IN THE SUPREME COURT OF MISSISSIPPI

Mississippi Supreme Court Case No. 2023-DR-01076-SCT

WILLIE JEROME MANNING, ***Petitioner***

v.

STATE OF MISSISSIPPI, ***Respondent***

STATE OF Virginia
COUNTY OF Spotsylvania

AFFIDAVIT OF WILLIAM A. TOBIN

I, William Tobin, having been duly sworn, depose and state that I am over the age of twenty-one and am competent to make the statements contained in this affidavit. All statements are within my personal knowledge.

TABLE OF CONTENTS

A. EXECUTIVE SUMMARY

In 2013, I was asked by Willie Jerome Manning's legal team to provide a scientific explanation of a statement contained in a May 6, 2013 letter from the Department of Justice, which states: "The science regarding firearms examinations does not permit

Exhibit 4

examiner testimony that a specific gun fired a specific bullet to the exclusion of all other guns in the world." My 2013 Affidavit is attached at **Exhibit A** to this 2023 Affidavit.

In the 2013 Affidavit, I opined that the mainstream scientific community had concluded that firearms identification, a subset of toolmark examinations, and the conclusions drawn therefrom, lack validation or acceptable indicia of reliability. Subsequently, several landmark studies have conclusively established that the forensic practice of firearms identification is not only without foundational validity, it is so egregiously flawed that the three key indicia of reliability (accuracy, repeatability and reproducibility) are patently (and shockingly) unacceptable for forensic utility as evidence of guilt.

Firearms identification is a subset of forensic toolmarks identification practice. In this discipline, the firearm is the tool and the toolmarks (striations and/or impressions) are imparted to ammunition components during the cycling of a firearm. In the practice of firearms identification, the striations and/or impressions generated during forced contact with the firearm components during the cycling are generally compared between and among questioned and known bullets and/or cartridge cases, ostensibly to determine source attribution.

Historically, the discipline of firearms/toolmarks analysis has been practiced by crime lab firearms examiners who assumed, but never really questioned, that the characteristics (striations and impressions) that they use for comparisons and to "identify" a specific firearm are unique and belong only to that firearm. Not only are they *assuming* the existence of uniqueness, but also they are *assuming* that firearm examiners have the ability to discern that uniqueness ("discernible uniqueness") in casework. Over the past two decades, the legal and scientific communities have challenged the assumptions inherent within the discipline and found them to be unfounded. As of the date of this Affidavit, the two required premises of uniqueness and discernible uniqueness have never been established to exist.

Further, appending opinions of individualization with any probabilistic statement of certainty is objectionable and highly prejudicial. The letter from the U.S. Department of Justice dated May 6, 2013, the first "crack in the dam" at the time, recognized that the extreme probabilistic statement "to the exclusion of all other guns in the world" is patently unacceptable. Subsequently, the scientific community, and very reluctantly, the practitioner community, have determined that the letter did not go far enough in excluding any expression of certainty, and should have included "scientific certainty", "ballistic certainty", "practical certainty", or any similar. The practitioner community eventually conceded admonitions from the scientific community and, as of the date of this Affidavit, firearm/toolmark examiners no longer express any degree of certainty to their opinions. Additionally, the 2013 USDOJ letter incorrectly characterized the forensic practice as "science". There are seven reasons the practice cannot be considered to be a science, typically a prejudicial characterization for jurors. These developments were a direct result of the mainstream scientific community's rejection of claims that bullets can be matched to guns with any scientifically, empirically, or even heuristically acceptable degree of certainty using the methods employed by the analysis in Manning's case. **Ex. A.**

In 2023, I was contacted by the Mississippi Office of Capital Post-Conviction Counsel and asked to opine on the foundational validity[1] and established reliability/validity, *vel non*, of the forensic practice of firearms/toolmarks identification based on new and additional research since my 2013 Affidavit. I have reviewed my 2013 Affidavit, and I was provided, and reviewed, the following materials by counsel:

- Department of Justice letter dated May 6, 2013, with attachments;
- Transcript testimony of John Lewoczko, firearms examiner;
- Mississippi Crime Laboratory and FBI reports concerning ballistics evidence;
- The opening and closing statements of the prosecution and defense counsel in Manning's trial.

Since my Affidavit submitted in 2013, there has been very significant and revelatory new research, effecting a paradigm shift, adopted by the scientific community revealing that there is *no* demonstrable basis with scientific, empirical, or even heuristic foundational validity underlying the opinions of the forensic firearms identification expert at Manning's trial, nor supporting claims presented to the *Manning* jury.

The first landmark post-*Manning* study, published September 2016, was a report by one of the two most respected authoritative voices of the relevant scientific community, the President's Council of Advisors on Science and Technology (PCAST). Among its myriad findings was that firearms/toolmarks identification forensic practice is without foundational validity.

[1] Foundational validity is defined herein as the property of a practice or process whereby the practice has been established to exhibit acceptable metrics of the three critical indicia of reliability: accuracy, repeatability, and reproducibility, either by scientific, empirical (by experimentation), or heuristic ("training and experience") means.

The second landmark study was even more alarming. A recent research Report of the Ames National Laboratory (known as 'Ames II') dated October 10, 2020 exposes rates of error and indicia of reliability for firearms identification methodology that was used in *Manning* that are egregiously unacceptable, even for what are called 'gun-recovered' cases; the *Manning* matter is known as a 'no-gun-recovered' case, which is so problematic that some crime labs do not allow examiners to individualize cartridge cases or bullets to specific guns in 'no-gun-recovered' cases. This admonition[2] of requiring recovery of a firearm was well-known and articulated in the AFTE literature generally in the 2003-2004 firearms identification literature and promulgated at the annual AFTE meetings nationwide.

