 13.4% are Black-only, and 5.9% are Asian-only, and 18% are Hispanic. These respondents are more educated than the United States general public. In the present sample, 87.2% have graduated high school, nearly the same as the 87.3% of the general public. However, approximately 52% had earned a four-year degree or higher in the sample compared to 31% in the entire United States. Twenty respondents (12.9% of the sample) had served on a jury, with 65% (13 respondents) of these being involved in a case that included forensic evidence.

The survey utilized in the current study is a modified version of the Ribeiro et al. [29] study (see Ribeiro et al. [29] for how to access their survey).

2.2. Forensic science investigation process

To understand public perceptions of the likelihood of an error occurring during the forensic science investigation process, we asked respondents “how likely is it that an error could occur” at each stage. The six stages of the forensic science investigation process are: collection, storage, testing, analysis, reporting, and presenting. The respondents’ answers were on a slider from 0 to 100 with the default position set at 50.³ Respondents

³ Analyses were conducted in a separate pilot study to determine whether a default anchor of 0, 50, or 100 would affect participant responses. Results indicated that responses between the three anchors were similar on average, thus suggesting respondents were not influenced by the initial position of the anchor.

were required to select a value to proceed to the next question, even if they selected the value of 50. For each process, respondents were also asked “to what extent does the [process] involve human judgement?” with a 7-point Likert scale answer from *None at all* (1) to *Entirely* (7).

2.3. Forensic science techniques

Respondents were then asked how accurate they perceive each of 10 forensic science techniques to be and whether there was significant human judgement involved.⁴ As with the forensic science investigation process questions, the accuracy was measured on a slider from 0 to 100 with the default position set to 50. We included 10 techniques or analyses in this survey: bloodstain pattern, brain imaging, DNA, dental, fingerprint, firearm and toolmark, footwear, gunshot residue, toxicology (e.g. urine, drugs), and voice analysis.

Eight of these techniques (all except for brain imaging and footwear analysis) were studied by Ribeiro et al. [29], allowing for a comparison of perceptions between US and Australian populations. In addition to the eight techniques shared with Ribeiro et al. [29], we included footwear analysis, since it is one of the primary methods in feature-comparison and is commonly used in forensic laboratories, and brain imaging because it has been used as evidence during criminal cases as a method of demonstrating defendants’ mental states and capabilities. We decided not to include some of the techniques studied in Ribeiro et al. [29] (anthropological, document, faces, fire/explosives, geological materials, image, materials, and wildlife) because they were not included in reports that review the state of forensic science [35] and in the interest of focusing more heavily on feature-comparison methods.

Human judgement was measured by asking whether they believed there to be “key procedures that involve significant human judgement” in that forensic science technique. Respondents could answer *No*, *Yes*, or *Not Sure*.

2.4. CSI effect

The popularity of TV shows depicting forensic science such as *CSI* and *Law & Order* has led to concerns about a “CSI effect” where watchers believe that the shows accurately depict forensic science and use standards based on the show’s inaccurate depictions as their basis for judging the validity of the techniques [29,5]. These shows often depict forensic science as infallible, nearly instantaneous, and entirely objective. If jurors do indeed base their opinion of forensic science on what is depicted on these shows, they may conclude that a piece of forensic evidence is more powerful than it actually is. Conversely, the lack of forensic evidence - which is found in nearly all crime scenes on these shows - may be seen as evidence that the defendant is not guilty.

Past studies of this topic primarily use TV viewing habits to measure whether watching these shows affects perceptions of forensic evidence [29,32,26]. This method has a number of limitations as it is unclear whether watching more of these shows reflects merely that the respondents watch more TV overall, if they are particularly interested in forensic evidence - and what other material they use to learn about forensic evidence - and only indirectly measures how watching these shows affects perceptions of forensic evidence. In this study we attempt to address the CSI effect directly by asking respondents how accurate they believe the “most accurate fictional show” and the “average fictional show” is in depicting forensic science. Respondents could choose from a 4-point Likert-scale from *Not Accurate at all* to *Very Accurate*, as well as *Not Sure*. As these shows are largely fictitious or a gross exaggeration of real forensic evidence techniques, asking respondents directly how accurate they believe these shows to be allows for a better measure of the CSI effect than previously evaluated [15].

⁴ We did not define any of the forensic techniques to avoid biasing responses. As such, the results should be interpreted as baseline knowledge.

2.5. Importance of forensic evidence during criminal cases

Jurors may believe that there are substantial flaws in the accuracy of individual techniques or the forensic science investigation process yet may still be willing to accept forensic evidence presented at trial if they believe that only the strongest evidence - that which has avoided the concerns that they have for the evidence - will be presented. To assess this, we asked respondents how strongly they agreed with four statements about the usability and importance of evidence in criminal trials. These questions come from the Forensic Evidence Evaluation Bias Scale (FEEBS), a questionnaire designed and validated by Smith and Bull [32–33], to evaluate people’s perceptions of forensic evidence.

1. Forensic evidence always provides a conclusive answer.
2. Forensic evidence always identifies the guilty person.
3. If no forensic evidence is recovered from a crime scene, then the prosecutor should drop the case.
4. If forensic evidence suggests a defendant is guilty, this should be enough to convict even if other evidence (e.g., eyewitness testimony, alibi) suggest otherwise.

3. Results

3.1. Forensic science investigation process

3.1.1. Estimates of error

Table 1 shows how prone to error respondents believe the forensic process to be. Columns (1–2) show the results from the current study with Column (1) showing the percent likelihood of an error occurring and Column (2) showing the cumulative chance of an error occurring at each consecutive stage of the process. Columns (4–5) follow this same pattern and show results from Ribeiro et al.’s [29] study of the general public in Australia. To allow easy comparison between the US and Australian results, the final three columns are the difference between US and Australian values.

At each stage in the forensic science investigation process, respondents believe there to be a high chance of an error occurring. The first stage, collection, was perceived to be the riskiest stage with a 56% chance of an error occurring. The least risky stage, reporting, fared a little better with a perceived 44% chance of an error occurring. The forensic science investigation process is considered to be rife with possibilities for errors, with respondents perceiving that an error could occur about half the time at each stage. The Australian sample believed that an error would occur about 40% of the time on average, approximately 10 percentage points lower than the American sample. For each stage, American respondents believed that an error was more likely to occur - with differences ranging from +2.82 for presenting to +13.26 for collection - than Australian respondents did.

3.1.2. Human judgement

For each stage in the forensic process, respondents were asked how much human judgement was involved in that stage. This question used a seven-point Likert-scale from *None at all* (1) to *Entirely* (7). Column (3) of Table 1 shows the mean respondent score. Respondents believed that there was a high level of human judgement involved at each stage, with all except two stages - storage at 4.65 and testing at 4.78 - having a score above 5. Because variables were nonnormally distributed, Kendall’s tau-b correlations were run to examine the association between the likelihood of an error and the level of human judgement involved for each stage of the forensic process. There was a positive correlation between how likely an error could occur and how much human judgement was involved for all six stages: collection ($\tau_b = 0.363$, $p < .001$), storage ($\tau_b = 0.412$, $p < .001$), testing ($\tau_b = 0.289$, $p < .001$), analysis ($\tau_b = 0.229$, $p < .001$), reporting ($\tau_b = 0.350$, $p < .001$), and presentation ($\tau_b = 0.218$, $p < .001$). These correlational results suggest that respondents believe that people involved in

Table 1
Perceived Accuracy and Level of Human Judgement for Each Stage of the Forensic Science Process.

Process Stage	US Sample			Australian Sample			US – Australian Difference		
	Error	Cumulative Error	Human Judgement	Error	Cumulative Error	Human Judgement	Error	Cumulative Error	Human Judgement
Collection	55.74 (27.37)	55.74	5.39 (1.47)	42.48 (27.12)	42.48	5.55 (1.60)	13.26	13.26	–0.16
Storage	48.45 (26.29)	104.19	4.65 (1.67)	39.35 (28.11)	81.83	5.15 (1.66)	9.10	22.36	–0.50
Testing	45.26 (27.07)	149.45	4.78 (1.58)	39.27 (27.77)	121.10	4.94 (1.70)	5.99	28.35	–0.16
Analysis	52.45 (26.28)	201.90	5.57 (1.46)	44.55 (27.60)	165.65	5.25 (1.52)	7.90	36.25	0.32
Reporting	44.25 (27.38)	246.15	5.06 (1.71)	40.69 (26.87)	206.34	5.43 (1.53)	3.56	39.81	–0.37
Presenting	45.04 (26.97)	291.19	5.37 (1.63)	42.22 (29.64)	248.56	5.55 (1.53)	2.82	42.63	–0.18

Note: This table shows the mean and (standard deviation) for the perceived likelihood that an error could occur during each stage in the forensic science process. Error is measured on a scale from 0 to 100. Human judgement is measured on a seven-point scale from 1 to 7. A value of one indicates that no human judgement is involved in the process; a value of seven indicates that the process is entirely based on human judgement. Responses of “Not sure” for the amount of human judgement involved are excluded. The US sample is from the present study, the Australian sample is from Ribeiro et al.’s [29] study of 101 members of the public in Australia.

Table 2
Perceived Accuracy and Level of Human Judgement for Each Forensic Evidence Technique.

	US Sample		Australian Sample		US – Australian <i>t</i> value
Type of Forensic Evidence	Accuracy	Human Judgement	Accuracy	Human Judgement	Accuracy
DNA	83.09 (17.92)	58% (49%)	89.95 (15.85)	58% (49%)	3.13**
Fingerprints	78.62 (17.47)	54% (50%)	88.15 (17.66)	54% (50%)	4.25***
Toxicology (e.g. urine, drugs)	76.12 (18.21)	43% (50%)	86.66 (13.75)	43% (50%)	4.97***
Dental	75.88 (22.02)	41% (49%)	89.26 (12.04)	41% (49%)	5.58***
Firearms and toolmarks	68.15 (19.41)	82% (38%)	79.63 (16.77)	82% (38%)	4.87***
Gunshot residue	67.98 (19.66)	65% (48%)	78.87 (17.97)	65% (48%)	4.48***
Bloodstain pattern	64.28 (20.50)	85% (36%)	78.53 (19.03)	85% (36%)	5.59***
Brain imaging	60.74 (24.92)	58% (50%)	–	58% (50%)	–
Footwear	56.98 (23.44)	82% (39%)	–	82% (39%)	–
Voice	55.30 (22.25)	86% (35%)	71.47 (19.16)	86% (35%)	6.00***

Note: This table shows the mean and (standard deviation) for perceived accuracy of each forensic science technique. Accuracy is measured on a scale from 0 to 100. Human judgement asks respondents whether they believe each technique involves ‘key procedures that involve significant human judgement?’ Responses shown are the percent the responded ‘Yes’, excluding those who responded ‘Not Sure’. The US sample is from the present study, the Australian sample is from Ribeiro et al.’s [29] study of 101 members of the public in Australia. The final column shows the *t*-value from a *t*-test comparing US responses to Australian responses from Ribeiro et al. [29].

**p* < 0.05.
 ***p* < 0.01.
 ****p* < 0.001.

the forensic science investigation process are liable to make mistakes that reduce the accuracy of the evidence. US respondents believe that there is slightly less human judgement than the general public in Australia (Column 6) do.

3.2. Forensic evidence techniques

3.2.1. Estimates of accuracy

Table 2 assesses how accurate respondents believe each of the 10 forensic techniques examined are. Column (1) shows how accurate respondents believe each technique to be, from 0 to 100. Based on the perceived accuracy, the most accurate to least accurate technique are: DNA, fingerprints, toxicology, dental, firearms/toolmarks, gunshot residue, bloodstain pattern, brain imaging, footwear, and voice.

Respondents believe that DNA analysis is the most accurate forensic technique at 83% accurate, followed by fingerprint analysis at 79%. DNA analysis is the only technique considered above 80% accurate, with most within the range of 65–75% accurate. Two analyses are considered below 60% accurate: voice analysis is considered to be 55% accurate and footwear analysis is considered to be 57% accurate.

For a comparison to Ribeiro et al.’s [29] Australian sample, Column (3) show the accuracy rate among their participants. Column (4) shows the *t*-value from a *t*-test comparing the current study’s responses to Ribeiro et al.’s [29] Australian sample. For each type of forensic

evidence, there is a statistically significant (*p* < 0.01) difference between each sample’s perceptions of accuracy. Relative to the Australian sample studied by Ribeiro et al. [29], American respondents viewed forensic techniques as less accurate. For the eight techniques studied which overlap with Ribeiro et al. [29], US respondents believed that the techniques were on average 12 percentage points less accurate than Australians did.⁵ For every comparable technique, US respondents rated it as less accurate than Australian respondents did. In six of the eight comparable techniques, US respondents perceived it to be around 10 percentage points less accurate than Australian respondents.⁶ These results may suggest that Americans are less trusting of forensic science overall, though they have relatively similar perceptions of the accuracy of forensic techniques relative to each other.

3.2.2. Comparison between survey responses and levels of accuracy from reports

Table 3 shows the comparison of accuracy rankings between the

⁵ Bloodstain pattern, DNA, dental, fingerprints, firearm and toolmarks, gunshot residue, toxicology, and voice analysis overlapped with the Ribeiro et al. [29] study. Brain imaging and footwear analysis were examined in this study but not Ribeiro et al.’s [29] study.

⁶ The two exceptions are DNA at 6.86% less accurate and fingerprints at 9.53% less accurate.

Table 3

PCAST report conclusions about foundational validity, which requires a method to be repeatable, reproducible, and accurate, of forensic disciplines [35]. The conclusions derived from the PCAST report have been interpreted and summarized by the authors of this article.

Conclusion by PCAST authors	Discipline
Foundationally valid	DNA Fingerprints
Not foundationally valid yet	Dental*
	Firearms/toolmarks**
	Footwear***
Unranked	Bloodstain pattern
	Voice
	Gunshot residue
	Brain imaging
	Toxicology

* There are low prospects of developing bitemark analysis into a scientifically valid method, according to PCAST.

** There is one appropriate study so far, but more are needed to show the technique is reproducible.

*** Source identification was found to not be foundationally valid, but the validity of class characteristic identification was not evaluated by PCAST.

survey responses and the conclusions from reports (see Section 1.3).⁷ It is not possible to make a numerical comparison between these two sources, so instead we analyze the differences in ordering. Other researchers might have different opinions about the ordering of the levels of accuracy of the forensic disciplines.

