



**INNOCENCE
PROJECT**

The National Criminal Defense College and the Innocence Project
Present

A Forensic Cross-Examination Workshop: Drilling Down on Toolmark Evidence



Essential Reference Materials

READ before you arrive at Workshop

This page intentionally blank.

Table of Contents

GBI Firearm and Toolmarks Overview, 9 pages, 2020	1-8
<i>Summary: This summary appears on the GBI website.</i>	
National Academies of Science, National Research Council, "Strengthening Forensic Science in the United States: A Path Forward" 2009.....	11-20
<i>Summary: This is what is known as the NAS Report on Forensic Science. It was watershed at the time issued alerting the world to the fact that serious improvement is needed in the forensic sciences if they are to be used in court, particularly criticizing the lack of uniform standards within certain disciplines and even a lack of basic foundational validity of some disciplines. Several recommendations were made by the NAS Report to strengthen the forensic disciplines, particularly those that involve pattern matching, like toolmarks.</i>	
AFTE Response to NAS Report.....	21-26
<i>The Association of Firearms and Toolmarks Examiners unsurprisingly disagrees with much of the NAS commentary on their discipline. They explain why they disagree and respond to the recommendations.</i>	
PCAST Report Excerpt Relating to Firearms and Toolmarks Analysis, 14 pages, 2016.....	27-38
<i>Summary: PCAST stands for the President's Council of Advisors on Science and Technology. This Council was appointed by President Obama and tasked with further study of the problems identified in the NAS Report. Following a comprehensive review of literature and studies within the field, the PCAST finds that a number of the feature (or pattern) comparison forensic fields are deficient in their foundational validity, because of a lack of appropriate studies to evaluate how accurate or inaccurate the field may be. The PCAST further recommended that the opinions of examiners should be limited in scope and accompanied by appropriate jury instructions.</i>	
AFTE Response to PCAST Report.....	39-40
<i>The Association of Firearms and Toolmarks Examiners issued this response to the PCAST report's criticism saying that the "AFTE strongly disagrees with the premise that additional ongoing structured research strengthens the foundational and applied validity of firearm identification, as well as endeavors to reduce the effects of cognitive bias and subjectivity. However, we cannot overstate our disappointment in the PCAST's choice to ignore the research that has been conducted." This is an important read in order to understand what an expert is likely to say in response to defense points about the critique of the discipline.</i>	
Addendum to PCAST Report Rebuttal to AFTE Response.....	41-50
<i>PCAST Rebuttal to the AFTE Statement</i>	
United States v. Tibbs, 2019 WL 4359486 (2019)	51-80

This page intentionally blank.

Firearm and Toolmark Overview

The **Firearm and Toolmark discipline** is a versatile, well-equipped unit offering a number of services that can be useful to investigators.



Firearm and Toolmark Examiners provide reliable scientific support to all law enforcement personnel. Services are provided at both the investigation and trial-preparation stages of criminal cases involving the use of a firearm or other tool.

- The type of firearm that a bullet or cartridge case was fired from
- Whether a bullet was, or was not fired from a suspected firearm
- Whether a cartridge case was, or was not fired in a suspected firearm
- Whether a tool found in a suspect's possession was, or was not used to cut, scrape, pry, or pinch evidence material seized from a crime scene*
- The original serial number of a firearm or other metal object after the number has been obliterated
- If gunpowder is present on a victim's clothing or on other evidence that may have been the target of the suspect
- The distance from the muzzle of the firearm to the target at the time the firearm was fired**

Tools found at the scene of a crime that cannot be associated with a suspect **will not be examined.*

No muzzle-to-target distance tests can be done without the firearm that was involved in the shooting. **Note: It cannot be determined "how long" it has been since a firearm was fired.

Other miscellaneous examinations may be performed at the request of the customer. Examiners in the Firearm and Toolmark discipline may conduct other testing that is of special interest to an investigator. Such requests may be made at the time of evidence submissions or by phone.

Firearms Analysis

Basics of Firearms Comparisons



Inside the barrels of handguns and rifles are spiral impressions called rifling. The raised portions of the rifling are known as lands and the recessed portions are known as grooves. When a firearm is fired, these lands and grooves cut into the bullet, putting spin on it as it travels through the barrel of a firearm. Because bullets have an oblong shape, spin is necessary for accurate flight.



The impressions of lands and grooves are transferred to the bullet when it is fired.



Since rifling characteristics can differ from one firearm manufacturer to another, firearms examiners can determine the type of firearm that fired a bullet by examining the impressions of the lands and grooves on the bullet. They examine the width, the number, and the direction of the twist of the lands and grooves. For example, a 9mm pistol made by one company might have a barrel with 6 lands and grooves that twist to the right and another company's 9mm might have 6 that twist to the left. In addition, the width of the lands and grooves may differ.

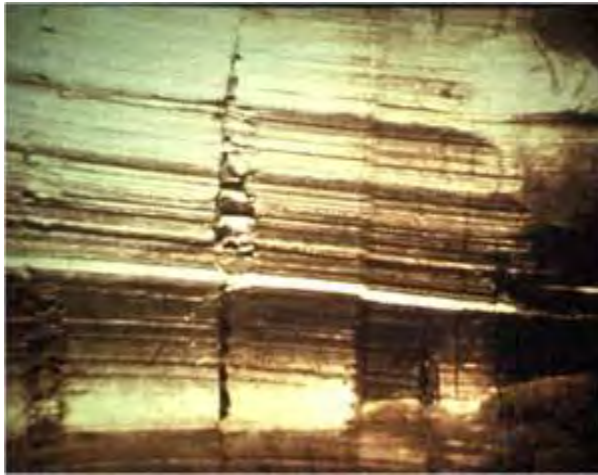
Because each barrel will have imperfections left by the manufacturing process that will leave unique marks on a bullet, firearms examiners can determine whether a bullet recovered from a crime scene or victim was fired from a firearm taken from a suspect.

Comparison Process



The first step in the **Comparison Process** is to test fire the firearm into a water tank in the lab.

Bullets



The second step involves using a comparison microscope to compare the test bullet to the bullet recovered from the victim or crime scene.

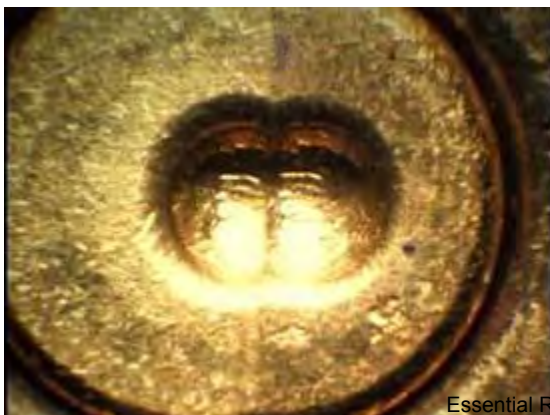
The photo on the left shows the split screen image the scientist sees using the comparison microscope. The right side of the photo shows the test bullet fired from the suspect's firearm into the water tank and the left side, the bullet recovered from the crime scene. The marks or striations on the evidence bullet were identified as being made by the suspect's firearm.

Cartridge Cases



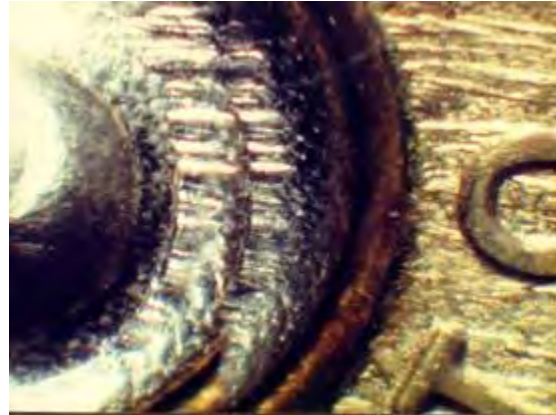
Since a firearm will also leave unique marks on cartridge cases; cartridge cases left at crime scenes can link a suspect's firearm to the crime.

The following photographs of split screen images from a comparison microscope show three different types of markings left on cartridge cases that firearms examiners can use in determining if the cartridge cases were fired from the same firearm.



Firing pin impressions - When a firearm's trigger is pulled, the firing pin will move forward striking the primer cup located at the rear center of the cartridge.

Breech face marks- These marks come from the breech face area of the firearm. This is the portion of the firearm that supports the cartridge when loaded in the chamber. After the cartridge powder is ignited by the firing pin striking the primer cup, tremendous pressure is exerted in the chamber of the firearm, forcing the back of the cartridge case against the breech face.



Extractor marks - After a semi-automatic pistol has been fired, an extractor pulls the cartridge from the chamber and ejects it from the pistol.

IBIS/NIBIN

The Integrated Ballistic Identification System (IBIS) is used to potentially associate evidence in previously unlinked crimes. IBIS is a highly technical, computerized image analysis system that records images from bullets and cartridge cases and compares them to a national database of images called the National Integrated Ballistic Information Network (NIBIN). The GBI Firearms section only has the capability to enter cartridge cases for search against the NIBIN database.

Cartridge cases recovered from crime scenes are imaged into the system and searched to find potential associations with other evidence in the database or from test fires from firearms that are submitted to the laboratory for testing.

The images from the test fires and evidence are correlated automatically against the database in a matter of minutes; an impossible task for a firearms examiner using conventional procedures. The images are correlated by the system and given a score as to a possible association. The results of this correlation are called NIBIN leads. Agencies will be notified of any NIBIN Leads within 24 hours of discovery.

The firearms examiner makes a final determination by conducting a microscopic examination of the evidence generating the NIBIN Lead. If an association is

confirmed, it becomes a NIBIN Hit. When a Hit has occurred, the involved law enforcement agencies are notified, and they can then take the appropriate investigative and legal actions.

Evidence cartridge cases and test fires remain in the system to be searched indefinitely. The images from the test fires and evidence are correlated automatically against the entire database for the NIBIN region encompassing the host site. Manual searches outside the region can be made at the request of the submitting agency.

In order to provide timely investigative lead information, the GBI Firearms section has implemented a screening process for all submitted evidence to the laboratory.

- For cases in which two (2) or more cartridge cases and/or shotshell cases are submitted:
 - Submitted evidence will be screened and grouped based on similar characteristics
 - At least one (1) cartridge case and/or shotshell case representing each group shall be imaged.
 - Appropriate NIBIN system searches will be conducted.
- For cases in which only one cartridge case and/or shotshell is submitted, appropriate NIBIN system searches will be conducted.

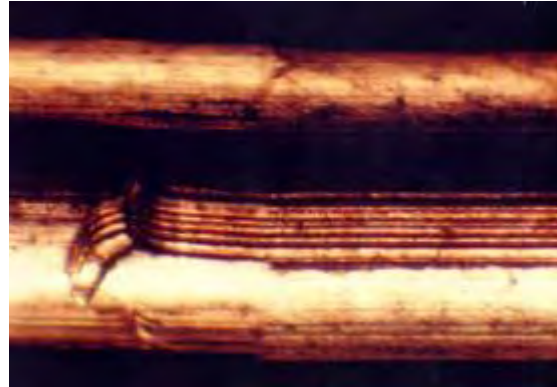
All submitted evidence will be screened on entry into the laboratory system and submitted to the NIBIN database. Further microscopic comparisons will be performed on evidence from generated NIBIN leads. If no NIBIN lead is generated following the completion of this initial screening, only evidence for the following offenses will remain at the laboratory for further examination and comparative analysis: Homicide, Crimes against Children, and other expedited requests. Evidence screened for all other offenses will be returned to the agency after screening completion.

Toolmark Examination

The Firearm and Toolmark discipline accepts tools suspected of being associated with a crime scene. Cutting, pinching, prying, and striking tools are all examined by this discipline. Great care must be taken by the officer to protect the marking surfaces on the tool. If the entire tool cannot be wrapped, the tool's marking surface should be protected using whatever materials are available to the officer. Tools found at the crime scene that cannot link a person to the scene will not be examined. Tools must be associated with a suspect by means of DNA, Latent Prints, or recovery on their person, in their home, vehicle, or other limited access location. Materials recovered from the crime scene that are suspected of being marked by the recovered tool should be carefully removed from the scene so that the marks are not disrupted. Any cuts made by the officer to remove the suspect areas should be clearly marked so as not to be confused with the suspect areas. Any suspected tool collected **SHOULD NOT** be used for this purpose, as this can

damage or change the characteristics of the tool. Each item should be packaged separately and submitted to the laboratory.

Tools also leave unique marks that can provide important clues in linking a suspect to a crime. The photo on the right is a split screen comparison of copper wires cut by a wire cutter found in the possession of a burglary suspect. The test cut on the right matches the evidence cut from the crime scene.



A maul recovered from a burglary suspect is compared against the indentation left in the victim's doorjamb.

The maul fits the indentation perfectly, providing local investigators with evidence to charge the suspect in the burglary.



Evidence Submissions

The following are general requirements for officers and other law enforcement personnel in collecting and submitting evidence for Firearm and Toolmark analysis. As in all cases, common sense should be used when attempting to protect the integrity of evidence.

Firearms

Firearms should be collected carefully so that no parts of the firearm are damaged. Officers should make sure that nothing comes in contact with either the inside of the barrel or the breech face, which is where the head of the cartridge rests before firing. All firearms **MUST** be unloaded prior to submission to the laboratory. If the firearm cannot be unloaded prior to submission, you **MUST** contact DOFS personnel for instructions **prior to submitting the firearm**. Firearms should be made secure in gun boxes so that it can be easily confirmed that the firearm is unloaded. When securing firearms in the box, please follow these guidelines for placement of zip ties.



The use of zip ties in the position(s) pictured will allow lab staff to see that the weapon is unloaded in addition to securing it to the box. Effective **June 1, 2023**, all firearms (handguns and long guns) will be required to be submitted in commercially available gun boxes **with clear windows** that will allow the viewing of the contents. Additionally, it will be required to fasten a plastic zip tie through the gun in the manner shown. See Operations Bulletin 2023-01 – New Submission Requirements for Firearms Evidence.

There is no specific required vendor for these gun boxes and you may use boxes from the vendor of your choice. Some vendors that offer boxes with windows are listed below (we are not endorsing any particular vendor).

Handgun boxes:

<https://www.copsplus.com/evi-paq-window-gun-boxes-pack-of-20>

<https://www.idtechnologies.com/products/window-gun-boxes-pack-of-20>

https://forensicssource.com/products/window-gun-boxes-pack-of-20-1008209?variant=34495782682669¤cy=USD&utm_medium=product_sync&utm_source=google&utm_content=sag_organic&utm_campaign=sag_organic

Long gun boxes:

<https://www.idtechnologies.com/products/window-rifle-boxes-pack-of-20>

<https://forensicssource.com/collections/evidence-packaging-boxes/products/window-rifle-boxes-pack-of-20-1008212>

Firearms Recovered From Water

Firearms removed from **fresh** water **must** be placed in the same water for submission to the laboratory. Small lunch coolers work very well for handguns. If a rifle or shotgun is removed from a lake or pond but cannot fit into a container, the firearm should be sprayed heavily with WD-40 or other lubricant and taken to the laboratory immediately. The slightest bit of rust to the inside of the barrel will alter the individual characteristics necessary to make identification. If the firearm is removed from the water, it must be oiled, making sure that the inside of the barrel is sprayed or filled with oil. This will slow the oxidation process. Firearms removed from **saltwater** should be rinsed, heavily oiled and brought to the crime laboratory immediately. Heavily bloodstained firearms should be packaged in boxes with a biohazard label.

Protecting the Firearm from Damage

Do not insert foreign objects into any part of the firearm such as the barrel or ejection port. In the event the firearm will be dusted for fingerprints or super glued, block both ends of the barrel gently with tape. This will prevent residue from building up inside of the barrel. Do not dry fire or work the action of any firearm that is to be submitted to the crime lab. Leaving empty cartridge cases in the chamber of a revolver when submitted might assist the examiner in determining from which chamber the round was fired.

Bullets, Cartridge Cases and Cartridges

When bullets and cartridge cases are submitted for analysis, they should be individually packaged in their own canister, envelope, or other small container.

- Do not mark or engrave any surface on a bullet or cartridge case as this may damage individual characteristics. If evidence must be marked, mark the container itself.
- Do not let any metal object such as forceps, knives or screwdrivers come into contact with a bullet. Metal objects will scratch the surface and alter the markings used for identification.
- Numerous cartridge cases recovered from the same area may be packaged together in one container to save time and supplies.

- Be sure to collect any wads or pellets in cases involving a shotgun. Under certain circumstances a wad can be compared to the barrel of a shotgun, especially if it has a sawed-off barrel.
- Film canisters or pillboxes make excellent containers for packaging bullets and cartridge cases.
- Please submit bullets in separate packaging from any collected cartridge cases or firearms. This will help expedite the NIBIN screening process.

Clothing

Clothing submitted to the laboratory for distance determination should be packaged in a paper bag or box. Do not package wet or bloody clothes until they have air-dried, taking care to not cross contaminate items. Wet clothes will mold, making them difficult to examine. Inform the firearms examiner of how the layers of clothing were worn in order to assist in determining the path of the bullet. This information should be written on the evidence bag or the submission form.

Analysis for Muzzle to Target Distance Determination cases will only be performed after consultation and approval of the Firearms Discipline manager or designee. For this service to be performed, a firearm identified as being used in the shooting must also be submitted. If no firearm has been submitted, no distance determination can be rendered unless there is a clear indication of a contact gunshot hole of entry.

STRENGTHENING
**FORENSIC
SCIENCE**
IN THE UNITED STATES

A PATH FORWARD

Committee on Identifying the Needs of the Forensic Science Community

Committee on Science, Technology, and Law
Policy and Global Affairs

Committee on Applied and Theoretical Statistics
Division on Engineering and Physical Sciences

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

Preface

Recognizing that significant improvements are needed in forensic science, Congress directed the National Academy of Sciences to undertake the study that led to this report. There are scores of talented and dedicated people in the forensic science community, and the work that they perform is vitally important. They are often strapped in their work, however, for lack of adequate resources, sound policies, and national support. It is clear that change and advancements, both systemic and scientific, are needed in a number of forensic science disciplines—to ensure the reliability of the disciplines, establish enforceable standards, and promote best practices and their consistent application.

In adopting this report, the aim of our committee is to chart an agenda for progress in the forensic science community and its scientific disciplines. Because the work of forensic science practitioners is so obviously wide-reaching and important—affecting criminal investigation and prosecution, civil litigation, legal reform, the investigation of insurance claims, national disaster planning and preparedness, homeland security, and the advancement of technology—the committee worked with a sense of great commitment and spent countless hours deliberating over the recommendations that are included in the report. These recommendations, which are inexorably interconnected, reflect the committee’s strong views on policy initiatives that must be adopted in any plan to improve the forensic science disciplines and to allow the forensic science community to serve society more effectively.

The task Congress assigned our committee was daunting and required serious thought and the consideration of an extremely complex and decentralized system, with various players, jurisdictions, demands, and limitations. Throughout our lengthy deliberations, the committee heard testimony

from the stakeholder community, ensuring that the voices of forensic practitioners were heard and their concerns addressed. We also heard from professionals who manage forensic laboratories and medical examiner/coroner offices; teachers who are devoted to training the next generation of forensic scientists; scholars who have conducted important research in a number of forensic science fields; and members of the legal profession and law enforcement agencies who understand how forensic science evidence is collected, analyzed, and used in connection with criminal investigations and prosecutions. We are deeply grateful to all of the presenters who spoke to the committee and/or submitted papers for our consideration. These experts and their work served the committee well.

In considering the testimony and evidence that was presented to the committee, what surprised us the most was the consistency of the message that we heard:

The forensic science system, encompassing both research and practice, has serious problems that can only be addressed by a national commitment to overhaul the current structure that supports the forensic science community in this country. This can only be done with effective leadership at the highest levels of both federal and state governments, pursuant to national standards, and with a significant infusion of federal funds.

The recommendations in this report represent the committee's studied opinion on how best to achieve this critical goal.

We had the good fortune to serve as co-chairs of the committee entrusted with addressing Congress' charge. The committee, formed under the auspices of the National Academies' Committee on Science, Technology, and Law and Committee on Applied and Theoretical Statistics, was composed of many talented professionals, some expert in various areas of forensic science, others in law, and still others in different fields of science and engineering. They listened, read, questioned, vigorously discussed the findings and recommendations offered in this report, and then worked hard to complete the research and writing required to produce the report. We are indebted to our colleagues for all the time and energy they gave to this effort. We are also most grateful to the staff, Anne-Marie Mazza, Scott Weidman, Steven Kendall, and the consultant writer, Kathi Hanna, for their superb work and dedication to this project; to staff members David Padgham and John Sislin, and editor, Sara Maddox, for their assistance; and to Paige Herwig, Laurie Richardson, and Judith A. Hunt for their sterling contributions in checking source materials and assisting with the final production of the report.

Harry T. Edwards and Constantine Gatsonis
Committee Co-chairs

work in laboratories that conduct hundreds or thousands of evaluations of impression evidence develop useful experience and judgment, it is difficult to assert that the field has enough collective judgment about the variabilities in lip prints and ear prints based on tens of examinations. The community simply does not have enough data about the natural variability of those less frequent impressions, absent the presence of a clear deformity or scar, to infer whether the observed degree of similarity is significant.

Most of the research in the field is conducted in forensic laboratories, with the results published in trade journals, such as the *Journal of Forensic Identification*. With regard to reporting, SWGTREAD is moving toward the use of standard language to convey the conclusions reached.⁵⁸ But neither IAI nor SWGTREAD addresses the issue of what critical research should be done or by whom, critical questions that should be addressed include the persistence of individual characteristics, the rarity of certain characteristic types, and the appropriate statistical standards to apply to the significance of individual characteristics. Also, little if any research has been done to address rare impression evidence. Much more research on these matters is needed.

TOOLMARK AND FIREARMS IDENTIFICATION

Toolmarks are generated when a hard object (tool) comes into contact with a relatively softer object. Such toolmarks may occur in the commission of a crime when an instrument such as a screwdriver, crowbar, or wire cutter is used or when the internal parts of a firearm make contact with the brass and lead that comprise ammunition. The marks left by an implement such as a screwdriver or a firearm's firing pin depend largely on the manufacturing processes—and manufacturing tools—used to create or shape it, although other surface features (e.g., chips, gouges) might be introduced through post-manufacturing wear. Manufacturing tools experience wear and abrasion as they cut, scrape, and otherwise shape metal, giving rise to the theory that any two manufactured products—even those produced consecutively with the same manufacturing tools—will bear microscopically different marks. Firearms and toolmark examiners believe that toolmarks may be traced to the physical heterogeneities of an individual tool—that is, that “individual characteristics” of toolmarks may be uniquely associated with a specific tool or firearm and are reproduced by the use of that tool and only that tool.

The manufacture and use of firearms produces an extensive set of

⁵⁸ SWGTREAD. 2006. *Standard Terminology for Expressing Conclusions of Forensic Footwear and Tire Impression Examinations*. Available at www.theiai.org/guidelines/swgtread/terminology_final.pdf.

specialized toolmarks. Gun barrels typically are rifled to improve accuracy, meaning that spiral grooves are cut into the barrel's interior. The process of cutting these grooves into the barrel leaves marks and scrapes on the relatively softer metal of the barrel.⁵⁹ In turn, these markings are transferred to the softer metal of a bullet as it exits the barrel. Over time, with repeated use (and metal-to-metal scraping), the marks on a barrel (and the corresponding "stria" imparted to bullets) may change as individual imperfections are formed or as cleanliness of the barrel changes. The brass exterior of cartridge cases receive analogous toolmarks during the process of gun firing: the firing pin dents the soft primer surface at the base of the cartridge to commence firing, the primer area is forced backward by the buildup of gas pressure (so that the texture of the gun's breech face is impressed on the cartridge), and extractors and ejectors leave marks as they expel used cartridges and cycle in new ammunition.

Firearms examination is one of the more common functions of crime laboratories. Even small laboratories with limited services often perform firearms analysis. In addition to the analysis of marks on bullets and cartridges, firearms examination also includes the determination of the firing distance, the operability of a weapon, and sometimes the analysis of primer residue to determine whether someone recently handled a weapon. These broader aspects are not covered here.

Sample and Data Collection

When a tool is used in a crime, the object that contains the tool marks is recovered when possible. If a toolmark cannot be recovered, it can be photographed and cast. Test marks made by recovered tools can be made in a laboratory and compared with crime scene toolmarks.

In the early 1990s, the FBI and the Bureau of Alcohol, Tobacco, Firearms, and Explosives (ATF) developed separate databases of images of bullet and cartridge case markings, which could be queried to suggest possible matches. In 1996, the National Institute of Standards and Technology (NIST) developed data exchange standards that permitted the integration of the FBI's DRUGFIRE database (cartridge case images) and the ATF's CEASEFIRE database (then limited to bullet images). The current National Integrated Ballistic Information Network (NIBIN) includes images from both cartridge cases and bullets that are associated with crime scenes and is maintained by the ATF.

Periodically—and particularly in the wake of the Washington, DC,

⁵⁹ Although the metal and initial rifling are very similar, the cutting of the individual barrels, the finishing machining, and the cleaning and polishing begin the process of differentiation of the two sequentially manufactured barrels.

sniper attacks in 2002—the question has been raised of expanding the scope of databases like NIBIN to include images from test firings of newly manufactured firearms. In concept, this would permit downstream investigators who recover a cartridge case or bullet at a crime scene to identify the likely source firearm. Though two states (Maryland and New York) instituted such reference ballistic image databases for newly manufactured firearms, proposals to create such a database at the national level did not make substantial progress in Congress. A recent report of the National Academies, *Ballistic Imaging*, examined this option in great detail and concluded that “[a] national reference ballistic image database of all new and imported guns is not advisable at this time.”⁶⁰

Analyses

In both firearm and toolmark identification, it is useful to distinguish several types of characteristics that are considered by examiners. “Class characteristics” are distinctive features that are shared by many items of the same type. For example, the width of the head of a screwdriver or the pattern of serrations in the blade of a knife may be class characteristics that are common to all screwdrivers or knives of a particular manufacturer and/or model. Similarly, the number of grooves cut into the barrel of a firearm and the direction of “twist” in those grooves are class characteristics that can filter and restrict the range of firearms that match evidence found at a crime scene. “Individual characteristics” are the fine microscopic markings and textures that are said to be unique to an individual tool or firearm. Between these two extremes are “subclass characteristics” that may be common to a small group of firearms and that are produced by the manufacturing process, such as when a worn or dull tool is used to cut barrel rifling.