Following severe criticism of the discipline by the 2008 and 2009 National Academies of Science reports, the Association of Firearm and Toolmark Examiners (AFTE), the trade association of firearm toolmark examiners, performed a number of purported 'validation studies' in an attempt to validate the practice and claim that reliable error rates existed. The 2016 President's Council of Advisors on Science and Technology (PCAST) Report, and its subsequent addendum, dismissed all but one of those validation

[2] For example, see Nichols, R., "The Scientific Foundations of Firearms and Toolmark Examinations – A Response to Recent Challenges" (2004), "…there is not one conscientious firearms and toolmarks examiner who would suggest that personal familiarity with tool finishing processes and their effects on tool surfaces is anything but vital to the proper understanding of subclass characteristics. Without such knowledge and appreciation of manufacturing techniques examiners would have no way of ascertaining if subclass characteristics could exist." See also, Nies, R.,"Anvil Marks of the Ruger MKII Target Pistol – An Example of Subclass Characteristics", 35 *AFTE J.* 1 (2003) at 78: "Direct examination of the tool working surface responsible for producing the questioned toolmark must be done and the surface evaluated for potential subclass influence prior to making a final opinion that the questioned toolmark was produced by a particular tool to the exclusion of all other similarly marking tools… Knowledge of how the firearm is manufactured, as well as the manner in which the responsible working surface is applied to produce the toolmark, will be critical in determining whether the toolmark is truly individualistic", among many others.

studies as seriously flawed.[3] However, even that one study cited by PCAST failed to show that examiners could reliably reach the correct result because, among other defects, it artificially deflated the error rate by counting every answer of 'inconclusive' (*i.e.*, 'I don't know') as a correct response, among many other flaws. As a recent court decision explained, an answer of 'inconclusive' where the correct answer is either 'identification' or 'exclusion' because ground truth is known, is a failure of the discipline and should be counted as incorrect, not correct. When correcting for that misclassification, the error rates surge. *See U.S. v. Tibbs*, No. 2016-CF1-19431, 2019 WL 4359486, at *1 (D.C. Super. Sep. 05, 2019). Additionally, the study was not double-blind nor even blind. Recruitment for the study involved a self-selected respondent pool (typically only the most confident examiners sign up for a study), and notwithstanding, had a high survivorship bias (approximately 23.2% dropout rate). When respondents know they are being tested, the rate of invoking "inconclusive" (33.7% in the Ames I Study) skyrockets because examiners know they are not counted as incorrect responses. That practice is analogous to students being allowed to choose to answer only questions they're most confident in answering on tests.

In the shocking Ames-FBI Study[4] (aka, 'Ames II' Study) released October 2020, details to be articulated *infra*, the three indicia of reliability for foundational validity of the forensic practice were evaluated: accuracy, repeatability, and reproducibility.

[3] The one study that PCAST characterized as *properly designed*, known as 'Ames I', however, was seriously flawed and cannot be represented as having 'external validity' to be applied to judicial proceedings or represent actual rates of error in forensic case work.

[4] It is colloquially known as the 'Ames-FBI' study because of the FBI's involvement in the design of experiments. In true empirical science, participation by the potential respondent pool (practitioner community being tested) is highly objectionable. The FBI rejected the initial Ames National Laboratory design of experiments and proceeded to heavily influence the eventual methodology to its liking.

Approximately 36.4% of the time, respondents (examiners) would disagree on claimed 'matches'; a stunning 59.7% would disagree on 'non-matches'.

The National Academy of Sciences (NAS) and the President's Council of Advisors on Science and Technology (PCAST) **Exhibit C**, and the findings of the very recent Ames-FBI Report (aka, 'Ames II') **Exhibit D** , indicate that, in *Manning*, the jury and court would have been better served with a coin toss to assess firearm source attribution.

The only scientifically and forensically defensible opinion that Firearms Examiner Lewoczko could have offered at trial, and even today, is that a specific firearm could not be eliminated as the firing platform for the questioned bullets; in other words, that *in his opinion*,[5] it's possible that the bullets were fired from the same firearm. Because no gun was recovered in *Manning* for direct comparison, Lewoczko's opinion would not have even reached trial stage in some jurisdictions because it would have been unacceptable. The forensic practice, for seven reasons, cannot be called a science;[6] numerous courts have so observed and concluded. Allowing it to be called a 'science' has been personally observed by Affiant to have a serious prejudicial effect on jurors.

The PCAST report concluded that firearms identification is without foundational validity. Based on the finding that firearms identification is without

[5] This *caveat* is essential because once the appropriate sample pool is identified (all .38 caliber-capable firearms in *Manning*), the practice is purely (100%) subjective. His opinion, with no firearm recovered for direct comparison, is just that and, without foundationally valid underpinnings, basically intuited. The Scientific Method was developed to eliminate all subjectivity in empirical experimentation as much as humanly possible; a practice with 100% subjectivity is antithetical to the Scientific Method.

[6] The six critical cornerstones of the Scientific Method are: falsifiability, scientifically acceptable protocol, parameters of detection, rules of parameter application, repeatability, and reproducibility. Missing any one, a practice cannot be considered comporting to the Scientific Method. Firearms identification is devoid of ALL six. The seventh reason the practice cannot be called a science is that the Scientific Method was devised to eliminate subjectivity. Firearms identification practice is 100% subjective once the appropriate sample pool is identified.

foundational validity, PCAST recommended to the U.S. Department of Justice that such evidence not be offered as evidence in criminal proceedings.[7]

B. OVERVIEW OF QUALIFICATIONS

My background is described in my 2013 Affidavit attached here as **Exhibit A** and my CV attached as **Exhibit B**, but a summary is also provided herein, indicating a Bachelor of Science degree in Metallurgy from Case Institute of Technology[8] in Cleveland, Ohio, and graduate studies in metallurgy and materials science at Ohio State University and the University of Virginia.

I also have 24 years of experience as a forensic metallurgist/materials scientist with the FBI Laboratory in Washington, D.C. From 1986 until my retirement in 1998, I was personally responsible for virtually all forensic metallurgical examinations requested of the FBI by all local, state, federal (including military), and foreign agencies.

I am very familiar with the current practice and methodology of forensic firearms/toolmarks examinations. As a forensic metallurgist at the FBI Laboratory, I frequently conducted toolmark comparisons using the same methodology (although with scientifically acceptable opinions) and comparison microscopy instrumentation, and was periodically asked in that capacity to assist firearms/toolmarks examiners by explaining phenomena and material behavior that they encountered during their firearms/toolmarks examinations.

I am also very familiar with the scientific processes involved in establishing scientific, empirical, and heuristic validity and reliability of forensic practices, as well as

[7] PCAST Report at page 141.