Toxicology, gunshot residue, bloodstain pattern analysis, brain imaging, and voice analysis were unranked by PCAST, so it is not surprising that they are scattered in the survey responses (they are in places 3, 6, 7, 8, 10, respectively in the survey responses).

Of the techniques that are ranked, the top two disciplines in the survey responses (DNA and fingerprints) are also the only two that are considered foundationally valid by PCAST. It is notable that dental analysis scored high (4 out of 10) in the survey since it is considered not foundationally valid by PCAST. Indeed, PCAST found that “available scientific evidence strongly suggests that examiners not only cannot identify the source of bitemark with reasonable accuracy, they cannot even consistently agree on whether an injury is a human bitemark” [35]. In fact, dental scored higher than firearms and toolmarks, even though PCAST found that firearms and toolmarks was almost shown to be foundationally valid, but it was not yet because there was only one appropriate study of scientific validity instead of multiple, which are required to show reproducibility.

Similar to Ribeiro et al.’s [29] study, we did not separate the DNA analysis into different types (single-source, simple mixture, complex mixture) for the survey, but PCAST did make this important distinction. It would be interesting to study whether the general public is aware of these differences and whether it considers some more accurate than others, but that is left as future work. Thus, for our comparison in Table 3, we refer to any type of DNA evidence as just “DNA”. Moreover, the survey asks about firearms/toolmarks, but most of the current research about the accuracy of these methods is about firearms, not toolmarks in general, such as the marks left by screwdrivers or wire cutters. It is common to present firearms and toolmarks as a single category, since imprints on a used bullet or cartridge (considered marks) were made by the firearm (considered a tool). These are issues for future research on forensic techniques to consider.

3.2.3. Human judgement

To judge how objective respondents believed each technique to be, we asked whether they believed there to be “key procedures” in the technique involving human judgement. The percent of respondents who

answered *Yes* are shown in Column (3) of Table 2, excluding those who responded *Not sure*.⁸ Respondents believe that there is a high level of human judgement involved in each technique. Over 50% of respondents believe that human judgement is involved in the forensic technique for all except for toxicology (43% of respondents) and dental analysis (41% of respondents). Even for the two most trusted analyses, DNA and fingerprints, over half of respondents believe that human judgement is involved in “key procedures” for that analysis with 58% and 54% reporting so, respectively. Because responses were non-normally distributed, Mann-Whitney *U* tests were conducted to examine differences in perception of accuracy between those who perceived the technique to involve human judgement or not. Individuals who believed no human judgement was involved in brain imaging (mean rank = 71.17) thought that this technique was more accurate than those who believed brain imaging involved human judgement (mean rank = 57.93), $U = 1528, p = .044$. Similarly, respondents who believed no human judgement (mean rank = 79.15) was involved in toxicology thought this technique was more accurate than individuals who believed the technique involved with human judgement (mean rank = 62.39), $U = 1914.5, p = .017$. For all other techniques, there were no significant differences in perception of accuracy between those who believed human judgement was involved and those who did not.

3.3. CSI effect

Table 4 shows the percent of respondents who chose each answer for the two questions used to measure the CSI effect. Column (1) shows the responses for the “most accurate fictional show” while Column (2) shows responses for the “average fictional show” that depicts forensic science. In both cases the vast majority of respondents believe that the shows are between slightly and moderately accurate. For the “most accurate” show, 43% of respondents believe it to be “moderately accurate,” more than the 26% who say the “average” show is “moderately accurate.” Approximately 10% of respondents believe that these shows are “very accurate.” For the “most accurate show,” the same number of respondents believe it to be “not at all accurate” as to be “very accurate.” For the “average show,” however, nearly twice as many (18%) of respondents believe it to be “not at all accurate.”

When asked whether watching these shows changed their interest in forensic science, nearly three-quarters of respondents (99 of 135 respondents; 20 respondents in the sample did not watch these shows) claimed they are “Much more interested” or “Somewhat more interested” in forensic science as a result of these shows.

3.4. Importance of forensic evidence during criminal cases

Table 5 shows the responses to the four questions regarding the importance and reliability of forensic evidence during the criminal justice process. Each row is a single question and Columns (1–5) show the percent of respondents who choose each answer. Respondents could select if they strongly or somewhat agree or disagree, or if they are not sure.

Row (1) shows responses to the statement that “forensic evidence always provides a conclusive answer” and the majority of respondents (52%) somewhat or strongly agree. A smaller amount, 41%, agree that “forensic evidence always identifies the guilty person” while the majority of respondents (55%) somewhat or strongly disagreed (Row (2)). These results seem contradictory to previous sections which showed that the forensic science investigation process and many forensic science techniques were perceived to have high levels of human judgement involved and to be relatively inaccurate. It is unclear why

⁷ The conclusions from reports are summarized by the authors of this article and are not a consensus that exists in the forensic science community.

⁸ Ribeiro et al. [29] also assessed the degree of human judgement for each forensic technique. However, their question was a Likert-scale question, preventing a comparison from our *Yes-No* question.

Table 4
Perceived accuracy of fictional TV shows that depict forensic science.

	Most Accurate Show	Average Show
Very accurate	9.68	9.68
Moderately accurate	43.23	26.45
Slightly accurate	33.55	41.94
Not accurate at all	9.68	18.06
Not sure	3.87	3.87

Note: Respondents were asked “How accurate do you think the [most accurate/average] fictional show is in depicting forensic science?” This table shows the percent of respondents who gave each answer to the questions. Column percentages may not total to 100 due to rounding.

respondents appear to be more supportive of “forensic evidence” abstractly yet hold relatively negative views of each specific technique or stage of the forensic science investigation process.

Row (3) demonstrates the extent to which respondents agree that prosecutors should drop a case if there is no forensic evidence collected at the crime scene. Nearly a third of respondents (29%) somewhat or strongly agreed with this statement while 65% disagreed and 6.5% were not sure. This suggests that, even though overall forensic evidence is considered to be relatively inaccurate, a nontrivial number of respondents would be unwilling to convict a defendant without it. As this study did not assess perceptions of other forms of evidence, such as eyewitness testimony, it is unclear whether this group believes that forensic evidence itself is particularly strong or that other forms of evidence are less valid. Finally, Row (4) reflects how strongly respondents agree that if forensic evidence suggests that the defendant is guilty, they should convict that defendant even if other evidence suggests that the defendant is not guilty. Here, 37% of respondents either somewhat or strongly agreed with this statement. These results indicate that while overall respondents believe there to be serious flaws in forensic evidence, an appreciable portion are willing to make decisions on the defendant’s guilt based solely on forensic evidence.

4. Discussion

This study sought to understand public perceptions of forensic science by surveying members of the general public in the United States. Overall, our hypotheses in general were not supported. While we expected respondents to have a high level of confidence in the forensic science investigation process and for the accuracy of each forensic science technique (Hypothesis 1), our results suggest that members of the US public hold significant doubts about the accuracy of forensic techniques and believe that each technique contains high levels of human judgement. The technique perceived to be most accurate was DNA evidence at 83% accuracy, while voice analysis at 55% and footwear analysis at 57% were perceived to be least reliable. Most forensic techniques were considered to be in the range of 65–75% accurate. Our results align with prior work indicating that DNA is often perceived to be among the most accurate forensic techniques, though our study yields lower perceptions of accuracy for DNA than reported elsewhere [18]. Additionally, respondents indicated that they believed there was a substantial risk of error at each stage of the forensic science process, and that each stage involves a large amount of human judgement. Relative to Ribeiro et al.’s [29] study in Australia, our sample reported a higher likelihood of error at every stage, especially in the collection, storage, and analysis stages.

Our second hypothesis reflected our expectation that respondents would overestimate the accuracy of forensic evidence. When comparing the accuracy rankings between the survey responses and the conclusions from reports, it was notable that the top two disciplines in the survey responses (DNA and fingerprints) were also the only two that were considered foundationally valid by the relevant literature [35]. Furthermore, dental analysis ranked 4th most accurate in the survey,

although it is considered not foundationally valid by PCAST. In fact, PCAST considers that it is far from being so as examiners “cannot even consistently agree on whether an injury is a human bitemark.” In fact, dental analysis scored higher than firearms and toolmarks in the survey, even though PCAST found that firearms and toolmarks was almost shown to be foundationally valid.⁹ Several techniques that were ranked in the survey (toxicology, gunshot residue, bloodstain pattern analysis, brain imaging, and voice analysis) were not in the PCAST report, thus, we could not compare their rankings. Overall, there was mixed support for Hypothesis 2.

We also hypothesized that respondents would believe fictional forensic science television shows would be highly accurate (Hypothesis 3). Ribeiro et al. [29] used the number of hours of forensic science-related TV shows that a respondent watched as a measure of their interest in the field and examined the correlations between this measure and respondents’ attitudes toward the likelihood of an error in the forensic science investigation process and for individual techniques. They found that there was no significant relationship between the number of hours watched and opinions on the likelihood of an error to occur. In this study we attempted to address the *CSI* effect directly by asking respondents how accurate they believe the “most accurate fictional show” and the “average fictional show” is in depicting forensic science. Our findings indicate that respondents believed that the average forensic science shows were only slightly accurate, and that even the “most accurate fictional show” was only moderately accurate. Arguably, a *CSI* effect would have been contingent on individuals believing what they see in forensic science-related TV shows (i.e., having most people report a *Very Accurate* rating), but the current results suggest that people do not blindly believe the accuracy of these shows. Respondents generally believe that such shows are slightly to moderately accurate at best. These results thus did not seem to indicate a *CSI* effect, and did not support our hypothesis. While this study measured the *CSI* effect in a different way than Ribeiro et al.’s [29] did, our findings are similar as neither study found support for a *CSI* effect.

Finally, we expected that respondents would give great weight to forensic evidence in criminal trials such that the evidence would be considered a decisive factor in whether a defendant is considered guilty or not guilty (Hypothesis 4). Results partially support this hypothesis as nearly 30% of respondents believe that the absence of forensic evidence is sufficient for a prosecutor to drop the case and almost 40% believed that the presence of forensic evidence, even if other forms of evidence suggest the defendant is not guilty, is enough to convict the defendant.

While the current study provides insights into public perceptions of forensic science, the impact of the current study may be limited in scope. In the US criminal justice system, jurors hold immense power during trials, determining whether the defendant is guilty of the crimes they are accused of committing. The Sixth Amendment to the United States Constitution guarantees that defendants the right to be judged by an “impartial jury” consisting of members of the public. In practice, however, juries only impact a small number of criminal cases as in nearly all but the most serious cases, the defendant pleads guilty or the case is dismissed before trial [17,4,28,3]. For the crime of murder, however, nearly 40% of cases do proceed to trial, where jury perceptions of the usefulness and validity of forensic science techniques can play an outsized role in determination of guilt. In the vast majority of murder cases at least one form of forensic evidence was collected by investigators at the scene [22].

However, juries are not presented only with forensic evidence during a trial. Their decision is likely based on other evidence involved in the case, personal biases, and how these factors interact with the forensic evidence presented. Therefore, asking respondents to rate the

⁹Firearms and toolmarks are not considered foundationally valid as there is only one appropriate study of scientific validity instead of multiple, which are required to show reproducibility.

Table 5
Importance of Forensic Evidence in Determining Guilt in a Criminal Trial.

	Strongly Agree	Somewhat Agree	Somewhat Disagree	Strongly Disagree	Not Sure
Forensic evidence always provides a conclusive answer.	16.13	36.13	28.39	16.13	3.23
Forensic evidence always identifies the guilty person.	10.32	30.32	37.42	17.42	4.52
If no forensic evidence is recovered from a crime scene, then the prosecutor should drop the case.	10.32	18.71	29.68	34.84	6.45
If forensic evidence suggests a defendant is guilty, this should be enough to convict even if other evidence (e.g., eyewitness testimony, alibi) suggest otherwise.	10.32	27.10	37.42	19.35	5.81

Note: This table shows the percent of respondents who gave each answer to the questions. Row percentages may not total to 100 due to rounding.

accuracy and degree of human judgement involved in each step on the forensic process or for each type of forensic science technique only captures some of the factors that potential jurors consider when deciding on a verdict. Future research may consider interviewing members of a jury whose case involved forensic science to determine how that piece of evidence influenced their decision. Additional research could use a vignette-design to simulate a juror’s experience in a case and vary the forensic science technique involved to measure how much each technique influences their decision and what other variables matter in such a decision.

This study did not define any of the forensic science techniques, allowing the respondent to respond based on what knowledge they already have on the topic. While most of the techniques are self-explanatory, the interpretation of dental analysis may have needed to be clarified. It is unclear whether participants interpreted this as bite mark analysis, as was intended, or if they believed this item to refer to the identification of human remains based on teeth examination. This is a limitation that should be considered and clarified in future studies. In a trial, both the prosecution and the defense would likely explain to the jury what the technique is and argue about its accuracy and relevance. Therefore, this study measures people’s baseline beliefs about each forensic technique rather than beliefs at the time that a juror must render a verdict. These results may be useful to attorneys who argue in front of a jury as it provides a guide on the techniques the jurors will expect to be accurate and those that prompt more skepticism. Lawyers may use these results to argue more forcefully for or against certain evidence with the knowledge that jurors already have certain beliefs about these techniques. In addition to its impact on lawyers, these results may be useful to investigative teams who can prioritize techniques that are both based in evidence and have a high degree of support by the public.

This study used data from 155 participants during late June 2019 through Mechanical Turk. Having a larger sample size and utilizing additional recruitment sources may provide more representative responses. The results of the current study may be a reflection of the characteristics of the sample and methods employed, thus replication is needed to assess the ecological validity of the current findings. Moreover, during the past several years the rise of movements such as Black Lives Matters and the election of progressive prosecutors in a number of major cities in the United States reflects a shift in attention towards negative aspects of the criminal justice system such as racial bias and miscarriages of justice. While a majority of those in the US overall remain confident in the police, a growing number – 14% in 2018 – report “very little” confidence [11]. Among Blacks and Hispanics in the US, groups which are over-represented in the criminal justice system, confidence in the police has fallen significantly with fewer than half of Hispanic people and fewer than a third of Black people having a “great deal or quite a lot” of confidence in police [24]. This attention towards negative aspects of the criminal justice system may have affected our results if respondents with low trust of the police cause low trust in the forensic evidence process - or in the people tasked at each stage of the forensic evidence process. A longitudinal study of this topic could detect whether perceptions of forensics change over

time and if there is any relationship between trust in the criminal justice system and beliefs towards forensic evidence.