Bullets and cartridge cases are first examined to determine which class characteristics are present. If these differ from a comparison bullet or cartridge, further examination may be unnecessary. The microscopic markings on bullets and cartridge cases and on toolmarks are then examined under a comparison microscope (made from two compound microscopes joined by a comparison bridge that allows viewing of two objects at the same time). The unknown and known bullet or cartridge case or toolmark surfaces are compared visually by a firearms examiner, who can evaluate whether a match exists.

⁶⁰ National Research Council. 2008. *Ballistic Imaging*. Washington, DC: The National Academies Press, p. 5.

Scientific Interpretation

The task of the firearms and toolmark examiner is to identify the individual characteristics of microscopic toolmarks apart from class and subclass characteristics and then to assess the extent of agreement in individual characteristics in the two sets of toolmarks to permit the identification of an individual tool or firearm.

Guidance from the Association of Firearm and Tool Mark Examiners (AFTE)⁶¹ indicates that an examiner may offer an opinion that a specific tool or firearm was the source of a specific set of toolmarks or a particular bullet striation pattern when “sufficient agreement” exists in the pattern of two sets of marks. The standards then define agreement as significant “when it exceeds the best agreement demonstrated between tool marks known to have been produced by different tools and is consistent with the agreement demonstrated by tool marks known to have been produced by the same tool.”⁶²

Knowing the extent of agreement in marks made by different tools, and the extent of variation in marks made by the same tool, is a challenging task. AFTE standards acknowledge that these decisions involve subjective qualitative judgments by examiners and that the accuracy of examiners’ assessments is highly dependent on their skill and training. In earlier years, toolmark examiners relied on their past casework to provide a foundation for distinguishing between individual, class, and subclass characteristics. More recently, extensive training programs using known samples have expanded the knowledge base of examiners.

The emergence of ballistic imaging technology and databases such as NIBIN assist examiners in finding possible candidate matches between pieces of evidence, including crime scene exhibits held in other geographic locations. However, it is important to note that the final determination of a match is always done through direct physical comparison of the evidence by a firearms examiner, not the computer analysis of images. The growth of these databases also permits examiners to become more familiar with similarities in striation patterns made by different firearms. Newer imaging techniques assess toolmarks using three-dimensional surface measurement data, taking into account the depth of the marks. But even with more training and experience using newer techniques, the decision of the toolmark examiner remains a subjective decision based on unarticulated

⁶¹ Theory of identification, range of striae comparison reports and modified glossary definitions—An AFTE Criteria for Identification Committee report. 1992. *Journal of the Association of Firearm and Tool Mark Examiners* 24:336-340.

⁶² *Ibid.*, p. 336.

standards and no statistical foundation for estimation of error rates.⁶³ The National Academies report, *Ballistic Imaging*, while not claiming to be a definitive study on firearms identification, observed that, “The validity of the fundamental assumptions of uniqueness and reproducibility of firearms-related toolmarks has not yet been fully demonstrated.” That study recognized the logic involved in trying to compare firearms-related toolmarks by noting that, “Although they are subject to numerous sources of variability, firearms-related toolmarks are not completely random and volatile; one can find similar marks on bullets and cartridge cases from the same gun,” but it cautioned that, “A significant amount of research would be needed to scientifically determine the degree to which firearms-related toolmarks are unique or even to quantitatively characterize the probability of uniqueness.”⁶⁴

Summary Assessment

Toolmark and firearms analysis suffers from the same limitations discussed above for impression evidence. Because not enough is known about the variabilities among individual tools and guns, we are not able to specify how many points of similarity are necessary for a given level of confidence in the result. Sufficient studies have not been done to understand the reliability and repeatability of the methods. The committee agrees that class characteristics are helpful in narrowing the pool of tools that may have left a distinctive mark. Individual patterns from manufacture or from wear might, in some cases, be distinctive enough to suggest one particular source, but additional studies should be performed to make the process of individualization more precise and repeatable.

⁶³ Recent research has attempted to develop a statistical foundation for assessing the likelihood that more than one tool could have made specific marks by assessing consecutive matching striae, but this approach is used in a minority of cases. See A.A. Biasotti. 1959. A statistical study of the individual characteristics of fired bullets. *Journal of Forensic Sciences* 4:34; A.A. Biasotti and J. Murdock. 1984. “Criteria for identification” or “state of the art” of firearms and tool marks identification. *Journal of the Association of Firearms and Tool Mark Examiners* 16(4):16; J. Miller and M.M. McLean. 1998. Criteria for identification of tool marks. *Journal of the Association of Firearms and Tool Mark Examiners* 30(1):15; J.J. Masson. 1997. Confidence level variations in firearms identification through computerized technology. *Journal of the Association of Firearms and Tool Mark Examiners* 29(1):42. For a critique of this area and a comparison of scientific issues involving toolmark evidence and DNA evidence, see A. Schwartz. 2004-2005. A systemic challenge to the reliability and admissibility of firearms and tool marks identification. *Columbia Science and Technology Law Review* 6:2. For a rebuttal to this critique, see R.G. Nichols. 2007. Defending the scientific foundations of the firearms and tool mark identification discipline: Responding to recent challenges. *Journal of Forensic Sciences* 52(3):586-594.

⁶⁴ All quotes from National Research Council. 2008. *Ballistic Imaging*. Washington, DC: The National Academies Press, p. 3.

A fundamental problem with toolmark and firearms analysis is the lack of a precisely defined process. As noted above, AFTE has adopted a theory of identification, but it does not provide a specific protocol. It says that an examiner may offer an opinion that a specific tool or firearm was the source of a specific set of toolmarks or a bullet striation pattern when “sufficient agreement” exists in the pattern of two sets of marks. It defines agreement as significant “when it exceeds the best agreement demonstrated between tool marks known to have been produced by different tools and is consistent with the agreement demonstrated by tool marks known to have been produced by the same tool.” The meaning of “exceeds the best agreement” and “consistent with” are not specified, and the examiner is expected to draw on his or her own experience. This AFTE document, which is the best guidance available for the field of toolmark identification, does not even consider, let alone address, questions regarding variability, reliability, repeatability, or the number of correlations needed to achieve a given degree of confidence.

Although some studies have been performed on the degree of similarity that can be found between marks made by different tools and the variability in marks made by an individual tool, the scientific knowledge base for toolmark and firearms analysis is fairly limited. For example, a report from Hamby, Brundage, and Thorpe⁶⁵ includes capsule summaries of 68 toolmark and firearms studies. But the capsule summaries suggest a heavy reliance on the subjective findings of examiners rather than on the rigorous quantification and analysis of sources of variability. Overall, the process for toolmark and firearms comparisons lacks the specificity of the protocols for, say, 13 STR DNA analysis. This is not to say that toolmark analysis needs to be as objective as DNA analysis in order to provide value. And, as was the case for friction ridge analysis and in contrast to the case for DNA analysis, the specific features to be examined and compared between toolmarks cannot be stipulated *a priori*. But the protocols for DNA analysis do represent a precisely specified, and scientifically justified, series of steps that lead to results with well-characterized confidence limits, and that is the goal for all the methods of forensic science.

ANALYSIS OF HAIR EVIDENCE

The basis for hair analyses as forensic evidence stems from the fact that human and animal hairs routinely are shed and thus are capable of being

⁶⁵ J.E. Hamby, D.J. Brundage, and J.W. Thorpe. 2009. The identification of bullets fired from 10 consecutively rifled 9mm Ruger pistol barrels—A research project involving 468 participants from 19 countries. Available online at <http://www.fti-ibis.com/DOWNLOADS/Publications/10%20Barrel%20Article-%20a.pdf>.

transferred from an individual to the crime scene, and from the crime scene to an individual. Forensic hair examiners generally recognize that various physical characteristics of hairs can be identified and are sufficiently different among individuals that they can be useful in including, or excluding, certain persons from the pool of possible sources of the hair. The results of analyses from hair comparisons typically are accepted as class associations; that is, a conclusion of a “match” means only that the hair could have come from any person whose hair exhibited—within some levels of measurement uncertainties—the same microscopic characteristics, but it cannot uniquely identify one person. However, this information might be sufficiently useful to “narrow the pool” by excluding certain persons as sources of the hair.

Although animal hairs might provide useful evidence in certain cases (e.g., animal poaching), animal hair analysis often can lead to an identification of only the type of animal, not the specific breed⁶⁶; consequently, most (90 to 95 percent) of hair analyses refer to analyses of human hair. Human hairs from different parts of the body have different characteristics; Houck cautions strongly against drawing conclusions about hairs from one part of the body based on analyses of hairs from a different body part.⁶⁷

Houck and Bisbing recommend as minimal training for hair examiners a bachelor’s degree in a natural or applied science (e.g., chemistry, biology, forensic science), on-the-job training programs, and an annual proficiency test.⁶⁸

Sample Data and Collection

Sample hairs received for analysis initially are examined macroscopically for certain broad features such as color, shaft form (e.g., straight, wavy, curved, kinked), length, and overall shaft thickness (e.g., fine, medium, coarse).

In the second stage of analysis, hairs are mounted on microscopic slides using a mounting medium that has the same refractive index (about 1.54) as the hair, to better view the microscopic features (see next section). One hair or multiple hairs from the same source may be mounted on a glass microscope slide with an appropriate cover slip, as long as each mounted hair is clearly visible. It is most important that questioned and known hairs are mounted in the same type of mounting medium.

During this examination, the hair analyst attempts to identify the part of the body from which the hair might have come, based on certain de-

⁶⁶ P.D. Barnett and R.R. Ogle. 1982. Probabilities and human hair comparison. *Journal of Forensic Sciences* 27(2):272-278.

⁶⁷ M.M. Houck and R.E. Bisbing. 2005. Forensic human hair examination and comparison in the 21st century. *Forensic Science Review* 17(1):7.

⁶⁸ *Ibid.*, p. 12.

The Response of the Association of Firearm and Tool Mark Examiners[1] to the February 2009 National Academy of Science Report “Strengthening Forensic Science in the United States: A Path Forward.”[2]

June 22, 2009

By: The AFTE Committee for the Advancement of the Science of Firearm and Tool Mark Identification

Keywords: AFTE Response to the 2009 NAS Report, NAS Report, National Academy of Science, Daubert, NAS Recommendations, Strengthening Forensic Science in the United States: A Path Forward

ABSTRACT

The National Academy of Science Report, “Strengthening Forensic Science in the United States: A Path Forward” made 13 general recommendations regarding Forensic Science. Six of these recommendations directly relate to AFTE. Activities conducted by AFTE and SWGGUN already meet certain conditions of these six recommendations and are fully described in this response. The NAS report briefly critiqued firearm and toolmark identification directly; however, as stated on page S-5 of the report, a detailed evaluation by the NAS was not feasible. The critiques are addressed in this response even though it is evident that the NAS did not look critically at the scientific underpinning of firearm and toolmark identification despite having been provided with hundreds of relevant references.

In February 2009, the National Academy of Science (NAS) issued a report authored by its Committee on Identifying the Needs of the Forensic Science Community (herein referred to as the NAS Committee) entitled, “Strengthening Forensic Science in the United States: A Path Forward.” The aim of the NAS Committee, as stated on page P-1 of the report, was “to chart an agenda for progress in the forensic science community and its scientific disciplines,” including firearm and toolmark identification. Pursuant to this goal, the report offers 13 recommendations that represent the Committee’s studied opinion on how best to achieve its agenda.

The Association of Firearm and Tool Mark Examiners (AFTE) acknowledges what a tremendous undertaking it must have been to report on the needs of the forensic science community outside of the discipline of DNA analysis. Our review of the thirteen recommendations made by the NAS Committee found that six of them, numbers 2, 3, 6, 7, 8 and 9, directly relate to AFTE. We are pleased to report that activities conducted by AFTE and the Scientific Working Group for Firearms and Toolmarks (SWGGUN) already meet certain requirements or expectations of these six recommendations. These recommendations and our responses are as follows:

Recommendation 2 (page S-16):

The National Institute of Forensic Science (NIFS), after reviewing established standards such as ISO 17025, and in consultation with its advisory board, should establish standard terminology to be used in reporting on and testifying about the results of forensic science investigations. Similarly, it should establish model laboratory reports for different forensic science disciplines and specify the minimum information that should be included. As part of the accreditation and certification processes, laboratories and forensic scientists should be required to utilize model laboratory reports when summarizing the results of their analyses.

AFTE response to Recommendation 2:

In 1980, AFTE established an extensive glossary of terms and definitions covering all phases of firearm and toolmark examinations. This document, which is periodically revised as necessary, has served to establish standardized terminology and statements that can be rendered as conclusions in reports.

Recommendation 3 (pages S-16 and S-17):

Research is needed to address issues of accuracy, reliability, and validity in the forensic science disciplines. The National Institute of Forensic Science (NIFS) should competitively fund peer-reviewed research in the following areas:

(a) Studies establishing the scientific bases demonstrating the

Date Received: July 27, 2009
Peer Review Completed: July 28, 2009

validity of forensic methods.

(b) The development and establishment of quantifiable measures of the reliability and accuracy of forensic analyses. Studies of the reliability and accuracy of forensic techniques should reflect actual practice on realistic case scenarios, averaged across a representative sample of forensic scientists and laboratories. Studies also should establish the limits of reliability and accuracy that analytic methods can be expected to achieve as the conditions of forensic evidence vary. The research by which measures of reliability and accuracy are determined should be peer reviewed and published in respected scientific journals.

(c) The development of quantifiable measures of uncertainty in the conclusions of forensic analyses.

(d) Automated techniques capable of enhancing forensic technologies.

AFTE response to Recommendation 3:

There is an extensive body of research, extending back over one hundred years, which establishes the accuracy, reliability, and validity of conclusions rendered in the field of firearm and toolmark identification. A list of some of this pertinent research has been compiled by SWGGUN and is easily accessible on their website [3]. Since its inception in 1969, AFTE has emerged as a leading forensic organization and represents the relevant scientific community for the publication and dissemination of research in firearm and toolmark identification. In this role, AFTE actively encourages collaboration with educational institutions and governmental agencies.

Recommendation 6 (page S-18):

To facilitate the work of the National Institute of Forensic Science (NIFS), Congress should authorize and appropriate funds to NIFS to work with the National Institute of Standards and Technology (NIST), in conjunction with government laboratories, universities, and private laboratories, and in consultation with Scientific Working Groups, to develop tools for advancing measurement, validation, reliability, information sharing, and proficiency testing in forensic science and to establish protocols for forensic examinations, methods, and practices. Standards should reflect best practices and serve as accreditation tools for laboratories and as guides for the education, training, and certification of professionals. Upon completion of its work, NIST and its partners should report findings and recommendations to NIFS for further dissemination and implementation.

AFTE response to Recommendation 6:

AFTE facilitates the exchange of information between its members by holding annual training seminars and by the quarterly publication of a peer-reviewed, scientific journal. AFTE has also adopted documentation standards [4] and collaborates with SWGGUN in not only the development of examination protocols but also the periodic review of established ones. Finally, AFTE has had a comprehensive training program since 1982. This program has been frequently updated.

Recommendation 7 (page S-19):

Laboratory accreditation and individual certification of forensic science professionals should be mandatory, and all forensic science professionals should have access to a certification process. In determining appropriate standards for accreditation and certification, the National Institute of Forensic Science (NIFS) should take into account established and recognized international standards, such as those published by the International Organization for Standardization (ISO). No person (public or private) should be allowed to practice in a forensic science discipline or testify as a forensic science professional without certification. Certification requirements should include, at a minimum, written examinations, supervised practice, proficiency testing, continuing education, recertification procedures, adherence to a code of ethics, and effective disciplinary procedures. All laboratories and facilities (public or private) should be accredited, and all forensic science professionals should be certified, when eligible, within a time period established by NIFS.

AFTE response to Recommendation 7:

AFTE, through the assistance of a National Institute of Justice (NIJ) grant, developed and implemented a certification program in firearms, toolmarks, and gunshot residue examination and identification in 1999 [5]. This program includes all of the minimum requirements for a certification program recommended above.

Recommendation 8 (page S-19):

Forensic laboratories should establish routine quality assurance and quality control procedures to ensure the accuracy of forensic analyses and the work of forensic practitioners. Quality control procedures should be designed to identify mistakes, fraud, and bias; confirm the continued validity and reliability of standard operating procedures and protocols; ensure that best practices are being followed;

and correct procedures and protocols that are found to need improvement.

AFTE response to Recommendation 8:

AFTE endorses the quality assurance and quality control (QA/QC) requirements of accreditation inspections conducted by the American Society of Crime Lab Directors-Laboratory Accreditation Board (ASCLD-LAB), as well as the QA guidelines recommended by SWGGUN. Furthermore, SWGGUN has recently developed training and quality assurance recommendations that, if followed, help ensure accurate examination results.

Recommendation 9 (page S-19):

The National Institute of Forensic Science (NIFS), in consultation with its advisory board, should establish a national code of ethics for all forensic science disciplines and encourage individual societies to incorporate this national code as part of their professional code of ethics. Additionally, NIFS should explore mechanisms of enforcement for those forensic scientists who commit serious ethical violations. Such a code could be enforced through a certification process for forensic scientists.

AFTE response to Recommendation 9:

For many years, AFTE has had a comprehensive ethics code (adopted in 1980) and an equally comprehensive enforcement process.

However, AFTE is disappointed about what appears to be a deliberate oversight, expressed by the NAS Committee on page S-5:

The committee decided early in its work that it would not be feasible to develop a detailed evaluation of each discipline in terms of its scientific underpinning, level of development, and ability to provide evidence to address the major types of questions raised in criminal prosecutions and civil litigation.

By approaching their stated task with this self-imposed limitation in mind, the NAS Committee in effect chose to ignore extensive research supporting the scientific underpinnings of the identification of firearm and toolmark evidence, despite having been provided with documentation referencing many of these studies as early as June 2008.

The NAS report specifically addresses the subject of firearm and toolmark examination on pages 5-18 through 5-21. However, the Committee's discussion of the discipline is

inconsistent at times. For example, after stating on page 5-21, "because not enough is known about the variabilities among individual tools and guns, we are not able to specify how many points of similarity are necessary for a given level of confidence in the result," the Committee goes on to comment, "individual patterns from manufacture or from wear might, in some cases, be distinctive enough to suggest one particular source, but additional studies should be performed to make the process of individualization more precise and repeatable."

The NAS report also cites several statements critical of firearm and toolmark identification that appear in the National Research Council (NRC) 2008 report on ballistic imaging, while not referencing the AFTE response [6] to these statements, dated August 20, 2008. This AFTE response was sent to NRC Chairman, Dr. John Rolph, NAS Director-Committee of Law and Justice, Carol Petrie, and NAS Media Relations Officer, Sara Frueh. Additionally, in May 2008, Dr. Rolph wrote an affidavit to correct some misconceptions surrounding the critical comments contained in the NRC report for a court proceeding regarding the admissibility of firearms-related evidence. Both the AFTE response and Dr. Rolph's affidavit should have been readily available to the NAS Committee for review prior to publication of their February 2009 report.

In their report, the NAS Committee painted an incomplete and inaccurate portrait of the field of firearm and toolmark identification using a very broad brush, and in doing so did not consider the appropriate scientific principles on which our discipline was founded. AFTE is confident that the majority of its members can dispel the limitations and inaccuracies portrayed in the NAS report through well-prepared court testimony, which gives us the opportunity to explain and defend the identification of firearms and toolmarks using what we feel will be perceived as a compelling justification for our conclusions. A partial listing of relevant literature articles summarizing some of the foundational scientific research that has been conducted in the discipline of firearm and toolmark identification is provided below. [7-15]

Unfortunately, some firearm and toolmark examiners performing casework today are clearly outside the mainstream of forensic consciousness and do not conform or adhere to the current protocols and standards recommended by AFTE when conducting such examinations. These examiners take few case notes or other forms of documentation and are not familiar with the extensive amount of empirical and theoretical research that serves as the scientific basis of firearm and toolmark identification. Some of these examiners have been responsible for judicial rulings wherein their testimony has been limited in some way by the court due to their

nonconformity to accepted forensic protocols. Those of us in the mainstream of our profession are working very hard to overcome the cloud of suspicion that has formed over all of us by the shallow court presentations of a few. Justice cannot be served if the results of well-documented firearm and toolmark comparisons are precluded from American courts. Forensic casework performed by trained and competent examiners not only has the potential to identify the responsible firearm used in a crime, but may also quickly exclude a suspected firearm as having any association with a shooting incident. Either of these determinations can be of critical importance to the administration of justice.

The NAS report states that firearm and toolmark examinations have “a heavy reliance on the subjective findings of examiners rather than on the rigorous quantification and analysis of sources of variability” (page 5-21). However, the NAS report again does not address the relevant scientific literature that demonstrates a concerted effort by researchers to achieve a statistical foundation for the conclusions rendered in firearm and toolmark casework. [16-20] There was no apparent attempt by the Committee to acknowledge either existing research or that which is ongoing at various academic institutions across the country in order to formulate statistical foundations for toolmark identifications. [21, 22] This research has the potential to further support the validity and reliability of firearm and toolmark identifications and provide quantitative data to supplement the many years of empirical research that has been conducted in the field.

In closing, regardless of whether or not the NAS Committee’s vision of the formation of a National Institute of Forensic Science (NIFS) ultimately comes to fruition, AFTE remains committed to the advancement of the field of firearm and toolmark identification and looks forward to diligently working with whatever entity may eventually become responsible for the forensic enterprise in the United States. The stakes are too high to do anything less.

References

- [1] The AFTE Committee for the Advancement of the Science of Firearm and Tool Mark Identification. Committee members include: John Murdock (Chair), Brandon Giroux, Lucien Haag, James Hamby, Ph.D., Andy Smith, and Peter Striupaitis.
- [2] A copy of the NAS report is currently available on the internet at: <http://www.nap.edu>.
- [3] Internet source: <http://www.swggun.org/resource/resourcekit.htm>
- [4] “Standardization of Comparison Documentation” – an AFTE policy statement adopted at the business meeting of the 2005 AFTE Training Seminar, Indianapolis, IN – AFTE Journal, Vol. 38, No. 1, Winter 2006, pp. 72-73.
- [5] Kowalski, Ken, “Summary Report on the Development of Certification Examinations for Practicing Firearms and Toolmark Examiners,” AFTE Journal, Vol. 32, No. 4, Fall 2000, pp. 373-379.
- [6] AFTE Committee for the Advancement of the Science of Firearm & Toolmark Identification, “The Response of the Association of Firearm and Tool Mark Examiners to the National Academy of Sciences 2008 Report Assessing the Feasibility, Accuracy, and Technical Capability of a National Ballistics Database August 20, 2008,” AFTE Journal, Vol. 40, No. 3, Summer 2008, pp. 234-244.
- [7] Biasotti, A.A., “A Statistical Study of the Individual Characteristics of Fired Bullets,” Journal of Forensic Sciences, Vol. 4, No. 1, January 1959, pp. 34-50.
- [8] Biasotti, A.A. and Murdock, J.E., “‘Criteria for Identification’ or ‘State of the Art’ of Firearm and Toolmark Identification,” AFTE Journal, Vol. 16, No. 4, October 1984, pp. 16-34.
- [9] Nichols, R.G., “Firearm and Toolmark Identification Criteria: A Review of the Literature,” (Part I) Journal of Forensic Sciences, Vol. 42, No. 3, May 1997, pp. 466-474.
- [10] Bonfanti, M.S. and DeKinder, J., “The Influence of Manufacturing Processes on the Identification of Bullets and Cartridge Cases – A Review of the Literature,” Science and Justice, Vol. 39, No. 1, 1999, pp. 3-10.
- [11] Nichols, R.G., “Firearm and Toolmark Identification Criteria: A Review of the Literature, Part II,” Journal of Forensic Sciences, Vol. 48, No. 2, March 2003, pp. 318-327.
- [12] Miller, J., “An Examination of the Application of the Conservative Criteria for Identification of Striated Toolmarks Using Bullets Fired from Ten Consecutively Rifled Barrels,” AFTE Journal, Vol. 33, No. 2, Spring 2001, pp. 125-132.
- [13] Grzybowski, R., Miller, J., Moran, B., Murdock, J., Nichols, R., and Thompson, R., “Firearm/Toolmark Identification: Passing the Reliability Test under Federal and State Evidentiary Standards,” AFTE Journal, Vol. 35, No. 2, Spring 2003, pp. 209-241. *Of special note in this article is Appendix No. 2 (pp.234-240), which addresses the application of the scientific method to firearm and toolmark examination.
- [14] Nichols, R.G., “Defending the Scientific Foundations of the Firearms and Toolmark Identification Discipline: Responding to Recent Challenges,” Journal of Forensic Sciences, Vol. 52, No. 3, May 2007, pp. 586-594.
- [15] Nichols, R.G., “Firearm and Toolmark Identification: The Scientific Reliability and Validity of the AFTE Theory of Identification Discussed Within the Framework of a Study of Ten Consecutively Manufactured Extractors,” AFTE Journal, Vol. 36, No. 1, Winter 2004, pp. 67-88, and Vol. 36, No. 2, Spring 2004, pp. 124.

[16] Stone R.S., "How 'Unique' Are Impressed Toolmarks?," AFTE Journal, Vol. 35, No. 4, Fall 2003, pp. 376-383.

[17] Collins, E.R., "How 'Unique' Are Impressed Toolmarks? - An Empirical Study of 20 Worn Hammer Faces," AFTE Journal, Vol. 37, No. 4, Fall 2005, pp. 252-295.

[18] Neel, M. and Wells, M., "A Comprehensive Statistical Analysis of Striated Tool Mark Examinations, Part 1: Comparing Known Matches and Known Non-Matches," AFTE Journal, Vol. 39, No. 3, Summer 2007, pp. 176-192, and Vol. 39, No. 4, Fall 2007, p. 264.

[19] Howitt, D., Tulleners, F., Cebra, K., and Chen, S., "A Calculation of the Theoretical Significance of Matched Bullets," Journal of Forensic Sciences, Vol. 53, No. 4, July 2008, pp. 868-875.

[20] Biasotti, A. and Murdock, J., Chapter 23, "Firearms and Toolmark Identification" from Modern Scientific Evidence: The Law and Science of Expert Testimony, Volume 2, West Pub. Co., 1997 (1st edition), pp. 124-155; and currently: Biasotti, A., Murdock, J., and Moran, B., Chapter 34, "Firearms and Toolmark Identification" in Modern Scientific Evidence: The Law and Science of Expert Testimony, Volume 4, St. Paul: Thompson-West, 2008-2009 edition, pp. 573-631.

[21] Faden, D., Kidd, J., Craft, J., Chumbley, L.S., Morris, M., Genalo, L., Kreiser, J., and Davis, S., "Statistical Confirmation of Empirical Observations Concerning Tool Mark Striae," AFTE Journal, Vol. 39, No. 3, Summer 2007, pp. 211-216.