[8] Now known as Case Western University.

the recent developments in the understanding of the validity of extant feature comparison methods such as firearm identification. I have authored or coauthored twenty papers on forensic matters. Of the papers involving toolmark and related methodological considerations, including design of experiments and hypothesis testing processes, the four most relevant directly relate to firearms/toolmarks identification issues similar to those involved in the Manning matter. The scientific principles and issues articulated in many of my papers directly apply to the premises underlying firearms identification forensic practice in general and to the unfounded claims of specific source attribution of the firearms examiners in this case The most relevant and noteworthy papers are "Hypothesis Testing of the Critical Underlying Premise of Discernible Uniqueness in Firearms Toolmarks Forensic Practice," W. Tobin and P. Blau, 53 Jurimetrics J. 121-146 (Winter 2013); "Analysis of Experiments in Firearms/Toolmarks Practice Offered as Support for Low Rates of Practice Error and Claims of Inferential Certainty," C. Spiegelman and W. Tobin, 12 Law, Probability & Risk 115-133 (2013), doi:10.1093/lpr/mgs028; "Absence of Statistical and Scientific Ethos: The Common Denominator in Deficient Forensic Practices," an 'Editor's Choice' award selection for 2017 in Journal of Statistics and Public Policy, and Simon A. Cole, *et al.*, *A Retail Sampling Approach to Assess Impact of Geographic Concentrations on Probative Value of Comparative Bullet Lead Analysis*, 4 LAW, PROBABILITY & RISK 199, 202 (2005).

I was invited, and served, as a scientific editorial reviewer for the National Academy of Sciences' National Resource Council 2004 final Report of the Committee Weighing Bullet Lead and am so acknowledged in several locations of the final published report. As indicated in my *curriculum vitae*, I have been qualified as an expert in 302 courts in 46 states/jurisdictions (including D.C. and Puerto Rico) and in testimonies before U.S. Senate Subcommittees on the Judiciary and Court Oversight. I have testified in firearms/toolmarks identification matters over 58 times throughout the United States.

C. RELIABILITY TESTING OF FIREARMS IDENTIFICATION

The Ames-FBI Study (aka, 'Ames II') released in October 2020, presented myriad opportunities for analyses of forensic firearms identification reliability, to include evaluations of various scenarios for calculation of practitioner rates of error, repeatability (retesting of same-examiner with same-samples over periods of time), and reproducibility (how much specific opinion agreement or disagreement exists between and among examiners). The principal difference in the scenarios is analytical treatment of an 'inconclusive' opinion. Seminally, the consideration that an 'inconclusive' is a correct response increases the denominator in the calculations of rate of error thus diluting (reducing) the calculated overall rate of error. This is quite acceptable in case work, as inconclusives would be considered non-incriminating, as would false negatives (Type II error). However, ground truth is not known in case work; in proficiency tests and purported "validation" studies, ground truth IS known.

It is quite notable that the Ames II study was not blinded; respondents knew they were being tested. Prior studies have demonstrated that when respondents know they are being tested, use of the opinion 'inconclusive' dramatically increases for various reasons. This, in essence, allows a respondent to pick and choose which questions he/she feels most confident in answering. As one of the two most cited scholars at the intersection of science and the law points out, such allowance is tantamount to permitting law students taking the 200-question MBE to avoid answering any of the questions they find to be too difficult, too ambiguous, or too "inconclusive," then calculating percentage correct only on the basis of the questions they did answer, or worse, including those questions not answered as correct responses. In his own words, "An examinee who gets to choose which questions to answer is likely to do very well indeed on the test. Such a testing protocol, of course, would be absurd. It is similarly absurd as a research design."[9] The overwhelming majority of purported "validation" studies reviewed by Affiant and colleagues, as well as the PCAST, have incorporated the same flawed design.

It is additionally noted that the Ames II study is associated with a high survivorship bias (32.4% dropouts in the 1st phase alone), surprisingly high given that the respondent pool was self-selected, meaning that only examiners most confident in their abilities volunteered as respondents. They likely do not constitute representative sampling of the firearms identification field at large as to confidence, competence, and skillsets. Because of the self-selected high-confidence respondent pool, it would be expected that the consequent data for rates of error, repeatability and reproducibility should have been quite low, high, and high, respectively.

The resulting data in the Ames-FBI Study do not support that expectation. According to Prof. Faigman's calculations, the error rate for comparing bullets in the Ames II study "…was as much as a whopping 53%."[10] For comparing cartridge cases, the error rate was "…as a similarly eye-popping 44%."[11] Thus, in the Ames II study, "…a controlled black-

[9] David L. Faigman, Chancellor & Dean and John F. Digardi Distinguished Professor of Law at UC-Hastings, Professor of Medicine in the Dept. of Psychiatry at UC-San Francisco, and leading scholar on the subject of the use of scientific research in legal decision making, in affidavit *re State v. Abruquah*, No. CT121375X, Prince George's County Circuit Court, Criminal Division, Prince George's County, Md., filed May 19, 2021, attached as Exhibit E.
[10] *Ibid*, para. 74.
[11] *Ibid*, para. 76.

box study where ground truth is known, examiners are worse than flipping a coin in making bullet comparisons and only slightly better than flipping a coin in making cartridge case comparisons."[12]

As if the data for error rates weren't sufficiently alarming, the data for repeatability and reproducibility are equally disturbing. Calculations by two scholars of experimental design[13] have revealed that, for repeatability,

- ~ **21%** of the time, same examiner would disagree with prior opinion for 'matches',
- ~ **35.3%** of the time, same examiner would disagree with prior opinion for non-matches
- ~ **50%** of the time disagreement (same examiner) if grouped with 'Inconclusive C' ("leaning toward elimination").

As for reproducibility, the same statistician calculated;

- ~ **36.4%** of the time different examiners would disagree on 'matches',
- A stunning **59.7%** would disagree on 'non-matches'.

Additional observation on the repeatability data: as bad as the data are, actual error rates are likely worse than the data on repeatability would imply. It is noted that a respondent could have scored poorly (0%, incorrect response) in the first phase of the study, then poorly (0%, incorrect response) in the second phase or subsequent, thus scoring highly in the *reliability* indicator of 'repeatability'; the examiner was reliable, just reliably wrong.

[12] *Ibid*, para. 77.