4.1. Implications and future directions

Based on our findings, US respondents believe that there is less human judgement but more errors at each stage of the forensic science process than their counterparts in Australia. It is unclear why this is the case, but this may suggest that US respondents believe that the science itself is more prone to error. Future research should investigate precisely which aspects of each stage is considered at risk of an error occurring. They should also continue to examine perceptions in different countries to better understand how people from different cultures understand and evaluate forensic evidence.

Our results also indicate that while fictional shows depicting forensic science are considered relatively accurate, the vast majority of US respondents do not believe that they are a perfect, or even near-perfect, representation of forensic science practices. The large difference in perceptions of accuracy between the “most accurate” and the “average” shows also indicate that people believe that they have enough knowledge of the field of forensic science to make this distinction between shows. Further studies of this topic should examine this question further, helping to distinguish how accurate these shows truly are and which specific features people believe to be accurate. While the *CSI* effect has been hypothesized to change viewers’ opinions on forensic science because they believe that the shows are accurate, it may be that people already interested in forensic science are more likely to watch these shows. Watching shows may also change a person’s belief in forensic science if they decide to look up the techniques that they see on the show to read more about them. In the current study, most respondents (99 of 135) acknowledged that their interest in forensic science increased as a result of forensic science-related shows. While this study did not ask if respondents did any research on the forensic science they saw, it does offer avenues for future research to examine if there was a behavioral change as a result of these shows.

5. Conclusion

This study found that US respondents believe that there is a high degree of human judgement involved and high risk of an error occurring at each stage of the forensic science process. When considering forensic science techniques specifically, those in the US hold a skeptical view of the vast majority of techniques, viewing some of them as little more accurate than a coin flip, and no technique more than 84% accurate. When compared to their counterparts in Australia, as studied by Ribeiro et al. [29], members of the US general public have a similar though more negative view of the field of forensic science than Australians.

Inaccurate perceptions of jurors towards forensic techniques likely has a severe and detrimental effect on the criminal justice system as it may influence their decisions of guilt or innocence. As the use of forensic science becomes more common in criminal cases that go before juries, it is increasingly important that we understand preconceptions

that jurors hold towards this field to better reduce biases during trials. Juries during criminal cases, however, are rare in the US justice system. The vast majority of criminal cases, over 90%, are settled through plea bargains, causing an outsized role of prosecutors in the criminal justice system [8]. However, little is known about prosecutors' perceptions of forensic science or how they use the evidence collected during the plea-bargaining process. It is important, therefore, for research in this field to continue to examine perceptions among members of the general public, who decide guilt for a small number of serious cases, and among prosecutors, whose decisions affect nearly all cases in the criminal justice system.

References

- [1] A.O. Amankwaa, Forensic DNA retention: Public perspective studies in the United Kingdom and around the world, *Science & Justice* 58 (6) (2018) 455–464.
- [2] C. Call, A.K. Cook, J.D. Reitzel, R.D. McDougle, Seeing is believing: The CSI effect among jurors in malicious wounding cases, *Journal of Social, Behavioral, and Health Sciences* 7 (1) (2013) 52–66.
- [3] T.H. Cohen, B.A. Reaves, Felony defendants in large urban counties, 2002. Bureau of Justice Statistics (2006).
- [4] T.H. Cohen, B.A. Reaves, Felony defendants in large urban counties, 2006. Bureau of Justice Statistics (2010).
- [5] S.A. Cole, R. Dioso-Villa, CSI and its effects: Media, juries, and the burden of proof, *New England Law Review* 41 (3) (2007) 435–470.
- [6] D.W. Denno, The myth of the double-edged sword: An empirical study of neuroscience evidence in criminal cases, *Boston College Law Review* 56 (2) (2015) 493–551.
- [7] D.W. Denno, How prosecutors and defense attorneys differ in their use of neuroscience evidence, *Fordham Law Review* 85 (2) (2016) 453–480.
- [8] L. Devers, Plea and charge bargaining. Bureau of Justice Statistics (2011).
- [9] G. Edmond, J. Vuille, Comparing the use of forensic science evidence in Australia, Switzerland, and the United States: Transcending the adversarial-nonadversarial dichotomy, *Jurimetrics* 54 (3) (2014) 221–276.
- [10] B.A.J. Fisher, D.R. Fisher, *Techniques of Crime Scene Investigation*, CRC Press, 2003.
- [11] Gallup. Crime. (2019). Retrieved from <https://news.gallup.com/poll/1603/crime.aspx>.
- [12] B.L. Garrett, P.J. Neufeld, Invalid forensic science testimony and wrongful convictions, *Virginia Law Review* 95 (1) (2009) 1–97.
- [13] L.M. Gaudet, G.E. Marchant, Under the radar: Neuroimaging evidence in the criminal courtroom, *Drake Law Review* 64 (3) (2016) 577–662.
- [14] J. Goldstein, Police agencies are assembling records of DNA. *New York Times*. (2013). Retrieved from <https://www.nytimes.com/2013/06/13/us/police-agencies-are-assembling-records-of-dna.html>.
- [15] M.M. Houck, CSI: The reality, *Scientific American* 295 (1) (2006) 84–89.
- [16] S.M. Kassir, I.E. Dror, J. Kukucka, The forensic confirmation bias: Problems, perspectives, and proposed solutions, *Journal of Applied Research in Memory and Cognition* 2 (1) (2013) 42–52.
- [17] T. Kyckelhahn, T.H. Cohen, Felony defendants in large urban counties, 2004. Bureau of Justice Statistics. (2008).
- [18] J.D. Lieberman, C.A. Carrell, T. Miethe, D.A. Krauss, Gold versus platinum: Do jurors recognize the superiority and limitations of DNA evidence compared to other types of forensic evidence? *Psychology, Public Policy, and Law* 14 (1) (2008) 27–62.
- [19] H. Machado, S. Silva, What influences public views on forensic DNA testing in the criminal field? A scoping review of quantitative evidence, *Human Genomics* 13 (1) (2019) 23.
- [20] L. Mannix, Top judge worried forensic evidence putting innocent people behind bars. *The Age*. (2019). Retrieved from <https://www.theage.com.au/national/top-judge-worried-forensic-evidence-putting-innocent-people-behind-bars-20190823-p52k3l.html>.
- [21] P. Marcus, V. Wayne, Australia and the United States: Two common criminal justice systems uncommonly at odds, *Tulane Journal of International and Comparative Law* 12 (2004) 27–116.
- [22] T. McEwen, *The Role and Impact of Forensic Evidence in the Criminal Justice System*, Final Report, National Institute of Justice, 2011.
- [23] National Registry of Exonerations. Exonerations in 2018. (2019). Retrieved from <https://www.law.umich.edu/special/exoneration/Documents/Exonerations%20in%202018.pdf>.
- [24] J. Norman, Confidence in police back at historic average. Gallup. (2017). Retrieved from <https://news.gallup.com/poll/213869/confidence-police-back-historical-average.aspx>.
- [25] K. Podlas, “The CSI effect”: Exposing the media myth, *Fordham Intellectual Property, Media and Entertainment Law Journal* 16 (2) (2005) 429–466.
- [26] K. Podlas, The CSI effect and other forensic fictions, *Loyola of Los Angeles Entertainment Law Review* 27 (2) (2006) 87–126.
- [27] The Innocence Project. *Overturing wrongful convictions involving misapplied forensics*. (2019). Retrieved from <https://www.innocenceproject.org/overturing-wrongful-convictions-involving-flawed-forensics/>.
- [28] B.A. Reaves, Felony Defendants in Large Urban Counties, 2009 - Statistical Tables, Bureau of Justice Statistics, 2013.
- [29] G. Ribeiro, J.M. Tangen, B.M. McKimmie, Beliefs about error rates and human judgment in forensic science, *Forensic Science International* 297 (2019) 138–147.
- [30] J.K. Roman, S.E. Reid, A.J. Chalfin, C.R. Knight, The DNA field experiment: A randomized trial of the cost-effectiveness of using DNA to solve property crimes, *Journal of Experimental Criminology* 5 (2009) 345–369.
- [31] E. Smith, A.J. Hattery, Race, wrongful conviction & exoneration, *Journal of African American Studies* 15 (1) (2011) 74–94.
- [32] L.L. Smith, R. Bull, Identifying and measuring juror pre-trial bias for forensic evidence: Development and validation of the Forensic Evidence Evaluation Bias Scale, *Psychology, Crime & Law* 18 (9) (2012) 797–815.
- [33] L.L. Smith, R. Bull, Validation of the factor structure and predictive validity of the Forensic Evidence Evaluation Bias Scale for robbery and sexual assault trial scenarios, *Psychology, Crime & Law* 20 (5) (2014) 450–466.
- [34] The National Research Council, *Strengthening Forensic Science in the United States: A Path Forward*, National Academies Press, 2009.
- [35] The 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. Executive Office of the President. (2016). Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/peast_forensic_science_report_final.pdf.
- [36] The Scientific Literature Working Group, *Speaker Recognition Subcommittee. Foundational scientific literature for forensic speaker recognition*, 1st edition. (2019). Retrieved from <https://www.nist.gov/topics/forensic-science/speaker-recognition-subcommittee>.
- [37] V. Wayne, P. Marcus, Australia and the United States: Two common criminal justice systems uncommonly at odds, Part 2, *Tulane Journal of International and Comparative Law* 18 (2) (2010) 335–402.

**UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF COLUMBIA**

UNITED STATES OF AMERICA, :
 :
 v. : Criminal Action No.: 19-358 (RC)
 :
 DEMONTRA HARRIS, : Re Document No.: 22
 :
 Defendant. :

MEMORANDUM OPINION

**DENYING DEFENDANT’S MOTION IN LIMINE TO EXCLUDE EXPERT TESTIMONY AS TO
FIREARM EXAMINATION TESTING**

I. INTRODUCTION

Defendant Demontra Harris is charged with unlawful possession of a firearm as a person previously convicted of a felony, assault with a dangerous weapon, and possession of a firearm during a crime of violence. Superseding Indictment at 1–2, ECF No. 39. On July 24, 2019, the D.C. Metropolitan Police Department (“MPD”) responded to a report of gunshots and recovered four 9mm shell casings from the incident scene, which were then entered into the National Integrated Ballistic Information Network (“NIBIN”). A witness later provided MPD with a video filmed that night that allegedly shows Mr. Harris holding and then discharging a firearm in the location where the shell casings were later discovered. No firearm was recovered at the time. Roughly six weeks later on September 8, 2019, during a response to a call for service for a person with a weapon, MPD recovered a Glock 17 Gen4 9x19 pistol (“Glock 17”). This recovered firearm was test-fired and the resulting casings were entered into the NIBIN, where a match was identified with the casings recovered on the night of July 24, 2019. The Government then submitted the relevant evidence to an independent firearms examiner for forensic examination. Chris Monturo, a tool mark examiner who operates the Ohio-based forensic

services firm Precision Forensic Testing, examined the evidence and concluded in a report that he believed the four recovered casings from the July 24, 2019 incident scene were fired by the recovered Glock 17. *See* March 14, 2020 Report of Chris Monturo (“Monturo Report”), ECF No. 22-2. The Government intends to call Mr. Monturo to testify regarding these findings at the upcoming trial in this matter.

This opinion addresses Mr. Harris’s *motion in limine* to Exclude Expert Testimony as to Firearm Examination Testing (“Def.’s Mot.”), ECF No. 22, pursuant to *Daubert v. Merrell Dow Pharm. Inc.*, 509 U.S. 579 (1993), Federal Rule of Evidence 702, and Federal Rule of Evidence 403. Def.’s Mot. at 1–2. The motion has been fully briefed, with both parties also filing supplemental motions. *See generally* Def.’s Mot.; Govt.’s Opp’n to Def.’s Mot. to Excl. Firearm and Toolmark Testimony (“Govt. Opp’n”), ECF No. 28; Def.’s Supp. Mot. to Excl. Expert Testimony as to Firearm Exam. Testing (“Def.’s Supp. Mot.”), ECF No. 32; Govt.’s Opp’n to Def.’s Supp. to Excl. Firearm and Toolmark Testimony (“Govt. Supp. Opp’n”), ECF No. 33. In addition, the Court conducted a *Daubert* hearing on October 15, 2020 to consider this issue, taking the testimony of Todd Weller, an expert in the field. A jury trial in this matter is currently scheduled to begin on November 12, 2020.

Mr. Harris argues that the field of firearm and toolmark identification lacks a reliable scientific basis and is not premised on sufficient facts or data, is not the product of reliable principles and methods, and was not applied properly by Mr. Monturo to the facts of the case. Def.’s Mot. at 1–2. The Court disagrees, and will admit Mr. Monturo’s testimony to the extent it falls within the Department of Justice’s Uniform Language for Testimony of Reports for the Forensic Firearms/Toolmarks Discipline – Pattern Matching Examination (“DOJ ULTR”). While Mr. Harris raises important issues as to the reliability of firearm and toolmark

identification, memorialized most notably by the 2016 President’s Council of Advisors on Science and Technology Report (“PCAST Report”), these issues are for cross-examination, not exclusion, as recent advancements in the field in the four years since the PCAST Report address many of Mr. Harris’s concerns. Mr. Harris also remains free to have his own expert examine the firearm and ballistics evidence and contradict the Government’s case.

II. ANALYSIS

A. Legal Standard

“Motions *in limine* are designed to narrow the evidentiary issues at trial.” *Williams v. Johnson*, 747 F. Supp. 2d 10, 14 (D.D.C. 2010). “While neither the Federal Rules of Civil Procedure nor the Federal Rules of Evidence expressly provide for motions *in limine*, the Court may allow such motions ‘pursuant to the district court’s inherent authority to manage the course of trials.’” *Barnes v. District of Columbia*, 924 F. Supp. 2d 74, 78 (D.D.C. 2013) (quoting *Luce v. United States*, 469 U.S. 38, 41 n.4 (1984)).

Federal Rule of Evidence 702 provides that qualified expert testimony is admissible if “(a) the expert’s scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue; (b) the testimony is based on sufficient facts or data; (c) the testimony is the product of reliable principles and methods; and (d) the expert has reliably applied the principles and methods to the facts of the case.” Fed. R. Evid. 702. “In general, Rule 702 has been interpreted to favor admissibility.” *Khairkhwa v. Obama*, 793 F. Supp. 2d 1, 10 (D.D.C. 2011) (citing *Daubert v. Merrell Dow Pharm., Inc.*, 509 U.S. 579, 587 (1993); Fed. R. Evid. 702 advisory committee’s note to 2000 amendment (“A review of the caselaw after *Daubert* shows that the rejection of expert testimony is the exception rather than the rule.”)). Indeed, the Supreme Court has clarified that it is not exclusion, but rather “vigorous

cross-examination, presentation of contrary evidence, and careful instruction on the burden of proof” that “are the traditional and appropriate means of attacking shaky but admissible evidence.” *Daubert*, 509 U.S. at 596.