[22] Petraco, N., Petraco, N.D.K., Faber, L., and Pizzola, P., "Preparation of Tool Mark Standards with Jewelry Modeling Waxes," Journal of Forensic Sciences, Vol. 54, No. 2, March 2009, pp. 353-358.

September 2016, PCAST Report Excerpt

5.5 Firearms Analysis

Methodology

In firearms analysis, examiners attempt to determine whether ammunition is or is not associated with a specific firearm based on toolmarks produced by guns on the ammunition.^{310,311} (Briefly, gun barrels are typically rifled to improve accuracy, meaning that spiral grooves are cut into the barrel's interior to impart spin on the bullet. Random individual imperfections produced during the tool-cutting process and through "wear and tear" of the firearm leave toolmarks on bullets or casings as they exit the firearm. Parts of the firearm that come into contact with the cartridge case are machined by other methods.)

The discipline is based on the idea that the toolmarks produced by different firearms vary substantially enough (owing to variations in manufacture and use) to allow components of fired cartridges to be identified with particular firearms. For example, examiners may compare "questioned" cartridge cases from a gun recovered from a crime scene to test fires from a suspect gun.

Briefly, examination begins with an evaluation of class characteristics of the bullets and casings, which are features that are permanent and predetermined before manufacture. If these class characteristics are different, an elimination conclusion is rendered. If the class characteristics are similar, the examination proceeds to identify and compare individual characteristics, such as the striae that arise during firing from a particular gun. According to the Association of Firearm and Tool Mark Examiners (AFTE) the "most widely accepted method used in conducting a toolmark examination is a side-by-side, microscopic comparison of the markings on a questioned material item to known source marks imparted by a tool."³¹²

Background

In the previous section, PCAST expressed concerns about certain foundational documents underlying the scientific discipline of firearm and tool mark examination. In particular, we observed that AFTE's "Theory of Identification as it Relates to Toolmarks"—which defines the criteria for making an identification—is circular.³¹³ The "theory" states that an examiner may conclude that two items have a common origin if their marks are in "sufficient agreement," where "sufficient agreement" is defined as the examiner being convinced that the items are extremely unlikely to have a different origin. In addition, the "theory" explicitly states that conclusions are subjective.

³¹⁰ Examiners can also undertake other kinds of analysis, such as for distance determinations, operability of firearms, and serial number restorations as well as the analyze primer residue to determine whether someone recently handled a weapon.

³¹¹ For more complete descriptions, see, for example, National Research Council. *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press. Washington DC. (2009), and archives.fbi.gov/archives/about-us/lab/forensic-science-communications/fsc/july2009/review/2009_07_review01.htm.

³¹² See: Foundational Overview of Firearm/Toolmark Identification tab on afte.org/resources/swggun-ark (accessed May 12, 2016).

³¹³ Association of Firearm and Tool Mark Examiners. "Theory of Identification as it Relates to Tool Marks: Revised," *AFTE Journal*, Vol. 43, No. 4 (2011): 287.

Much attention in this scientific discipline has focused on trying to prove the notion that every gun produces “unique” toolmarks. In 2004, the NIJ asked the NRC to study the feasibility, accuracy, reliability, and advisability of developing a comprehensive national ballistics database of images from bullets fired from all, or nearly all, newly manufactured or imported guns for the purpose of matching ballistics from a crime scene to a gun and information on its initial owner.

In its 2008 report, an NRC committee, responding to NIJ’s request, found that “the validity of the fundamental assumptions of uniqueness and reproducibility of firearms-related toolmarks” had not yet been demonstrated and that, given current comparison methods, a database search would likely “return too large a subset of candidate matches to be practically useful for investigative purposes.”³¹⁴

Of course, it is not necessary that toolmarks be unique for them to provide useful information whether a bullet may have been fired from a particular gun. However, it is **essential** that the accuracy of the method for comparing them be known based on empirical studies.

Firearms analysts have long stated that their discipline has near-perfect accuracy. In a 2009 article, the chief of the Firearms-Toolmarks Unit of the FBI Laboratory stated that “a qualified examiner will rarely if ever commit a false-positive error (misidentification),” citing his review, in an affidavit, of empirical studies that showed virtually no errors.³¹⁵

With respect to firearms analysis, the 2009 NRC report concluded that “sufficient studies have not been done to understand the reliability and reproducibility of the methods”—that is, that the foundational validity of the field had not been established.³¹⁶

The Scientific Working Group on Firearms Analysis (SWGgun) responded to the criticisms in the 2009 NRC report by stating that:

The SWGgun has been aware of the scientific and systemic issues identified in this report for some time and has been working diligently to address them. . . . [the NRC report] identifies the areas where we must fundamentally improve our procedures to enhance the quality and reliability of our scientific results, as well as better articulate the basis of our science.³¹⁷

³¹⁴ National Research Council. *Ballistic Imaging*. The National Academies Press. Washington DC. (2008): 3-4.

³¹⁵ See: www.fbi.gov/about-us/lab/forensic-science-communications/fsc/july2009/review/2009_07_review01.htm.

³¹⁶ The report states that “Toolmark and firearms analysis suffers from the same limitations discussed above for impression evidence. Because not enough is known about the variabilities among individual tools and guns, we are not able to specify how many points of similarity are necessary for a given level of confidence in the result. Sufficient studies have not been done to understand the reliability and repeatability of the methods. The committee agrees that class characteristics are helpful in narrowing the pool of tools that may have left a distinctive mark.” National Research Council. *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press. Washington DC. (2009): 154.

³¹⁷ See: www.swggun.org/index.php?option=com_content&view=article&id=37&Itemid=22.

Non-black-box studies of firearms analysis: Set-based analyses

Because firearms analysis is at present a subjective feature-comparison method, its foundational validity can only be established through multiple independent black box studies, as discussed above.

Although firearms analysis has been used for many decades, only relatively recently has its validity been subjected to meaningful empirical testing. Over the past 15 years, the field has undertaken a number of studies that have sought to estimate the accuracy of examiners' conclusions. While the results demonstrate that examiners can under some circumstances identify the source of fired ammunition, 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.

Specifically, many of the studies employ “set-based” analyses, in which examiners are asked to perform all pairwise comparisons within or between small samples sets. For example, a “within-set” analysis involving n objects asks examiners to fill out an $n \times n$ matrix indicating which of the $n(n-1)/2$ possible pairs match. Some forensic scientists have favored set-based designs because a small number of objects gives rise to a large number of comparisons. The study design has a serious flaw, however: the comparisons are not independent of one another. Rather, they entail internal dependencies that (1) constrain and thereby inform examiners' answers and (2) in some cases, allow examiners to make inferences about the study design. (The first point is illustrated by the observation that if A and B are judged to match, then every additional item C must match either both or neither of them—cutting the space of possible answers in half. If A and B match one another but do not match C, this creates additional dependencies. And so on. The second point is illustrated by “closed-set” designs, described below.)

Because of the complex dependencies among the answers, set-based studies are not appropriately-designed black-box studies from which one can obtain proper estimates of accuracy. Moreover, analysis of the empirical results from at least some set-based studies (“closed-set” designs) suggest that they may substantially underestimate the false positive rate.

The Director of the Defense Forensic Science Center analogized set-based studies to solving a “Sudoku” puzzle, where initial answers can be used to help fill in subsequent answers.³¹⁸ As discussed below, DFSC's discomfort with set-based studies led it to fund the first (and, to date, only) appropriately designed black-box study for firearms analysis.

We discuss the most widely cited of the set-based studies below. We adopt the same framework as for latent prints, focusing primarily on (1) the 95 percent upper confidence limit of the false positive rate and (2) false positive rates based on the proportion of conclusive examinations, as the appropriate measures to report (see p. 91).

³¹⁸ PCAST interview with Jeff Salyards, Director, DFSC.

Within-set comparison

Some studies have involved within-set comparisons, in which examiners are presented, for example, with a collection of samples and asked them to determine which samples were fired from the same firearm. We reviewed two often-cited studies with this design.^{319,320} In these studies, most of the samples were from distinct sources, with only 2 or 3 samples being from the same source. Across the two studies, examiners identified 55 of 61 matches and made no false positives. In the first study, the vast majority of different-source samples (97 percent) were declared inconclusive; there were only 18 conclusive examinations for different-source cartridge cases and no conclusive examinations for different-source bullets.³²¹ In the second study, the results are only described in brief paragraph and the number of conclusive examinations for different-source pairs was not reported. It is thus impossible to estimate the false positive rate among conclusive examinations, which is the key measure for consideration (as discussed above).

Set-to-set comparison/ closed set

Another common design has been **between**-set comparisons involving a “closed set.” In this case, examiners are given a set of questioned samples and asked to compare them to a set of known standards, representing the possible guns from which the questioned ammunition had been fired. In a “closed-set” design, the source gun is

³¹⁹ Smith, E. “Cartridge case and bullet comparison validation study with firearms submitted in casework.” *AFTE Journal*, Vol. 37, No. 2 (2005): 130-5. In this study from the FBI, cartridges and bullets were fired from nine Ruger P89 pistols from casework. Examiners were given packets (of cartridge cases or bullets) containing samples fired from each of the 9 guns and one additional sample fired from one of the guns; they were asked to determine which samples were fired from the same gun. Among the 16 same-source comparisons, there were 13 identifications and 3 inconclusives. Among the 704 different-source comparisons, 97 percent were declared inconclusives, 2.5 percent were declared exclusions and 0 percent false positives.

³²⁰ DeFrance, C.S., and M.D. Van Arsdale. “Validation study of electrochemical rifling.” *AFTE Journal*, Vol. 35, No. 1 (2003): 35-7. In this study from the FBI, bullets were fired from 5 consecutively manufactured Smith & Wesson .357 Magnum caliber rifle barrels. Each of 9 examiners received two test packets, each containing a bullet from each of the 5 guns and two additional bullets (from the different guns in one packet, from the same gun in the other); they were asked to perform all 42 possible pairwise comparisons, which included 37 different-source comparisons. Of the 45 total same-source comparisons, there were 42 identifications and 3 inconclusives. For the 333 total different-source comparisons, the paper states that there were no false positives, but does not report the number of inconclusive examinations.

³²¹ Some laboratory policies mandate a very high bar for declaring exclusions.

always present. We analyzed four such studies in detail.^{322,323,324,325} In these studies, examiners were given a collection of questioned bullets and/or cartridge cases fired from a small number of consecutively manufactured firearms of the same make (3, 10, 10, and 10 guns, respectively) and a collection of bullets (or casings) known to have been fired from these same guns. They were then asked to perform a matching exercise—assigning the bullets (or casings) in one set to the bullets (or casings) in the other set.

This “closed-set” design is simpler than the problem encountered in casework, because the correct answer is always present in the collection. In such studies, examiners can perform perfectly if they simply match each bullet to the standard that is ~~closest~~. By contrast, in an open-set study (as in casework), there is no guarantee that the correct source is present—and thus no guarantee that the closest match is correct. Closed-set comparisons would thus be expected to underestimate the false positive rate.

Importantly, it is not necessary that examiners be told explicitly that the study design involves a closed set. As one of the studies noted:

The participants were not told whether the questioned casings constituted an open or closed set. However, from the questionnaire/answer sheet, participants could have assumed it was a closed set and that every questioned casing should be associated with one of the ten slides.³²⁶

³²² Stroman, A. “Empirically determined frequency of error in cartridge case examinations using a declared double-blind format.” *AFTE Journal*, Vol. 46, No. 2 (2014):157-175. In this study, bullets were fired from three Smith & Wesson guns. Each of 25 examiners received a test set containing three questioned cartridge cases and three known cartridge cases from each gun. Of the 75 answers returned, there were 74 correct assignments and one inconclusive examination.

³²³ Brundage, D.J. “The identification of consecutively rifled gun barrels.” *AFTE Journal*, Vol. 30, No. 3 (1998): 438-44. In this study, bullets were fired from 10 consecutively manufactured 9 millimeter Ruger P-85 semi-automatic pistol barrels. Each of 30 examiners received a test set containing 20 questioned bullets to compare to a set of 15 standards, containing at least one bullet fired from each of the 10 guns. Of the 300 answers returned, there were no incorrect assignments and one inconclusive examination.

³²⁴ Fadul, T.G., Hernandez, G.A., Stoiloff, S., and S. Gulati. “An empirical study to improve the scientific foundation of forensic firearm and tool mark identification utilizing 10 consecutively manufactured slides.” *AFTE Journal*. Vol. 45, No. 4 (2013): 376-93. An empirical study to improve the scientific foundation of forensic firearm and tool mark identification utilizing 10 consecutively manufactured slides. In this study, bullets were fired from 10 consecutively manufactured semi-automatic 9mm Ruger pistol slides. Each of 217 examiners received a test set consisting of 15 questioned casings and two known cartridge cases from each of the 10 guns. Of the 3255 answers returned, there were 3239 correct assignments, 14 inconclusive examinations and two false positives.

³²⁵ Hamby, J.E., Brundage, D.J., and J.W. Thorpe. “The identification of bullets fired from 10 consecutively rifled 9mm Ruger pistol barrels: a research project involving 507 participants from 20 countries.” *AFTE Journal*, Vol. 41, No. 2 (2009): 99-110. In this study, bullets were fired from 10 consecutively rifled Ruger P-85 barrels. Each of 440 examiners received a test set consisting of 15 questioned bullets and two known standards from each of the 10 guns. Of the 6600 answers returned, there were 6593 correct assignments, seven inconclusive examinations and no false positives.

³²⁶ Fadul, T.G., Hernandez, G.A., Stoiloff, S., and S. Gulati. “An empirical study to improve the scientific foundation of forensic firearm and tool mark identification utilizing 10 consecutively manufactured slides.” *AFTE Journal*, Vol. 45, No. 4 (2013): 376-93.

Moreover, as participants find that many of the questioned casings have strong similarities to the known casings, their surmise that matching knowns are always present will tend to be confirmed.

The issue with this study design is not just a theoretical possibility: it is evident in the results themselves. Specifically, the closed-set studies have inconclusive and false-positives rate that are dramatically lower (by more than 100-fold) than those for the partly open design (Miami-Dade study) or fully open, black-box designs (Ames Laboratory) studies described below (Table 2).³²⁷

In short, the closed-set design is problematic in principle and appears to underestimate the false positive rate in practice.³²⁸ The design is not appropriate for assessing scientific validity and measuring reliability.

Set-to-set comparison/party open set ('Miami Dade study')

One study involved a set-to-set comparison in which a few of the questioned samples lacked a matching known standard.³²⁹ The 165 examiners in the study were asked to assign a collection of 15 questioned samples, fired from 10 pistols, to a collection of known standards; two of the 15 questioned samples came from a gun for which known standards were not provided. For these two samples, there were 188 eliminations, 138 inconclusives and 4 false positives. The inconclusive rate was 41.8 percent and the false positive rate among conclusive examinations was 2.1 percent (confidence interval 0.6-5.25 percent). The false positive rate corresponds to an estimated rate of 1 error in 48 cases, with upper bound being 1 in 19.

As noted above, the results from the Miami-Dade study are sharply different than those from the closed-set studies: (1) the proportion of inconclusive results was 200-fold higher and (2) the false positive rate was roughly 100-fold higher.

Recent black-box study of firearms analysis

In 2011, the Forensic Research Committee of the American Society of Crime Lab Directors identified, among the highest ranked needs in forensic science, the importance of undertaking a black-box study in firearms analysis analogous to the FBI's black-box study of latent fingerprints. DFSC, dissatisfied with the design of previous studies of firearms analysis, concluded that a black-box study was needed and should be conducted by an independent testing laboratory unaffiliated with law enforcement that would engage forensic examiners as

³²⁷ Of the 10,230 answers returned across the three studies, there were there were 10,205 correct assignments, 23 inconclusive examinations and 2 false positives.

³²⁸ Stroman (2014) acknowledges that, although the test instructions did not explicitly indicate whether the study was closed, their study could be improved if "additional firearms were used and knowns from only a portion of those firearms were used in the test kits, thus presenting an open set of unknowns to the participants. While this could increase the chances of inconclusive results, it would be a more accurate reflection of the types of evidence received in real casework."

³²⁹ Fadul, T.G., Hernandez, G.A., Stoiloff, S., and S. Gulati. "An empirical study to improve the scientific foundation of forensic firearm and tool mark identification utilizing consecutively manufactured Glock EBIS barrels with the same EBIS pattern." National Institute of Justice Grant #2010-DN-BX-K269, December 2013.
www.ncjrs.gov/pdffiles1/nij/grants/244232.pdf.

participants in the study. DFSC and Defense Forensics and Biometrics Agency jointly funded a study by the Ames Laboratory, a Department of Energy national laboratory affiliated with Iowa State University.³³⁰

Independent tests/ open ('Ames Laboratory study')

The study employed a similar design to the FBI's black-box study of latent fingerprints, with many examiners making a series of **independent** comparison decisions between a questioned sample and one or more known samples that may or may not contain the source. The samples all came from 25 newly purchased 9mm Ruger pistols.³³¹ Each of 218 examiners³³² was presented with 15 **separate** comparison problems—each consisting of one questioned sample and three known test fires from the same known gun, which might or might not have been the source.³³³ Unbeknownst to the examiners, there were five same-source and ten different-source comparisons. (In an ideal design, the proportion of same- and different-source comparisons would differ among examiners.)

Among the 2178 different-source comparisons, there were 1421 eliminations, 735 inconclusives and 22 false positives. The inconclusive rate was 33.7 percent and the false positive rate among conclusive examinations was 1.5 percent (upper 95 percent confidence interval 2.2 percent). The false positive rate corresponds to an estimated rate of 1 error in 66 cases, with upper bound being 1 in 46. (It should be noted that 20 of the 22 false positives were made by just 5 of the 218 examiners—strongly suggesting that the false positive rate is highly heterogeneous across the examiners.)

The results for the various studies are shown in Table 2. The tables show a striking difference between the closed-set studies (where a matching standard is always present by design) and the non-closed studies (where there is no guarantee that any of the known standards match). Specifically, the closed-set studies show a dramatically lower rate of inconclusive examinations and of false positives. With this unusual design, examiners succeed in answering all questions and achieve essentially perfect scores. In the more realistic open designs, these rates are much higher.

³³⁰ Baldwin, D.P., Bajic, S.J., Morris, M., and D. Zamzow. "A study of false-positive and false-negative error rates in cartridge case comparisons." Ames Laboratory, USDOE, Technical Report #IS-5207 (2014) afte.org/uploads/documents/swggun-false-positive-false-negative-usdoe.pdf.

³³¹ One criticism, raised by a forensic scientist, is that the study did not involve **consecutively** manufactured guns.

³³² Participants were members of AFTE who were practicing examiners employed by or retired from a national or international law enforcement agency, with suitable training.

³³³ Actual casework may involve more complex situations (for example, many different bullets from a crime scene). But, a proper assessment of foundational validity must **start** with the question of how often an examiner can determine whether a questioned bullet comes from a specific known source.

Table 2: Results From Firearms Studies*

Study Type	Results for different-source comparisons				
	Raw Data	Inconclusives	False positives among conclusive exams ³³⁴		
	Exclusions/ Inconclusives/ False positives		Freq. (Confidence Bound)	Estimated Rate	Bound on Rate
Set-to-set/closed (four studies)	10,205/23/2	0.2%	0.02% (0.06%)	1 in 5103	1 in 1612
Set-to-set/partly open (Miami-Dade study)	188/138/4	41.8%	2.0% (4.7%)	1 in 49	1 in 21
Black-box study (Ames Laboratory study)	1421/735/22	33.7%	1.5% (2.2%)	1 in 66	1 in 46

* “Inconclusives”: Proportion of total examinations that were called inconclusive. “Raw Data”: Number of false positives divided by number of conclusive examinations involving questioned items without a corresponding known (for set-to-set/slightly open) or non-mated pairs (for independent/open). “Freq. (Confidence Bond)”: Point estimate of false positive frequency, with the upper 95 percent confidence bounds. “Estimated”: The odds of a false positive occurring, based on the observed proportion of false positives. “Bound”: The odds of a false positive occurring, based on the upper bound of the confidence interval—that is, the rate could reasonably be as high as this value.

Conclusions

The early studies indicate that examiners can, under some circumstances, associate ammunition with the gun from which it was fired. However, as described above, most of these studies involved designs that are not appropriate for assessing the scientific validity or estimating the reliability of the method as practiced. Indeed, comparison of the studies suggests that, because of their design, many frequently cited studies seriously underestimate the false positive rate.

At present, there is only a single study that was appropriately designed to test foundational validity and estimate reliability (Ames Laboratory study). Importantly, the study was conducted by an independent group, unaffiliated with a crime laboratory. Although the report is available on the web, it has not yet been subjected to peer review and publication.

The scientific criteria for foundational validity require appropriately designed studies by more than one group to ensure reproducibility. Because there has been only a single appropriately designed study, the current evidence falls short of the scientific criteria for foundational validity.³³⁵ There is thus a need for additional, appropriately designed black-box studies to provide estimates of reliability.

³³⁴ The rates for all examinations are, reading across rows: 1 in 5115; 1 in 1416; 1 in 83; 1 in 33; 1 in 99; and 1 in 66.

³³⁵ The DOJ asked PCAST to review a recent paper, published in July 2016, and judge whether it constitutes an additional appropriately designed black-box study of firearms analysis (that is, the ability to associate ammunition with a particular gun). PCAST carefully reviewed the paper, including interviewing the three authors about the study design. Smith, T.P.,

Finding 6: Firearms analysis

Foundational validity. PCAST finds that firearms analysis currently falls short of the criteria for foundational validity, because there is only a single appropriately designed study to measure validity and estimate reliability. The scientific criteria for foundational validity require more than one such study, to demonstrate reproducibility.

Whether firearms analysis should be deemed admissible based on current evidence is a decision that belongs to the courts.

If firearms analysis is allowed in court, the scientific criteria for validity as applied should be understood to require clearly reporting the error rates seen in appropriately designed black-box studies (estimated at 1 in 66, with a 95 percent confidence limit of 1 in 46, in the one such study to date).

Smith, G.A., and J.B. Snipes. "A validation study of bullet and cartridge case comparisons using samples representative of actual casework." *Journal of forensic sciences* Vol. 61, No. 4 (2016): 939-946.

The paper involves a novel and complex design that is unlike any previous study. Briefly, the study design was as follows: (1) six different types of ammunition were fired from eight 40 caliber pistols from four manufacturers (two Taurus, two Sig Sauer, two Smith and Wesson, and two Glock) that had been in use in the general population and obtained by the San Francisco Police Department; (2) tests kits were created by randomly selecting 12 samples (bullets or cartridge cases); (3) 31 examiners were told that the ammunition was all recovered from a single crime scene and were asked to prepare notes describing their conclusions about which sets of samples had been fired from the same gun; and (4) based on each examiner's notes, the authors sought to re-create the logical path of comparisons followed by each examiner and calculate statistics based on this inferred numbers of comparisons performed by each examiner.

While interesting, the paper clearly is not a black-box study to assess the reliability of firearms analysis to associate ammunition with a particular gun, and its results cannot be compared to previous studies. Specifically: (1) The study employs a ~~within-set~~ comparison design (interdependent comparisons within a set) rather than a black-box design (many independent comparisons); (2) The study involves only a small number of examiners; (3) The central question with respect to firearms analysis is whether examiners can associate spent ammunition with a particular gun, not simply with a particular make of gun. To answer this question, studies must assess examiners' performance on ammunition fired from different guns of the same make ("within-class" comparisons) rather than from guns of different makes ("between-class" comparison); the latter comparison is much simpler because guns of different makes produce marks with distinctive "class" characteristics (due to the design of the gun), whereas guns of the same make must be distinguished based on "randomly acquired" features of each gun (acquired during rifling or in use). Accordingly, previous studies have employed only within-class comparisons. In contrast, the recent study consists of a mixture of within- vs. between-class comparisons, with the substantial majority being the simpler between-class comparisons. To estimate the false-positive rate for within-class comparisons (the relevant quantity), one would need to know the number of independent tests involving different-source within-class comparisons resulting in conclusive examinations (identification or elimination). The paper does not distinguish between within- and between-class comparisons, and the authors noted that they did not perform such analysis.

PCAST's comments are not intended as a criticism of the recent paper, which is a novel and valuable research project. They simply respond to DOJ's specific question: the recent paper does not represent a black-box study suitable for assessing scientific validity or estimating the accuracy of examiners to associate ammunition with a particular gun.

Validity as applied. If firearms analysis is allowed in court, validity as applied would, from a scientific standpoint, require that the expert:

- (1) has undergone rigorous proficiency testing on a large number of test problems to evaluate his or her capability and performance, and discloses the results of the proficiency testing; and
- (2) discloses whether, when performing the examination, he or she was aware of any other facts of the case that might influence the conclusion.

The Path Forward

Continuing efforts are needed to improve the state of firearms analysis—and these efforts will pay clear dividends for the criminal justice system.

One direction is to continue to improve firearms analysis as a subjective method. With only one black-box study so far, there is a need for additional black-box studies based on the study design of the Ames Laboratory black-box study. As noted above, the studies should be designed and conducted in conjunction with third parties with no stake in the outcome (such as the Ames Laboratory or research centers such as the Center for Statistics and Applications in Forensic Evidence (CSAFE)). There is also a need for more rigorous proficiency testing of examiners, using problems that are appropriately challenging and publically disclosed after the test.

A second—and more important—direction is (as with latent print analysis) to convert firearms analysis from a subjective method to an objective method.

This would involve developing and testing image-analysis algorithms for comparing the similarity of tool marks on bullets. There have already been encouraging steps toward this goal.³³⁶ Recent efforts to characterize 3D images of bullets have used statistical and machine learning methods to construct a quantitative “signature” for each bullet that can be used for comparisons across samples. A recent review discusses the potential for surface topographic methods in ballistics and suggests approaches to use these methods in firearms examination.³³⁷ The authors note that the development of optical methods have improved the speed and accuracy of capturing surface topography, leading to improved quantification of the degree of similarity.

³³⁶ For example, a recent study used data from three-dimensional confocal microscopy of ammunition to develop a similarity metric to compare images. By performing all pairwise comparisons among a total of 90 cartridge cases fired from 10 pistol slides, the authors found that the distribution of the metric for same-gun pairs did not overlap the distribution of the metric for different-gun pairs. Although a small study, it is encouraging. Weller, T.J., Zheng, X.A., Thompson, R.M., and F. Tulleners. “Confocal microscopy analysis of breech face marks on fired cartridge cases from 10 consecutively manufactured pistol slides.” *Journal of Forensic Sciences*, Vol. 57, No. 4 (2012): 912-17.