[13] Dorfman, A. & Valliant, R., "Inconclusives, errors, and error rates in forensic firearms analysis: Three statistical perspectives", Forensic Science International: Synergy 5 (2022) 100273. *See also*, Alan H. Dorfman & Richard Valliant. *A Re-analysis of Repeatability and Reproducibility in the Ames-USDOE-FBI Study*. Statistics and Public Policy (to appear) (2022) *http://arxiv.org/abs/2204.08889*

As to survivorship bias, the high dropout rate of 32.4% for the first phase of the study testing rose to an ultimate 68.8% for the study as a whole, meaning only approximately 31% of the initial respondents completed the study. The Report indicated that the principal reason cited by the dropouts was ostensibly the fact that they were overloaded in their normal workload. It should be noted that it is quite likely that the rates of error and poor repeatability and reproducibility rates would be even worse (1) for the general population of examiners, and (2) that those dropping out of the study could well be those most likely to make errors in the crush of heavy workloads, if indeed the rationale for dropping out is as represented.

D. REVELATORY ALERT FROM PRACTITIONER DOMAIN

Since the trial in *Manning*, firearms examiners outside the U.S. seem to be awakening to what the scientific community, primarily metallurgists and materials scientists, have known for decades: every other firearm in the same production lot and distributed in the region of a crime, would present as analytically indistinguishable in casework requiring inductive reasoning (which is all of them except the very isolated *Waco*- and *Ruby Ridge*-type cases).

A relatively recent paper by a pair of Israeli and Russian coauthors published an alarming expose of "subclass carryover" in their 2020 paper, "The Problem of Subclass Features in Forensic Identification", cautioning of misattributions without knowing, at a minimum, how each firearm component was manufactured (not possible in no-gun recovered cases as in *Manning*). Subclass carryover is the term used to denote characteristics that appear on bullets and cartridge cases from the manufacturing process and belonging to a potentially very large number of firearms in the same production lot, at a minimum. They are not "individual" in nature as described by Lewoczko in his *Manning* trial testimony. Worse, as observed by the PCAST and contained in my paper addressing the subject 4 years earlier[14], there exists no known method by which a forensic examiner can discern or differentiate between subclass characteristics (from manufacturing) from purportedly "individual" characteristics.[15] The number of virtually identical firearms distributed in the region of a crime is never known by litigants; thus, jurors have no information by which to interpret the meaning of a claimed "match". Smith & Wesson makes approximately 2,000 9mm S&W semi-automatic pistols *per day*; they are boxed, palleted, and shipped to distributors and retail outlets as such.

Demonstrating the high risk of misattribution from subclass carryover, the authors present numerous examples. In each of the following photos, two different samples are displayed side-by-side in split-screen images (delineated by a dark vertical line in the center of each photograph). Each of the samples in the sample pairs was fired in/from a completely different firearm. They are virtually indistinguishable and quite likely to have been subject to misattribution in case work where no other firearms from the same production lot were available for comparison.

[14] Tobin, W.A. & Blau, P.J., "Hypothesis Testing of the Critical Underlying Premise of Discernible Uniqueness in Firearms-Toolmarks Forensic Practice," 53 *Jurimetrics J.* 121-146. (2013).

[15] PCAST report at 64, citing OSAC Research Needs Assessment Form. "Assessment of Examiners' Toolmark Categorization Accuracy", issued October 2015 (Approved January 2016). Available at: www.nist.gov/forensics/osac/upload/FATM-Research-Needs-Assessment_Blackbox.pdf .

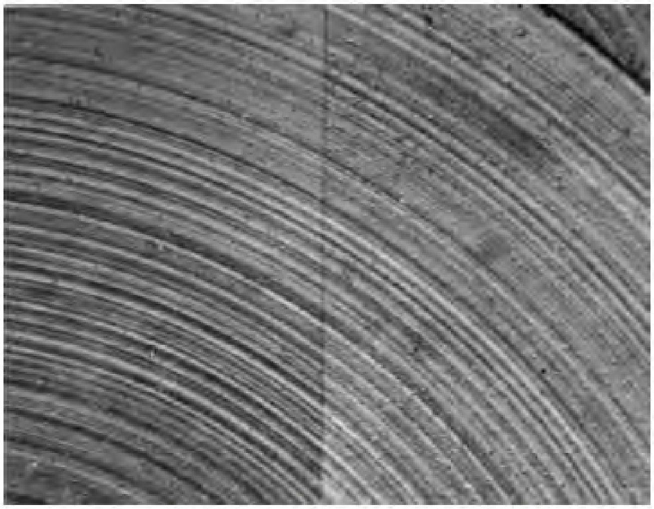


Fig. 2. Matching of surface microrelief of consecutively made breeches of Ruger M77 Mark II rifles (the left is the first and the right is the sixth) [6]

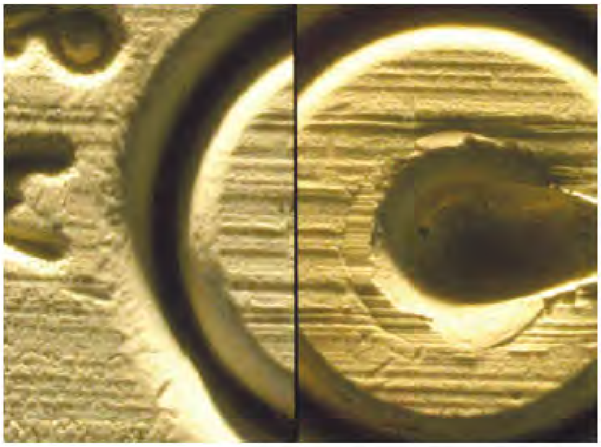


Fig. 4. Matching of the surface microrelief of cartridge cases dischardged from pistol PBV7152 (left) and pistol PBV7164 (right) [8]

Fig. 2 & **Fig. 4** photos from "The Problem of Subclass Features in Firearms Identification" (2020), pp. 111 & 112, respectively, available at https://doi.org/10.30764/1819-2785-2020-1-109-117

E. SUMMARY OPINIONS

Without recovery of a firearm, the firearms examiner should not have rendered an opinion of specific source attribution (individualization) in the case at bar. However, even had a firearm been recovered, the only scientifically and forensically defensible opinion that the examiner (Lewoczko) in *Manning* could have rendered is that a specific firearm could not be eliminated as the firing platform, in other words that it was possible that the same firearm fired the questioned bullets. The most respected voices of the relevant scientific community are in consensus that the practice is without foundational validity and should not have been presented as evidence of guilt in a criminal trial. That has been my opinion both prior to the NAS and PCAST findings as evidenced in my published papers, and it remains my opinion today. Scientific studies post-*Manning* have established that the methodology is egregiously unreliable and that examiners cannot do what they claim to do.