When considering the admissibility of expert evidence under Federal Rule of Evidence 702, district courts are required to “assume a ‘gatekeeping role,’ ensuring that the methodology underlying an expert’s testimony is valid and the expert’s conclusions are based on ‘good grounds.’” *Chesapeake Climate Action Network v. Export-Import Bank of the U.S.*, 78 F. Supp. 3d 208, 219 (D.D.C. 2015) (quoting *Daubert*, 509 U.S. at 590–97). This gatekeeping analysis is “flexible,” and “the law grants a district court the same broad latitude when it decides how to determine reliability as it enjoys in respect to its ultimate reliability determination.” *Kumho Tire Co. v. Carmichael*, 526 U.S. 137, 141–42 (1999) (emphasis omitted). While district courts may apply a variety of different factors to assess reliability, in *Daubert* the Supreme Court provided a non-exhaustive list of five factors to guide the determination, including: (1) whether the technique has been or can be tested; (2) whether the technique has a known or potential rate of error; (3) if the technique has been subject to peer review and publishing; (4) the existence of controls that govern the technique’s operation; and (5) whether the technique has been generally accepted within the relevant scientific community. *See Daubert*, 509 U.S. at 593–94. In contrast, expert testimony “that rests solely on ‘subjective belief or unsupported speculation’ is not reliable.” *Groobert v. President & Directors of Georgetown Coll.*, 219 F. Supp. 2d 1, 6 (D.D.C. 2002) (citing *Daubert*, 509 U.S. at 590).

“The burden is on the proponent of [expert] testimony to show by a preponderance of the evidence that . . . the testimony is reliable.” *Sykes v. Napolitano*, 634 F. Supp. 2d 1, 6 (D.D.C. 2009) (citing *Meister v. Med. Eng’g Corp.*, 267 F.3d 1123, 1127 n.9 (D.C. Cir. 2001)). Even if

the proposed expert testimony is reliable, the Court may nonetheless exclude it “if its probative value is substantially outweighed by a danger of one or more of the following: unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time, or needlessly presenting cumulative evidence.” Fed. R. Evid. 403; *see Bazarian Int’l Fin. Assocs., LLC v. Desarrollos Aerohotelco, C.A.*, 315 F. Supp. 3d 101, 128 (D.D.C. 2018) (analyzing expert testimony under Rule 403).

B. Firearm and Toolmark Identification

1. Firearm and Toolmark Identification Science

Mr. Harris’s motion challenges the reliability of the Government’s proposed use of firearm toolmark identification as a discipline for expert testimony. Firearm identification began as a forensic discipline in the 1920s, *see* James E. Hamby, *The History of Firearm and Toolmark Identification*, 31 Ass’n of Firearm and Tool Mark Examiners J. 266, 266–284 (1999), and “for decades” has been routinely admitted as appropriate expert testimony in district courts. *United States v. Taylor*, 663 F. Supp. 2d 1170, 1175 (D.N.M. 2009); *see also United States v. Brown*, 973 F.3d 667, 704 (7th Cir. 2020) (noting firearm and toolmark identification has been “almost uniformly accepted by federal courts”) (citations omitted).

Firearm and toolmark identification “is used to determine whether a bullet or casing was fired from a particular firearm.” *Brown*, 973 F.3d at 704. A firearm and toolmark examiner will make this determination “by looking through a microscope to see markings that are imprinted on the bullet or casing by the firearm during the firing process,” which will include marks left on the bullet by the firing pin as well as scratches that occur when the bullet travels down the barrel. *Id.*

A firearm examiner is trained to observe and classify these marks into three types of characteristics during a firearm toolmark examination, which include:

(1) Class characteristics: i.e., the weight or caliber of the bullet, the number of lands and grooves, the twist of the lands and grooves, and the width of the lands and grooves, that appear on all bullet casings fired from the same type of weapon and are predetermined by the gun manufacturer;

(2) Individual characteristics: unique, microscopic, random imperfections in the barrel or firing mechanism created by the manufacturing process and/or damage to the gun post-manufacture, such as striated and/or impressed marks, unique to a single gun; and

(3) Subclass characteristics: characteristics that exist, for example, within a particular batch of firearms due to imperfections in the manufacturing tool that persist during the manufacture of multiple firearm components mass-produced at the same time.

Ricks v. Pauch, No. 17-12784, 2020 WL 1491750, at *8–9 (E.D. Mich. Mar. 23, 2020).

A qualified examiner can conclude that casings were fired by the particular firearm by “comparatively examining bullets and determining whether ‘sufficient agreement’ of toolmarks exist,” which occurs when the class and individual characteristics match. *Id.* at *9; *see also Brown*, 973 F.3d at 704. The methodology of determining when sufficient agreement is present is detailed by the Association of Firearm Toolmark Examiners (“AFTE method”), and is “the field’s established standard.” *United States v. Ashburn*, 88 F. Supp. 3d 239, 246 (E.D.N.Y. 2015). Under the governing AFTE theory, no two firearms will bear the same microscopically identical toolmarks due to differences in individual characteristics. *United States v. Otero*, 849 F. Supp. 2d 425, 427 (D.N.J. 2012).

In recent years three scientific reports have examined the underlying scientific validity of firearm and toolmark identification. They include the 2008 Ballistic Imaging Report, Def.’s Supp. Mot. Ex. 1, ECF No. 32-1, the 2009 National Academy of Science Report, Def.’s Supp. Mot. Ex. 2, ECF No. 32-2, and the 2016 President’s Council of Advisors on Science and Technology Report (“PCAST Report”), Def.’s Supp. Mot. Ex. 3, ECF No. 32-3. Mr. Harris argues that these reports “reject the claim that firearms identification is a valid and reliable

science.” Def.’s Supp. Mot. at 2–3. The Court is generally convinced by the Government’s arguments and ample citations to case law that the 2008 Ballistic Imaging Report and the 2009 National Academy of Science Report are both “outdated by over a decade” due to intervening scientific studies and as a result have been repeatedly rejected by courts as a proper basis to exclude firearm and toolmark identification testimony. Govt. Supp. Opp’n at 2–4 (collecting cases holding firearms identification evidence admissible after considering these reports). The PCAST Report provides better support for Mr. Harris’s arguments, given its more recent origin and use in recent opinions that have interrogated the danger of subjectivity in this discipline. *See, e.g., United States v. Tibbs*, No. 2016-CF1-19431, 2019 WL 4359486 (D.C. Super. Ct. Sept. 5, 2019).

The PCAST Report ultimately concluded that firearm and toolmark identification fell “short of the criteria for foundational validity,” after raising a number of critiques of the science. PCAST Report at 11. Chief among them was that the report concluded that “foundational validity can only be established through multiple independent black-box studies”¹ and at the time the report was published in 2016, there had only been one black-box study conducted on the discipline to date. Def.’s Supp. Mot. at 4 (citing PCAST Report at 106, 111). In response, the Government has put forth sworn affidavits from researchers that speak to post-PCAST Report scientific studies that they argue contradicts the PCAST Report’s conclusions. The Government’s Daubert hearing expert, Todd Weller, devoted much of his testimony to

¹ The PCAST report defined a black-box study as “an empirical study that assesses a subjective method by having examiners analyze samples and render opinions about the origin or similarity of samples.” PCAST Report at 48. Mr. Weller added at the Evidentiary Hearing that a black-box study is one in which there are “question samples [given to examiners] that have a matching known, and question samples that do not have a matching known, and also that each of those comparisons is independent from each other.” October 15, 2020 Evidentiary Hearing Tr. (“Evid. Hr’g Tr.”) 49:6-12.

discussing the scientific advances that have occurred since the PCAST Report was published in 2016, all of which he posited affirms the discipline's validity. *See generally* Evid. Hr'g Tr.

2. Mr. Monturo's Report Methodology

Mr. Harris's *motion in limine* specifically challenges the proposed testimony of the Government's firearm and ballistics expert Chris Monturo, who examined the firearms evidence at issue in this case. In creating his report for the Government, Mr. Monturo first test fired the Glock 17 and found it to be operable. Monturo Report at 2. He then used the Glock 17 to create test-fired cartridge cases. *Id.* Mr. Monturo then microscopically compared his test-fired cartridge cases to the cartridge cases recovered from the crime scene on July 26, 2019, and found the two sets of cartridges "to have corresponding individual characteristics." *Id.* These results were then verified that same day by Calissa Chapin, another qualified firearm and ballistics expert from Mr. Monturo's lab. March 14, 2020 Report of Chris Monturo Notes ("Monturo Report Notes") at 3, ECF No. 22-3. As a result, Mr. Monturo is expected to testify that "[b]ased upon these corresponding individual characteristics. . . namely aperture sheer marks,"² "along with Mr. Monturo's training and experience, [he] is of the opinion that the Glock firearm fired" the cartridge casings recovered from the July 26, 2019 crime scene. Govt. Opp'n at 11-12.

C. The Subject Matter of Mr. Monturo's Testimony Meets Rule 702's Standards

Mr. Harris argues that the Government's proposed expert must be excluded under Rule 702 and *Daubert* because the underlying firearm and toolmark identification discipline "is based

² As defined in the AFTE Glossary, 6th Edition, a firing pin aperture shear is "[s]triated marks caused by the rough edges of the firing pin aperture scraping the primer metal during unlocking of the breech." Govt. Supp. Opp'n, Ex. 15, ECF No. 33-15. It is these individual characteristics Mr. Monturo used to classify the cartridge cases at issue.

not upon science but rather ‘subjectivity.’”³ Def.’s Supp. Mot. at 2. To address Mr. Harris’s concerns about the admission of Mr. Monturo’s expert testimony, the Court will undertake a factor-by-factor analysis of the discipline’s reliability, using *Daubert* as a guide. Complicating this process is the fact that Mr. Harris did not specifically address the *Daubert* criteria in his briefing on this topic, so the Court will instead rely on the implications raised by the PCAST Report and other scientific reports he has brought to the Court’s attention.

1. Whether the methodology has been tested

As previously noted, the first *Daubert* factor asks whether the technique in question has been or can be tested. *See Daubert*, 509 U.S. at 593–94. This “testability” inquiry, as articulated in the Advisory Committee Notes to Rule 702, concerns “whether the expert’s theory can be challenged in some objective sense, or whether it is instead simply a subjective, conclusory approach that cannot be reasonably assessed for reliability.” Fed. R. Evid. 702 advisory committee’s note to 2000 amendment. Mr. Harris argues that firearm and toolmark identification is “unavoidably subjective,” and also cites to the 2008 Ballistics Imaging Report which expressed concerns about “the fundamental assumptions of uniqueness and reproducibility of firearms-related toolmarks.” Def.’s Supp. Mot. at 2–3. In response, the Government has put forth evidence to show “[f]irearms and toolmark identification has been thoroughly tested with

³ Based on remarks such as these and his citation to *United States v. Glynn*, Mr. Harris appears to be peripherally raising the point that firearm and toolmark identification cannot “fairly be called ‘science,’” *United States v. Glynn*, 578 F. Supp. 2d 567, 570 (S.D.N.Y. 2008), a preliminary inquiry some courts have investigated before proceeding to the *Daubert* analysis. The Court does not believe such an inquiry is required here, given that, as other courts have also found, firearm and toolmaking identification is “clearly is technical or specialized, and therefore within the scope of Rule 702.” *United States v. Hunt*, No. CR-19-073-R, 2020 WL 2842844, at *3 n.2 (W.D. Okla. June 1, 2020) (citing *United States v. Willock*, 696 F. Supp. 2d 536, 571 (D. Md. 2010), *aff’d sub nom. United States v. Mouzone*, 687 F.3d 207 (4th Cir. 2012)).

ground-truth experiments designed to mimic casework.” Govt. Opp’n at 1. The Court agrees with the Government that this factor supports admissibility.

A number of courts have examined this factor in depth to conclude that firearm toolmark identification can be tested and reproduced. *See, e.g., Otero*, 849 F. Supp. 2d at 432 (“The literature shows that the many studies demonstrating the uniqueness and reproducibility of firearms toolmarks have been conducted.”); *Taylor*, 663 F. Supp. 2d at 1175–76 (noting studies “demonstrating that the methods underlying firearms identification can, at least to some degree, be tested and reproduced.”); *United States v. Diaz*, No. CR 05-00167, 2007 WL 485967, at *6 (N.D. Cal. Feb. 12, 2007) (holding that “the theory of firearms identification, though based on examiners’ subjective assessment of individual characteristics, has been and can be tested.”). Indeed, even Judge Edelman in the *Tibbs* opinion relied on by Mr. Harris concluded that “virtually every court that has evaluated the admissibility of firearms and toolmark identification has found the AFTE method to be testable and that the method has been repeatedly tested.” *Tibbs*, 2019 WL 439486 at *7 (collecting cases).