³³⁷ Vorburger, T.V., Song, J., and N. Petraco. “Topography measurements and applications in ballistics and tool mark identification.” *Surface topography: Metrology and Properties*, Vol. 4 (2016) 013002.

In a recent study, researchers used images from an earlier study to develop a computer-assisted approach to match bullets that minimizes human input.³³⁸ The group’s algorithm extracts a quantitative signature from a bullet 3D image, compares the signature across two or more samples, and produces a “matching score,” reflecting the strength of the match. On the small test data set, the algorithm had a very low error rate.

There are additional efforts in the private sector focused on development of accurate high-resolution cartridge casing representations to improve accuracy and allow for higher quality scoring functions to improve and assign match confidence during database searches. The current NIBIN database uses older (non-3D) technology and does not provide a scoring function or confidence assignment to each candidate match. It has been suggested that a scoring function could be used for blind verification for human examiners.

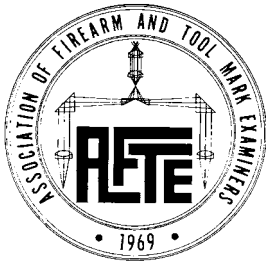
Given the tremendous progress over the past decade in other fields of image analysis, we believe that fully automated firearms analysis is likely to be possible in the near future. However, efforts are currently hampered by lack of access to realistically large and complex databases that can be used to continue development of these methods and validate initial proposals.

NIST, in coordination with the FBI Laboratory, should play a leadership role in propelling this transformation by creating and disseminating appropriate large datasets. These agencies should also provide grants and contracts to support work—and systematic processes to evaluate methods. In particular, we believe that “prize” competitions—based on large, publicly available collections of images³³⁹—could attract significant interest from academic and industry.

³³⁸ Hare, E., Hofmann, H., and A. Carriquiry. “Automatic matching of bullet lands.” Unpublished paper, available at: arxiv.org/pdf/1601.05788v2.pdf.

³³⁹ On July 7, 2016 NIST released the NIST Ballistics Toolmark Research Database (NBTRD) as an open-access research database of bullet and cartridge case toolmark data (tsapps.nist.gov/NBTRD). The database contains reflectance microscopy images and three-dimensional surface topography data acquired by NIST or submitted by users.





Association of Firearm and Tool Mark Examiners

Response to PCAST Report on Forensic Science October 31, 2016

In September, 2016 the President's Council of Advisors on Science and Technology (PCAST) issued a report titled "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods." As the leading professional organization for practitioners of forensic firearm identification, the Association of Firearm and Tool Mark Examiners (AFTE) acknowledges the challenge faced by the PCAST to understand the scientific field of comparative sciences from their stated brief review of the literature. AFTE strongly agrees with the premise that additional ongoing structured research strengthens the foundational and applied validity of firearm identification, as well as endeavors to reduce the effects of cognitive bias and subjectivity. However, we cannot overstate our disappointment in the PCAST's choice to ignore the research that has been conducted.

Decades of validation and proficiency studies have demonstrated that firearm and toolmark identification is scientifically valid, and that despite the subjective nature of the final comparison stage of analysis, competent examiners employing standard, validated procedures will rarely, if ever, commit false identifications or false eliminations. The foundational literature of the science has been presented to bodies such as the PCAST and the National Academy of Science (NAS) on multiple occasions and can be found at these links on the AFTE website: <https://afte.org/resources/afte-position-documents> , <https://afte.org/resources/swggun-ark>. The PCAST report is highly critical of any research that is not considered a "black box" study; and while this type of research is valuable and should be utilized more going forward, AFTE believes it is not the sole standard by which good science is measured.

The PCAST report references one such black box study conducted in 2014 by the Midwest Forensics Resource Center (MFRC) at the Ames Laboratory, Iowa State University, as the solitary study that can be utilized to accurately determine the error rate for firearm identification. The results of the Ames study were consistent with previous research demonstrating a very low error rate among properly trained examiners. However, the PCAST recommendation that any and all court testimony should refer to this one study as the singular foundational research of firearm and tool mark examination is irresponsible and inaccurate, and suggests a fundamental lack of understanding about the range of analyses done in this forensic discipline. While a global and numerically precise average of accuracy (error rate) would be useful in evaluating the value of an analytical technique, of greater relevance is the performance of the individual examiner as demonstrated by their participation in proficiency testing and similar testing. It should be noted that when foundational black-box type studies have been conducted in the past, the reported errors tend to be clustered among individuals or small groups

of participants rather than generally distributed amongst the population of all examiners participating in the study. Moreover, the technical and quality review processes utilized by laboratories for casework are not applied in these studies.

The PCAST report's assessment of the AFTE Theory of Identification as circular further illustrates the lack of adequate investigation and understanding on the part of the PCAST. First, the Theory of Identification has been in existence since 1992, not 2011 as cited.^(p.59) Second, the report erroneously defines sufficient agreement as "the examiner being convinced that the items are extremely unlikely to have a different origin."^(p.104) This characterization is utterly incorrect. The AFTE Theory of Identification clearly defines for the examiner when sufficient agreement does exist and how it is related to the significant duplication of random toolmarks. Only after sufficient agreement has been established does an examiner conclude that the two items are extremely unlikely to have a different origin. It has been consistently demonstrated that when the AFTE Theory of Identification is properly applied, examiners are able to conduct quality, accurate analysis.

Finally, the PCAST insistence on independent inquiry of our field in validation studies and matters of peer review implies a fatal limitation or bias within our community that can only be cured by an outside source. It is true that the majority of past research has been conducted by AFTE members, because while DNA and fingerprints have applications outside of forensics (such as medicine and biometrics), firearm identification has few profit-making applications and does not garner research attention from the private sector. Fortunately, in recent years a great diversity of academics, scientific professionals and agencies have joined in research on firearm and tool mark examination, but they require the input and participation of qualified forensic practitioners. We welcome the attention and ongoing collaboration of such organizations as the National Institute of Standards and Technology (NIST) and the newly-formed Center for Statistics and Applications in Forensic Evidence (CSAFE) in current and future research. Meanwhile, AFTE remains dedicated to the exchange of information, methods and best practices, and the furtherance of research in support of its members world-wide.

AN ADDENDUM TO THE PCAST REPORT ON FORENSIC SCIENCE IN CRIMINAL COURTS

On September 20, 2016, PCAST released its unanimous report to the President entitled “*Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*.” This new document, approved by PCAST on January 6, 2017, is an addendum to the earlier report developed to address input received from stakeholders in the intervening period.

Background

PCAST’s 2016 report addressed the question of when expert testimony based on a forensic feature-comparison method should be deemed admissible in criminal courts.¹ We briefly summarize key aspects of the previous report.

Forensic feature-comparison methods

PCAST chose to focus solely on forensic feature-comparison methods. These methods seek to determine whether a questioned sample is likely to have come from a known source based on shared features in certain types of evidence. Specific methods are defined by such elements as:

- (i) the type of evidence examined (e.g., DNA, fingerprints, striations on bullets, bite marks, footwear impressions, head-hair);
- (ii) the complexity of the sample examined (e.g., a DNA sample from a single person vs. a three-person mixture in which a person of interest may have contributed only 1%); and
- (iii) whether the conclusion concerns only “class characteristics” or “individual characteristics” (e.g., whether a shoeprint was made by a pair of size 12 Adidas Supernova Classic running shoes vs. whether it was made by a *specific* pair of such running shoes).

The U.S. legal system recognizes that scientific methods can assist the quest for justice, by revealing information and allowing inferences that lie beyond the experience of ordinary observers. But, precisely because the conclusions are potentially so powerful and persuasive, the law requires scientific testimony be based on methods that are scientifically valid and reliable.²

Requirement for empirical testing of subjective methods

In its report, PCAST noted that the *only* way to establish the scientific validity and degree of reliability of a *subjective* forensic feature-comparison method—that is, one involving significant human judgment—is to test it *empirically* by seeing how often examiners actually get the right answer. Such an empirical test of a subjective forensic feature-comparison method is referred to as a “black-box test.” The point reflects a central tenet underlying all science: *an empirical claim cannot be considered scientifically valid until it has been empirically tested*.

If practitioners of a subjective forensic feature-comparison method claim that, through a procedure involving substantial human judgment, they can determine with reasonable accuracy whether a particular type of evidence came from a particular source (e.g., a specific type of pistol or a specific pistol), the claim cannot be considered scientifically valid and reliable until one has tested it by (i) providing an adequate number of examiners with an adequate number of test problems that resemble those found in forensic practice and (ii) determining whether they get the right answer with acceptable

¹ As noted in the report, PCAST did not address the use of forensic methods in criminal *investigations*, as opposed to in criminal prosecution in courts.

² See discussion of the Federal Rules of Evidence in Chapter 3 of PCAST’s report.

frequency for the intended application.³ While scientists may debate the precise design of a study, there is no room for debate about the absolute requirement for empirical testing.

Importantly, the test problems used in the empirical study define the specific bounds within which the validity and reliability of the method has been established (e.g., is a DNA analysis method reliable for identifying a sample that comprises only 1% of a complex mixture?).

Evaluation of empirical testing for various methods

To evaluate the empirical evidence supporting various feature-comparison methods, PCAST invited broad input from the forensic community and conducted its own extensive review. Based on this review, PCAST evaluated seven forensic feature-comparison methods to determine whether there was appropriate empirical evidence that the method met the threshold requirements of “scientific validity” and “reliability” under the Federal Rules of Evidence.

- In two cases (DNA analysis of single-source samples and simple mixtures; latent fingerprint analysis), PCAST found that there was clear empirical evidence.
- In three cases (bitemark analysis; footwear analysis; and microscopic hair comparison), PCAST found *no empirical studies whatsoever* that supported the scientific validity and reliability of the methods.
- In one case (firearms analysis), PCAST found only one empirical study that had been appropriately designed to evaluate the validity and estimate the reliability of the ability of firearms analysts to associate a piece of ammunition with a specific gun. Because scientific conclusions should be shown to be reproducible, we judged that firearms analysis currently falls short of the scientific criteria for scientific validity.
- In the remaining case (DNA analysis of complex mixtures), PCAST found that empirical studies had evaluated validity within a limited range of sample types.

Responses to the PCAST Report

Following the report’s release, PCAST received input from stakeholders, expressing a wide range of opinions. Some of the commentators raised the question of whether empirical evidence is truly needed to establish the validity and degree of reliability of a forensic feature-comparison method.

The Federal Bureau of Investigation (FBI), which clearly recognizes the need for empirical evidence and has been a leader in performing empirical studies in latent-print examination, raised a different issue. Specifically, although PCAST had received detailed input on forensic methods from forensic scientists at the FBI Laboratory, the agency suggested that PCAST may have failed to take account of some relevant empirical studies. A statement issued by the Department of Justice (DOJ) on September 20, 2016 (the same day as the report’s release) opined that:

The report does not mention numerous published research studies which seem to meet PCAST’s criteria for appropriately designed studies providing support for foundational validity. That omission discredits the PCAST report as a thorough evaluation of scientific validity.

Given its respect for the FBI, PCAST undertook a further review of the scientific literature and invited a variety of stakeholders—including the DOJ—to identify any “published . . . appropriately designed

³ The size of the study (e.g., number of examiners and problems) affects the strength of conclusions that can be drawn (e.g., the upper bound on the error rate). The acceptable level of error rate depends on context.

studies” that had not been considered by PCAST and that established the validity and reliability of any of the forensic feature-comparison methods that the PCAST report found to lack such support. As noted below, DOJ ultimately concluded that it had no additional studies for PCAST to consider.

PCAST received written responses from 26 parties, including from Federal agencies, forensic-science and law-enforcement organizations, individual forensic-science practitioners, a testing service provider, and others in the US and abroad.⁴ Many of the responses are extensive, detailed and thoughtful, and they cover a wide range of topics; they provide valuable contributions for advancing the field. PCAST also held several in-person and telephonic meetings with individuals involved in forensic science and law enforcement. In addition, PCAST reviewed published statements from more than a dozen forensic-science, law-enforcement and other entities.⁵ PCAST is deeply grateful to all who took the time and effort to opine on this important topic.

In what follows, we focus on three key issues raised.

[Issue: Are empirical studies truly necessary?](#)

While forensic-science organizations agreed with the value of empirical tests of subjective forensic feature-comparison methods (that is, black-box tests), many suggested that the validity and reliability of such a method could be established *without* actually empirically testing the method in an appropriate setting. Notably, however, none of these respondents identified any *alternative* approach that could establish the validity and reliability of a subjective forensic feature-comparison method.

PCAST is grateful to these organizations because their thoughtful replies highlight the fundamental issue facing the forensic sciences: *the role of empirical evidence*. As noted in PCAST’s report, forensic scientists rightly point to several elements that provide critical foundations for their disciplines. However, there remains confusion as to whether these elements can suffice to establish the validity and degree of reliability of particular methods.

- (i) The forensic-science literature contains many papers describing variation among features. In some cases, the papers argue that patterns are “unique” (e.g., that no two fingerprints, shoes or DNA patterns are identical if one looks carefully enough). Such studies can provide a valuable *starting point* for a discipline, because they suggest that it may be worthwhile to attempt to develop reliable methods to identify the source of a sample based on feature comparison. However, such studies—no matter how extensive—can *never* establish the validity or degree of reliability of any particular method. Only empirical testing can do so.
- (ii) Forensic scientists rightly cite examiners’ experience and judgment as important elements in their disciplines. PCAST has great respect for the value of examiners’ experience and judgment: they are critical factors in ensuring that a scientifically valid and reliable method is practiced correctly. However, experience and judgment alone—no matter how great—can *never* establish the validity or degree of reliability of any particular method. Only empirical testing of the method can do so.⁶

⁴ www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_2016_additional_responses.pdf.

⁵ www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_2016_public_comments.pdf.

⁶ Some respondents, such as the Organization of Scientific Area Committees’ Friction Ridge Subcommittee, suggested that forensic science should be considered as analogous to medicine, in which physicians often treat patients on the basis of experience and judgment even in the absence of established empirical evidence. However, the analogy is inapt. Physicians act with a patient’s consent for the patient’s benefit. There is no legal requirement, analogous to the requirement imposed upon expert testimony in court by the Federal Rules of Evidence, that physician’s actions be based on “reliable principles and methods.” Physicians may rely on hunches; experts testifying in court about forensic feature-comparison methods may not.

- (iii) Forensic scientists cite the role of professional organizations, certification, accreditation, best-practices manuals, and training within their disciplines. PCAST recognizes that such practices play a critical role in any professional discipline. However, the existence of good professional practices alone—no matter how well crafted—can *never* establish the validity or degree of reliability of any particular method. Only empirical testing of the method can do so.

PCAST does not diminish in any way the important roles of prior research and other types of activities within forensic science and practice. Moreover, PCAST expresses great respect for the efforts of forensic practitioners, most of whom are devoted public servants. It is important to emphasize, however, contrary to views expressed by some respondents, that there is no “hierarchy” in which empirical evidence is simply the best way to establish validity and degree of reliability of a subjective feature-comparison method. In science, empirical testing is the only way to establish the validity and degree of reliability of such an empirical method.

Fortunately, empirical testing of empirical methods is feasible. There is no justification for accepting that a method is valid and reliable in the absence of appropriate empirical evidence.

Issue: Importance of other kinds of studies

In its response to PCAST’s call for further input, the Organization of Scientific Area Committees’ Friction Ridge Subcommittee (OSAC FRS), whose purview includes latent-print analysis, raised a very important issue:

While the OSAC FRS agrees with the need for black box studies to evaluate the overall validity of a particular method, the OSAC FRS is concerned this view could unintentionally stifle future research agendas aimed at dissecting the components of the black box in order to transition it from a subjective method to an objective method. If the PCAST maintains such an emphasis on black box studies as the *only* means of establishing validity, the forensic science community could be inundated with predominantly black box testing and potentially detract from progress in refining other foundational aspects of the method, such as those previously outlined by the OSAC FRS, in an effort to identify ways to emphasize objective methods over subjective methods (see www.nist.gov/topics/forensic-science/osac-research-development-needs.) Given the existing funding limitations, this will be especially problematic and the OSAC FRS is concerned other foundational research will thus be left incomplete.

PCAST applauds the work of the friction-ridge discipline, which has set an excellent example by undertaking both (i) path-breaking black-box studies to establish the validity and degree of reliability of latent-fingerprint analysis, and (ii) insightful “white-box” studies that shed light on how latent-print analysts carry out their examinations, including forthrightly identifying problems and needs for improvement. PCAST also applauds ongoing efforts to transform latent-print analysis from a subjective method to a fully objective method. In the long run, the development of objective methods is likely to increase the power, efficiency and accuracy of methods—and thus better serve the public.

In the case of subjective methods whose validity and degree of reliability have already been established by appropriate empirical studies (such as latent-print analysis), PCAST agrees that continued investment in black-box studies is likely to be less valuable than investments to develop fully objective methods. Indeed, PCAST’s report calls for substantial investment in such efforts.

The situation is different, however, for subjective methods whose validity and degree of reliability has not been established by appropriate empirical studies. If a discipline wishes to offer testimony based on a subjective method, it must first establish the method's validity and degree of reliability—which can only be done through empirical studies. However, as the OSAC FRS rightly notes, a discipline could follow an alternative path by abandoning testimony based on the subjective method and instead developing an objective method. Establishing the validity and degree of reliability of an objective method is often more straightforward. PCAST agrees that, in many cases, the latter path will make more sense.

Issue: Completeness of PCAST's evaluation

Finally, we considered the important question, raised by the DOJ in September, of whether PCAST had failed to consider “numerous published research studies which seem to meet PCAST's criteria for appropriately designed studies providing support for foundational validity.”

PCAST re-examined the five methods evaluated in its report for which the validity and degree of reliability had not been fully established. We considered the more than 400 papers cited by the 26 respondents; the vast majority had already been reviewed by PCAST in the course of the previous study. At the suggestion of John Butler of the National Institute of Standards and Technology (NIST), we also consulted INTERPOL's extensive summary of the forensic literature to identify additional potentially relevant papers.⁷ Although our inquiry was undertaken in response to the DOJ's concern, DOJ informed PCAST in late December that it had no additional studies for PCAST to consider.

Bitemark analysis

In its report, PCAST stated that it found no empirical studies whatsoever that establish the scientific validity or degree of reliability of bitemark analysis as currently practiced. To the contrary, it found considerable literature pointing to the unreliability of the method. None of the respondents identified any empirical studies that establish the validity or reliability of bitemark analysis. (One respondent noted a paper, which had already been reviewed by PCAST, that studied whether examiners agree when measuring features in dental casts but did not study bitemarks.) One respondent shared a recent paper by a distinguished group of biomedical scientists, forensic scientists, statisticians, pathologists, medical examiners, lawyers, and others, published in November 2016, that is highly critical of bitemark analysis and is consistent with PCAST's analysis.

Footwear analysis

In its report, PCAST considered feature-comparison methods for associating a shoeprint with a specific shoe based on randomly acquired characteristics (as opposed to with a class of shoes based on class characteristics). PCAST found no empirical studies whatsoever that establish the scientific validity or reliability of the method.

The President of the International Association for Identification (IAI), Harold Ruslander, responded to PCAST's request for further input. He kindly organized a very helpful telephonic meeting with IAI member Lesley Hammer. (Hammer has conducted some of the leading research in the field—including a 2013 paper, cited by PCAST, that studied whether footwear examiners reach similar conclusions when they are presented with evidence in which the identifying features have already been identified.)

⁷ The INTERPOL summaries list 4232 papers from 2010-2013 and 4891 papers from 2013-2016, sorted by discipline, see www.interpol.int/INTERPOL-expertise/Forensics/Forensic-Symposium.

Hammer confirmed that no empirical studies have been published to date that test the ability of examiners to reach correct conclusions about the source of shoeprints based on randomly acquired characteristics. Encouragingly, however, she noted that the first such empirical study is currently being undertaken at the West Virginia University. When completed and published, this study should provide the first actual empirical evidence concerning the validity of footwear examination. The types of samples and comparisons used in the study will define the bounds within which the method can be considered reliable.

Microscopic hair comparison

In its report, PCAST considered only those studies on microscopic hair comparison cited in a recent DOJ document as establishing the scientific validity and reliability of the method. PCAST found that none of these studies provided any meaningful evidence to establish the validity and degree of reliability of hair comparison as a forensic feature-comparison method. Moreover, a 2002 FBI study, by Houck and Budowle, showed that hair analysis had a stunningly high error rate in practice: Of hair samples that FBI examiners had found in the course of actual casework to be microscopically indistinguishable, 11% were found by subsequent DNA analysis to have come from different individuals.

PCAST received detailed responses from the Organization of Scientific Area Committees' Materials Subcommittee (OSAC MS) and from Sandra Koch, Fellow of the American Board of Criminalistics (Hairs and Fibers). These respondents urged PCAST not to underestimate the rich tradition of microscopic hair analysis. They emphasized that anthropologists have published many papers over the past century noting differences in average characteristics of hair among different ancestry groups, as well as variation among individuals. The studies also note intra-individual differences among hair from different sites on the head and across age.

While PCAST agrees that these empirical studies describing hair differences provide an encouraging starting point, we note that the studies do not address the validity and degree of reliability of hair comparison as a forensic feature-comparison method. What is needed are empirical studies to assess how often examiners incorrectly associate similar but distinct-source hairs (i.e., false-positive rate). Relevant to this issue, OSAC MS states: "Although we readily acknowledge that an error rate for microscopic hair comparison is not currently known, this should not be interpreted to suggest that the discipline is any less scientific." In fact, this is the central issue: the acknowledged lack of any empirical evidence about false-positive rates indeed means that, as a *forensic feature-comparison method*, hair comparison lacks a scientific foundation.

Based on these responses and its own further review of the literature beyond the studies mentioned in the DOJ document, PCAST concludes that there are no empirical studies that establish the scientific validity and estimate the reliability of hair comparison as a forensic feature-comparison method.

Firearms analysis

In its report, PCAST reviewed a substantial set of empirical studies that have been published over the past 15 years and discussed a representative subset in detail. We focused on the ability to associate ammunition not with a class of guns, but with a specific gun within the class.

The firearms discipline clearly recognizes the importance of empirical studies. However, most of these studies used flawed designs. As described in the PCAST report, "set-based" approaches can inflate examiners' performance by allowing them to take advantage of internal dependencies in the data. The

most extreme example is the “closed-set design”, in which the correct source of each questioned sample is always present; studies using the closed-set design have underestimated the false-positive and inconclusive rates by more than 100-fold. This striking discrepancy seriously undermines the validity of the results and underscores the need to test methods under appropriate conditions. Other set-based designs also involve internal dependencies that provide hints to examiners, although not to the same extent as closed-set designs.

To date, there has been only one appropriately designed black-box study: a 2014 study commissioned by the Defense Forensic Science Center (DFSC) and conducted by the Ames Laboratory, which reported an upper 95% confidence bound on the false-positive rate of 2.2%.⁸

Several respondents wrote to PCAST concerning firearms analysis. None cited additional appropriately designed black-box studies similar to the recent Ames Laboratory study. Stephen Bunch, a pioneer in empirical studies of firearms analysis, provided a thoughtful and detailed response. He agreed that set-based designs are problematic due to internal dependencies, yet suggested that certain set-based studies could still shed light on the method if properly analyzed. He focused on a 2003 study that he had co-authored, which used a set-based design and tested a small number of examiners (n=8) from the FBI Laboratory’s Firearms and Toolmarks Unit.⁹ Although the underlying data are not readily available, Bunch offered an estimate of the number of truly independent comparisons in the study and concluded that the 95% upper confidence bound on the false-positive rate in his study was 4.3% (vs. 2.2% for the Ames Laboratory black-box study).

The Organization of Scientific Area Committee’s Firearms and Toolmarks Subcommittee (OSAC FTS) took the more extreme position that all set-based designs are appropriate and that they reflect actual casework, because examiners often start their examinations by sorting sets of ammunition from a crime-scene. OSAC FTS’s argument is unconvincing because (i) it fails to recognize that the results from certain set-based designs are wildly inconsistent with those from appropriately designed black-box studies, and (ii) the key conclusions presented in court do not concern the ability to sort collections of ammunition (as tested by set-based designs) but rather the ability to accurately associate ammunition with a specific gun (as tested by appropriately designed black-box studies).

Courts deciding on the admissibility of firearms analysis should consider the following scientific issues:

- (i) There is only a single appropriate black-box study, employing a design that cannot provide hints to examiners. The upper confidence bound on the false-positive rate is equivalent to an error rate of 1 in 46.
- (ii) A number of older studies involve the seriously flawed closed-set design, which has dramatically underestimated the error rates. These studies do not provide useful information about the actual reliability of firearms analysis.
- (iii) There are several studies involving other kinds of set-based designs. These designs also involve internal dependencies that can provide hints to examiners, although not to the same extent that closed-set designs do. The large Miami-Dade study cited in the PCAST report and the small studies cited by Bunch fall into this category; these two studies have upper confidence bounds corresponding to error rates in the range of 1 in 20.

From a scientific standpoint, scientific validity should require at least two properly designed studies to ensure reproducibility. The issue for judges is whether one properly designed study, together with

⁸ PCAST also noted that some studies combine tests of both class characteristics and individual characteristics, but fail to distinguish between the results for these two very different questions.

⁹ PCAST did not select the paper for discussion in the report owing to its small size and set-based design, although it lists it.

ancillary evidence from imperfect studies, adequately satisfies the legal criteria for scientific validity. Whatever courts decide, it is essential that information about error rates is properly reported.

DNA analysis of complex mixtures

In its report, PCAST reviewed recent efforts to extend DNA analysis to samples containing complex mixtures. The challenge is that the DNA profiles resulting from such samples contain many alleles (depending on the number of contributors) that vary in height (depending on the ratios of the contributions), often overlap fully or partially (due to their “stutter patterns”), and may sometimes be missing (due to PCR dropout). Early efforts to interpret these profiles involved purely subjective and poorly defined methods, which were not subjected to empirical validation. Efforts then shifted to a quantitative method called combined probability of inclusion (CPI); however, this approach also proved seriously problematic.¹⁰

Recently, efforts have focused on an approach called probabilistic genotyping (PG), which uses mathematical models (involving a likelihood-ratio approach) and simulations to attempt to infer the likelihood that a given individual’s DNA is present in the sample. PCAST found that empirical testing of PG had largely been limited to a narrow range of parameters (number and ratios of contributors). We judged that the available literature supported the validity and reliability of PG for samples with three contributors where the person of interest comprises at least 20% of the sample. Beyond this approximate range (i.e. with a larger number of contributors or where the person of interest makes a lower than 20% contribution to the sample), however, there has been little empirical validation.¹¹

A recent controversy has highlighted issues with PG. In a prominent murder case in upstate New York, a judge ruled in late August (a few days before the approval of PCAST’s report) that testimony based on PG was inadmissible owing to insufficient validity testing.¹² Two PG software packages (STRMix and TrueAllele), from two competing firms, reached differing¹³ conclusions about whether a DNA sample in the case contained a tiny contribution (~1%) from the defendant. Disagreements between the firms have grown following the conclusion of the case.