William A. Tobin
WILLIAM A. TOBIN

SWORN TO AND SUBSCRIBED BEFORE ME, this 26th day of September 2023.

CHRISHAUN DEE'L OAKES
NOTARY PUBLIC
REG # 7677900
MY COMMISSION EXPIRES 8/31/2024
COMMONWEALTH OF VIRGINIA

Chrishaun Dee'l Oakes
NOTARY PUBLIC

My Commission Expires:
8/31/2024
(Seal)

IN THE SUPREME COURT OF MISSISSIPPI

No. 2013-DR-00491-SCT

WILLIE MANNING	PETITIONER
V.	
STATE OF MISSISSIPPI	RESPONDENT

AFFIDAVIT OF WILLIAM A. TOBIN

I, William Tobin, declare as follows:

Table of Contents:

Exhibit A

- -

A. Case-Specific Documents and Request for Scientific Review

1. I was provided the following documents pertaining to this case:

   (a) Department of Justice letter dated May 6, 2013, with attachments;
   (b) transcript of testimony of John Lewoczko, FBI firearms examiner;
   (c) Mississippi Crime Laboratory and FBI reports concerning ballistics evidence.
   (d) State's "Response To Supplemental To Motion To Stay Execution…" filed May 17, 2013.

I have been asked to provide a scientific explanation of a statement contained in the May 6, 2013 letter from the Department of Justice to all parties in this matter, which states, "The science regarding firearms examinations does not permit examiner testimony that a specific gun fired a specific bullet to the exclusion of all other guns in the world." I have also reviewed the transcript of testimony given by the FBI firearms analyst in the 1994 trial of this case. The analyst testified that he was able to determine that seven bullets were all fired "from the exact same firearm . . . to the exclusion of every other firearm in the world."

I have personally and collaboratively reviewed virtually all of the purported "validation studies" existing in the firearms identification domain, and of the findings of the National Research Council (NRC) of the National Academy of Sciences (NAS) with regard to such studies undertaken in recent years, to determine the validity of conclusions drawn from firearms identification examinations. Additionally, I have coauthored several scientific papers addressing our observations and conclusions as to the scientific foundations, *vel non*, of the purported "validation studies" in the domain.[1] As discussed below, those studies have led the true (mainstream) scientific community to conclude that firearms examinations and the conclusions drawn from them lack validation or reliability. The letter from the Department of Justice dated May 6, 2013, reflects the mainstream scientific community's rejection of claims that bullets can be matched to guns with any scientifically acceptable degree of certainty using the methods employed by the analyst in Manning's case.

B. Background Overview as Materials Scientist / Metallurgist

2. I have a Bachelor of Science degree in Metallurgy from Case Institute of Technology[2] in Cleveland, Ohio, and graduate studies in metallurgy and materials science at Ohio State University and the University of Virginia. While in graduate school, I accepted an offer of employment by the Federal Bureau of Investigation (FBI) as a Special Agent in 1971. After serving approximately 3½ years as a "street

[1] See "Hypothesis Testing of the Critical Underlying Premise of Discernible Uniqueness in Firearms/Toolmarks Forensic Practice", 53 *Jurimetrics J.* 121-142 (Winter 2013). See also, "Analysis of Experiments in Forensic Firearms/Toolmarks Practice Offered as Support for Low Rates of Practice Error and Claims of Inferential Certainty", *Law, Probability & Risk* (Oxford University Press), doi:10.1093/lpr/mgs028 (forthcoming).

[2] Case Institute of Technology is now known as Case Western Reserve University.

Agent" and because of my academic and professional experience, I was assigned to the FBI Laboratory in Washington, D.C., as a forensic metallurgist, where I remained until my retirement as the manager of forensic metallurgy operations in 1998. During my career at the FBI Laboratory, I undertook additional graduate studies in materials science (metallurgy) at the University of Virginia, and also studies for a Master of Arts in Special Studies at George Washington University (GWU), a program sponsored and instructed by both the Forensic Science Department and Law School at GWU.

By congressional mandate, the FBI Laboratory is charged with providing "assistance to all duly-authorized law enforcement agencies" throughout the U.S. Because no forensic metallurgy component existed in any state, local, or other federal law enforcement entity in the United States, or even in most federal regulatory (non-law enforcement) entities such as the Occupational Safety and Health Administration (OSHA), Food and Drug Administration (FDA), or Department of State, *inter alia*, the FBI Metallurgy Unit provided requested assistance for all federal, state and local criminal, civil and non-legal matters, and periodically for the international community in foreign police cooperation matters. From the retirement of the former FBI Chief Forensic Metallurgist in 1986 until my own retirement in 1998, my unit was personally responsible for virtually all forensic metallurgical examinations requested of the FBI by all local, state, federal, and foreign agencies. Such assistance included my participation with the National Transportation Safety Board (NTSB) in determination of the causes of the TWA 800 midair explosion disaster over Long Island, N.Y., the nation's worst rail disaster (the "Sunset Limited" in Mobile, AL), the nation's second worst environmental disaster (oil spill by the "Emily S./Morris J. Berman"), and numerous other high profile incidents. Because of the volume of high profile cases for which I was responsible, my scientific work product has been subject to substantial public scrutiny in the United States and internationally throughout my career as a forensic metallurgist/materials scientist.

Included in my academic background are various courses typical of a metallurgy/materials science curriculum, at both an undergraduate (U) and graduate (G) level. Most directly or indirectly relate to production and functioning of firearm components and cartridge cases (not all inclusive and generally in reverse chronological order):

a. Manufacturing Processes & Materials (G)
b. Statistics for Scientists & Engineers (G)
c. Structure & Properties of Materials (G)
d. Shaping & Forming of Metals (G)
e. Engineering Metallurgy (G)
f. Physical Metallurgy (1 G, 1 U)
g. Advanced Materials Laboratory (U)
h. Properties of Materials (U)
i. Engineering & Mechanical Properties of Materials (U)
j. Relaxation Properties of Solids (U)

k. Engineering Applications of Materials (U)
l. Foundry Metallurgy (U)
m. Diffusion Processes Laboratory (U)
n. Diffusion Principles (U)
o. Plastic Flow Laboratory (U)
p. Dislocation & Plastic Flow (U)
q. Metallurgical Processes Laboratory (U)
r. Fundamental Metallurgical Processes (U)
s. Behavior of Materials (U)
t. Production Metallurgy (U)
u. Thermodynamics (U)
v. Heat & Mass Transfer (U)
w. Structure of Crystals (U)
x. Introduction to Materials (U)

It should be noted that the term 'plastic' in the above listing does not refer to the common usage as the synthetic amorphous polymer solid, but rather describes the non-reversible behavior (deformation) of metals and materials reacting to applied stresses.