The fact that there are subjective elements to the firearm and toolmark identification methodology is not enough to show that the theory is not “testable.” Indeed, studies have shown that “the AFTE theory is testable on the basis of achieving consistent and accurate results.” *Otero*, 849 F. Supp. 2d at 433; *see also* July 7, 2017 Decl. of Todd Weller (“Weller I”) at 2–6, ECF No. 28-5 (describing various studies that support the reproducibility of the AFTE identification theory). This conclusion has only been further strengthened in recent years due to advances in three-dimensional imaging technology, which has allowed the field to interrogate the process and sources of “subjectivity” behind firearm and toolmark examiners’ conclusions. For example, Mr. Weller testified regarding a study which used 3D image technology to assess the

process used by trained firearm examiners when identifying casings to a particular firearm. *See* Sept. 19, 2019 Decl. of Todd Weller (“Weller II”) at 15–16 (citing Pierre Duez et al., *Development and Validation of a Virtual Examination Tool for Firearm Forensics*, 63 J. Forensic Sci, 1069–84 (2018), (“Heat Map Study”)), ECF No. 28-6. The Heat Map Study indicated that firearm examiners from fifteen different laboratories, all conducting an independent assessment, were “mostly using the same amount and same location of microscopic marks when concluding identification.” Weller II at 16. Critically, the trained examiners also correctly reported 100% of known matches while reporting no false positives or false negatives. *Id.*

It is also important to note that the testability criticism leveled at the firearm and toolmark field in the PCAST Report—that at the time of publishing “there [was] only a single appropriately designed study to measure validity and estimate reliability”—appears to now be out of date. PCAST Report at 112. As previously discussed, the PCAST Report only considered studies that were a “black-box” or “open-set” design, disregarding hundreds of validation studies in the process. *See* Evid. Hr’g Tr. 48:9-17 (noting that PCAST only evaluated nine of the hundreds of studies that were submitted for review). Setting aside for the moment the utility of this “black-box” requirement— which goes beyond what is required by Rule 702— the Government has provided to the Court three recent scientific studies that meet the PCAST’s black-box model requirements and demonstrate the reliability of the firearm and toolmark identification method. These include one of the tests administered during the Heat Map Study detailed above, *see* Weller II at 16 n. 84, along with another recent black box study testing the identification of fired casings, which resulted in a .433% false positive error rate from three errors among 693 total comparisons. *See* Lilien et al., *Results of the 3D Virtual Comparison*

Microscopy Error Rate (VCMER) Study for Firearm Forensics, J. of Forensic Sci. Oct. 1, 2020 (“Lilien Study”) at 1, ECF No. 41. A third post-PCAST Report study also followed the PCAST recommended black-box model and found that of 1512 possible identifications tested, firearms examiners correctly identified 1508 casings to the firearm from which the casing was fired. Keisler et. al., *Isolated Pairs Research Study*, Ass’n of Firearm and Tool Mark Examiners J. 56, 58 (2018) (“Keisler Study”), ECF No. 33-9; *see also* Evid. Hr’g Tr. 65:3-11. This evidence indicates that even under the PCAST’s stringent black-box only criteria, firearm and toolmark identification can be tested and reasonably assessed for reliability.

A final factor demonstrating the strength of the testability prong is that firearm and toolmark examiners are required, as Mr. Monturo has done here, to document their results and findings through written reports and photo documentation, and have these results validated by another qualified examiner. These elements “ensure sufficient testability and reproducibility to ensure that the results of the technique are reliable.” *Diaz*, 2007 WL 485967 at *5 (citing *United States v. Monteiro*, 407 F.Supp.2d 351, 369 (D. Mass. 2006)).⁴ For all of these reasons, the Court concludes that the testability factor supports admissibility of Mr. Monturo’s testimony.

2. The known or potential error rate

The second *Daubert* factor inquires as to whether the technique has a known or potential rate of error. *See Daubert*, 509 U.S. at 594. The PCAST Report concluded that non-black box

⁴ Mr. Harris’s only explicit acknowledgement of this *Daubert* factor is an assertion in a parenthetical that the court in *United States v. Green* found that “ballistic evidence fails to meet *Daubert* criteria regarding . . . testability.” Def.’s Mot. at 7 (citing *United States v. Green*, 405 F. Supp. 2d 104, 120–22 (D. Mass. 2005)). But the facts at issue in *Green* were quite different than the instant case. *Green*’s holding that the methods at issue could not be tested rested on an absence of notes and photographs from the initial examination that “made it difficult, if not impossible” for another expert to verify the examination. *Green*, 405 F. Supp. 2d at 120. In contrast, Mr. Monturo documented his work in addition to having it verified that same day by another certified firearms analyst. Accordingly, reproducibility is not at issue here.

studies had “inconclusive and false-positives rate that are dramatically lower (by more than 100-fold)” compared to partly black-box or fully black-box designed studies. PCAST Report at 109. The Government counters that “collectively, th[e] body of scientific data demonstrate[s] a low rate of error” for firearm and toolmark identification, and provides several recently published studies to refute the PCAST Report’s finding of differences in rate of error tied to study design. Govt. Opp’n at 2; Govt. Supp. Opp’n at 13–14.

First, as the Government argues and this Court agrees, the critical inquiry under this factor is the rate of error in which an examiner makes a false positive identification, as this is the type of error that could lead to a conviction premised on faulty evidence. *See Otero*, 849 F. Supp. 2d at 434 (noting, “the critical validation analysis has to be the extent to which false positives occur”).⁵ Mr. Weller testified that “over the past couple of decades in research” he had seen a rate of false positives in research studies ranging from 0-1.6 percent. Evid. Hr’g. Tr. 84:19–22. To support this assertion, the Government provided the false positive error rates for nineteen firearm and toolmark validation studies conducted between 1998 and 2019, of which eleven studies had a false positive error rate of zero percent, and the highest false positive error rate calculated was 1.6%. Govt. Opp’n at 27–29. Other federal courts have also recognized that validation studies as a whole show a low rate of error for firearm and toolmark identification. *See, e.g., United States v. Romero-Lobato*, 379 F. Supp. 3d 1111, 1119 (D. Nev. 2019) (“[T]he studies cited by [the firearms examiner] in his testimony and by other federal courts examining the issue universally report a low error rate for the AFTE method.”); *Taylor*, 663 F. Supp. 2d at 1177 (“[T]his number [less than 1%] suggests that the error rate is quite low”).

⁵ Perhaps the false negative rate could be important in a case where a defendant asserts his co-defendant (or a third party) was the culprit and examination of that person’s firearm tested negative. But that situation does not apply here.

As was the case under the testability prong of the *Daubert* analysis, here too recent studies have resolved some of the concerns raised by the PCAST Report. Mr. Weller described for the Court how three black box studies that post-date the PCAST Report all have extremely low rates of error. Govt. Supp. Opp'n at 14, Evid. Hr'g Tr. 65:2-77:8. The Heat Map and Keisler studies both had an overall error rate of zero percent, and the Lilien study produced a false positive rate of only 0.433%. Govt. Supp. Opp'n at 14. Because the evidence shows that error rates for false identifications made by trained examiners is low—even under the PCAST's black-box study requirements—this factor also weighs in favor of admitting Mr. Monturo's expert testimony.

3. Whether the methodology has been subject to peer review and publication

The third *Daubert* factor concerns if the methodology has been subject to peer review and published in scientific journals, a component the Supreme Court emphasized as critical to “good science” since “it increases the likelihood that substantive flaws in methodology will be detected.” *See Daubert*, 509 U.S. at 593–94. The Government contends that scientific data concerning firearms and toolmark identification “have been published in a multitude of scientific peer-reviewed journals,” Govt. Opp'n at 1, and Mr. Weller presented evidence to this effect at the evidentiary hearing, describing the variety of scientists from different disciplines who have published on the topic in several different peer-reviewed journals. *See Weller I* at 9–10. The Court agrees with the Government that this factor weighs in favor of admissibility.

Much of the literature in this discipline has been published in the AFTE Journal, a peer-reviewed journal that “publishes articles, studies and reports concerning firearm and toolmark evidence.” *United States v. McCluskey*, No. CR 10-2734 JCH, 2013 WL 12335325, at *6 (D.N.M. Feb. 7, 2013). The AFTE Journal uses a formal process for article submissions,

including “specific instructions for writing and submitting manuscripts, assignment of manuscripts to other experts within the scientific community for a technical review, returning of manuscripts to other experts within the scientific community for clarification or re-write, and a final review by the Editorial Committee.” *Id.* (quoting Richard Grzybowski, et al., *Firearm/Toolmark Identification: Passing the Reliability Test Under Federal and State Evidentiary Standards*, 35 AFTE J. 209, 220 (2003)).

Other courts have examined the scientific credibility of the AFTE Journal. Notably, the court in *Tibbs* concluded that the AFTE Journal’s lack of a double-blind peer review process along with the fact that it is published by the group of practicing firearms and toolmark examiners could create an “issue in terms of quality of peer review.” *Tibbs*, 2019 WL 4359486, at *10. In response, the Government asserts, citing to testimony from Dr. Bruce Budowle, “the most published forensic DNA scientist in the world,” that there is far from consensus in the scientific community that double-blind peer review is the only meaningful kind of peer review. Govt. Supp. Opp’n at 23; *see also* Affidavit of Bruce Budowle at 2, ECF No. 33–17. To this point, Mr. Weller described the various advantages and disadvantages of each type of peer review. *Weller II* at 22–24. Compellingly, the Government also refuted the allegation by Judge Edelman in *Tibbs* that the AFTE Journal does not provide “meaningful” review, by bringing to the Court’s attention a study that was initially published in the AFTE Journal, and then was subsequently published in the *Journal of Forensic Science* with no further alterations. Govt. Supp. Opp’n at 27. Because the *Journal of Forensic Science* employs a double-blind peer review process, this indicates that at least in this instance, the open peer review process of the AFTE Journal led to the same outcome as a double-blind peer review. *Id.* In addition, numerous courts have concluded that publication in the AFTE Journal satisfies this prong of the *Daubert*

admissibility analysis. *See, e.g., Romero-Lobato*, 379 F. Supp. 3d at 1119; *United States v. Johnson*, No. 16 Cr. 281, 2019 WL 1130258, at *16 (S.D.N.Y. Mar. 11, 2019); *Ashburn*, 88 F. Supp. 3d at 245–46; *Otero*, 849 F. Supp. 2d at 433; *Taylor*, 663 F. Supp. 2d at 1176; *Monteiro*, 407 F. Supp. 2d at 366–67. The Court queries whether excluding certain journals from consideration based on the type of peer review the journal employs goes beyond a court’s appropriate gatekeeping function under *Daubert*.

And even if the Court were to discount the numerous peer-reviewed studies published in the AFTE Journal, Mr. Weller’s affidavit also cites to forty-seven other scientific studies in the field of firearm and toolmark identification that have been published in eleven other peer-reviewed scientific journals. *Weller II* at Ex. A. This alone would fulfill the required publication and peer review requirement.

Because the toolmark identification methodology used by Mr. Monturo has been subject to peer review and publication, the Court finds this *Daubert* factor to also weigh in favor of admission.

4. The existence and maintenance of standards to control the methodology’s operation

The fourth *Daubert* factor inquires as to whether there are proper standards and controls to govern the operation of the technique in question. *See Daubert*, 509 U.S. at 594. Mr. Harris argues that there are insufficient objective standards in place, citing to the PCAST Report to claim that the AFTE’s “sufficient agreement” analysis that is used by examiners to reach their conclusions is subjective and impermissibly based on the “personal judgment” of each examiner. *Def.’s Supp. Mot.* at 4 (citing PCAST Report at 47, 60, 104, 113). In opposition, the Government argues that “the firearms community has implemented standards,” citing to a

number of industry guidebooks and regulations. Govt. Opp'n at 2. While a close call, the Court finds that the lack of objective standards ultimately means this factor cannot be met.⁶

The Government identifies a number of what they refer to as “standards for professional guidance” for the firearm and toolmark profession, Govt. Opp'n at 32–33, but the primary standard that governs the discipline is the AFTE Theory of Identification, which describes the methodology examiners should undertake when “pattern matching” between firearms and cartridges. *See, e.g.*, Govt. Opp'n at 8 (explaining that Theory of Identification was created “to explain the basis of opinion of common origin in toolmark comparisons”). According to the AFTE Theory of Identification, examiners can conclude that a firearm and cartridges have a common origin when a comparison of toolmarks shows there is “sufficient agreement” between “the unique surface contours of two toolmarks.” The Association of Firearm and Tool Mark Examiners, *AFTE Theory of Identification as It Relates to Toolmarks*, <https://afte.org/about-us/what-is-afte/afte-theory-of-identification> (last visited November 4, 2020). This theory of identification dictates that “sufficient agreement” between two toolmarks exists only when “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 Court finds this standard to be generally vague, and indeed, the AFTE Theory acknowledges that “the interpretation of individualization/identification is subjective in nature, founded on scientific principles and based on the examiner’s training and experience.” *Id.* As other courts have found, under this method “matching two tool marks essentially comes down to the examiner's subjective judgment based on his training, experience, and knowledge of firearms.”

⁶ This *Daubert* factor is, as the Government concedes, “the only *Daubert* factor that some courts have found lacking” in firearm toolmark identification. Govt. Opp'n at 33. This makes it all the more puzzling that the Government fails entirely to address this factor in its reply.

Romero-Lobato, 379 F. Supp. 3d at 1121; *Glynn*, 578 F. Supp. 2d at 572 (“[T]he standard defining when an examiner should declare a match – namely ‘sufficient agreement’ – is inherently vague.”).

Accordingly, it is evident and hardly disputed that the “AFTE theory lacks objective standards.” *Ricks*, 2020 WL 1491750, at *10. The entire process of reaching a conclusion regarding the “sufficient agreement in individual characteristics” is one that relies wholly on the examiner’s judgment, without any underlying numerical standards or guideposts to direct an examiner’s conclusion. *See Evid. Hr’g Tr.* 37:16–38:25 (noting the absence at this time of objective standards to guide an examiner’s findings). And as Mr. Weller testified, even in contrast to other subjective disciplines such as fingerprint analysis, firearm toolmark identification does not provide objective standards even as a quality control measure, such as a baseline to trigger further verification. *See Evid. Hr’g Tr.* 112:18-113:17 (explaining that while fingerprint testing does not have an agreed-upon standard for the number of matching points required for an identification, it does use matching points as a quality control measure that triggers further verification if below a certain threshold). While Mr. Monturo’s additional use of “basic scientific standards” through taking contemporaneous notes, documenting his comparison with photographs, and the use of a second reviewer for verification surely assist in maintaining reliable results, without more the Court cannot conclude this *Daubert* factor is met.

It should be noted, however, that even if this factor cannot be met, a partially subjective methodology is not inherently unreliable, or an immediate bar to admissibility. Rule 702 “does not impose a requirement that the expert must reach a conclusion via an objective set of criteria or that he be able to quantify his opinion with a statistical probability. *Romero-Lobato*, 379 F. Supp. 3d at 1120. And indeed, “all technical fields which require the testimony of expert

witnesses engender some degree of subjectivity requiring the expert to employ his or her individual judgment, which is based on specialized training, education, and relevant work experience.” *Johnson*, 2019 WL 1130258 at *18 (citations omitted); *see also* Evid. Hr’g Tr. at 30:14–31:6 (Mr. Weller testified that “all science involves some level of interpretation,” and went on to describe subjective components to both drug testing and DNA interpretation). Accordingly, this factor weighs against the admission of Mr. Monturo’s testimony, but does not disqualify it.