PCAST convened a meeting with the developers of the two programs (John Buckleton and Mark Perlin), as well as John Butler from NIST, to discuss how best to establish the range in which a PG software program can be considered to be valid and reliable. Buckleton agreed that empirical testing of PG software with different kinds of mixtures was necessary and appropriate, whereas Perlin contended that empirical testing was unnecessary because it was mathematically impossible for the likelihood-ratio approach in his software to incorrectly implicate an individual. PCAST was unpersuaded by the latter argument. While likelihood ratios are a mathematically sound concept, their application requires

¹⁰ Just as the PCAST report was completed, a paper was published that proposed various rules for the use of CPI. See Bieber, F.R., Buckleton, J.S., Budowle, B., Butler, J.M., and M.D. Coble. “Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion.” *BMC Genetics*. bmcbgenet.biomedcentral.com/articles/10.1186/s12863-016-0429-7. While PCAST agreed that these rules are *necessary*, PCAST did not review whether these rules were sufficient to ensure reliability and took no position on this question.

¹¹ The few studies that have explored 4- or 5-person mixtures often involve mixtures that are derived from only a few sets of people (in some cases, only one). Because the nature of overlap among alleles is a key issue, it is critical to examine mixtures from various different sets of people. In addition, the studies involve few mixtures in which a sample is present at an extremely low ratio. By expanding these empirical studies, it should be possible to test validity and reliability across a broader range.

¹² See McKinley, J. “Judge Rejects DNA Test in Trial Over Garrett Phillips’s Murder.” *New York Times*, August 26, 2016, www.nytimes.com/2016/08/27/nyregion/judge-rejects-dna-test-in-trial-over-garrett-phillips-murder.html. The defendant was subsequently acquitted.

¹³ Document updated on January 17, 2017.

making a set of assumptions about DNA profiles that require empirical testing.¹⁴ Errors in the assumptions can lead to errors in the results. To establish validity with a range of parameters, it is thus important to undertake empirical testing with a *variety* of samples in the relevant range.¹⁵

PCAST received thoughtful input from several respondents. Notably, one response¹⁶ suggested that the relevant category for consideration should be expanded from “complex mixtures” (defined based on the number of contributors) to “complex samples” (defined to include also samples with low amounts of template, substantial degradation, or significant PCR inhibition, all of which will also complicate interpretation). We agree that this expansion could be useful.

The path forward is straightforward. The validity of specific PG software should be validated by testing a diverse collection of samples within well-defined ranges. The DNA analysis field contains excellent scientists who are capable of defining, executing, and analyzing such empirical studies.

When considering the admissibility of testimony about complex mixtures (or complex samples), judges should ascertain whether the published validation studies adequately address the nature of the sample being analyzed (e.g., DNA quantity and quality, number of contributors, and mixture proportion for the person of interest).

Conclusion

Forensic science is at a crossroads. There is growing recognition that the law requires that a forensic feature-comparison method be established as scientifically valid and reliable before it may be used in court and that this requirement can only be satisfied by actual empirical testing. Several forensic disciplines, such as latent-print analysis, have clearly demonstrated that actual empirical testing is feasible and can help drive improvement. A generation of forensic scientists appears ready and eager to embrace a new, empirical approach—including black-box studies, white-box studies, and technology development efforts to transform subjective methods into objective methods.

PCAST urges the forensic science community to build on its current forward momentum. PCAST is encouraged that NIST has already developed an approach, subject to availability of budget, for carrying out the functions proposed for that agency in our September report.

In addition, progress would be advanced by the creation of a cross-cutting Forensic Science Study Group—involving leading forensic and non-forensic scientists in equal measure and spanning a range of feature-comparison disciplines—to serve as a scientific forum to discuss, formulate and invite broad input on (i) empirical studies of validity and reliability and (ii) approaches for new technology development, including transforming subjective methods into objective methods. Such a forum would complement existing efforts focused on developing best practices and informing standards and might strengthen connections between forensic disciplines and other areas of science and technology. It might be organized by scientists in cooperation with one or more forensic and non-forensic science organizations—such as DFSC, NIST, IAI, and the American Association for the Advancement of Science.

¹⁴ Butler noted that one must make assumptions, for each locus, about the precise nature of reverse and forward stutter and about the probability of allelic dropout.

¹⁵ Butler noted that it is important to consider samples with different extents of allelic overlap among the contributors.

¹⁶ This response was provided by Keith Inman, Norah Rudin and Kirk Lohmueller.

2019 WL 4359486 (D.C.Super.) (Trial Order)
Superior Court of the District of Columbia.
Criminal Division - Felony Branch

UNITED STATES,

v.

Marquette TIBBS.

No. 2016-CF1-19431.

September 5, 2019.

Memorandum Opinion

Jessica Willis, Esq., jwillis@pdsdc.org, for defendant.


Prescott Loveland, Esq., ploveland@pdsdc.org, for defendant.

[Michael Ambrosino](#), Esq., michael.ambrosino@usdoj.gov, Assistant United States Attorney.

[Charles Willoughby, Jr.](#), Esq., charles.willoughby2@usdoj.gov, Assistant United States Attorney.

Lindsey Merikas, Esq., lindsey.merikas@usdoj.gov, Assistant United States Attorney.

[Todd E. Edelman](#), Judge.

***1** In this case, the defense raised and extensively litigated its objection to the government's proffer of expert testimony regarding firearms and toolmark identification, a species of specialized opinion testimony that judges have routinely admitted in criminal trials. Specifically, the government sought to introduce the testimony of the firearms and toolmark examiner who used a high-powered microscope to compare a cartridge casing found on the scene of the charged homicide with casings test-fired from a firearm allegedly discarded by a fleeing suspect. According to the government's proffer, this analysis permitted the examiner to identify the recovered firearm as the source of the cartridge casing collected from the scene. The defense argued that such a conclusion does not find support in reliable principles and methods, and thus must be excluded pursuant to the standard set by the District of Columbia Court of Appeals in [Motorola Inc. v. Murray](#), 147 A.3d 751 (D.C. 2016) (en banc); by the United States Supreme Court in  [Daubert v. Merrell Dow Pharms., Inc.](#), 509 U.S. 579 (1993); and by [Federal Rule of Evidence 702](#).

Courts across the country have regularly admitted such source attribution statements from firearms and toolmark examiners, without restriction, for several decades. However, on the heels of several major reports emanating from outside of the judiciary calling into question the foundations of the firearms and toolmark identification discipline, recent decisions of the District of Columbia Court of Appeals have imposed significant limitations on the conclusions that an expert in this field can render in court.

After conducting an extensive evidentiary hearing in this case—one that involved detailed testimony from a number of distinguished expert witnesses, review of all of the leading studies in the discipline, pre- and post-hearing briefing, and lengthy arguments by skilled and experienced counsel—this Court ruled on August 8, 2019 that application of the *Daubert* factors requires substantial restrictions on specialized opinion testimony in this area. Based largely on the inability of the published studies in the field to establish an error rate, the absence of an objective standard for identification, and the lack of acceptance

U.S. v. Tibbs, 2019 WL 4359486 (2019)

of the discipline's foundational validity outside of the community of firearms and toolmark examiners, the Court precluded the government from eliciting testimony identifying the recovered firearm as the source of the recovered cartridge casing. Instead, the Court ruled that the government's expert witness must limit his testimony to a conclusion that, based on his examination of the evidence and the consistency of the class characteristics and microscopic toolmarks, the firearm cannot be excluded as the source of the casing. The Court issues this Memorandum Opinion to further elucidate the ruling it made in open court.

I. BACKGROUND

A. Firearms and Toolmark Identification: The Basics






Numerous reports and court decisions have described in detail the theory and methodology behind the forensic discipline of firearms and toolmark identification. See, e.g., *United States v. Johnson*, (S5) 16 Cr. 281 (PGG), 2019 U.S. Dist. LEXIS 39590, at *16-21, 2019 WL 1130258, at *5-7 (S.D.N.Y. Mar. 13, 2019); *United States v. Simmons*, Case No. 2:16cr130, 2018 U.S. Dist. LEXIS 18606, at *5-11, 2018 WL 1882827, at *2-3 (E.D. Va. Jan. 12, 2018); *United States v. Otero*, 849 F. Supp. 2d 425, 427-28 (D.N.J. 2012); *United States v. Monteiro*, 407 F. Supp. 2d 351, 359-61 (D. Mass. 2006); *United States v. Green*, 405 F. Supp. 2d 104, 110-12 (D. Mass. 2005); Nat'l Res. Council, Nat'l Academies, *Strengthening Forensic Science in the United States: A Path Forward* 150-51, 152-53 (2009) [hereinafter *2009 NRC Report*]. In short, this field endeavors to match the components of spent ammunition, i.e., bullets and cartridge casings, to a particular firearm. See *Monteiro*, 407 F. Supp. 2d at 359. Firearms and toolmark identification is a specialized area of forensic toolmark identification, a discipline concerned with matching toolmarks to the specific tools that made them. *Otero*, 849 F. Supp. 2d at 427. Forensic toolmark identification rests on the notion that manufacturing processes leave behind "toolmarks" when a hard object, the tool, comes into contact with the relatively softer manufactured object. *2009 NRC Report* at 150.

*2 The discipline of firearms and toolmark identification derives from the theory that the tools used in the manufacture of firearms leave distinct markings on the internal components of a firearm, such as the barrel, breech face, and firing pin. *Otero*, 849 F. Supp. 2d at 427. These distinct markings, sometimes referred to as "individual characteristics," are said to result from the cutting, drilling, grinding, and hand-filing involved in the firearm manufacturing process. *Monteiro*, 407 F. Supp. 2d at 359. Such markings are supposedly individualized to each particular firearm as a result of the changes undergone by the tool being used to manufacture the firearm each time it cuts and scrapes metal to produce a new weapon. *Otero*, 849 F. Supp. 2d at 427. According to the theory, no two firearms, even those consecutively produced on the same production line, should bear microscopically identical toolmarks. See *id.*

When a firearm discharges a round of ammunition, the components of that ammunition come into contact with the internal components of the firearm. *Monteiro*, 407 F. Supp. 2d at 359-60. According to the proponents of firearms and toolmark identification, the tool markings on the firearm then transfer to the ammunition's components. *Id.* at 360. The theory underlying firearms and toolmark identification ultimately hypothesizes that "no two firearms should produce the same microscopic features on bullets and cartridge cases such that they could be falsely identified as having been fired from the same firearm." *Id.* at 361 (citation omitted). Stated more simply, firearms and toolmark examiners believe they can trace the toolmarks left on spent ammunition back to a particular firearm and that firearm only. See *2009 NRC Report* at 150.

Trained firearms examiners generally follow a particular methodology in attempting to reach conclusions as to the source of a bullet or cartridge casing. By using a comparison microscope to examine the markings on ammunition test fired from

U.S. v. Tibbs, 2019 WL 4359486 (2019)

a particular firearm and those on spent ammunition recovered from a crime scene, trained firearms examiners attempt to determine whether the spent ammunition was fired from that particular firearm. See  [Monteiro, 407 F. Supp. 2d at 361](#). When making these comparisons, examiners observe three types of characteristics of the ammunition—class, subclass, and individual characteristics.  [Otero, 849 F. Supp. 2d at 428](#). “Class characteristics are gross features common to most if not all bullets and cartridge cases fired from a *type* of firearm,” such as caliber and the number of lands and grooves on a bullet. *Id.* (emphasis added). These characteristics are predetermined at manufacture, [Simmons, 2018 U.S. Dist. LEXIS 18606, at *8, 2018 WL 1882827, at *2](#), and have been described as “family resemblances,”  [Monteiro, 407 F. Supp. 2d at 360](#). Subclass characteristics appear on a smaller subset of a particular make and model of firearm, such as a group of guns produced together at a particular place and time. *Id.* They are produced incidental to manufacture, sometimes as the result of being manufactured by the same irregular tool.  [Otero, 849 F. Supp. 2d at 428](#). Individual characteristics are microscopic markings produced during manufacture by the random and constantly-changing imperfections of tool surfaces as well as by subsequent use or damage to the firearm. *Id.* These are the markings purported to be unique to a particular firearm and that permit an individualized source determination—in other words, a conclusion that a particular firearm discharged a particular component of ammunition. See  [United States v. Taylor, 663 F. Supp. 2d 1170, 1174 \(D.N.M. 2009\)](#).

The forensic examination begins with the identification of class characteristics. *2009 NRC Report* at 152. If the observable class characteristics differ between the recovered and test fired ammunition, the examiner can immediately eliminate the recovered firearm as the source of the recovered ammunition. President's Council of Advisors on Sci. and Tech., Executive Off. of the President, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature- Comparison Methods* 104 (2016) [hereinafter PCAST Report]. If the class characteristics match, the examiner will use the comparison microscope to identify and compare the individual characteristics in both samples. *Id.* Under the theory of identification promulgated by the Association of Firearm and Tool Mark Examiners (“AFTE”) and discussed in detail *infra* at Section III(D), an examiner may declare the two samples to be of common origin (i.e., fired from the same gun) if she finds “sufficient agreement” between their individual characteristics. See *2009 NRC Report* at 153. Dissimilarities in observed subclass and/or individual characteristics can allow an examiner to exclude or eliminate the firearm as the source of the questioned sample of ammunition. The examiner may also render an inconclusive determination when there is agreement between the two samples' class characteristics but insufficient agreement or disagreement between their individual characteristics to make an identification or exclusion determination. See [Johnson, 2019 U.S. Dist. LEXIS 39590, at *9, 2019 WL 1130258, at *3](#).

B. Proffered Firearms and Toolmark Evidence in this Case, and the Defendant's Motion to Exclude

*3 Mr. Tibbs is charged with one count of first degree murder while armed as well as other related offenses. According to the government, a .40 caliber Smith & Wesson cartridge casing from a semi-automatic weapon was recovered from the scene of the homicide on November 11, 2016. The government alleges that a police officer observed Mr. Tibbs discarding a .40 caliber Smith & Wesson semi-automatic pistol shortly after the homicide occurred. On December 21, 2016, District of Columbia Department of Forensic Sciences Examiner Christopher Coleman prepared a report of examination, which indicated the recovered cartridge casing “was microscopically examined and identified as having been fired in [the recovered pistol], based on breechface marks and firing pin aperture shear marks.” Christopher Coleman, D.C. Dep't of Forensic Sci., *Report of Examination: Firearms Examination Unit Report* 1 (Dec. 21, 2016), Def's Mot. Ex. A, at 3 (Dec. 18, 2018).

Through his counsel, Mr. Tibbs challenged the admissibility of Mr. Coleman's opinion testimony with regard to firearms and toolmark identification. Specifically, the Defendant filed his Motion to Exclude the Testimony of Government's Proposed Expert Witness in Firearms Examination (“Defendant's Motion”) on December 18, 2018. The government filed its Opposition to Defendant's Motion on January 24, 2019; the Defendant filed a Reply on March 23, 2019, to which the government filed

U.S. v. Tibbs, 2019 WL 4359486 (2019)

a Surreply on April 15, 2019. The defense supplemented its pleadings with affidavits from Professor David Faigman and Dr. Nicholas Scurich, while the government submitted a declaration from Todd J. Weller, a report by Dr. Nicholas Petraco, and an affidavit from Dr. Bruce Budowle.

The Court conducted an extensive hearing on Defendant's Motion during the week of May 13, 2019, hearing lengthy testimony from Dr. Petraco, Mr. Weller, Dr. Scurich, and Professor Faigman. The parties' arguments on these issues spanned several days and finally concluded on June 10, 2019. Subsequent to the conclusion of the hearing, the Court provided the parties with the opportunity to file supplemental pleadings on the effect of the District of Columbia Court of Appeals' June 27, 2019 decision in *Williams v. United States (Williams II)*, 210 A.3d 734 (D.C. 2019), on the Court's resolution of Defendant's Motion; the parties each filed such a brief on July 10, 2019.¹

In his written pleadings, the Defendant asked the Court to exclude all testimony regarding firearms examination and identification in this case. In the alternative, he requested that the Court preclude Mr. Coleman from testifying that the recovered pistol fired the recovered cartridge casing, and limit his testimony to a conclusion that he could not exclude the recovered firearm as the source of the recovered cartridge casing. At the hearing, Mr. Tibbs proposed alternative restrictions on Mr. Coleman's proposed testimony but ultimately conceded that Mr. Coleman should at least be permitted to testify about his comparison of class characteristics between the recovered and test fired cartridge casings.

II. LEGAL STANDARD

A. Daubert and Rule 702: General Principles

In 2016, the District of Columbia Court of Appeals, sitting *en banc*, abandoned this jurisdiction's previous standard for the admissibility of expert opinion testimony. *Motorola*, 147 A.3d at 756-57. That standard, commonly referred to as the *Frye/Dyas* test, was originally developed by the United States Court of Appeals for the District of Columbia, and held that a scientific technique or principle could serve as the subject of expert testimony to the extent it had been “generally] accept[ed]” within its field of origin. See *Frye v. United States*, 293 F. 1013, 1014 (D.C. Cir. 1923). See generally *Dyas v. United States*, 376 A.2d 827, 831-32 (D.C. 1977). In *Motorola*, the Court of Appeals adopted the admissibility standard announced by the United States Supreme Court in *Daubert*—the same standard that has been applied in federal courts for over twenty years and that now appears in *Federal Rule of Evidence 702*. See *Motorola*, 147 A.3d at 756-57.





*4 *Daubert* itself repudiated *Frye* by holding its standard had been “superseded by the adoption of the Federal Rules of Evidence” and, in particular, by *Rule 702*. See 509 U.S. at 58789. The Supreme Court stated that trial judges considering the admissibility of proffered expert opinion testimony must conduct a “preliminary assessment of whether the reasoning or methodology underlying the testimony is scientifically valid and of whether that reasoning or methodology properly can be applied to the facts in issue.” *Id.* at 592-93. Thus, under *Daubert* and *Rule 702*, the admissibility of proffered expert opinion testimony does not exclusively rest on the acceptance of the opinion's underlying theory or methodology within a community of scientists or practioners. See *id.* at 594-95. Nor does it turn on the trial judge's view on the ultimate accuracy of the offered conclusion. See *id.* at 595. Instead, the admissibility inquiry focuses on whether reliable principles and methods support the proposed testimony and on whether those principles and methods were reliably applied in the case at hand. *Id.* at 594-95; see also *Motorola*, 147 A.3d at 754. *Rule 702* articulates the elements of the *Daubert* inquiry:

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:

(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;



U.S. v. Tibbs, 2019 WL 4359486 (2019)

- (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.

In changing the standard for the admissibility of expert opinion testimony, *Daubert* also modified the judge's role in making the admissibility determination. A judge must serve as a gatekeeper to “ensure that any and all scientific testimony or evidence admitted is not only relevant, but reliable.”  *Daubert*, 509 U.S. at 589. Indeed, *Daubert*, its progeny, and subsequent amendments to Rule 702 “gave to the courts a more significant gatekeeper role with respect to the admissibility of scientific and technical evidence than courts previously had played.”  *United States v. Glynn*, 578 F. Supp. 2d 567, 569 (S.D.N.Y. 2008). *Daubert* noted that such an assessment would involve the examination of a diverse set of factors. See  509 U.S. at 593. Envisioning a flexible inquiry, the Supreme Court did “not presume to set out a definitive checklist or test.”  *Id.* at 593-94. It did, however, enumerate five factors that would generally guide a trial court's admissibility inquiry:

- (1) whether a theory or technique can be (and has been) tested;²
- (2) whether the theory or technique has been subjected to peer review and publication;
- (3) the theory's or technique's known or potential rate of error;
- (4) the existence and maintenance of standards controlling the technique's operation; and
- (5) whether the theory or technique is generally accepted within the relevant scientific community.

Id.; see also *Motorola*, 147 A.3d at 754.

The proponent of the expert testimony bears the burden of proving its reliability by a preponderance of the evidence. *Cf.*  *Daubert*, 509 U.S. at 592 n.10. Our Court of Appeals has consistently held that admissibility determinations are within the discretion of the trial court. See, e.g.,  *Johnson v. United States*, 960 A.2d 281, 296 (D.C. 2008) (citing *Dockery v. United States*, 853 A.2d 687, 697 (D.C. 2004); *Smith v. United States*, 686 A. 2d 537, 542 (D.C. 1996))

B. Daubert and Firearms and Toolmark Identification

1. Mr. Tibbs's Daubert challenge


Mr. Tibbs raised a general challenge to the reliability of the principles and methods underlying firearms and toolmark identification. See generally Def.'s Mot. Accordingly, he at times moved to exclude all such evidence. At other points in his


U.S. v. Tibbs, 2019 WL 4359486 (2019)

pleadings and arguments, however, he offered a series of concessions and alternative proposals as well. As described in the Court's August 8, 2019 oral ruling, the undersigned found it useful to conceptualize Mr. Tibbs's challenge in several different ways. The Court could have analyzed the issues raised in Defendant's Motion by first determining whether the discipline of firearms and toolmark identification generally employs reliable principles and methods—such that it is admissible under *Daubert*, *Motorola*, and [Rule 702](#)—and subsequently, whether *Daubert* requires any limitations on the proffered testimony. Alternatively, the Court could have treated Mr. Tibbs's challenge as requiring two separate *Daubert* inquiries: (1) whether the Court could characterize the underlying theory of firearms and toolmark identification—the theory that manufacturing tools leave certain unique marks on firearms, and that firearms therefore leave unique and/or identifiable marks on bullets and cartridge casings—as reliable; and (2) whether the Court could conclude that a firearms examiner's opinion that she can compare bullets or cartridge casings and make an accurate source attribution statement (that is, a conclusion that a particular firearm fired a particular bullet or cartridge casing) finds support in reliable principles and methods. Regardless of the framework under which Mr. Tibbs's challenge was to be evaluated, Defendant's Motion ultimately required the Court to determine what type of opinion, if any, can be rendered with respect to firearms and toolmark evidence.

2. The limited persuasive value of existing case law

*5 Judges across the United States have considered similar challenges to firearms and toolmark identification evidence. Of course, “for many decades ballistics testimony was accepted almost without question in most federal courts in the United States.”

 [Glynn](#), 578 F. Supp. 2d at 569. Based on the pleadings in this case, as well as the Court's own research, there do not appear to be any reported cases in which this type of evidence has been excluded in its entirety. Earlier this year, the United States District Court for the District of Nevada also surveyed the relevant case law and concluded that no federal court had found the method of firearms and toolmark examination promoted by AFTE—the method generally used by American firearms examiners and employed by Mr. Coleman in this case—to be unreliable. [United States v. Romero-Lobato](#), 379 F. Supp. 3d 1111, 1117 (D. Nev. 2019); see also [Simmons](#), 2018 U.S. Dist. LEXIS 18606, at *28, 2018 WL 1882827, at *9 (“Defendants concede, as they must, that no court has ever *totally* rejected firearms and toolmark examination testimony.”); [State v. DeJesus](#), 7 Wn. App. 2d 849, 864 (2019) (“[T]he judicial decisions uniformly conclude toolmark and firearms identification is generally accepted and admissible at trial.”).

In evaluating the persuasive weight of these decisions, however, the undersigned could not help but note that, despite the enhanced gatekeeping role demanded by  [Daubert](#), see 509 U.S. at 589, the overwhelming majority of the reported *post-Daubert* cases regarding this type of expert opinion testimony have not engaged in a particularly extensive or probing analysis of the evidence's reliability. In 2009, the National Research Council (“NRC”) specifically criticized the judiciary's treatment of issues relating to the admissibility of firearms and toolmark evidence and the judiciary's failure to apply *Daubert* in a meaningful fashion. In the NRC's view, “[t]here is little to indicate that courts review firearms evidence pursuant to *Daubert*'s standard of reliability.” 2009 NRC Report at 107 n.82. The NRC observed that trial judges

... often affirm admissibility citing earlier decisions rather than facts established at a hearing. Much forensic evidence—including, for example, bite marks and firearm and toolmark identification—is introduced in criminal trials without any meaningful scientific validation, determination of error rates, or reliability testing to explain the limits of the discipline.

Id. at 107-08 (footnote and internal quotation marks omitted). Without disparaging the work of other courts, the NRC's critique of our profession rings true, at least to the undersigned: many of the published *post-Daubert* opinions on firearms and toolmark

U.S. v. Tibbs, 2019 WL 4359486 (2019)





identification involved no hearing on the admissibility of the evidence or only a cursory analysis of the relevant issues. Our Court of Appeals has noted that “[t]here is no ‘grandfathering’ provision in [Rule 702](#).” [Motorola](#), 147 A.3d at 758. Yet, the case law in this area follows a pattern in which holdings supported by limited analysis are nonetheless subsequently deferred to by one court after another. This pattern creates the appearance of an avalanche of authority; on closer examination, however, these precedents ultimately stand on a fairly flimsy foundation. The NRC credited Professor David Faigman—one of the defense experts who testified at the *Daubert* hearing in this matter—with the observation that trial courts defer to expert witnesses; appellate courts then defer to the trial courts; and subsequent courts then defer to the earlier decisions. *See 2009 NRC Report* at 108 n.85.