During my metallurgy studies and my tenure as an FBI forensic metallurgist, I visited many metal manufacturing and processing plants throughout the United States and Taiwan to observe a wide variety of industrial manufacturing practices in detail. I also served as a plant metallurgist in both the copper and aluminum industries, and as a research metallurgist in the field of aerospace and nuclear metallurgy. My *curriculum vitae* is attached as Exhibit E-1.

I was asked, and accepted, to serve as a scientific editorial reviewer for the draft final report of the 2004 National Research Council of the National Academy of Sciences (NAS) Committee on Bullet Lead Analysis.

C. Specific Qualifications Applicable to Forensic & Firearms/Toolmarks Issues

3. The domain of metallurgists and materials scientists includes material behavior in virtually every phase in the life of a metal, regardless of form, from its extraction as an ore to the use and functioning of a finished product. Each stage of product development, including for firearms and consumer tools, involves important metallurgical considerations, from material selection and process design to bulk metal forming, shaping, heat treatment, finishing, and related production processes. In scientifically evaluating the characteristics used by firearms examiners in 'firearms identification' practice as it is called, *it is imperative that the underlying scientific phenomena affecting material behavior and tribological interactions with, for example, forming tools and dies, in various conditions and environments of both production and consumer use, are understood.* The need to understand the scientific principles governing material behavior and their interactions also patently extends beyond production processes. Clearly, interactions of both the product with its environment, and of product components with each other in service (ultimate consumer use), are important metallurgical design considerations. Knowledge of

the material behavior resulting from the effects of friction, lubrication, and wear, is critical to evaluating the manifestations of tribological interaction (striations and impressions) for efficacy of product function and for failure analysis both in production and in user service. It is also important in scientifically evaluating the pattern-matching practice of firearms examiners in their forensic comparisons.

The heart of virtually every metal forming/shaping operation is the tool/die responsible for changing the shape of the metal work piece under pressure (forced contact). This is true regardless of the actual product produced, such as the barrel, ejector, extractor, firing pin, breech face of a firearm, ammunition cartridge cases, screwdrivers, aerospace components, wire, tubing, *etc.*, or the function that the product is intended to serve in the consumer market.

A critical aspect of production continuity, and a seminal issue for forensic tool marks comparisons, is the material behavior of both the metal product/component and the tool/die during metal-to-metal contact under pressure (defined as force per unit area) during production. Material responses to applied stresses during fabrication frequently result in formation of striations and/or impressions on the work piece component surface from forced contact with the forming tool (die), characteristics used as the basis for firearms/toolmarks comparisons. The formation of these striations and impressions depends on numerous parameters including, but not limited to, manner of fabrication, regime of tribological interaction, cleanliness of lubrication system operative, component alloy, mechanical properties (*e.g.*, tensile strength), temper, speed of processing, temperature of process, *inter alia*.

Tribology is the science of friction, lubrication and wear, of [primarily] metals in contact and in relative motion.[3] It is such an important consideration during all metal-to-metal contact that it is a sub-discipline of metallurgy/materials science and is included in various academic metallurgical, materials science and mechanical engineering studies/courses. Metal-metal contact involving tribological considerations is patently unavoidable during production and/or consumer use of most wrought[4] metal products. Such forced contact in relative motion occurs both in production (in the metal forming and shaping processes for firearm components), and in service use (a bullet traveling under pressure against the lands and grooves of a barrel, cartridge case against breech face, firing pin contact with primer cup, extractor and ejector contact with the cartridge case, *inter alia*). The most appropriate and relevant true scientific discipline to address issues of metal-to-metal interactions, such as occur during formation and transfer of striae and impressions during production of firearms, ammunition and production tooling, and as also occur

[3] See *McGraw Hill Dictionary of Scientific and Technical Terms*, Fourth Edition, Sybil P. Parker, Ed.-in-Chief, McGraw Hill Book Company (1989), ISBN 0-07-045270-9, at 1965.

[4] Generally, a metal alloy or product that was not formed to final shape by casting. The term 'wrought' is used to denote a metal that has been mechanically worked/shaped after the alloy was originally cast into raw form. See *McGraw Hill, ibid*, at 2071.

during actual product use (such as cycling a bullet and cartridge through a firearm), is metallurgy/materials science.

Part of my responsibilities as a plant metallurgist included evaluating tribological regimes operative during production, and toolmarks imparted by tools and dies during fabrication and production, in efforts to insure efficacy of operations and production continuity, while reducing product variability and breakdown of production tooling. Additionally, I am very familiar with the current practice and methodology of firearm and toolmark examinations inasmuch as I used the same methodology and comparison microscopy instrumentation in my capacity as a forensic metallurgist. I have also functioned as a consultant in the ammunition manufacturing industry.

D. Forensic Individualization Defined and Consensus in Scientific and Scholarly Forensic Communities

4. Forensic individualization, also known as specific source attribution, is the process by which a questioned item is purportedly associated with a specific source. Source attribution may or may not be enhanced with probabilistic language like "to the exclusion of all others," "infinitesimal chance" of a coincidental match, "unique signature," *inter alia*. With or without this enhancing language, the examiner is still individualizing, or attributing markings to a specific source. In individualizing, the forensic examiner rules out – by subjective belief as it turns out – all other possible

sources for the combination of characteristics ('striations' or 'striae', and/or 'impressions', for firearms/toolmarks examinations) observed in the questioned marking, including the vast universe of possible sources he has never examined. Before getting into specifics, it should be noted that it is a strong consensus among my colleagues, distinguished members of the relevant scientific and scholarly forensic communities, that individualization in firearms/toolmarks is without scientific merit or foundation, but rather is "based on anecdote, intuition and speculation rather than on a scientific foundation. Consequently, individualizations in casework rely on a 'leap of faith'."[5] In fact, the belief by firearms/toolmarks examiners that their practice can render specific source attributions with any scientific basis is considered a fallacy.[6]

5. Individualization in forensic firearms/toolmarks practice is rejected by a unanimous consensus of my colleagues and collaborators, most with scientific backgrounds and/or specializing in forensic evidence, with whom I frequently or periodically interact. Individuals expressly rejecting the practice as scientifically invalid, listed with permission, with their areas of specialty are:

> **David L. Faigman**, Distinguished Professor of Law, UC-Hastings, Consortium on Law, Science & Health Policy, specializing in scientific

[5] Saks, Michael J., Koehler, Jonathan J., "The Individualization Fallacy in Forensic Science Evidence", *Vanderbilt LR* 6:1, at 10.