5. Whether the methodology has achieved general acceptance in the relevant community

Finally, the fifth and last *Daubert* factor asks whether the technique has been generally accepted within the relevant scientific community, reasoning that “a known technique which has been able to attract only minimal support within the community, may properly be viewed with skepticism.” *See Daubert*, 509 U.S. at 594. The Court finds that the Government has put forth more than sufficient evidence to show that the AFTE theory as used by Mr. Monturo enjoys widespread scientific acceptance. *See* Govt. Opp’n at 2; Govt. Supp. Opp’n at 28.

Mr. Weller testified that firearm and toolmark identification is practiced by accredited laboratories in the United States and throughout the world, including England (Scotland Yard), New Zealand, Canada, South Africa, Australia, Germany, Sweden, Greece, Turkey, China, Mexico, Singapore, Malaysia, Belgium, Netherlands, and Denmark. *See* Weller II at 30. In the United States alone, there are 233 accredited firearm and toolmark laboratories, that often operate within a larger forensic laboratory providing chemistry, DNA, and fingerprint identification, and scientists from a variety of disciplines author studies within the area of firearms and toolmark identification. *Id.*

The criticism contained in the PCAST Report does not undermine this factor, as “techniques do not need to have universal acceptance before they are allowed to be presented before a court.” *Romero-Lobato*, 379 F. Supp. 3d at 1122. Even courts that have been critical of the validity of the discipline have conceded that it does enjoy general acceptance as a reliable methodology in the relevant scientific community of examiners. *See Otero*, 849 F. Supp. 2d at 435 (collecting cases). Furthermore, as Mr. Weller noted at the evidentiary hearing, the committee responsible for the PCAST Report did not include any firearm and toolmark examiners or researchers in the field, *see Evid. Hr’g Tr.* 47:18-23, thus raising the question of whether the PCAST Report criticism would even constitute a lack of acceptance from the “relevant scientific community.” For all of these reasons, this factor weighs in favor of admitting Mr. Monturo’s testimony.

6. The *Daubert* Analysis Urges Admission of Mr. Monturo’s Testimony

Balancing all five *Daubert* factors, the Court finds that the Government’s proposed expert testimony of Mr. Monturo is reliable and admissible, though subject to what the Court considers prudent limitations, discussed in detail below. The only factor that does not favor admissibility is the lack of objective criteria under the fourth *Daubert* factor, but as discussed, “the subjectivity of a methodology is not fatal under Rule 702 and *Daubert*.” *Ashburn*, 88 F. Supp. 3d at 246. And as other courts have also found, this deficiency “is countered by the method’s relatively low rate of error, widespread acceptance in the scientific community, testability, and frequent publication in scientific journals.” *Romero-Lobato*, 379 F. Supp. 3d at 1122. Accordingly, the Court will allow the admission of Mr. Monturo’s expert testimony as to his firearm and toolmark identification analysis, subject to certain limitations.

D. Federal Rule of Evidence 702(d)

Federal Rule of Evidence 702(d) provides that qualified expert testimony is admissible only when “the expert has reliably applied the principles and methods to the facts of the case.” Fed. R. Evid. 702. Mr. Harris challenges the admission of Mr. Monturo’s testimony, asserting that he “has not applied the principles and methods reliably to the facts of the case.” Def.’s Mot. at 1. However, he provides no evidence or further analysis to flesh out this conclusory claim. Accordingly, the Court finds this argument to be without merit.

As previously described, Mr. Monturo detailed the firearm and toolmark examination he conducted in his report, providing both a description of his process and photo documentation. *See generally* Monturo Report. Mr. Monturo’s findings were then verified by another qualified examiner the same day. Monturo Report Notes at 2. In contrast, Mr. Harris has not put forth any evidence to suggest that Mr. Monturo applied the firearm and toolmarking methodology in an unreliable manner. Mr. Monturo also appears to be well-qualified, with the Government noting that he “has significant training and experience, has not failed any proficiency exams, and has designed consecutively manufactured firearms test kits for training other firearms examiners,” information that they plan to elicit at trial during qualification of his testimony and also set out in his curriculum vitae. Govt. Opp’n at 35. In light of his failure to identify any unreliability on Mr. Monturo’s part, and also because Mr. Harris will have the ability to question Mr. Harris regarding his analysis during cross examination, the Court is convinced exclusion on this ground is not warranted. *See Daubert*, 509 U.S. at 596 (“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.”). If Mr. Harris has lingering concerns about Mr. Monturo’s application of the firearm and toolmark methodology in this case,

he is welcome to retain an independent expert to review Mr. Monturo's work, or have an independent examination of his own performed.

E. Federal Rule of Evidence 403

Next, Mr. Harris argues that even if the proposed testimony of Mr. Monturo is admissible pursuant to *Daubert* and Federal Rule of Evidence 702, it is inadmissible under Federal Rule of Evidence 403. Def. Mot. at 2. In support of this claim, Mr. Harris argues that Mr. Monturo's "conclusions appear to extend beyond his claimed expertise and are not reliable since they are not based on objective standards but rather his subjective observations and conclusions." *Id.* "The prejudice to Mr. Harris is simple, a connection to a firearm, a connection to a shell casing, all premised on analysis that at its best can only conclude that it 'may' be correct." Def. Supp. Mot. at 2.

Under Rule 403, a Court may exclude otherwise probative testimony if its value is substantially outweighed by unfair prejudice, confusing the issues, misleading the jury, undue delay, a waste of time, or cumulative evidence. Fed. R. Evid. 403. Mr. Harris's concern under Rule 403 appears to be that the value of Mr. Monturo's testimony will be substantially outweighed by the risk of him potentially misleading the jury through his reliance on a methodology Mr. Harris does not believe is sufficiently reliable. First, Mr. Harris's concerns about the reliability of the firearm and toolmarking methodology have already been analyzed, and the Court has found the underlying analysis sufficiently reliable such that Mr. Harris's concerns do not "substantially outweigh" the value of Mr. Monturo's testimony. Additionally, the Court believes that the risk of prejudice raised here can be alleviated through alternatives to exclusion. Cross-examination of Mr. Monturo's testimony, in conjunction with the appropriate limiting instruction governing the degree of certainty Mr. Monturo can express about his conclusions will sufficiently deter the risks of harm Mr. Harris has raised.

F. Limiting Instruction

In his final request, Mr. Harris asks that if the testimony of Mr. Monturo is not excluded, then the Court put in place limitations on his testimony. Def. Supp. Mot. at 6–7. Specifically, he requests that Mr. Monturo not “use the term ‘match’” but he “may be allowed to tell the jury that he could not exclude the gun as the weapon that produced a casing.” *Id.*

Limitations restricting the degree of certainty that may be expressed on firearm and toolmark expert testimony are not uncommon. *See, e.g., Romero-Lobato*, 379 F. Supp. 3d at 1117 (noting the “general consensus” of the courts “is that firearm examiners should not testify that their conclusions are infallible or not subject to any rate of error, nor should they arbitrarily give a statistical probability for the accuracy of their conclusions”); *Ashburn*, 88 F. Supp. 3d at 249 (limiting expressions of an expert’s conclusions to that of a “reasonable degree of ballistics certainty” or a “reasonable degree of certainty in the ballistics field.”); *Diaz*, 2007 WL 485967 at *1 (same).

With respect to Mr. Harris’s stated concerns, the Government has already agreed to a number of limitations on Mr. Monturo’s testimony, chief among them that he will not use terms such as “match,” he will “not state his expert opinion with any level of statistical certainty,” and he will not use the phrases when giving his opinion of “to the exclusion of all other firearms” or “to a reasonable degree of scientific certainty.” Govt. Opp’n at 12. These limitations are in accord with the Department of Justice Uniform Language for Testimony and Reports for the Forensic Firearms/Toolmarks Discipline—Pattern Matching Examination. *See* Govt. Opp’n, Ex. 4 (“DOJ ULTR”), ECF No. 28-4. The DOJ ULTR permits firearms examiners to conclude that casings were fired from the same firearm when all class characteristics are in agreement, and “the quality and quantity of corresponding individual characteristics is such that the examiner

would not expect to find that same combination of individual characteristics repeated in another source and has found insufficient disagreement of individual characteristics to conclude they originated from different sources.” *Id.* at 2–3. This Court believes, as other courts have also concluded, *see Hunt*, 2020 WL 2842844, at *8, that the testimony limitations as codified in the DOJ ULTR are reasonable and should govern the testimony at issue here. Accordingly, the Court instructs Mr. Monturo to abide by the expert testimony limitations detailed in the DOJ ULTR.

III. CONCLUSION

For the foregoing reasons, Defendant’s Motion to Exclude Expert Testimony as to Firearm Examination Testing, ECF No. 22, is DENIED. An order consistent with this Memorandum Opinion is separately and contemporaneously issued.

Dated: November 4, 2020

RUDOLPH CONTRERAS
United States District Judge

2020 WL 2842844

United States District Court, W.D. Oklahoma.

UNITED STATES of America, Plaintiff,


v.

Dominic Eugene HUNT, Defendant.

Case No. CR-19-073-R

Signed 06/01/2020


Synopsis

Background: Defendant was charged with being a felon in possession of ammunition. Defendant moved in limine to exclude ballistic evidence, or alternatively, for  *Daubert* hearing.

Holdings: The District Court, [David L. Russell](#), Senior District Judge, held that:

[1] expert testimony derived from Association of Firearms and Toolmark Examiners (AFTE) methodology was reliable and therefore admissible;

[2] experts reliably applied AFTE method;

[3] formal  *Daubert* hearing in advance of qualifying expert was not required; and

[4] experts could testify that their conclusions were reached to reasonable degree of ballistic certainty.

Motion denied.

Procedural Posture(s): Pre-Trial Hearing Motion.

West Headnotes (13)

[1] **Criminal Law**  [Subjects of Expert Testimony](#)

When it comes to the admissibility of expert evidence, a district court maintains the role of gatekeeper. [Fed. R. Evid. 702](#).

[2] **Criminal Law**  [Knowledge, Experience, and Skill](#)

Criminal Law  [Necessity and sufficiency](#)

A court assesses proffered expert testimony to ensure it is both relevant and reliable; to do this, the court generally first determines whether the expert is qualified, and if the expert is sufficiently qualified, the court then determines whether the expert's opinion is reliable. [Fed. R. Evid. 702](#).

[3] **Criminal Law**  [Hearing, ruling, and objections](#)

When faced with a party's objection to proffered expert testimony, a court must adequately demonstrate by specific findings on the record that it has performed its duty as gatekeeper, although it has discretion in how it performs its gatekeeping function. [Fed. R. Evid. 702](#).

[4] **Criminal Law**  [Preliminary evidence as to competency](#)

The proponent of expert testimony bears the burden of showing that its proffered expert's testimony is admissible. [Fed. R. Evid. 702](#).

[5] **Criminal Law**  [Necessity and sufficiency](#)

A court assesses the reasoning and methodology underlying the expert's opinion to determine reliability. [Fed. R. Evid. 702](#).

[6] **Criminal Law**  [Necessity and sufficiency](#)

The proponent has to show a court only that its expert opinion is reliable, not that it is substantively correct, because the reliability standard is lower than the merits standard of correctness.

[7] **Criminal Law**  [Necessity and sufficiency](#)


The reliability inquiry for expert testimony is specific to the case and facts: no one factor is dispositive or always applicable, and the goal

remains ensuring that an expert employs the same level of intellectual rigor in the courtroom that characterizes the practice of an expert in the relevant field. [Fed. R. Evid. 702](#).

[8] Criminal Law 🔑 Identification of persons, things, or substances

Expert testimony derived from Association of Firearms and Toolmark Examiners (AFTE) methodology was reliable and therefore admissible in defendant's trial on felon in possession charges; although AFTE's processes were subjective and some peer review was unfavorable, method had been tested, it had been reviewed by peers and subject to publication, it had been found to have potential low rate of error, and it had been widely accepted in relevant community. [Fed. R. Evid. 702](#).


[9] Criminal Law 🔑 Necessity and sufficiency

 *Daubert* does not mandate a technique, such as a black-box study, to satisfy its error rate element.

[10] Criminal Law 🔑 Identification of persons, things, or substances


Experts reliably applied Association of Firearms and Toolmark Examiners (AFTE) method, as required for expert testimony to be admissible in defendant's trial on felon in possession charges, where experts wrote detailed reports explaining their analysis, those reports were reviewed by other examiners in field, experts' examination reports detailed what case-specific facts of which they were aware when drawing their conclusions, and they demonstrated their experience, certifications, and continued training. [Fed. R. Evid. 702\(d\)](#).

[11] Criminal Law 🔑 Hearing, ruling, and objections

Formal  *Daubert* hearing in advance of qualifying expert on Association of Firearms and

Toolmark Examiners (AFTE) method was not required for expert testimony to be admissible in defendant's trial on felon in possession charges, since reliability of government's expert testimony was sufficiently addressed on the briefs. [Fed. R. Evid. 702](#).

[12] Criminal Law 🔑 Hearing, ruling, and objections

A court is not required to hold a formal  *Daubert* hearing in advance of qualifying an expert. [Fed. R. Evid. 702](#).

[13] Criminal Law 🔑 Identification of persons, things, or substances

In defendant's trial on felon in possession charges, experts on Association of Firearms and Toolmark Examiners (AFTE) method could testify that their conclusions were reached to reasonable degree of ballistic certainty, reasonable degree of certainty in field of firearm toolmark identification, or any other version of that standard, but they could not assert that two toolmarks originated from same source to exclusion of all other sources, assert that examinations conducted in forensic firearms-toolmarks discipline were infallible or had zero error rate, provide conclusion that included statistic or numerical degree of probability except when based on relevant and appropriate data, or cite number of examinations conducted in forensic firearms-toolmarks discipline performed in his or her career as direct measure for accuracy of proffered conclusion. [Fed. R. Evid. 702](#).


[1 Cases that cite this headline](#)

Attorneys and Law Firms

Jacquelyn M. Hutzell, US Attorney's Office, Oklahoma City, OK, for Plaintiff.

ORDER

DAVID L. RUSSELL, UNITED STATES DISTRICT JUDGE


*1 Before the Court is Defendant Dominic Hunt's Motion in Limine to Exclude Ballistic Evidence, or Alternatively, for a  *Daubert* Hearing. Doc. No. 67. The Government has responded in opposition to the motion. Doc. No. 81. Upon review of the parties' submissions, the Court denies Defendant's motion.