It is difficult to avoid the conclusion that, despite the criticisms of the NRC and other bodies, the judicial branch has demonstrated an aversion to meaningful hearings on this issue. In 2005, Judge Nancy Gertner of the United States District Court for the District of Massachusetts commented, “every single court post-*Daubert* has admitted [firearms identification] testimony, sometimes without any searching review, much less a hearing.” [Green](#), 405 F. Supp. 2d at 108 (emphasis omitted). Indeed, in 2012, the United States District Court for the Eastern District of New York could identify only four federal cases in which a judge had conducted a *Daubert* hearing on the admissibility of firearms and toolmark evidence. [United States v. Sebborn](#), 10 Cr. 87 (SLT), 2012 U.S. Dist. LEXIS 170576, at *17-18, 2012 WL 5989813, at *6 (E.D.N.Y. Nov. 30, 2012). Since then, few other federal courts have held similar hearings.³ *See Romero-Lobato*, 379 F. Supp. 3d at 1114; *Johnson*, 2019 U.S. Dist. LEXIS 39590, at *4-5, 2019 WL 1130258, at *2; *Simmons*, 2018 U.S. Dist. LEXIS 18606, at *3, 2018 WL 1882827, at *1; *United States v. Wrensford*, Criminal No. 2013-0003, 2014 U.S. Dist. LEXIS 102446, at *2, 2014 WL 3715036, at *1 (D. V.I. July 28, 2014). In most cases, courts resolved the objection to firearms and toolmark identification testimony without conducting any hearing at all. *See, e.g., United States v. Hylton*, Case No. 2:17-cr-00086-HDM-NJK, 2018 U.S. Dist. LEXIS 188817, at *6, 2018 WL 5795799, at *3 (D. Nev. Nov. 5, 2018); *United States v. White*, 17 Cr. 611 (RWS), 2018 U.S. Dist. LEXIS 163258, at *5, 2018 WL 4565140, at *2 (S.D.N.Y. Sept. 24, 2018); *United States v. Johnson*, Case No. 14-cr-00412-TEH, 2015 U.S. Dist. LEXIS 111921, at *11, 2015 WL 5012949, at *4 (N.D. Cal. Aug. 24, 2015); [United States v. Ashburn](#), 88 F. Supp. 3d 239, 244 (E.D.N.Y. 2015). Even in the few cases in which a *Daubert* hearing was conducted, it most often consisted only of the testimony of the examiner who worked on the case at issue, rather than of experts with a broader understanding of the foundational validity of the field.⁴ *See Romero-Lobato*, 379 F. Supp. 3d at 1115; *Johnson*, 2019 U.S. Dist. LEXIS 39590, at *3-5, 2019 WL 1130258, at *1-2; *Simmons*, 2018 U.S. Dist. LEXIS 18606, at *3, 2018 WL 1882827, at *1. The Court does not suggest that these decisions represent an abuse of discretion by the judges who issued them. The seemingly perfunctory nature of many of these written decisions does, however, lessen the persuasive weight of what would have otherwise been afforded to a near unanimous set of judicial opinions.

3. Judicial restrictions on firearms and toolmark identification testimony

*6 Although, as stated *supra*, no trial court has entirely excluded firearms and toolmark evidence in its entirety, some judges admitting firearms and toolmark evidence have recently restricted the conclusions examiners can render before a jury. *See Romero-Lobato*, 379 F. Supp. 3d at 1117; *DeJesus*, 7 Wn. App. 2d at 864 (“Courts have considered scholarly criticism of the methodology, and occasionally placed limitations on the opinions experts may offer based on the methodology.”). For example, at least one judge has precluded the sponsor of such evidence from referring to it as a “science.” [Glynn](#), 578 F. Supp. 2d at 568-69. Other courts have prohibited examiners from stating their conclusions to an absolute or statistical certainty. *See, e.g., Monteiro*, 407 F. Supp. 2d at 372. Some of these judges have permitted examiners to state their opinions only to a “reasonable degree of ballistic certainty” or a “reasonable degree of certainty in the ballistics field,” *see Ashburn*, 88 F. Supp. 3d at 249; [Monteiro](#), 407 F. Supp. 2d at 372; *Simmons*, 2018 U.S. Dist. LEXIS 18606, at *30, 2018 WL 1882827, at *10, while others have precluded any reference to the concept of “certainty,” regardless of what modifiers the examiner may attach, *see White*,


U.S. v. Tibbs, 2019 WL 4359486 (2019)

2018 U.S. Dist. LEXIS 163258, at *7, 2018 WL 4565140, at *3;  *United States v. Willock*, 696 F. Supp. 2d 536, 549 (D. Md. 2010);  *Glynn*, 578 F. Supp. 2d at 568-69. A number of courts have prevented examiners from stating that recovered ballistics evidence can be matched to a firearm to the exclusion of all other firearms. See  *Taylor*, 663 F. Supp. 2d at 1180;  *Green*, 405 F. Supp. 2d at 124.

Other judges have gone further in limiting expert opinion testimony regarding firearms and toolmark examination. In *Glynn*, a United States District Court Judge permitted a firearms examiner to state his conclusions of the match between the recovered ammunition and recovered firearm in terms of “more likely than not, but nothing more.”  578 F. Supp. 2d at 575 (internal quotation marks omitted). And in *State v. Terrell*, a state trial court judge referenced a case in which he had limited an examiner “to describing the similarities and dissimilarities between the known and unknown shell casings” and allowed her to conclude only that “the casings were consistent with having been fired from the subject hand gun.” CR170179563, 2019 Conn. Super. LEXIS 827, at *19, 2019 WL 2093108, at *5 (Mar. 21, 2019). Nonetheless, despite the handful of judges that have imposed these restrictions, “limitations on firearm and toolmark expert testimony [have been] the exception rather than the rule.” *Romero-Lobato*, 379 F. Supp. 3d at 1117.


The District of Columbia Court of Appeals, in a series of cases, has similarly restricted the conclusions firearms examiners may offer in court. See *WilliamsII*, 210 A.3d at 738; *Gardner v. United States*, 140 A.3d 1172, 1184 (D.C. 2016); *Jones v. United States*, 27 A.3d 1130, 1139 (D.C. 2011). Although, as discussed in Section IV *infra*, some ambiguity exists as to the state of the law post-*Williams II*, there can be no dispute that these authorities preclude firearms examiners from stating their conclusions with absolute or 100% certainty. See, e.g., *Gardner*, 140 A.3d at 1177. Nor can these expert witnesses identify a particular firearm as the source of spent ammunition to the exclusion of all other firearms. *Id.* Furthermore, it is unlikely examiners are even able to state their conclusions “with a reasonable degree of certainty.” See *id.* at 1184 n.19 (“[W]e have doubts as to whether trial judges in this jurisdiction should permit toolmark experts to state their opinions with a reasonable degree of certainty.” (internal quotation marks omitted)). None of these precedents, however, entirely control the *Daubert* challenge posed by Defendant's Motion. *Jones*, *Gardner*, and *WilliamsII* addressed the reliability of an examiner's conclusion, but all three were decided prior to the Court of Appeals' decision in *Motorola*—when the *Frye/Dyas* test still governed the admissibility of expert opinion testimony in the District of Columbia. None of them explicitly evaluated the admissibility of firearms and toolmark evidence under *Daubert* and Rule 702. And, while providing some examples of what firearms examiners *cannot* say in court, none of these cases provide definitive guidance as to what these witnesses *can* say.




4. Conclusion

*7 Granted, the precedents from other jurisdictions do provide at least some amount of guidance as to the challenge presented, and the Court of Appeals' recent opinions do have some bearing on the Court's present decision. However, particularly in light of the absence of any District of Columbia authority applying *Daubert* to firearms and toolmark identification testimony and the lack of any particularly persuasive authority from other jurisdictions, Defendant's Motion posed an issue of first impression. Accordingly, the Court undertook to determine the admissibility of the proffered testimony under *Daubert*, *Motorola*, and Rule 702. As explained by Judge Gertner, “*Daubert* plainly raised the standard for existing, established fields, inviting a reexamination even of generally accepted venerable, technical fields. Refusing to do so would be equivalent to grandfathering old irrationality.”  *Green*, 405 F. Supp. 2d at 118 (internal citations and quotation marks omitted).

III. APPLICATION OF THE DAUBERT FACTORS TO FIREARMS AND TOOLMARK ANALYSIS




A. Can and has the technique been tested?

The first of the *Daubert* factors—whether the technique or process in question can and has been tested—represents a “key question” in determining whether expert testimony should be admitted. *Romero-Lobato*, 379 F. Supp. 3d at 1118. As described in the Advisory Committee Notes to Rule 702, the “testability” of a theory refers to “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.” As *Daubert* itself noted, “generating hypotheses and testing them to see if they can be falsified ... is what distinguishes science from other fields of human inquiry.”  *Daubert*, 509 U.S. at 593 (citation omitted).

“There appears to be little dispute that toolmark identification is testable as a general matter.” *Johnson*, 2019 U.S. Dist. LEXIS 39590, at *44, 2019 WL 1130258, at *15. Indeed, 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. *See, e.g., Romero-Lobato*, 379 F. Supp. 3d at 1118-19; *Simmons*, 2018 U.S. Dist. LEXIS 18606, *18, 2018 WL 1882827, at *6;  *Ashburn*, 88 F. Supp. 3d at 245;  *Otero*, 849 F. Supp. 2d at 433. Although the NRC and PCAST reports have levied significant criticism against firearms and toolmark analysis, courts have found that such reports do not affect the method's testability. *See, e.g., Romero-Lobato*, 379 F. Supp. 3d at 1119; *see also*  *Otero*, 849 F. Supp. 2d at 433 (“Though the methodology of comparison and the AFTE ‘sufficient agreement’ standard inherently involves the subjectivity of the examiner's judgment as to matching toolmarks, the AFTE theory is testable on the basis of achieving consistent and accurate results.”). Additionally, some courts have cited annual proficiency testing undergone by firearms and toolmark examiners as further evidence of the method's testability. *See Johnson*, 2019 U.S. Dist. LEXIS 39590, at *45-46, 2019 WL 1130258, at *15 (citing *United States v. Diaz*, No. CR 05-000167 WHA, 2007 U.S. Dist. LEXIS 13152, at *15, 2007 WL 485967, at *5 (N.D. Cal. Feb. 12, 2007)); *United States v. Johnson*, 2015 U.S. Dist. LEXIS 111921, at *9, 2015 WL 5012949, at * 3.

Here, the propositions advanced by the government in support of its proffer of the expert testimony at issue—namely, that firearms leave discernible toolmarks on bullets and cartridge casings fired from them, and that trained examiners can conduct comparisons to determine whether a particular gun has fired particular ammunition—can be, and have been, tested. The Defendant's written pleadings and oral argument did not specifically contest this particular point, and the government met its burden with respect to testability.

B. Has the theory or technique been subjected to peer review and publication?

*8 The second of the *Daubert* factors considers whether the theory or technique “has been subjected to peer review and publication.” *Motorola*, 147 A.3d at 754 (quoting  *Daubert*, 509 U.S. at 593-94). As the Supreme Court emphasized in *Daubert*, “submission to the scrutiny of the scientific community is a component of ‘good science,’ in part because it increases the likelihood that substantive flaws in methodology will be detected.”  509 U.S. at 593. While the existence of peer reviewed literature can help determine a methodology's reliability under *Daubert*, the “fact of publication (or lack thereof) in a peer reviewed journal” is not dispositive. *Id.*; *see also Romero-Lobato*, 379 F. Supp. 3d at 1119;  *United States v. Mouzone*, 696 F. Supp. 2d 536, 571 (D. Md. 2009).

Evidence presented at the hearing demonstrated that studies assessing the foundational validity and reliability of the type of firearms pattern matching evidence proffered here—that is, studies that attempt to show whether trained firearms examiners can accurately attribute a particular firearm as the source of a particular cartridge casing or bullet—have been published and subjected to varying types of review. Two of the studies in this area, the 2019 study by James E. Hamby et al., *A Worldwide Study*

U.S. v. Tibbs, 2019 WL 4359486 (2019)

of Bullets Fired from 10 Consecutively Rifled 9MM RUGER Pistol Barrels—Analysis of Examiner Error Rate, 64 J. Forensic Sci. 551 (2019) [hereinafter 2019 Hamby Study], and the 2016 study by Tasha P. Smith et al., *A Validation Study of Bullet and Cartridge Case Comparisons Using Samples Representative of Actual Casework*, 61 J. Forensic Sci. 692 (2016) [hereinafter 2016 Smith Study], were published in the *Journal of Forensic Sciences*, and thus have undergone meaningful peer review. The *Journal of Forensic Sciences* employs “double-blind” peer review, a type of review process used throughout many scientific disciplines and designed to limit various types of bias by requiring that neither the study's authors nor the journal's reviewers know the identity of the other. Scurich Test. May 15, 2019, 37:3-7; Expert Report of Nicholas Scurich, PhD, 6 [hereinafter Scurich Report] (citing Author Guidelines, https://onlinelibrary.wiley.com/page/journal/1556_4029/homepage/forauthors.html (last visited August 28, 2019)). Further, this particular publication is an independent journal, unaffiliated with AFTE, any crime lab, or any individual with a financial or professional interest in the validation of the field of firearms and toolmark analysis.

However, most of the other studies in this field—including the vast majority of those relied upon by the government and the expert witnesses it presented at the *Daubert* hearing—have been published in the *AFTE Journal*, a publication produced by the Association of Firearm and Toolmark Examiners. The government's experts, Mr. Weller and Dr. Petraco, contended that the studies published in the *AFTE Journal* are subjected to both pre- and post-publication peer review. Prior to publication, articles submitted to the *AFTE Journal* are reviewed by AFTE members; the *AFTE Journal* utilizes an “open” pre-publication peer review process in which the author and the reviewers know each other's identity and may communicate directly during the review period. Scurich Report 7 (citing AFTE Peer Review Process - August 2009, <https://afte.org/afte-journal/afte-journal-peer-review-process> (last visited Aug. 28, 2019)). Both government experts primarily focused on post-publication peer review, and characterized letters to the editor in response to a published study as part of the *AFTE Journal's* peer review process. Suppl. Decl. of Todd J. Weller 7-8 [hereinafter Weller Suppl. Decl.]; Report of Dr. Nicholas Petraco 1-2 [hereinafter Petraco Report]; Petraco Test. May 13, 2019, 20:7-18. Further, Dr. Petraco also discussed the publication of “counter studies” as part of the peer review process. Petraco Report at 2.

*9 Other courts considering challenges to this discipline under *Daubert* have concluded that publication in the *AFTE Journal* satisfies this prong of the admissibility analysis. See, e.g., *Romero-Lobato*, 379 F. Supp. 3d at 1119 (citing *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); *Mouzone*, 696 F. Supp. 2d at 571. It is striking, however, that these courts devote little attention to the sufficiency of this journal's peer review process or to the issues stemming from a review process dominated by financially and professionally interested practitioners, and instead, mostly accept at face value the assertions regarding the adequacy of the journal's peer review process. See, e.g., *Romero-Lobato*, 379 F. Supp. 3d at 1119; *Johnson*, 2019 U.S. Dist. LEXIS 39590, at *49-50, 2019 WL 1130258, at *16-17; *Ashburn*, 88 F. Supp. 3d at 245-46; *Wrensford*, 2014 U.S. Dist. LEXIS 102446, at *43-44, 2014 WL 3715036, at *13; *Otero*, 849 F. Supp. 2d at 433; *Monteiro*, 407 F. Supp. 2d at 366-67.⁵

In the undersigned's view, three aspects of publication in the *AFTE Journal* make this journal's review process far less meaningful (and its published articles that much less reliable) than *Daubert* contemplates. First, as noted *supra*, the *AFTE Journal* peer review process itself is “open,” meaning that both the author and reviewer know the other's identity and may contact each other during the review process. Scurich Report 7 (citing AFTE Peer Review Process - August 2009, <https://afte.org/afte-journal/afte-journal-peer-review-process> (last visited Aug. 28, 2019)). This open process seems highly unusual for the publication of empirical scientific research, as Dr. Scurich testified and as Dr. Petraco admitted in his written report. Scurich Test. May 15, 2019, 28:17-18; Petraco Report at 2. The practice of double-blind peer review, by contrast, constitutes the standard among scientific publications and guards against personal and institutional biases by shielding both reviewer and author from the identity of the other. Mr. Weller, even while defending the *AFTE Journals* open process, acknowledged that the publication is now moving toward a blind peer review process. Weller Test. May 14, 2019 (1), 23:18; Weller Suppl. Decl. 8. While neither *Daubert*, *Motorola*, nor Rule 702 mandate any specific type of peer review process, the *AFTE Journals* use of a so-called “open”

U.S. v. Tibbs, 2019 WL 4359486 (2019)

process diminishes the extent to which proponents of firearms and toolmark identification evidence can claim that its articles have been subjected to meaningful, stringent peer review.



Second, AFTE does not make this publication generally available to the public or to the world of possible reviewers and commentators outside of the organization's membership. Of course, an interested party can receive the publication by joining AFTE, if such a person meets the organization's membership requirements, or can pay to access specific articles. Weller Test., May 14, 2019 (1), 18:16-21. But unlike other scientific journals, the *AFTE Journal* is not more broadly available and cannot even be obtained in university libraries. *Id.* 18:11-13. Such restricted access effectively forecloses the type of review of the journal's publications by a wider community of scientists, academics, and other interested parties that could serve as an important mechanism for quality assurance. Indeed, a National Commission on Forensic Science's (NCFS) publication listed among the criteria for "foundational, scientific literature supportive of forensic practice" that the articles be "published in a journal that is searchable using free, publicly available search engines (e.g. Pub Med, Google Scholar, National Criminal Justice Reference Service) that search major databases of scientific literature (e.g. Medline, National Criminal Justice Reference Service Abstracts Database, and Xplore)" and "published in a journal that is indexed in databases that are available through academic libraries and other services (e.g. JSTOR, Web of Science, Academic Search Complete, and SciFinder Scholar)." Nat'l Comm'n on Forensic Sci., *Scientific Literature in Support of Forensic Science and Practice*, 3 (2015), [justice.gov/archives/ncfs/file/786591/download](https://www.justice.gov/archives/ncfs/file/786591/download) [hereinafter NCFS Report].⁶ The *AFTE Journal*, by generally limiting the review of its publications and making them available only to its members or others who pay, avoids the scrutiny of scientists and academics outside the field of firearms and toolmark analysis. These limitations significantly diminish the stringency of the review that a study published in the *AFTE Journal* can be said to have undergone, even after its publication.

*10 Third, the very nature of AFTE impacts the meaningfulness of its review process. The *AFTE Journal* is published by the largest organization of practicing firearms and toolmark examiners, and its articles are reviewed by members of an editorial board composed entirely of members of AFTE. Scurich Report 7 (citing AFTE Peer Review Process - August 2009, <https://afte.org/afte-journal/afte-journal-peer-review-process> (last visited Aug. 28, 2019)). This oversight structure may create a threshold issue in terms of quality of peer review: as Dr. Scurich pointed out, those who review the *AFTE Journal's* articles may be trained and experienced in the field of firearms and toolmark examination, but do not necessarily have any specialized or even relevant training in research design and methodology. Scurich Report 7-8. Perhaps more importantly, members of the *Journal's* editorial board—those who review its articles prior to publication—have a vested, career-based interest in publishing studies that validate their own field and methodologies. In contrast with this particular publication's editorial structure, the National Commission on Forensic Science has specifically stated that foundational scientific literature should be "published in a journal that utilizes rigorous peer review with independent *external* reviewers to validate the accuracy in its publications and their overall consistency with scientific norms of practice." NCFS Report at 3 (emphasis added). The *AFTE Journal* is thus, in a sense, "comparable to talk within congregations of true believers" rather than an example of "the desired scientific practice of critical review and debate mentioned in *Daubert*." David H. Kaye, *How Daubert and its Progeny Have Failed Criminalistics Evidence and a Few Things the Judiciary Could Do About It*, 86 *Fordham L. Rev.* 1639, 1645 (2018). While the Court does not doubt the good faith of AFTE or those who serve on the editorial board of the *AFTE Journal*, neither can it ignore this intrinsic bias and lack of independence when analyzing the nature of peer review this journal utilizes.⁷ Discussing a similar journal within the field of handwriting analysis, Judge Jed. S. Rakoff of the United States District Court for the Southern District of New York highlighted the issue central to the question of whether publication in the *AFTE Journal* should qualify as peer reviewed publication under *Daubert*: the very meaning of the term "peer." As Judge Rakoff reasoned:

Of course, the key question here is what constitutes a 'peer,' because just as astrologers will attest to the reliability of astrology, defining 'peer' in terms of those who make their living through handwriting analysis would render this *Daubert* factor a charade. While some journals exist to serve the community of those who make their living through forensic document examination,

U.S. v. Tibbs, 2019 WL 4359486 (2019)

numerous courts have found that “[t]he field of handwriting comparison ... suffers from a lack of meaningful peer review” by anyone remotely disinterested.


 *Almeciga v. Ctr. for Investigative Reporting, Inc.*, 185 F. Supp. 3d 401, 420 (S.D.N.Y. 2016) (citation omitted). So, too, with the field of firearms and toolmark analysis: although studies analyzing error rates among practicing firearms and toolmark examiners have, on two occasions, been published in other journals utilizing double-blind peer review presumably performed by disinterested referees, the vast majority of published articles in the field have not undergone peer review by a “competitive, unbiased community of practitioners and academics, as would be expected in the case of a scientific field.” *Id.* (internal quotation marks omitted); see also  *United States v. Starzepyzel*, 880 F. Supp. 1027, 1037-38 (S.D.N.Y. 1995).

Overall, the *AFTE Journals* use of reviewers exclusively from within the field to review articles created for and by other practitioners in the field greatly reduces its value as a scientific publication, especially when considered in conjunction with the general lack of access to the journal for the broader academic and scientific community as well as its use of an open review process. Ultimately, the Court has seen only two meaningfully peer reviewed journal articles regarding the foundational validity of the field, as the vast majority of the studies are published in a journal that uses a flawed and suspect review process. While the implications of these conclusions arise again with respect to the third *Daubert* factor regarding the demonstrated rate of error, this factor on its own does not, despite the sheer number of studies conducted and published, work strongly in favor of admission of firearms and toolmark identification testimony.

C. Does the methodology have a known or potential rate of error?

The parties focused most of their attention on the third *Daubert* factor—“the known or potential rate of error.” And with good reason: determining the error rate for a particular methodology appears essential to determining its ultimate reliability. On this question, the undersigned agrees with one of the essential premises of the 2016 PC AST Report:

***11** Scientific validity and reliability require that a method has been subjected to empirical testing, under conditions appropriate to its intended use, that provides valid estimates of how often the method reaches an incorrect conclusion. For subjective feature-comparison methods, appropriately designed black-box studies are required, in which many examiners render decisions about many independent tests (typically, involving “questioned” samples and one or more “known” samples) and the error rates are determined. Without appropriate estimates of accuracy, an examiner’s statement that two samples are similar - or even indistinguishable - is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact. Nothing - not training, personal experience nor professional practices - can substitute for adequate empirical demonstration of accuracy.

PCAST Report at 46. Likewise, an expert witness’s ability to explain the methodology’s error rate—in other words, to describe the limitations of her conclusion—is essential to the jury’s ability to appropriately weigh the probative value of such testimony. As Judge Rakoff stated in *United States v. Glynn*: “The problem is how to admit [ballistics comparison evidence] into evidence without giving the jury the impression - always a risk where forensic evidence is concerned - that it has greater reliability than its imperfect methodology permits.”  578 F. Supp. 2d at 574.

U.S. v. Tibbs, 2019 WL 4359486 (2019)

Courts considering this issue have rather uniformly weighed this third *Daubert* factor in favor of admissibility. A few courts have characterized the calculation of an error rate for firearms and toolmark pattern matching evidence as an impossible or exceedingly difficult task and acknowledged that an error rate is “presently unknown.” *Johnson*, 2019 U.S. Dist. LEXIS 39590, at *55, 2019 WL 1130258, at *18 (citing *Ashburn*, 88 F. Supp. 3d at 246; *Diaz*, 2007 U.S. Dist. LEXIS 13152, at *27, 2007 WL 485967, at *9); *Romero-Lobato*, 379 F. Supp. 3d at 1119 (quoting *Monteiro*, 407 F. Supp. 2d at 367); *Ashburn*, 88 F. Supp. 3d at 246. The vast majority of courts have nonetheless accepted the notion that existing studies support the conclusion that the discipline's error rate is quite low—between one and two percent. *Romero-Lobato*, 379 F. Supp. 3d at 1119-20; *Johnson*, 2019 U.S. Dist. LEXIS 39590, at *56-57, 2019 WL 1130258, at *18-19; *Johnson*, 2015 U.S. Dist. LEXIS 111921, at *10, 2015 WL 5012949, at *4 (citing *Otero*, 849 F. Supp. 2d at 433-34); *Ashburn*, 88 F. Supp. 3d at 246. Indeed, one court ratified the assertion that the error rate for this discipline is “almost zero.” *Wrensford*, 2014 U.S. Dist. LEXIS 102446, at *56-57, 2014 WL 3715036, at *17.

In spite of the court system's widespread acceptance of the discipline's assertion that it enjoys low error rates, several extensive reports originating from institutions independent of the judiciary have recently taken a different view of the sufficiency of the existing studies in establishing an error rate and in validating the discipline in general. Two National Research Council reports have directly addressed the sufficiency of the published studies purporting to show a low error rate in the field of firearms and toolmark identification. In the first report, the NRC commented:



The validity of the fundamental assumptions of uniqueness and reproducibility of firearms-related toolmarks has not yet been fully demonstrated.... A significant amount of research would be needed to scientifically determine the degree to which firearms-related toolmarks are unique or even to quantitatively characterize the probability of uniqueness.

Nat'l Research Council, *Ballistics Imaging* 3 (2008) [hereinafter *2008 NRC Report*], Similarly, the NRC's second report noted, “[sufficient studies have not been done to understand the reliability and repeatability of the methods.” *2009 NRC Report* at 154. Finally, and most recently, PCAST concluded that most of the studies

*12 involved designs that are not appropriate for assessing the scientific validity or estimating the reliability of the method as practiced. Indeed, comparison of the studies suggests that, because of their design, many frequently cited studies seriously underestimate the false positive rate The scientific criteria for foundational validity require appropriately designed studies by *more than one group* to ensure reproducibility. Because there has been only a single appropriately designed study [the Baldwin/Ames Laboratory study], the current evidence falls short of the scientific criteria for foundational validity. There is thus a need for additional, appropriately designed black-box studies to provide estimates of reliability.

PCAST Report at 111. Together, these reports raise significant questions as to the extent to which courts should rely on certain studies and the low error rates they claim when evaluating this evidence under *Daubert*.

As a general matter, those courts that have found low error rates for this discipline appear to have done so by simply accepting the conclusions of the studies as presented and without any analysis of the methodological or other issues presented in them. *See*,

e.g.,  *Otero*, 849 F. Supp. 2d at 434; *Romero-Lobato*, 379 F. Supp. 3d at 1119-20; *Johnson*, 2019 U.S. Dist LEXIS 39590, at *56-57, 2019 WL 1130258, at *18-19; *Johnson*, 2015 U.S. Dist LEXIS 111921, at *10, 2015 WL 5012949, at *4;  *Ashburn*, 88 F. Supp. 3d at 246.⁸ However, after extensive review of the testimony of the expert witnesses and of the studies about which those experts testified, the undersigned finds it difficult to conclude that the existing studies provide a sufficient basis to accept the low error rates for the discipline that these studies purport to establish. Although the Defendant and the government provided expert testimony and argument on a range of issues presented by these studies, three main problems with the design and interpretation of these studies provide the greatest cause for concern. First, most of the studies suffer from basic, threshold design flaws that undermine the value of their stated results. Second, the reliance of most of these studies on “closed” and/or “set-based” design structures substantially limit the reliability of the error rates claimed in these studies. Third, and perhaps most significantly, the studies permit participants to label toolmark comparisons as “inconclusive” without adequately assessing the impact of such inconclusive determinations on the results of the study as a whole.