[6] *Ibid. See also* Koehler, Jonathan J., and Saks, Michael J., "Individualization Claims in Forensic Science: Still Unwarranted", *Brooklyn LR* 75:4.

evidence, coauthor of 5-volume *Modern Scientific Evidence: The Law and Science of Expert Testimony*;

Clifford Spiegelman, Distinguished Professor of Statistics, Texas A&M, and former member of the National Academy of Sciences in comparative bullet lead analysis;

William C. Thompson, Professor of Criminology, UC-Irvine, criminal justice & decision making, specializing in forensic science and statistical treatment;

Alicia Carriquiry, Distinguished Professor of Statistics, Iowa State University, and former member of the National Academy of Sciences in Ballistic Imaging;

Michael J. Saks, Professor of Law, Arizona State University, Ph.D in experimental psychology, reviews empirical research methodology and statistics in forensic science, doctoral background for evaluating design of experiments;

Jonathan J. Koehler, Professor of Law and Ph.D, Arizona State University, specializing in statistical treatment of forensic evidence and the role of hypothesis in forensic science evidence;

Pradip N. Sheth, Associate Professor of Mechanical Engineering, deceased; based on personal interactions and affidavit furnished in criminal matter of *U.S. v. Willie Gayden*, D.C. Superior Court., Criminal Case No. 2006 CFI 27899;

Adina Schwartz, Associate Professor of Law and Ph.D, John Jay College of Criminal Justice, evidence law, law and science.

Peter J. Blau, PhD., Fellow-ASM International, Fellow-ASTM International, Fellow-Society of Tribologists and Lubrication Engineers (STLE), and Consultant in Tribology. Has held research and project management positions at Air Force Materials Laboratory (1973-76), National Bureau of Standards (1979-1987), Office of Naval Research (1986-87), and Oak Ridge National Laboratory (1987 – present).

Further, these individuals endorse the NRC committee's findings regarding the unvalidated and scientifically unsupported nature of individualization opinions, which I discuss in more detail later in this affidavit.

6. The fallacy of individualization "...arises when the forensic scientist rules out all other possible sources for the unknown marking, including the multitude he has not examined, once he has found a single object or person that matches the features of the unknown marking. The fallacy is deeply entrenched in forensic science practice, where most examiners say that their knowledge, training, and experience enable them to make the inferential leap from observed consistencies between markings and their putative source to a conclusion that no other object in the world could have produced those markings."[7] Further, Koehler and Saks indicate, "In our view, the existing and foreseeable scientific knowledge falls far short of providing criminalists with enough scientific support to claim that the objects that they study are either unique or discernibly unique. Certainly the uniqueness question cannot turn on the beliefs that forensic scientists have about this issue based on their training and experience."[8]

E. Metallurgical Origins of Toolmarks, Relevant Considerations of Formation, and Forensic Practice

[7] Koehler and Saks, *ibid*, (page unknown as publication was in press).

[8] *Ibid.*

7. The nature, quality, and number of characteristics imparted to a metal product are dependent upon the type(s) and magnitude(s) of stresses (among many other parameters such as process speed, tooling material, product material, lubrication regime, *inter alia*) during the fabrication process. For plant metallurgists, tribological considerations are important to production continuity, production costs, quality control, safety and, quite subtly in some cases, potential civil litigation against a metal product manufacturer. Accordingly, they are significant considerations for fabrication tool and die design, the heart of most metal manufacturing operations, and are the very tools (known in the industry as 'tooling') used to form the various components of a firearm that impart the toolmark characteristics used by firearms and toolmarks examiners for source attributions.

8. In their evaluations of forensic evidence submitted for examinations, firearms examiners rely on the markings ('toolmarks') left on bullets and cartridge casings during the contact described above (during production) and while in relative motion by firearm components such as the barrel, firing pin, extractor and/or breech face of a gun during operation ('cycling') of the weapon. For conclusions of individualization (specific source attribution), one of two crucial premises upon which firearms examiners rely is that each firearm imparts individual characteristics (generally striations or striae, and impressions) to bullets and cartridge cases cycled

through the firearm that are purportedly unique to that firearm.[9] Scientific acceptance of the uniqueness premise is problematic for reasons that will be discussed below.

9. First, based on exhaustive literature research and review, *I find no body of data, collective studies, or even single study, which is sufficiently meaningful and comprehensive as to warrant the premise of uniqueness status as a universal assumption in the field of forensic firearms/toolmarks practice.* A relatively recent report issued by the National Research Council of the National Academies of Science has also concluded that the premise of uniqueness has not been scientifically established, stating:

> 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.[10]

10. The Association of Firearms and Toolmarks Examiners (AFTE), is a membership-supported trade association formed to represent the interests of its members who are generally without significant scientific background. AFTE is the principal, and realistically sole, source of guidelines for firearms and toolmark

[9] I will refer to this premise herein as the "premise of uniqueness."

[10] *Ballistic Imaging*, Report of the National Research Council; National Academies Press, Wash., D.C. (2008), p3. I will refer to the report herein as the Ballistic Imaging report.

examiners. It is not a scientific body, and does not establish scientifically validated standards or protocols comporting to rigors of the scientific method.[11]

11. AFTE's "Theory of Identification" is the criterion by which examiners declare purported association between a particular firearm and a recovered bullet, spent shell casing, or both. There is little agreement in the forensic community of firearms/toolmarks examiners itself as to number, type, quality,[12] and characteristics of striae/impressions that must match before a source attribution can be claimed. Within the firearms/toolmark examiners' community, the AFTE Theory of Identification is the only guideline that examiners follow even though the theory provides no practical guidance. Instead, the AFTE theory provides only the vague and subjective benchmarks of "sufficient agreement," "best agreement," and "practical impossibility." In more expanded form it states:

> Agreement is significant when it exceeds the best agreement demonstrated between toolmarks known to have been produced by

[11] See *Paul M. Dougherty v. Lucian Haag, et al.*, Case No. 05CC06993, Statement of Decision (Hon. Daniel J. Didier), Super.Ct.of CA, County of Orange, Dec. 6, 2006, at 1, where defendant (AFTE) declares status as trade association.