I. Background






On November 6, 2019, a federal grand jury returned a nine-count, third superseding indictment charging Defendant with, as relevant here, two counts of being a felon in possession of ammunition. Doc. No. 41. The two counts—Counts Eight and Nine—stem from two shootings: One in January of 2019 and another in February of 2019. *Id.* During the Oklahoma Police Department's (OCPD) investigation at the scene of the first shooting, officers found a Blazer 9mm Luger cartridge casing—the basis for Count Eight. *Id.* at 5–6. During the OCPD's investigation at the scene of the second shooting, officers found a Blazer 9mm Luger cartridge casing and two Winchester 9mm Luger cartridge casings—the basis for Count Nine. *Id.* at 6. Ronald Jones, a firearm and toolmark examiner for the OCPD, examined the casings and concluded that all four casings were likely fired from the same unknown firearm, potentially a Smith & Wesson 9mm Luger caliber pistol. Doc. Nos. 81–1, 81–2. Howard Kong, a firearm and toolmark examiner for the Bureau of Alcohol, Tobacco, Firearms and Explosives' (ATF) Forensic Science Laboratory, found the same. Doc. No. 81–4. The Government anticipates calling Mr. Jones and Mr. Kong at trial to “testify regarding their training, experience, and qualifications, the basis for firearms identification, their methods of examination in this case, their findings, and the basis for those findings.” Doc. No. 81, pp. 4–5. Specifically, the Government intends its experts to testify that:

(1) the ammunition charged in Count Eight was not fired from the Springfield Armory 9mm Luger caliber pistol [the Defendant's brother] had on March 11, 2019; (2) the

ammunition charged in Count Eight was not fired from the Smith & Wesson .40 caliber pistol [the Defendant's cousin] was convicted of possessing on January 20, 2019; (3) the probability the ammunition charged in Count Nine were fired in different firearms is so small it is negligible; (4) the ammunition charged in Count Nine was not fired from [the] Smith & Wesson .40 caliber pistol ...; (5) the probability the ammunition charged in Counts Eight and Nine were fired in different firearms is so small it is negligible; and (6) the unknown firearm was likely a Smith & Wesson 9mm Luger caliber pistol.


Id. Defendant now moves to exclude the testimony of Mr. Jones and Mr. Kong, or alternatively, for a  *Daubert* hearing. Doc. No. 67.

II. Legal Standard

[1] [2] [3] [4] When it comes to the admissibility of expert evidence, district courts maintain the role of gatekeeper.  *Bitler v. A.O. Smith Corp.*, 400 F.3d 1227, 1232 (10th Cir. 2005). In that role, district courts must adhere to Federal Rule of Evidence 702, which demands that courts “assess proffered expert testimony to ensure it is both relevant and reliable.” *United States v. Avitia-Guillen*, 680 F.3d 1253, 1256 (10th Cir. 2012). To do this, “the district court generally must first determine whether the expert is qualified”  *United States v. Nacchio*, 555 F.3d 1234, 1241 (10th Cir. 2009) (en banc). If the expert is sufficiently qualified, then “the court must determine whether the expert's opinion is reliable”  *Id.* “Although a district court has discretion in how it performs its gatekeeping function, ‘when faced with a party's objection, [the court] must adequately demonstrate by specific findings on the record that it has performed its duty as gatekeeper.’ ” *Avitia-Guillen*, 680 F.3d at 1257 (quoting  *Goebel v. Denver & Rio Grande W. R.R. Co.*, 215 F.3d 1083, 1088 (10th Cir. 2000)). “The proponent of expert testimony bears the burden of showing that its proffered expert's testimony is admissible.”  *Nacchio*, 555 F.3d at 1241.




*2 Here, Defendant Hunt does not object to the relevancy of the experts' testimony nor to the experts' qualifications. Defendant objects only to the reliability of the experts' testimony. Doc. No. 67, pp. 11–18. Therefore, the Court need only address whether the experts' testimony is reliable. *See Avitia-Guillen*, 680 F.3d at 1257.

[5] [6] [7] “To determine reliability, courts assess the reasoning and methodology underlying the [experts'] opinion” *Thompson v. APS of Oklahoma, LLC*, No. CIV-16-1257-R, 2018 WL 4608505, at *4 (W.D. Okla. Sept. 25, 2018) (internal quotation marks and citation omitted). “The reliability standard is lower than the merits standard of correctness, and plaintiffs need only show the Court that their experts' opinions are reliable, not that they are substantively correct.” *Id.* (internal quotation marks and citation omitted).

In  *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 113 S.Ct. 2786, 125 L.Ed.2d 469 (1993), the Supreme Court provided a non-exhaustive list of factors to aid in this determination:

- (1) whether the particular theory can be and has been tested;
- (2) whether the theory has been subjected to peer review and publication;
- (3) the known or potential rate of error;
- (4) the existence and maintenance of standards controlling the technique's operation; and
- (5) whether the technique has achieved general acceptance in the relevant scientific or expert community.

United States v. Baines, 573 F.3d 979, 985 (10th Cir. 2009)

(citing  *Daubert*, 509 U.S. at 592–94, 113 S.Ct. 2786).¹ The reliability inquiry, however, is fact- and case-specific: no one factor is dispositive or always applicable, and the goal remains “ensuring that an expert ‘employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field.’ ”  *Bitler*, 400 F.3d at 1233 (quoting  *Kumho Tire Co. v. Carmichael*, 526 U.S. 137, 152, 119 S.Ct. 1167, 143 L.Ed.2d 238 (1999)).

III. Firearm Toolmark Identification

In his motion, Defendant challenges the Governments use of firearm toolmark identification. “Forensic toolmark identification is a discipline that is concerned with the matching of a toolmark to the specific tool that made it. Firearm identification is a specialized area of toolmark identification dealing with firearms, which involve a specific category of tools.” *United States v. McCluskey*, No. 10-2734, 2013 WL 12335325, at *3 (D.N.M. Feb. 7, 2013) (citation omitted). “Toolmark identification is based on the theory that tools used in the manufacture of a firearm leave distinct marks on various firearm components, such as the barrel, breech face, or firing pins ... [and] that the marks are individualized to a particular firearm through changes the tool undergoes each time it cuts and scrapes metal to create an item in the production of the weapon.” *Id.* at *4. The field of firearm toolmark examination is based on the theory that some of these markings will be transferred to a bullet fired from the gun. *Id.* In conducting a firearm toolmark examination, a firearms examiner observes three types of characteristics:






*3 (1) Class characteristics: i.e., the weight or caliber of the bullet, the number of lands and grooves, the twist of the lands and grooves, and the width of the lands and grooves, that appear on all bullet casings fired from the same type of weapon and are predetermined by the gun manufacturer;



(2) Individual characteristics: unique, microscopic, random imperfections in the barrel or firing mechanism created by the manufacturing process and/or damage to the gun post-manufacture, such as striated and/or impressed marks, unique to single gun; and

(3) Subclass characteristics: characteristics that exist, for example, within a particular batch of firearms due to imperfections in the manufacturing tool that persist during the manufacture of multiple firearm components mass-produced at the same time.

Ricks v. Pauch, No. 17-12784, 2020 WL 1491750, at *8–9 (E.D. Mich., 2020). Pursuant to the theory used by the Government's experts in this case—the Association of Firearms and Toolmark Examiners (AFTE) method—“a qualified examiner can determine whether two bullets were fired by the same gun by comparatively examining bullets and determining whether ‘sufficient agreement’ of toolmarks exist,” meaning that there is significant similarity in the individual markings found on each bullet. *Id.* at *9.

IV. *Daubert* Analysis



[8] The use of this type of firearm toolmark identification in criminal trials is “hardly novel.”  *United States v. Taylor*, 663 F. Supp. 2d 1170, 1175 (D.N.M. 2009). “For decades ... admission of the type of firearm identification testimony challenged by the defendant[] has been semi-automatic”  *United States v. Monteiro*, 407 F. Supp. 2d 351, 364 (D. Mass. 2006); *see also, e.g., United States v. Hicks*, 389 F.3d 514 (5th Cir. 2004);  *United States v. Johnson*, 875 F.3d 1265, 1281 (9th Cir. 2017). Indeed, no federal court has deemed such evidence wholly inadmissible. *See United States v. Romero-Lobato*, 379 F. Supp. 3d 1111, 1117 (D. Nev. 2019). Having been routinely admitted, “[c]ourts [are] understandably ... gun shy about questioning the reliability of [such] evidence,”  *Monteiro*, 407 F.Supp.2d at 364. However, because of the seriousness of the criticisms launched against the methodology underlying firearms identification by Defendant in this case, the Court will carefully assess the reliability of this methodology, using  *Daubert* as a guide. *See, e.g., Taylor*, 663 F. Supp. 2d at 1176.²


The first  *Daubert* factor asks whether the experts' particular theory can be and has been tested.  *Daubert*, 509 U.S. at 592–94, 113 S.Ct. 2786. Defendant argues—without citation—that the theory of firearm toolmark identification rests on an assumption that has not been properly tested. Doc. No. 67, pp. 13–14. The Government responds that its experts' testimony is based upon the theory and methodology developed by the Association of Firearms and Toolmark Examiners (AFTE), and that this theory has been well tested. Doc. No. 81, pp. 15–16. The Court agrees.






*4 Put simply, the theory of firearm toolmark identification can be and has been tested. *See, e.g.,* The Association of Firearm and Tool Mark Examiners, *Testability of the Scientific Principle* (last visited May 14, 2020), <https://tinyurl.com/yal3ja4t> (collecting studies). This conclusion is supported by other courts within the Tenth Circuit that have already addressed the issue at length, *see, e.g., United States v. Taylor*, 663 F. Supp. 2d 1170, 1176 (D.N.M. 2009) (“[T]he methods underlying firearms identification can, at least to some degree, be tested and reproduced”), in addition to a number of other courts outside the Circuit, *see, e.g., Romero-*

Lobato, 379 F. Supp. 3d at 1118–19 (collecting cases where “federal courts have held that the AFTE method can be and has been frequently tested” and holding the same).

Accordingly, this first  *Daubert* factor weighs in favor of admissibility.

The second  *Daubert* factor asks whether the technique has been subjected to peer review and publication.  *Daubert*, 509 U.S. at 593–94, 113 S.Ct. 2786. Defendant argues that there have not been enough studies done of firearm toolmark identification, and that the studies available have not been subject to peer review. Doc. No. 67, p. 14. The Government contends that analysis recently provided by federal courts tells a different story. The Court agrees.

In evaluating whether AFTE's method of firearm toolmark identification satisfies the second  *Daubert* factor, the United States District Court for the District of Nevada recently found that:

AFTE publishes its own journal, the appropriately named *ATFE Journal*, which is subject to peer review. According to AFTE's website, the *AFTE Journal*, “is dedicated to the sharing of information, techniques, and procedures,” and the papers published within “are reviewed for scientific validity, logical reasoning, and sound methodology.” [*What is the Journal?*, The Association of Firearm and Tool Mark Examiners, <https://afte.org/afte-journal/what-is-the-journal> (last visited May 1, 2019)]. Several published federal decisions have also commented on the *AFTE Journal*, with all finding that it meets the  *Daubert* peer review element. *See U.S. v. Ashburn*, 88 F.Supp.3d 239, 245–46 (E.D.N.Y. 2015) (finding that the AFTE method has been subjected to peer review through the *AFTE Journal*);  *U.S. v. Otero*, 849 F.Supp.2d 425, 433 (D.N.J. 2012) (describing the *AFTE Journal*'s peer reviewing process and finding that the methodology has been subjected to peer review);  *U.S. v. Taylor*, 663 F.Supp.2d 1170, 1176 (D.N.M. 2009) (finding that the *AFTE* method has been subjected to peer review through the *AFTE Journal* and two articles submitted by the government in a peer-reviewed journal about the methodology);  *U.S. v. Monteiro*, 407 F.Supp.2d 351, 366–67 (D. Mass. 2006) (describing the *AFTE Journal*'s peer reviewing process and finding that it meets the  *Daubert* peer review element).

And of course, the NAS and PCAST Reports themselves constitute peer review despite the unfavorable view the two reports have of the AFTE method.

Romero-Lobato, 379 F. Supp. 3d at 1119. The second *Daubert* factor thus weighs in favor of admissibility.

Defendant suggests that the studies mentioned above are insufficient because they were not “black-box” studies.³ Doc. No. 67, p. 14. Defendant then cites the PCAST Report, arguing that there has been only one black-box study on firearms identification and that this one study has never been subject to peer review. *Id.* The PCAST Report cited by Defendant “rejected studies that it did not consider to be blind, such as where the examiners knew that a bullet or spent casing matched one of the barrels included with the test kit...” However, “The PCAST Report did not reach a conclusion as to whether the AFTE method was reliable or not because there was only one study available that met its criteria.” *Id.* The Court does not similarly restrict its judicial review to techniques tested through black-box studies. The Court does, however, approve of the PCAST Report’s ultimate conclusion: “[W]hether firearms analysis should be deemed admissible based on the ‘current evidence’ is a decision that should be left to the courts.” *Id.*

*5 The third *Daubert* factor asks whether the technique has a known or potential rate of error. *Daubert*, 509 U.S. at 594, 113 S.Ct. 2786. Defendant contends that because there is only one black-box study, there is not enough information available to determine a known or potential rate of error in the field of firearm toolmark identification. Doc. No. 67, p. 14. The Government objects, citing federal cases discussing studies that evidence a low rate of error in firearms analysis. Doc. No. 81, pp. 17–18. Again, the Court agrees with the Government.

[9] As noted above, the Court declines Defendant’s invitation to restrict judicial review to techniques tested through black-box studies. “*Daubert* does not mandate such a prerequisite for a technique to satisfy its error rate element.” *Romero-Lobato*, 379 F. Supp. 3d at 1120. Still, the Government bears the burden to demonstrate that its experts’ methodology is reliable. See *Nacchio*, 555 F.3d at 1241. To that end, the Government cites federal cases that discuss a number of studies which report a low error rate for the AFTE method. Doc. No. 81, p. 17 (citing *Romero-Lobato*, 379 F.