1. Most of the studies in the field of firearms and toolmark analysis suffer from basic, threshold design flaws.

*13 Generally, studies published within the area of firearms and toolmark analysis are designed exclusively by toolmark examination professionals who have no experience or training in research methods or decision science. Though these professionals have varying levels of experience within the field of firearms and toolmark analysis, there is no indication that they have experience or training in human subjects research that would facilitate the design of studies that, for example, account for test-taking biases and achieve consistent results by providing specific and uniform procedures for test takers to follow. See *Scurich Test.*, May 14, 2019 (2), 79:20-22, 80:3-10.

Concerns with test-taking biases arise from the notion that a person being tested on her ability to perform a task will, consciously or not, perform differently while being monitored, either guessing the purpose of the test and responding accordingly, *Faigman Test.*, May 16, 2019, 84:23-85:6, or being influenced by a test designer's cues toward one response over another, *Angela Stroman, Empirically Determined Frequency of Error in Cartridge Case Examinations Using a Declared Double-Blind Format*, 46 AFTE J. 157, 157 (2014) [hereinafter 2014 Stroman Study]; see also 2009 NRCReport at 122-24. A test-taker may, consciously or not, try harder or behave more conservatively to avoid being wrong and thus appear to be performing the task better than she would under other circumstances. See 2016 Smith Study at 693 (noting possible “fear of answering incorrectly” when taking a test lacking anonymity). Mr. Weller, having personally participated in research studies in this field, testified that questions regarding test-taking bias need not concern the courts:

I think if you ask a human factor person that is always a concern; the concept of test taking bias; that decisions, there may be a subconscious thing that is going on. So, the test may not be completely reflective of true casework decisions. From my own perspective, I treated the case samples in the same way I would treat casework and I used the same methods and comparison techniques and my own criteria to reach those conclusions. So, I appreciate the concern. I don't know how tangible that concern is and how you rectify that potential problem.

Weller Test., May 14, 2019 (1), 30:20-31:7.⁹ The Court simply cannot accept the conclusion that a recognized bias-related concern should not be a concern at all because a person participating in a study did not himself perceive any impact of that bias. This is, of course, precisely the problem with biases, which have their greatest impact whenever and wherever they operate

U.S. v. Tibbs, 2019 WL 4359486 (2019)

completely unacknowledged. *See 2009 NRC Report* at 124. Based on the evidence adduced at the hearing, it appears that the studies relied upon by the government do not address the potential impact of such biases.

A more concrete study design concern stems from the lack of clarity in these studies as to how the test-takers were expected to perform the work, and the resulting lack of information about what practices and procedures the test-takers actually followed when participating in a study. Many of the studies failed to instruct their participants clearly on whether to follow the testing policies and protocols of their individual laboratories, or to conduct the comparisons in a particular manner in order to ensure uniformity. *See, e.g.*, 2014 Stroman Study at 169 (instructing examiners to follow their “normal” procedures); Mark A. Keisler et al., *Isolated Pairs Research Study*, 50 AFTE J. 56, 58 (2018) [hereinafter 2018 Keisler Study] (instructing examiners to complete the research study like they would casework, but noting it was “unclear if participants ... deviated from laboratory policy”); 2016 Smith Study at 698 (failing to instruct examiners but noting factors “such as a laboratory’s quality assurance program (which includes verifications and peer review), would influence error rates in casework”). This inconsistency poses a significant interpretive problem because different labs have different policies for how to conduct toolmark examinations. *Scurich Test.*, May 15, 2019, 53:12-19; *Faigman Test.*, May 16, 2019, 85:24-86:6. For example, some lab policies require a second examiner to verify a first examiner’s work while others do not; similarly, some labs have policies that prohibit rendering a conclusion of “exclusion” when class characteristics are all in common, while others do not have such a policy. *See, e.g.*, 2018 Keisler Study at 58. In other words, in many of the studies that the government and its experts rely on, it is unknown whether one or more of the test participants had a colleague verify his or her work, and whether reported “inconclusives” were only deemed inconclusive due to adherence with a policy demanding such a result rather than on an actual analysis of the patterns on a particular bullet or casing.¹⁰ These design issues prevent the Court from evaluating whether the test-takers in these studies *were even taking the same test*—as it cannot be determined what instructions each examiner followed in completing the comparisons—and thus reduce the ability of these studies to support the foundational validity of the field.

***14** Yet another study design issue relates to the manner in which the test administrators selected practicing examiners to participate in the studies. *Scurich Test.*, May 14, 2019 (2), 93:9-20, 93:22-94:1. Some studies provided no information regarding how their participants were selected and recruited, *see, e.g.*, 2018 Keisler Study, but those studies that did indicated that they had solicited volunteer participation from AFTE membership lists or from groups of employees in specific crime laboratories: one study, for example, used only examiners employed by a Federal Bureau of Investigation laboratory, Charles DeFrance and Michael D. Van Arsdale, *Validation Study of Electrochemical Rifling*, 35 AFTE J. 35, 36 (2003) [hereinafter 2003 DeFrance Study]; another engaged a third party to solicit volunteers from laboratories, 2016 Smith Study at 693; and two others recruited volunteers via email, using a list of AFTE members, Thomas G. Fadul, Jr., et al., *An Empirical Study to Improve the Scientific Foundation of Forensic Firearm and Tool Mark Identification Utilizing 10 Consecutively Manufactured Slides*, 45 AFTE J. 376, 379 (2013) [hereinafter 2013 Fadul Study]; Thomas G. Fadul, Jr., et al., *An Empirical Study to Improve the Scientific Foundation of Forensic Firearm and Tool Mark Identification Utilizing Consecutively Manufactured Glock EBIS Barrels with the Same EBIS Pattern*, Final Report on Award Number 2010-DN-BX-K269, 16 (2013) [hereinafter Miami- Dade Study]. Other studies simply report that they used volunteers from laboratories or AFTE membership lists without clarifying further as to how the participants were recruited. David P. Baldwin et al., *A Study of False-Positive and False-Negative Error Rates in Cartridge Case Comparisons*, 7 (2014), <https://www.ncirs.gov/pdf/files1/nii/249874.pdf> [hereinafter Ames Laboratory Study]; David J. Brundage, *The Identification of Consecutively Rifled Gun Barrels*, 30 AFTE J. 438, 440, 442 (1998) [hereinafter 1998 Brundage Study]; 2014 Stroman Study at 168. Still, others do not specifically describe their pool of participants, let alone how those participants were solicited to take part in the study. *See 2019 Hamby Study*; 2018 Keisler Study; Dennis J. Lyons, *The Identification of Consecutively Manufactured Extractors*, 41 AFTE J. 246 (2009). In spite of this vagueness in some of these articles, these studies generally appear to use a self-selected set of volunteers. While simply soliciting volunteers is obviously the easiest way to perform these experiments, use of volunteers for what amounts to a proficiency examination does not provide the clearest indication of the accuracy of the conclusions that would be reached by average toolmark examiners. *Scurich Test.*, May 14, 2019 (2), 93:19-20.

These design issues do not necessarily invalidate the results of these studies, and *Daubert* does not necessarily require the proponent of a theory or methodology to present only studies with the best possible design. Undoubtedly, experts with extensive training in research methods could likely find fault with the methodology of any study. But these threshold design issues—perhaps the result of their designers not securing the assistance of individuals with design science expertise—surely impact the validity of these studies' conclusions and limit their utility to some extent.

2. Because of their reliance on “closed” and “set-based” designs, the studies in the field of firearms and toolmark analysis do not provide reliable data regarding the ability of an examiner to match unknown and known samples.

In general, the firearms and toolmark identification field has produced two types of comparison studies—those that are referred to as “open” and “independent comparison” studies (also called “pairwise comparison” studies), and those that are referred to as “closed” and “set-based” studies. See PCAST Report at 106-10. In the “open” and “independent comparison” studies, participants are given an unknown sample and asked to determine whether it matches another specific sample. *Id.* at 110. Such a study may involve a series of separate comparisons, but each comparison presents as a separate problem. See *id.* Most importantly, not all of the unknown samples will have a matching known sample, so the participant will not have reason to know whether the correct match is present. See *id.* Based on the testimony at the hearing and the materials submitted by the parties, it appears that only two studies have been conducted using this approach: the 2014 Ames Laboratory study and the 2018 Keisler study. In the Ames Laboratory study, participants were given a test kit consisting of fifteen separate problem sets for comparison. Ames Laboratory Study at 10. Each set contained three cartridge casings designated as being from the same “known” firearm and one cartridge casing designated as the “unknown” or “questioned” sample; unknown to the participants, each test kit contained five same-source pairs and ten different-source pairs. *Id.* Participants were asked to approach each of the fifteen problems separately and to render a conclusion, and they were not told whether any of the questioned samples would match the known samples. *Id.* Similarly, the Keisler study provided participants with a test kit made up of twenty sets of two cartridge casings each, and unknown to the participants, each test kit contained twelve same-source pairs and eight different-source pairs. 2018 Keisler Study at 56. Participants were asked to examine each pair separately from any other pair and to render a conclusion as to each pair. *Id.*

*15 By contrast, virtually all studies published in this field utilize a “closed” universe, where a match is always present for each unknown sample, and a “set-based” design, where comparisons are made within a set of samples. See PCAST Report at 106. This methodology differs from the “open” and “independent comparison” studies because the comparisons are not divided up into individual problems for the participant to consider one at a time; instead, participants are either given a group of samples and asked to compare all of those samples to each other and to find matches, or participants are given a group of known samples and a group of unknown samples and asked to make comparisons between the two groups to find matches. See *id.* at 106-08. For example, the 2019 Hamby Study, using the same design and test kits as the 1998 Brundage Study and published incorporating all data from several iterations of Brundage's original study over the last twenty-one years, provided participants with fifteen questioned samples and ten pairs of known samples and asked the participants to make comparisons. 2019 Hamby Study at 556; 1998 Brundage Study at 440. Similarly, the two Fadul studies gave participants a quantity of questioned samples and a number of known samples and asked them to make comparisons between the two groups. 2013 Fadul Study at 380; Miami-Dade Study at 19. These studies, and others like them, often involved the use of an answer sheet to allow the participant to indicate the known sample to which an unknown sample could be matched. See, e.g., Miami-Dade Study at 19.

During the hearing, counsel and witnesses debated the question of whether one of the study types better mimics casework. The PCAST report concluded that the “closed” and “set-based” studies did not replicate casework. PCAST Report at 106. The government expert witnesses, Mr. Weller and Dr. Petraco, disagreed with this contention. Weller Test., May 13, 2019, 126:21-127:19; Petraco Test., May 13, 2019, 71:15-21, 71:24-72:5. While the Court presently lacks sufficient information

U.S. v. Tibbs, 2019 WL 4359486 (2019)

to resolve this empirical question, its answer would not provide much guidance for the *Daubert* question at issue here. As Dr. Scurich stated, the question of whether a study mimics real-world casework differs from the question of whether a study accurately measures the ability of examiners to make source determinations based on pattern matching. *See Scurich Test.*, May 15, 2019, 77:20-24.

Having reviewed the studies and considered both parties' arguments on the different study designs, the undersigned finds that the independent comparison studies, or "pairwise" studies, best test the validity of the assumptions underlying the firearms and toolmark analysis field and that the closed, set-based studies have inherent limitations that preclude them from providing substantial validation. This conclusion mirrors that of PCAST, which explained:

Specifically, many of the studies employ 'set-based' analyses, in which examiners are asked to perform all pairwise comparisons within or between small samples sets.... The study design has a serious flaw, however: the comparisons are not *independent* of one another. Rather, they entail internal dependencies that (1) constrain and thereby inform examiners' answers and (2) in some cases, allow examiners to make inferences about the study design.... Because of the complex dependencies among the answers, set-based studies are not appropriately- designed black-box studies from which one can obtain proper estimates of accuracy. Moreover, analysis of the empirical results from at least some set-based studies ('closed-set' designs) suggest that they may substantially underestimate the false positive rate.

PCAST Report at 106. Of course, the PCAST report is hardly beyond critique, and the government's experts stated many valid criticisms of it throughout the hearing: the Council did not include anyone from the firearms and toolmark examination community, *Id.* at v-ix; it criticized studies for lack of peer review but was not itself peer reviewed, *Petraco Test.*, May 13, 2019, 34:20-24; and the report apparently miscounted or omitted data from several studies, *Weller Test.*, May 13, 2019, 108:10-109:8. Despite these shortcomings, the Court finds the conclusions of PCAST (as echoed by Dr. Scurich at hearing) about the very limited utility of closed-set studies to have been essentially correct.

Closed, set-based studies have two significant problems that make them difficult to rely upon as evidence of the reliability of conclusions regarding toolmark evidence. First, a set-based study involves an unknown number of total comparisons that a participant makes in the process of matching samples to each other, which means that such a study cannot calculate a true error rate based on the total comparisons made. In other words, the total number of comparisons made remains unknown at the conclusion of the study because it is not known whether a participating examiner compared a particular unknown sample to only one other sample, or to a few of the other samples, or to all of the other samples before making a conclusion regarding that sample. One of the government's expert witnesses acknowledged this issue in his testimony and agreed that in closed, set-based studies, it is not possible to know the total number of true different source comparisons performed and that a false positive error rate thus cannot be calculated. *Weller Test.*, May 14, 2019 (2), 22:17-23.


*16 Second, and perhaps more importantly, the participants in a closed, set-based study can see all of the questioned samples and all of the known samples at once and can thus employ inferences gained from looking at one of the individual problems in order to solve other individual problems. In independent comparison studies, the examiner simply makes a one-to-one comparison, an exercise well-suited to gauge her ability to look at two items and, based only on the features of those two items, make a determination of match. PCAST likened closed, set-based studies, by contrast, to a Sudoku puzzle, "where initial answers can be used to help fill in subsequent answers." PCAST Report at 106. This puzzle analogy, which Dr. Scurich also employed to explain this pitfall of closed, set-based studies, identifies a substantial problem with the closed and set-based study design. Such a design allows participants to rely on their own decisions and inferences about some of the samples to make

U.S. v. Tibbs, 2019 WL 4359486 (2019)

decisions regarding the remaining samples, which the defense aptly characterized as the “interdependency problem.” Tr. June 10, 2019, 20:20. In other words, the participant can rely on other, unrelated parts of the puzzle—or even the puzzle as a whole—to solve an individual part of the puzzle, and thus a match determination for each of the individual problems evaluated would depend not simply on one-to-one comparisons but also on information and inferences gleaned from other individual problems (or from the set as a whole). Such a study design does not provide a reliable measure of the ability of firearms and toolmark examiners to make comparisons between known and unknown samples where such inferences are not available to be drawn.

Because of these significant limitations of the closed and set-based studies, the vast majority of studies that the field relies upon to establish its foundational validity simply do not provide an adequate basis to do so. Unfortunately, the only studies with the more appropriate design for assessing reliability—the Ames Laboratory study and the Keisler study—have not, as described *supra*, undergone meaningful, independent peer review prior to publication.¹¹

3. The large number of “inconclusive” results, and the studies’ failure to address them, undermines the reliability of the studies’ claimed error rates.


The final, and perhaps most substantial, issue related to the studies proffered to support the reliability of firearms and toolmark analysis relates to how the studies address—or fail to address—the “inconclusive” answers (hereinafter “inconclusives”) frequently given by the examiners participating in these studies, and how such answers affect the error rate. In field work, examiners analyzing bullets and cartridge casings recovered from a crime scene and comparing them to test fired samples from a recovered firearm can reach three possible conclusions: they can conclude that the samples match, and thus make an “identification”; they can conclude the samples do not match, and thus make an “elimination”; or they can characterize the comparison as “inconclusive.” Inconclusive appears to be a reasonable and acceptable conclusion in casework, possibly because the firearm may not have left sufficient marks for comparison, *see* Weller Test., May 13, 2019, 117:15-19, or because environmental factors may change or distort the soft metal of a cartridge casing or bullet. As Judge Rakoff described, “[t]he bullets and/or shell casings recovered from the crime scene may be damaged, fragmented, crushed or otherwise distorted in ways that create new markings or distort existing ones.”  *Glynn*, 578 F. Supp. 2d at 573.

Nevertheless, the methods used in the proffered laboratory studies make a compelling case that inconclusive should not be accepted as a correct answer in these studies. First and foremost, the study designers make efforts to control the effects of the environment on the samples. Rather than being fired such that the casings or bullets could roll, hit walls or cars, or be stepped on or exposed to the weather, these studies use samples collected under test fire conditions. In the Ames Laboratory study, for example, all of the test fired casings were collected in a brass catcher, and any that fell out of the catcher and hit the floor were discarded. Ames Laboratory Study at 12.

Additionally, most of the studies involved some quality assurance mechanism to ensure that the samples to be examined by the participants had sufficient markings for comparison purposes before the test kits were supplied to the examiners. For example, one study involved several test fires to account for a so-called “break-in period” to ensure that the newly- manufactured firearms were producing consistent markings, and the study designers checked the samples to ensure that the markings were then consistent. 2003 DeFrance Study at 35.¹² In the two Fadul studies, study designers personally inspected every tenth test set to ensure that the samples had sufficient markings for comparison purposes. 2013 Fadul Study at 382; Miami- Dade Study at 19. Another study involved a “pre-test” that was conducted to review the test sets before they were delivered to participants. 2009 Lyons Study at 250-51. The 2018 Keisler Study, at 57, and the 2016 Smith Study, at 694, also noted that the samples used for comparison had been deemed determinable.

U.S. v. Tibbs, 2019 WL 4359486 (2019)

*17 The government and its expert witnesses view the number of inconclusives given by examiners in these published studies as irrelevant to the ultimate issue before the Court. Based on the premise that declaring a comparison inconclusive has no probative value, the government argues that such an opinion would not be given in court, and thus need not be a factor in assessing the reliability of pattern matching within the field of firearms and toolmark analysis.

In other words, the government and its experts contend that only identifications—i.e., “match conclusions”—and a false positive error rate calculated based upon identifications combine to establish reliability. PCAST addressed inconclusives in this manner—by removing them entirely from analysis of the studies and their data, PCAST Report at 153—as did the United States District Court for the District of New Jersey in the only published opinion addressing this aspect of the studies, see  *Otero*, 849 F. Supp. 2d at 434.¹³

However, in laboratory testing situations, in which samples were collected using procedures to minimize environmental alterations and in which samples were checked by test administrators to ensure they contained sufficient marks suitable for comparison purposes, a conclusion by an examiner characterizing the comparison as inconclusive should not qualify as a correct answer. Dr. Scurich opines, based on principles of mathematics and statistics in particular, that such responses should be viewed as false positive errors (i.e., included among false identifications),¹⁴ but such a characterization fails to make logical sense: while under laboratory conditions such inconclusives are surely some type of error, it does not follow that inconclusives are functionally the same as a false conclusion by an examiner who attributes a cartridge casing to a gun that did not fire it. While the Court does not accept Dr. Scurich's inclusion of inconclusives in the false positive error rate, it agrees with his essential premise that such responses should represent an error by the examiner. Under these controlled circumstances, an examiner who looks at a casing collected in a laboratory test fire and that has been examined by a test administrator to make sure it has markings suitable for comparison, and who nonetheless describes her comparison as yielding inconclusive results, *is making an error of some kind*. In these published studies, at the very least, the test taker giving an answer of inconclusive may simply be avoiding the most difficult problem on the test. Or it may be that the examiner's failure to identify or exclude the sample constitutes a mistake in her analysis. Alternatively, there may be some ambiguity, discussed at length in the Ames Laboratory study, regarding why some examiners make a determination of inconclusive, and whether some of those determinations are the result of laboratory policies against declaring exclusions when class characteristics are the same. Ames Laboratory Study at 18-19.

*18 Based on the studies and the testimony of the government's expert witnesses, no adequate explanation has been offered regarding the reason for examiners returning inconclusives in these controlled circumstances. The government's experts insist that inconclusives should not be treated as any kind of error because inconclusive is not a conclusion at all. See Petraco Report at 3. Nevertheless, and again under these controlled circumstances, an inconclusive response *is* a conclusion, even if it is only a conclusion against making any other conclusion. In a recent article, Itiel Dror asserts that inconclusive determinations may be the result of “over-reliance” by forensic examiners on the option of “decid[ing] not to decide.” Itiel E. Dror & Glenn Langenburg, “*Cannot Decide*”: *The Fine Line Between Appropriate Inconclusive Determinations Versus Unjustifiably Deciding Not to Decide*, 64 J. Forensic Sci. 10, 11 (2019). Where there is sufficient information for concluding “identification” or “exclusion,” “[a]n inconclusive determination is an erroneous decision because the evidence does not support that decision.” *Id.* at 13. In the end, all that is known is that some examiners in these studies, taking these tests involving samples collected under carefully controlled circumstances, responded that the comparison was “inconclusive.”¹⁵


Viewing these inconclusives as an error of some type greatly affects the overall error rates produced by the studies. Focusing on the only two “open” studies, the Ames Laboratory study calculated a false positive error rate of 1.01%, while the Keisler study reported a false positive error rate of 0%. If the inconclusives are considered as errors, however, the Ames Laboratory study's error rate among different source comparisons soars to 34.76% while the Keisler study's error rate rises to 20.14%. Again, Dr. Scurich's approach of treating inconclusives as false positives does not appropriately address the issue presented

U.S. v. Tibbs, 2019 WL 4359486 (2019)


by inconclusives, but the large number of the inconclusives reported in the studies greatly reduces their persuasive force in establishing the ability of a firearms and toolmark examiner to make accurate source determinations. Indeed, even Dr. Petraco acknowledged that the number of inconclusives increased uncertainty about calculations of the error rate, Petraco Test., May 13, 2019, 26:10-12, while Mr. Weller testified that the questions surrounding inconclusives and the error rate calculation were, in his words, “not well studied,” Weller Test., May 14, 2019 (1), 53:25-54:1.

4. Conclusion

Based on the basic design of the studies, the prevalence of closed-set studies, and the uncertain relationship in the open studies between inconclusives and the ultimate error rates, the undersigned was unable to conclude that the field has established a known or potential error rate with regard to the ability of a firearms and toolmark examiner to make a source determination. Dr. Petraco testified, and the government repeated several times in argument, that no studies refute the proposition that “firearms examiners can identify bullets or fired cartridge casings to particular guns with a high degree of accuracy.” Petraco Test., May 13, 2019, 12:24-13:4. This formulation of the issue turns both the scientific method and the *Daubert* burden of proof on their heads: instead, the question before the Court turns on whether the government can establish the foundational validity of the discipline, not whether the opposing party can prove a negative.

With regard to the proffered discipline, most of the studies on which the government relies involved closed-set designs that cannot provide an accurate accounting of the error rate. While the two studies that employ an open, independent comparison design could yield an accurate error rate measurement, neither was subjected to meaningful peer review, and both were plagued by a large number of “inconclusive” responses. Under such circumstances, the Court cannot conclude that the government has established that this forensic discipline has established a “known or potential rate of error.” See *Motorola*, 147 A.3d at 754 (citing  *Daubert*, 509 U.S. at 593-94). While other studies being conducted now or in the future may change this conclusion, the Court finds that this factor currently weighs against the admission of source attribution statements made by a firearms and toolmark examiner.

D. Is there a standard controlling the technique's operation?

*19 The fourth *Daubert* factor requires an inquiry into “the existence and maintenance of standards controlling the technique's operation.” *Motorola*, 147 A.3d at 754 (quoting  *Daubert*, 509 U.S. at 593-94). As described *supra*, the operative standard for firearms and toolmark identification is known as the “AFTE theory of identification,” which states that the examiner can make a conclusion of common origin when microscopic surface contours of the toolmarks are in “sufficient agreement.” PCAST Report at 59-60 (citing Ass'n of Firearm and Tool Mark Examiners, *Theory of Identification as it Relates to Tool Marks: Revised* 43 AFTE J. 287 (2011)). Stated in full, the AFTE Theory of Identification reads as follows:

1. The theory of identification as it pertains to the comparison of toolmarks enables opinions of common origin to be made when the unique surface of two toolmarks are in “sufficient agreement.”





2. This “sufficient agreement” is related to the significant duplication of random toolmarks as evidenced by the correspondence of a pattern or combination of patterns of surface contours. Significance is determined by the comparative examination of two or more sets of surface contour patterns comprised of individual peaks, ridges and furrows. Specifically, the relative height or depth width, curvature and spatial relationship of the individual peaks, ridges and furrows within one set of surface contours are defined and compared to the corresponding features in the second set of surface contours. Agreement is significant when the agreement in individual characteristics exceeds the best agreement demonstrated between toolmarks known to have been produced by different tools and is consistent with agreement demonstrated by toolmarks known to have been produced by the same tool. The statement that “sufficient agreement” exists between two toolmarks means the agreement of individual

U.S. v. Tibbs, 2019 WL 4359486 (2019)

characteristics is of a quantity and quality that the likelihood another tool could have made the mark is so remote as to be considered a practical impossibility.

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

Id.

As other courts have noted, and as the Defendant argues here, one of the primary challenges to firearms and toolmark identification stems from the methodology's lack of objective criteria for examiners to use in determining a “match.” *See, e.g., Romero-Lobato*, 379 F. Supp. 3d at 1120. Courts that have admitted firearms and toolmark identification testimony in the face of a *Daubert* challenge have found the standard articulated in the AFTE theory of identification sufficient. *See Johnson*, 2019 U.S. Dist. LEXIS 39590, at *51, 2019 WL 1130258, at *17; *Johnson*, 2015 U.S. Dist. LEXIS 111921, at *10-11, 2015 WL 5012949, at *4;  *Ashburn*, 88 F. Supp. 3d at 247; *Wrensford*, 2014 U.S. Dist. LEXIS 102446, at *54-55, 2014 WL 3715036, at *16. However, the AFTE theory of identification has been sharply criticized by a number of other courts as “inherently vague,”  *Glynn*, 578 F. Supp. 2d at 572; “inherently subjective,” *Romero-Lobato*, 379 F. Supp. 3d at 1121; and “either tautological or wholly subjective,”  *Green*, 405 F. Supp. 2d at 114. As one United States District Court Judge noted, “the AFTE Theory appears to be more of a description of the process of firearm identification rather than a strictly followed charter for the field.”  *Monteiro*, 407 F. Supp. 2d at 371.

Both the NRC and PCAST lodged similar criticisms. The NRC focused its critique on this lack of an objective comparison standard:


*20 AFTE has adopted a theory of identification, but it does not provide a specific protocol The meaning of “exceeds the best agreement” and “consistent with” are not specified, and the examiner is expected to draw on his or her own experience. This AFTE document, which is the best guidance available for the field of toolmark identification, does not even consider, let alone address, questions regarding variability, reliability, repeatability, or the number of correlations needed to achieve a given degree of confidence.

2009 NRC Report at 155. Calling this a “fundamental problem with toolmark and firearm analysis,” *Id.*, the NRC further stated, “even with more training and experience using newer techniques, the decision of the toolmark examiner remains a subjective decision based on unarticulated standards and no statistical foundation for estimation of error rates.” *Id.* at 153-54. And, more recently, PCAST criticized the AFTE standard as “circular.” PCAST Report at 60.

In this case, the evidence supports—and the undersigned agrees with—all of these assessments of the AFTE theory of identification. By its own terms, it is a fundamentally subjective standard that can only be characterized as entirely tautological: an opinion of common origin can be rendered when the surfaces of the two examined items are in “sufficient agreement,” which exists not when some objective measure is satisfied, but when the examiner determines, based on her training and experience, that it would be a “practical impossibility” for the two items not to share a common origin. In other words, this protocol permits the ultimate finding of “sufficient agreement” whenever an individual examiner concludes that she would be hard pressed (for reasons not specified in the governing standard) to find such similar markings on casings or bullets fired by different firearms. Although AFTE has attempted to use terms like “sufficient agreement” to resemble terminology that one would find in an

U.S. v. Tibbs, 2019 WL 4359486 (2019)


objective or scientific standard, in the end it simply leaves the determination of common origin to the standardless, undefined judgment of an individual examiner. Therefore, under this so-called standard, the process for determining what constitutes a “match” lacks defined criteria; it is merely unconstrained subjectivity masquerading as objectivity.

Courts that have admitted this type of expert opinion testimony have responded to such criticisms about the standard's subjective nature by correctly noting that “[t]he mere fact that an expert's opinion is derived from subjective methodology does not render it unreliable.” *Romero- Lobato*, 379 F. Supp. 3d at 1120 (citing  *Ashburn*, 88 F. Supp. 3d at 246-47; *Cohen v. Trump*, Case No.: 3:13-cv-2519-GPC-WVG, 2016 U.S. Dist. LEXIS 117059, at *35, 2016 WL 4543481, at *11 (S.D. Cal. Aug. 29, 2016)). Even the *Romero-Lobato* court, which found the lack of objective criteria to weigh against admissibility, explained:

[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. Such requirements would, in most circumstances, exclude psychologists, physicians, and lawyers from testifying as expert witnesses. Of course, a litigant would be hard pressed to make a good faith argument that the methods used by mainstream medical and legal experts are unreliable under *Daubert*.

379 F. Supp. 3d at 1120.

Of course, expert witnesses in many fields testify to subjective opinions. For example, an assessor testifying about home values would provide a subjective opinion about the value of a particular piece of property, but that assessor would be able to describe the basis of her opinion in objective terms, premised on a comparison with other properties that are similar in certain defined ways (such as the number of bedrooms, total square footage, or specific location), or on a general change in home values in a particular neighborhood since the last time the house was sold. Such an opinion would ultimately be subjective in nature, but it would be grounded in objective criteria, the applicability of which can be analyzed, debated, and critiqued, and not simply on the assessor's judgment, based on her experience, as to what the property is worth. Similarly, an expert in a medical malpractice case testifying about whether a doctor satisfied a particular standard of care would base her subjective opinion on objective criteria in the form of promulgated and practiced nationwide standards of care within that medical specialty and not in her personal opinion, based on her own training and experience, as to what that standard should be.

*21 The AFTE theory of identification is more subjective than such other examples of subjective opinions. “[B]allistics comparison lacks defining standards to a degree that exceeds most other kinds of forensic expertise.”  *Glynn*, 578 F. Supp. 2d at 574. Unlike the standards underlying opinions in other fields, the AFTE theory provides no objective yardstick to support or explicate the expert's opinion; instead, the expert is left to rely on her own thoughts and conclusions based only on the vagaries of her own training and experience. An opinion that “the agreement in individual characteristics exceeds the best agreement demonstrated between toolmarks known to have been produced by different tools” and “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” relies entirely on subjective judgment, without any underlying objective criteria that the examiner must reference or apply. For all of these reasons, this fourth *Daubert* factor strongly militates against the admission of expert witness testimony in the field of firearms and toolmark analysis.

E. To what degree is the technique accepted within the scientific community?

The final enumerated *Daubert* factor—the “degree of acceptance within [a relevant] scientific community”—incorporates, at least to some extent, the *Frye/Dyas* principles that the general acceptance of theories speaks to their validity. See [Daubert](#), 509 U.S. at 594; see also [Motorola](#), 147 A.3d at 754. As stated in *Daubert*, “[widespread acceptance can be an important factor in ruling particular evidence admissible, and a known technique which has been able to attract only minimal support within the community may properly be viewed with skepticism.” 509 U.S. at 594 (citation and internal quotation marks omitted). Every published opinion evaluating the admissibility of firearms and toolmark evidence has found that the AFTE method enjoys general acceptance in the relevant community and that such acceptance weighs in favor of admissibility. See, e.g., [Romero-Lobato](#), 379 F. Supp. 3d at 1122; [Johnson](#), 2019 U.S. Dist. LEXIS 39590, at *58, 2019 WL 1130258, at *19; [Johnson](#), 2015 U.S. Dist. LEXIS 111921, at *11, 2015 WL 5012949, at *4; [Ashburn](#), 88 F. Supp. 3d at 247; [Wrensford](#), 2014 U.S. Dist. LEXIS 102446, at *45-46, 2014 WL 3715036, at *14; [Taylor](#), 663 F. Supp. 2d at 1178. However, these precedents have generally limited the scope of the so-called “relevant community” to the specific community of firearms and toolmark examiners, or to those generally operating within the field of criminal forensics. See [Romero-Lobato](#), 379 F. Supp. 3d at 1122; [Johnson](#), 2019 U.S. Dist. LEXIS 39590, at *58, 2019 WL 1130258, at *19; [Johnson](#), 2015 U.S. Dist. LEXIS 111921, at *11, 2015 WL 5012949, at *4; [Ashburn](#), 88 F. Supp. 3d at 247; [Otero](#), 849 F. Supp. 2d at 435.

In the undersigned's view, if *Daubert*, *Motorola*, and [Rule 702](#) are to have any meaning at all, courts must not confine the relevant scientific community to the specific group of practitioners dedicated to the validity of the theory—in other words, to those whose professional standing and financial livelihoods depend on the challenged discipline. As Judge Jon M. Alander of the Superior Court of Connecticut aptly stated, “[i]t is self evident that practitioners accept the validity of the method as they are the ones using it. Were the relevant scientific community limited to practitioners, every scientific methodology would be deemed to have gained general acceptance.” [Terrell](#), 2019 Conn. Super. LEXIS 827, at *14, 2019 WL 2093108, at *4. Indeed, in other forensic science fields, techniques and methods that had gained “general acceptance” among practitioners have been deemed unreliable and have been excluded as a result of *Daubert* challenges. See, e.g., [United States v. Saelee](#), 162 F. Supp. 2d 1097, 1101-05 (D. Alaska 2001) (forensic handwriting analysis).

Here, the government failed to show general acceptance outside of the field of firearms and toolmark practitioners of the theory that an examiner can microscopically analyze individual toolmarks on a cartridge casing or bullet and reach a reliable conclusion that a particular firearm fired that particular cartridge casing or bullet. The conclusions of the NRC and PCAST reports indicate that the wider academic and scientific community does not necessarily generally accept this theory. With the majority of studies published by and for the review of professional firearms and toolmark examiners, there is currently insufficient evidence that this methodology is generally accepted as proven, established, or validated—a factor that weighs against admissibility.

F. A balancing of these factors requires that the expert be constrained to testify only that the recovered firearm cannot be excluded as the source of the recovered casing.


*22 In weighing and applying these factors pursuant to *Daubert*, *Motorola*, and [Rule 702](#), the Court found that—particularly in light of the inability of the published studies to establish an error rate, the absence of an objective standard for identification, and the lack of general acceptance of the foundational validity of the field outside of the community of practitioners within the field—reliable principles and methods do not adequately support the theory that a firearms examiner can identify a particular firearm as having fired a particular bullet or cartridge casing. Accordingly, the Court will not permit Mr. Coleman, the firearms examiner who conducted the comparison in the above-captioned case, to testify in the form of such a source attribution statement. Again, in light of the state of the evidence presented here, a conclusion that a particular firearm was the source of a particular bullet or cartridge casing does not yet find support in sufficiently reliable principles and methods.

U.S. v. Tibbs, 2019 WL 4359486 (2019)

Such a conclusion, however, does not require the exclusion of all specialized opinion testimony in the area of firearms and toolmark examination, nor does it equate to a finding that the entire discipline lacks foundational reliability. As such, the Court denied Defendant's request to exclude Mr. Coleman's testimony in its entirety. The defense has not challenged the general theory that tools used to create firearms leave accidental or incidental toolmarks on the firearms, and that those toolmarks leave impressions that can be discerned on the contours of the bullets and cartridge casings discharged through the firearm; based on the evidence before it, the Court found that reliable principles support this theory, at least at that stated level of generality. Nor did the defense challenge the reliability of the basic method used by Mr. Coleman and other firearms examiners, i.e., the use of a comparison microscope to observe these marks on bullets and cartridge casings. In addition, reliable principles permit a conclusion that a firearm cannot be excluded as the source of a recovered casing or bullet; indeed, this limited conclusion is supported by the reliable principle that firearms leave toolmark impressions on discharged cartridge casings and the reliable method of viewing those impressions under a comparison microscope. As the defense acknowledges, such a conclusion does not imply a particular statistical weight, and furthermore, it does not stray into territory unsupported by reliable principles and methods, such as a conclusion that a firearm "matches" or was the source of a particular casing.

Accordingly, the Court ruled that the government's proffered expert, Mr. Coleman, may testify and give general specialized opinion testimony in this case. Mr. Coleman may describe the work he performed and the comparisons he made; he may describe the basis of his conclusion regarding the physical consistency of the toolmarks that he observed; and he may make, as the Defendant concedes, a comparison of the samples based on class characteristics. In sum, Mr. Coleman may conclude that based on his examination and the consistency of the class characteristics and microscopic toolmarks, the recovered firearm cannot be excluded as the source of the cartridge casing found on the scene of the alleged shooting—in other words, that the firearm *may* have fired the recovered casing. Mr. Coleman may not state an ultimate conclusion in stronger terms. Similarly, Mr. Coleman will be precluded at any point in his testimony from stating that individual marks are unique to a particular firearm or that observed individual characteristics can be used to "match" a firearm to a piece of ballistics evidence.

In fashioning this ruling, the Court found that the government's alternative proposals for expressing Mr. Coleman's opinion did not adequately address the concerns raised by the *Daubert* factors. The government's proffer that Mr. Coleman could testify that, based on his training and experience, he believes that the recovered cartridge casing was fired from the recovered gun, represents no improvement over a simply-stated opinion that a recovered casing was fired from a particular gun, even if Mr. Coleman also expressed his opinion with the limitations on certainty statements imposed by the Court of Appeals. In this alternative, the expert would be characterizing his opinion as his own personal opinion—as any expert must—but would still be making a source attribution statement not sufficiently supported by reliable principles and methods.

*23 Similarly, the Court strongly disagrees with the government that cross-examination could cure any reliability issues created by a source attribution statement. Of course, the *Daubert* decision recognized, and other courts have noted, that "[v]igorous cross-examination, presentation of contrary evidence, and careful instruction on the burden of proof are the traditional and appropriate means of attacking shaky but admissible evidence."  509 U.S. at 596; see also *Motorola Inc.*, 147 A.3d at 754.¹⁶ While cross-examination may often play such a role, this discipline and the disputes surrounding it seem far too complex for a series of questions on cross-examination to allow a full understanding of the limitations of the field. Indeed, a full exploration of the issues surrounding the reliability of this evidence in the present case required several days of testimony from multiple expert witnesses, close evaluation of numerous applied-science studies, exploration into the studies' design and methodology and the problems arising therefrom, and advocacy by counsel on each side specially tasked with litigating forensic science issues. It would be fanciful to conclude that the normal adversarial process would enable a lay jury to adequately understand these issues, and it is similarly unrealistic to conclude that the average attorney in the average trial would be able to raise these issues in front of the jury in this fashion, particularly when this issue would be one among many issues to be presented to the jury in a trial. Ultimately, Judge Rakoff's characterization in *Glynn* captures the essence of this issue:

U.S. v. Tibbs, 2019 WL 4359486 (2019)



[O]nce expert testimony is admitted into evidence, juries are required to evaluate the expert's testimony and decide what weight to accord it, but are necessarily handicapped in doing so by their own lack of expertise. There is therefore [sic] a special need in such circumstances for the Court, if it admits such testimony at all, to limit the degree of confidence which the expert is reasonably permitted to espouse.

 578 F. Supp. 2d at 571.

For all of these reasons, the government's expert may testify that based on his examination, the recovered firearm cannot be excluded as the source of the cartridge casing found on the scene of the alleged shooting. This formulation of the expert's opinion is limited to the principles and methodologies which the evidence supports as sufficiently reliable. Any statements by the expert involving more certainty regarding the relationship between a casing and a firearm would stray into territory not presently supported by reliable principles and methodology.

IV. COHERENCE WITH RECENT DISTRICT OF COLUMBIA COURT OF APPEALS PRECEDENTS

The Court of Appeals issued its opinion in *Williams II* after the *Daubert* hearing was held in this case. Upon request of the Court, both parties filed additional pleadings to address what, if any, effect *Williams II* should have on the Court's present determination.

After his conviction for first-degree felony murder while armed and other related offenses, Marlon Williams appealed his convictions—arguing, *inter alia*, that the trial court should not have permitted the government's firearms examiner to testify, based on patternmatching, that the gun recovered from Mr. Williams's apartment was the murder weapon. *Williams II*, 210 A.3d at 737 (citing  *Williams v. United States (Williams I)*, 130 A.3d 343, 345, 347 (D.C. 2016)). At trial, the examiner testified that he microscopically examined the markings on three bullets recovered from the decedent's vehicle and that they matched the markings on the bullets test fired from the gun recovered from Mr. Williams's apartment. *WilliamsII*, 210 A.3d at 738. The expert further opined, “these three bullets were fired from [the recovered] firearm.” *Id.* On re-direct, the examiner also testified that he had no “doubt in [his] mind” that the recovered bullets were fired from the recovered gun. *Id.* The Court of Appeals initially affirmed Mr. Williams's convictions, holding there had yet to be any precedent in the District of Columbia “limit[ing] a toolmark and firearms examiner's testimony about the certainty of his pattern-matching conclusions.”  *Williams I*, 130 A.3d at 347-48. On re-hearing, and relying on its intervening decision in *Gardner*, the Court of Appeals subsequently held it was error to allow the examiner to provide “unqualified opinion testimony that purports to identify a specific bullet as having been fired by a specific gun via toolmark pattern matching.” *Williams II*, 210 A.3d at 742-43. In *Gardner*, the Court of Appeals had held that “a firearms and toolmark expert may not give an unqualified opinion, or testify with absolute or 100% certainty, that based on ballistics pattern comparison matching a fatal shot was fired from one firearm, to the exclusion of all other firearms.” 140 A.3d at 1177. The Court of Appeals did note, however, its decision allowed examiners to “offer an opinion that a bullet or shell casing was fired by a particular firearm,” just not with “absolute or 100% certainty.” *Id.* at 1184 n.19.

*24 *Williams II* appears to extend, or at least clarify, the Court of Appeals' holding in *Gardner*, even if not resolving the apparent contradiction between the language that appears in the text and in footnote 19 of the earlier case. See *WilliamsII*, 210 A.3d at 740-43. Not only does *Williams II* prohibit source attribution statements made with certainty, but it also prohibits any statement that conveys a “match” without sufficient qualification. See *id.* at 742-43. In *Gardner*, the “unqualified opinion” admitted in error was simply that the bullet recovered from the decedent's body and cartridge casing recovered from the crime scene were fired from the recovered firearm. 140 A.3d at 1182. The testimony was, “[i]n essence,” that “[the recovered gun] was the murder weapon.” *Id.* On re-direct, the examiner reiterated his opinion by stating the recovered bullet “was fired from the pistol.” *Id.* Similarly, in *Williams II*, the examiner concluded, “these three bullets were fired from this firearm.” 210 A.3d at 738.¹⁷ The Court of Appeals disparaged the government's argument—repeated as one of the bases of the posthearing briefs filed in this

U.S. v. Tibbs, 2019 WL 4359486 (2019)

case—that *Gardner's* limitation on firearms and toolmark testimony only applies to certainty statements. See *Williams II*, 210 A.3d at 740. In sum, *Williams II* barred “unqualified” statements of “match” and source attribution. *Id.* at 742-43. The Court of Appeals failed, and thought it unnecessary, to address what type of qualification could make such a statement admissible. *Id.* at 741-42 (“We ultimately conclude that we need not resolve the ambiguity of *Gardner's* footnote 19 in this case where the firearms and toolmark examiner not only testified, like the examiner in *Gardner*, that a specific bullet could be matched to a specific gun, but also that he did not have “any doubt” about his conclusion.”). Judge Catharine Easterly indicated in her concurrence, however, that an examiner might be able to testify that a specific bullet was fired by a specific gun if he could “reliably qualify” his opinion with a “verifiable error rate.” *Id.* at 746 (Easterly, J., concurring).

The extent to which these cases should affect the Court's decision seems a bit unclear. *Williams II*, like *Gardner* before it, reviewed trials that occurred in the *pre-Motorola* era, but nonetheless invoked the language of reliability in a manner more consistent with *Daubert* and Rule 702 than *Frye* and *Dyas*. See *WilliamsII*, 210 A.3d at 742. Although the Court's present decision has been made pursuant to *Daubert* and Rule 702, it restricts the firearms examiner's testimony such that he may not make a source attribution statement connecting the firearm and cartridge casing. This ruling fully comports with, and may even be compelled by, the strictures imposed by *Williams II* and other relevant precedent.¹⁸

For these reasons, as well as any others stated on the record in open court on August 8, 2019, Defendant's Motion has been GRANTED IN PART and DENIED IN PART.

<<signature>>

Todd E. Edelman

Associate Judge

(Signed in Chambers)


Date: September 5, 2019




Footnotes

- 1 On June 27, 2019, the government also filed a Motion to Correct Factual Inaccuracies in the Record. The Defendant filed his Reply on August 2, 2019.
- 2 In *Kumho Tire. Co. v. Carmichael*, the United States Supreme Court held that the *Daubert* reliability standard applies not just to expert testimony based on “scientific” knowledge, but to testimony based on “technical” or “other specialized” knowledge as well. 526 U.S. 137, 149 (1999).
- 3 Because many decisions on evidentiary issues do not result in the issuance of a reported or written opinion, the weight of authority from other courts and jurisdictions cannot be precisely determined. See 2009 NRC Report at 97.
- 4 Some trial courts have conducted full evidentiary hearings on the admissibility of firearms and toolmark identification evidence. See *Wrensford*, 2014 U.S. Dist. LEXIS 102446, at *2, 2014 WL 3715036, at *1; *Monteiro*, 407 F. Supp. 2d at 355. Others have even considered the recent critiques of firearms and toolmark identification. See *Romero-Lobato*, 379 F. Supp. 3d at 1117-22. These three courts admitted testimony similar to that proffered in this case under the *Daubert* framework. See *Romero-Lobato*, 379 F. Supp. 3d at 1123; *Wrensford*, 2014 U.S. Dist. LEXIS 102446, at * 58, 2014 WL 3715036, at *18; *Monteiro*, 407 F. Supp. 2d at 372.
- 5 Indeed, one court has recently found that the PCAST and NRC Reports themselves—despite their negative treatment of the established validity of firearms and toolmark evidence—constitute relevant peer review of the articles published in the *AFTE Journal*. See *Romero-Lobato*, 379 F. Supp. 3d at 1119. If negative post-publication commentary from an external reviewing body can satisfy this

prong of the *Daubert* analysis, then the peer reviewed publication component would be more or less read out of *Daubert*, leaving behind only the requirement of *some type* of publication.

Although surely not what the NFCS's recommendations contemplate, AFTE's website indicates that the public may search its articles' abstracts and keywords in its own index available on the AFTE website. *See* What is the Journal?, <https://afte.org/afte-journal/what-is-the-journal> (last visited Aug. 28, 2019).

At least one other court has made similar observations regarding the *AFTE Journal's* lack of independence. *See*  *Green*, 405 F. Supp. 2d at 109 n.7.

To be sure, a few judges who have admitted firearms and toolmark identification testimony have addressed, at least in some fashion, various criticisms of the discipline related to the methodology's error rate and its calculation. *See* *Romero-Lobato*, 379 F. Supp. 3d at 1120;  *Ashburn*, 88 F. Supp. 3d at 246;  *Otero*, 849 F. Supp. 2d at 434;  *Taylor*, 663 F. Supp. 2d at 1177. In response to the PCAST Report's criticism regarding the general lack of adequately designed studies for firearms and toolmark validation, the United States District Court for the District of Nevada explained that it would not “adopt such a strict requirement for which studies are proper and which are not.” *Romero-Lobato*, 379 F. Supp. 3d at 1120. The court went on to find that “*Daubert* does not mandate such a prerequisite for a technique to satisfy its error rate element.” *Id.* The United States District Court for the Eastern District of New York rejected a separate criticism levied by the 2009 NRC Report—that “the lack of objective standards prevents a ‘statistical foundation for estimation of error rates’”—and argued that the “information derived from [] proficiency testing is indicative of a low error rate[.]”

 *Ashburn*, 88 F. Supp. 3d at 246 (first quoting 2009 NRC Report at 154; then quoting  *Otero*, 849 F. Supp. 2d at 434).

Mr. Weller's training and experience, which involves a Master of Science degree in Forensic Science as well as over ten years of training and casework experience in firearms and toolmark analysis, *see* Decl. of Todd J. Weller 1, does not include any training or experience in decision science.

In one frequently-cited study, the test designers simply did not make clear whether their participants were to follow their specific lab's policies. 2018 Keisler Study at 58; Faigman Test., May 16, 2019, 85:24-86:6. The same study recognized this concern and specifically asked participants what their labs' policies were with respect to not excluding samples with matching class characteristics. 2018 Keisler Study at 58. However, when analyzing its data, that study made no attempt to disaggregate that data by the different policies used. *Id.* at 57-58.

The 2014 Ames Laboratory Study was made available on the internet without having undergone any clear peer review process, while the 2018 Keisler Study was published in the *AFTE Journal*.

The notion of a “break-in period,” during which time a firearm does not make consistent markings, would seem to undercut the general premise underlying the entire field of firearms and toolmark analysis—that is, that firearms reliably leave unique markings on casings and bullets fired based on marks left during the manufacturing process.

The studies themselves have treated inconclusives differently. For example, the Ames Laboratory study included the inconclusives in the denominator of the error rate calculation, such that inconclusives counted toward the total number of comparisons made, likely underestimating the overall error rate. Ames Laboratory Study at 15. The Lyons study, by contrast, treated an inconclusive response as a correct response in its calculations. Lyons Study at 254-55; Scurich Test., May 14, 2019 (2), 100:13-17, 19-21.

The Court understands Dr. Scurich to reason as follows: (1) The only correct answers in laboratory studies are “identification” or “exclusion” because the samples are such that they can be identified, Scurich Test., May 14, 2019 (2) 102:13-24; (2) In such a scenario involving a binary question, the basic principles of mathematics mean that the rate of true exclusions (called “specificity”) and the rate of false identifications or false positives (called “1 minus specificity”) must sum to 100% (i.e., of all the bullets that are known *not to match*, the percent declared “excluded,” and the percent declared “match” must sum to 100%), Scurich Test., May 14, 2019 (2), 86:6-19, 87:11-16, 87:2188:2; and (3) Therefore, the false positive rate must equal 100% minus the percentage of correct exclusions, Scurich Test., May 14, 2019 (2), 87:12-16. For an example, out of all of the possible correct exclusions, if examiners correctly concluded “exclusion” 80% of the time, then it must be true that they reached incorrect conclusions the remaining 20% of the time. *See* Scurich Test., May 14, 2019 (2), 99:10-14.

Additionally, it is important to note that inconclusives appear more frequently in open studies compared to closed and set-based studies, *see* PCAST Report at 109, and more frequently when the compared samples are true exclusions. For example, the Ames Laboratory Study, at 16-17, reported 735 inconclusives for 2,178 true different- source comparisons compared to only eleven for 1,090 possible true same-source comparisons. The evidence and testimony presented in the hearing did not adequately account for these disparities.

U.S. v. Tibbs, 2019 WL 4359486 (2019)

- 16 Some cases have premised findings of the ability of cross examination to illuminate questions regarding the foundational validity of this discipline on the supposed simplicity of the issues involved. “These weaknesses [in the methodology of toolmark identification] are also not particularly complicated or difficult to grasp, and thus are likely to be understood by jurors if addressed on cross-examination.” *Johnson*, 2019 U.S. Dist LEXIS 39590, at *58, 2019 WL 1130258, at * *19; see also *Johnson*, 2015 U.S. Dist. LEXIS 111921, at *8, 2015 WL 5012949, at *3.
- 17 On re-direct, the examiner said more about the uniqueness of the markings of the recovered firearm, *Williams II*, 210 A.3d at 738, but the Court of Appeals' ruling did not turn on the examiner's additional statements, *cf. id.* (“[W]e conclude that it was error to admit the examiner's opinion testimony, based on pattern matching, that the gun recovered from Mr. Williams's apartment was the murder weapon.”).
- 18 Although not addressed by this Memorandum Opinion, Mr. Tibbs also challenges whether Mr. Coleman reliably applied the AFTE method in this case. Based on the Court's present understanding of this aspect of Defendant's argument, this challenge would only be appropriate if Mr. Coleman were permitted to testify to a “match” (*i.e.*, that the recovered cartridge casing was fired from the recovered firearm). That, of course, is not the case; Mr. Coleman is restricted to testifying to his work, his observations, and the ultimate conclusion that the recovered firearm cannot be excluded as the source of the cartridge casing. It is not evident to the Court that the Defendant's argument applies to Mr. Coleman's application of the methodology given the restriction on any ultimate conclusion he would render. Accordingly, Defendant's Motion as it relates to Mr. Tibbs's as-applied challenge is denied as moot.

End of Document

© 2019 Thomson Reuters. No claim to original U.S. Government Works.