Supp. 3d at 1117–18 and *United States v. Otero*, 849 F. Supp. 2d 425, 433–34 (D.N.J. 2012)). Those cases discuss, for example, a Miami-Dade Study that reported a potential error rate of less than 1.2% and an error rate by the participants of 0.07%, in addition to an Ames Study that reported a false positive rate of 1.52%. *Id.*



Other federal courts examining the AFTE method’s rate of error have likewise found it to be low. See, e.g., v. *Ashburn*, 88 F. Supp. 3d 239, 246 (E.D.N.Y. 2015) (“the error rate, to the extent it can be measured, appears to be low, weighing in favor of admission”); *United States v. Taylor*, 663 F. Supp. 2d 1170, 1177 (D.N.M. 2009) (“this number [less than 1%] suggests that the error rate is quite low”). Even courts that have found it impossible to calculate an absolute error rate for firearm toolmark identification, have ultimately concluded that the known error rate is not “unacceptably high.” *United States v. Monteiro*, 407 F. Supp. 2d 351, 367–68 (D. Mass. 2006). Defendant does not introduce any contradictory studies. See Doc. No. 67, p. 14. Based on the record before the Court, this third *Daubert* factor weighs in favor of admissibility.





The fourth *Daubert* factor asks whether there are standards that control the technique’s operation. *Daubert*, 509 U.S. at 113 S.Ct. 2786594. Defendant argues that there are no uniform standards controlling the AFTE method of firearm toolmark identification, and that instead, the AFTE method is based on subjective methodology. Doc. No. 67, p. 14. The Government argues that this subjectivity does not weigh against admissibility under the fourth *Daubert* factor. Doc. No. 81, p. 18. The Court disagrees.

A main criticism of the AFTE method is that firearm examiners do not reach their conclusions through objective criteria. See *Romero-Lobato*, 379 F. Supp. 3d at 1120-121. Instead, examiners use a high-powered microscope, in conjunction with their experience and training, to determine if there is “sufficient agreement” between the “unique surface contours” of two firearm toolmarks. *AFTE Theory of Identification*, The Association of Firearm and Tool Mark Examiners, available at <https://afte.org/about-us/what-is-afte/afte-theory-of-identification> (last visited May 14, 2020). “The statement that “sufficient agreement” exists between two toolmarks means that the agreement of individual characteristics is of a quantity and quality that the likelihood




another tool could have made the mark is so remote as to be considered a practical impossibility.”⁴ *Id.* Ultimately, the AFTE itself recognizes that their method is “is subjective in nature.” *Id.* So too have other courts. See *Romero-Lobato*, 379 F. Supp. 3d at 1121 (collecting cases). This fourth factor, unlike the previous three, weighs against admissibility.

*6 The fifth and final  *Daubert* factor asks whether the theory or technique enjoys general acceptance within the relevant community.  *Daubert*, 509 U.S. at 594, 113 S.Ct. 2786. Defendant argues that the limitations of firearm toolmark identification is recent and growing, and that because courts have not seriously considered all aspects of the field or tested its reliability since the PCAST Report was published, the fifth  *Daubert* factor is not satisfied here. Doc. No. 67, p. 15. The Government responds arguing that nearly every court to have addressed the issue has found that the AFTE method enjoys general acceptance within the relevant community—both before and after publication of the PCAST Report. Doc. No. 81, p. 19. The Court agrees.


The AFTE method easily satisfies this final factor. See *Romero-Lobato*, 379 F. Supp. 3d at 1122 (collecting cases finding the AFTE theory to be widely accepted in the relevant community and finding the same). In fact, the AFTE method used by the Government's experts here, is “the field's established standard.” See  *Ashburn*, 88 F. Supp. 3d at 246. That the NAS and PCAST Reports criticize the method does not undermine the Court's conclusion. “Techniques do not need to have universal acceptance before they are allowed to be presented before a court.” *Romero-Lobato*, 379 F. Supp. 3d at 1122 (citing  *Daubert*, 509 U.S. at 588–99, 113 S.Ct. 2786). Accordingly, this factor weighs in favor of admissibility.

Balancing the  *Daubert* factors, the Court finds that the Government's expert testimony, derived from the AFTE methodology, is reliable and therefore admissible—though subject to the limitations discussed below. The only factor that weighs against admissibility is the fourth  *Daubert* factor, which highlights the AFTE's subjective processes. But, “the subjectivity of a methodology is not fatal under *Rule 702* and  *Daubert*.”  *United States v. Ashburn*, 88 F. Supp. 3d 239, 246 (E.D.N.Y. 2015). By its terms, *Federal Rule of Evidence 702* permits an expert with sufficient knowledge, experience, or training to testify about a particular subject matter. See

Fed. R. Evid. 702; *Romero-Lobato*, 379 F. Supp. 3d at 1120.

 *Daubert* does not impose a rigid requirement that the expert reach a conclusion through an entirely objective set of criteria. See  *Daubert*, 509 U.S. at 594–595, 113 S.Ct. 2786. Here, the lack of objective criteria is overcome by the Government's introduction of evidence demonstrating that the method has been tested, reviewed by peers and subject to publication, found to have a potential low rate of error, and widely accepted in the relevant community. Moreover, Defendant has not cited a single case where a federal court has completely prohibited firearms toolmark identification testimony under  *Daubert*.

V. Federal Rules of Evidence 702(d)

[10] Next, Defendant argues that even if the expert testimony is admissible under  *Daubert*, the Government has not met its burden under *Rule 702(d)* to show that its experts reliably applied the AFTE method in this case. Under that *Rule*:

A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:

...

(d) the expert has reliably applied the principles and methods to the facts of the case.

Fed. R. Evid. 702(d). Here, Defendant makes four specific objections. He argues that the Government has not complied with *Rule 702(d)* because its experts failed to document the basis for their findings, that a second examiner did not verify or review the experts' work, and that the experts failed to comply with two “validity” requirements discussed by the PCAST Report. Doc. No. 67, p. 17. The Government denies the validity of each objection. Doc. No. 81, pp. 21–23.

*7 First, as the Government demonstrates, both Mr. Jones and Mr. Kong wrote detailed reports explaining their analysis. Doc. Nos. 81–9, 81–10. Second, those reports were reviewed by other examiners in the field. Doc. Nos. 81–1, 81–2, 81–3, 81–4. Finally, the two validity requirements discussed by the PCAST Report—that experts must provide evidence demonstrating their rigorous proficiency testing, in addition to whether they were aware of any facts of the case that might influence their conclusion—are not required under *Rule 702(d)*. Nevertheless, the Government has presented evidence demonstrating the experience, certifications, and

continued training of both experts. *See* Doc. Nos. 81–6, 81–7, 81–8; *cf.* Doc. No. 81–5. And both experts' examination reports detail what case-specific facts they were aware of when drawing their conclusions. *See* Doc. Nos. 81–1, 81–2. Accordingly, the Court finds that Defendant's objections are without merit.

VI. [Daubert](#) Hearing

[11] [12] As an alternative, Defendant requests a [Daubert](#) hearing to require the Government to prove that Mr. Jones's and Mr. Kong's testimony will be reliable before admitting their testimony. Doc. No. 17. Again, the Government objects. Doc. No. 81, pp. 24–25. Nothing requires the Court to hold a formal [Daubert](#) hearing in advance of qualifying an expert. *See* [Goebel v. Denver and Rio Grande Western RR Co.](#), 215 F.3d 1083, 1087 (10th Cir. 2000); *see also* [Kumho Tire](#), 526 U.S. at 152, 119 S.Ct. 1167 (“The trial court must have the ... latitude ... to decide whether or when special briefing or other proceedings are needed to investigate reliability”). Considering the parties' briefing, in addition to the [Daubert](#) and Rule 702 analysis above, the Court finds it unnecessary to conduct such a proceeding here. *See, e.g.*, [Ashburn](#), 88 F. Supp. 3d at 244 (finding [Daubert](#) hearing unnecessary). The reliability of the Government's expert testimony has been sufficiently addressed on the briefs. *See* [Goebel](#), 215 F.3d at 1087 (noting that a [Daubert](#) hearing “is not mandated” and that a district court may “satisfy its gatekeeper role when asked to rule on a motion in limine”).

VII. Expert Testimony Limitations

[13] In his penultimate argument, Defendant asks the Court to place limitations on the Government's firearm toolmark experts because the jury will be unduly swayed by the experts if not made aware of the limitations on their methodology. Doc. No. 67, p. 18. The Government responds that no limitation is necessary because Department of Justice guidance sufficiently limits a firearm examiner's testimony. Doc. No. 81, pp. 23–24.

Some federal courts have imposed limitations on firearm and toolmark expert testimony. *See, e.g.*, [Ashburn](#), 88 F. Supp. 3d at 249. However, many courts have continued to allow

unfettered testimony. *See, e.g.*, [Romero-Lobato](#), 379 F. Supp. 3d at 1117.

The general consensus is that firearm examiners should not testify that their conclusions are infallible or not subject to any rate of error, nor should they arbitrarily give a statistical probability for the accuracy of their conclusions. Several courts have also prohibited a firearm examiner from asserting that a particular bullet or shell casing could only have been discharged from a particular gun to the exclusion of all other guns in the world.

Id. (citing David H. Kaye, [Firearm-Mark Evidence: Looking Back and Looking Ahead](#), 68 Case W. Res. L. Rev. 723, 734 (2018)).

In accordance with recent guidance from the Department of Justice, *see* Doc. No. 81–11, the Government's firearm experts have already agreed to refrain from expressing their findings in terms of absolute certainty, and they will not state or imply that a particular bullet or shell casing could only have been discharged from a particular firearm to the exclusion of all other firearms in the world. Doc. No. 81, p. 24. The Government has also made clear that it will not elicit a statement that its experts' conclusions are held to a reasonable degree of scientific certainty. *Id.*

*8 The Court finds that the limitations mentioned above and prescribed by the Department of Justice are reasonable, and that the Government's experts should abide by those limitations. *See* Doc. No. 81–11, p. 3. To that end, the Governments experts:

[S]hall not [1] assert that two toolmarks originated from the same source to the exclusion of all other sources.... [2] assert that examinations conducted in the forensic firearms/toolmarks discipline are infallible or have a zero error rate.... [3] provide a conclusion that includes a statistic

or numerical degree of probability except when based on relevant and appropriate data.... [4] cite the number of examinations conducted in the forensic firearms/toolmarks discipline performed in his or her career as a direct measure for the accuracy of a proffered conclusion..... [5] use the expressions ‘reasonable degree of scientific certainty,’ ‘reasonable scientific certainty,’ or similar assertions of reasonable certainty in either reports or testimony unless required to do so by [the Court] or applicable law.

Id. As to the fifth limitation described above, the Court will permit the Government's experts to testify that their conclusions were reached to a reasonable degree of ballistic certainty, a reasonable degree of certainty in the field of firearm toolmark identification, or any other version of that standard. *See, e.g.*, [U.S. v. Ashburn](#), 88 F. Supp. 3d 239, 249 (E.D.N.Y. 2015) (limiting testimony to a “reasonable degree of ballistics certainty” or a “reasonable degree of certainty in the ballistics field.”); [U.S. v. Taylor](#), 663 F. Supp. 2d 1170, 1180 (D.N.M. 2009) (limiting testimony to a “reasonable degree of certainty in the firearms examination field.”). Accordingly, the Government's experts should not testify, for example, that “the probability the ammunition

charged in Counts Eight and Nine were fired in different firearms is so small it is negligible,” *see* Doc. No. 81, p. 5. To the extent Defendant wishes to question or clarify the experts' findings, he may do so through cross examination or through direct examination of his own firearm toolmark expert.

VIII. Additional Expert Information

Defendant's final objection is to the alleged lack of information relating to Mr. Jones's expert testimony. Doc. No. 67, p. 19. Defendant claims that the Government should be required to provide “a significantly more detailed summary of what it expects Mr. Jones will testify about.” *Id.* Notably, Defendant provides no support for his objection, and the Government has failed to respond in opposition. Upon review, the Court finds that the Government has provided sufficient information relating to Mr. Jones's expert testimony. *See* Doc. No. 81, pp. 4–5; Doc. Nos. 81–1, 81–6, 81–7, 81–9.

IX. Conclusion

For the forgoing reasons, the Court denies Defendant Hunt's Motion in Limine to Exclude Ballistic Evidence, or Alternatively, for a [Daubert](#) Hearing, Doc. No. 67.



IT IS SO ORDERED this 1st day of June 2020.

All Citations

--- F.Supp.3d ----, 2020 WL 2842844, 112 Fed. R. Evid. Serv. 901

Footnotes

- 1 [Daubert](#) itself was limited to scientific evidence, *see* [United States v. Baines](#), 573 F.3d 979, 985 (10th Cir. 2009), but in [Kumho Tire Co. v. Carmichael](#), 526 U.S. 137, 119 S.Ct. 1167, 143 L.Ed.2d 238 (1999), the Supreme Court made clear that the gatekeeping obligation of the district courts described in [Daubert](#) applies, not just to scientific testimony, but to all expert testimony. *Id.* at 141, 119 S.Ct. 1167.
- 2 Some Courts have analyzed whether firearm toolmark identification can fairly be called “science” before evaluating the [Daubert](#) factors. *See* [United States v. Glynn](#), 578 F. Supp. 2d 567, 570 (S.D.N.Y. 2008). The Court need not conduct such an analysis here. Though Defendant argues firearm toolmark identification is not a science, Doc. No. 67, p. 14, it is clearly “technical or specialized, and therefore within the scope of

Rule 702.”  *United States v. Willock*, 696 F. Supp. 2d 536, 571 (D. Md. 2010), *aff'd sub nom.*  *United States v. Mouzone*, 687 F.3d 207 (4th Cir. 2012).

3 A black-box study is a blind study where “many examiners are presented with many independent comparison problems—typically involving ‘questioned’ samples and one or more ‘known’ samples—and asked to declare whether the questioned samples came from the same sources as one of the known samples. The researchers then determine how often examiners reach erroneous conclusions.” President’s Council of Advisors on Science and Technology, Exec. Office of the President, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*, 49 (2016), available at <https://tinyurl.com/j29c5ua>.

4 The AFTE further details their methodology in the following manner:
“[S]ufficient agreement” is related to the significant duplication of random toolmarks as evidence by the correspondence of a pattern or combination of patterns of surface contours. Significance is determined by the comparative examination of two or more sets of surface contour patterns comprised of individual peaks, ridges and furrows. Specifically, the relative height or depth, width, curvature and spatial relationship of the individual peaks, ridges and furrows within one set of surface contours are defined and compared to the corresponding features in the second set of surface contours. Agreement is significant when the agreement in individual characteristics exceeds the best agreement demonstrated between toolmarks known to have been produced by different tools and is consistent with agreement demonstrated by toolmarks known to have been produced by the same tool.

AFTE Theory of Identification, The Association of Firearm and Tool Mark Examiners, available at <https://afte.org/about-us/what-is-afte/afte-theory-of-identification> (last visited May 14, 2020).