[12] Arguably the most subjective aspect of the entirely subjective practice is that of characteristic 'quality.' Difficulty of striation 'quality' assessment can be demonstrated in side-by-side partial overlay comparisons of randomly selected SKU UPCs of unrelated products. As in firearms/toolmarks practice, lines will quite commonly be found in the same position, but arbitrary assessment as to whether the 'quality' of each pair of aligned lines merits declaration as a 'match' is without benchmark, protocol, or even guideline, and is completely up to the discretion of the human observer. This is a subtle, but critical, consideration rendering examiners significantly more vulnerable to observer bias. It has been observed in AFTE literature that misattributions are most typically caused by examiners "ascribing too much significance to a small amount of matching striae that is actually achievable in known non-matches." See Biasotti, Murdock & Moran, "Scientific Issues", in 4 David L. Faigman, *et al.*, *Modern Scientific Evidence*, at 562.

> different tools and is consistent with the agreement demonstrated by toolmarks known to have been produced by the same tool. The statement that "sufficient agreement" exists between two toolmarks means that the agreement 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."[13]

These are subjective criteria, as the AFTE concedes.[14] In its most recent report, the National Research Council of the National Academy of Sciences (NAS) was critical of the AFTE theory because its methodology is based on such subjective and nonspecific criteria:

> 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 toolmarks known to have been produced by different tools and is consistent with the agreement demonstrated by toolmarks 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,

[13] *See* AFTE Criteria for Identification Committee, "Theory of Identification, Range of Striae Comparison Reports and Modified Glossary Definitions – an AFTE Criteria for Identification Committee Report." *AFTE Journal*, Vol. 24, No. 2, April 1992, 336-340.

[14] See "principle" 3 in the AFTE Theory of Identification: "The current interpretation of individualization/identification is subjective in nature, founded on scientific principles and based on the examiner's training and experience", available online at: http://www.ojp.usdoj.gov/nij/training/firearms-training/module13/fir_m13_t05_07.htm.

questions regarding variability, reliability, repeatability, or the number of correlations needed to achieve a given degree of confidence.[15]

12. In the pattern-matching process of evaluating toolmarks used as a basis for purported individualization (specific source attribution),[16] the forensic toolmark profession defines three groups of characteristics: class, subclass and individual. Class characteristics are considered common to every member of a relatively large group of items or product, typically originate in the design stage, and are deliberately imparted as part of the manufacturing process. Class characteristics include, for example, the number and direction of lands[17] and grooves on a bullet that are common to numerous weapons of similar or different models. Comparisons of class characteristics as an early stage phase of forensic evaluation serve to filter the universe of all possible products to a manageable population with general comparability.

Subclass characteristics are fortuitously produced during the manufacturing process by a tool that can leave virtually identical markings on a number of products produced, including firearms, during the tool's useful life in which it typically produces lots (groups or "batches") over many months, depending on the process

[15] "*Strengthening Forensic Science in the United States: A Path Forward*," National Research Council, National Academy of Science; National Academies Press (2009), p.155. I will refer to the report herein as the NAS Forensic Science Report.

[16] In the context herein, individualization would be purported association of a bullet(s) or cartridge case(s) with a specific firearm.

[17] Lands are the raised areas between grooves in the rifling of a firearm barrel.

and product. The number of products bearing the subclass characteristics can be very large and can exist across many production lots spanning months, as will be discussed below. However, that number is a subset (smaller) population of items or product within the class defined, hence the term 'subclass.'

Individual characteristics are, by definition in the AFTE community, unique to one firearm or tool.

13. It should be kept in mind that: examiners are seeing only a relatively small portion of a bullet surface in the field of view of between 5X and 40X stereomicroscopy; comparisons are typically based on combinations of non-unique characteristics (primarily lines); there is significant concordance in characteristics among known non-matches (as will be discussed below); human cognitive pattern retention (from previous cases and training) is limited, particularly of a mundane geometric pattern not lending itself to description or 'memory/recollection tags' (lines)); and significant concordance occurs in characteristics imparted from different firearms of the same manufacturer.[18] As an example, tests on six 'consecutively machined' rifle bolts found a "startlingly" high correspondence of

[18] *See* Biasotti, "A Statistical Study of the Individual Characteristics of Fired Bullets," 4:1 *J.For.Sci.* 34, 34-50 (1959), *inter alia*. *See also* Rivera, Gene C., "Subclass Characteristics in Smith & Wesson SW40VE Sigma Pistols," *AFTE Journal*, Vol. 39 No. 3 (Summer 2007) for more examples.

microscopic characteristics, according to one study.[19] In another, 51.7 percent of 'matching lines' was observed in known non-matches.[20] The issue of cognitive retention is particularly significant in view of the highly subjective nature of toolmarks examinations and the AFTE guideline that 'match' pronouncements are based exclusively on *recollection* of previous cases and training, for similarities to "known non-matches."

14. When two metals are in forced contact with each other, the 'softer' material typically acquires characteristics from the 'harder' material (although, generally unknown outside the metallurgy/materials science field, and counterintuitively, hardness is not always the sole metallurgical determinant; it is a general guideline). As previously alluded, such forced contact occurs during the cycling of a cartridge through a firearm when the cartridge case is impacted (struck) by the firing pin, the cartridge head is forced (in compression) against the breech face, the bullet is propelled through the barrel, the expended cartridge case is extracted from the chamber, and the case is ejected from the weapon. Comparisons of striations and/or impressions imparted during these forced contacts are the basis of examinations and conclusions by firearms examiners.

[19] See "All we want you to do is confirm what we already know": A *Daubert* Challenge to Firearms Identifications," Lisa J. Steele, 38 *Crim.L.Bull.* 465 (2002), citing "Consecutively Machined Ruger Bolt Faces," *AFTE J.*19 (Winter 2000).

[20] Miller, J., and Neel, M., "Criteria For Identification Of Toolmarks Part III: Supporting The Conclusion," *AFTE Journal*, Winter 2004, Vol. 36 No. 1. at p.9.

15. It is sometimes claimed, with various phrasing, that cartridges[21] are "...cycled through a gun the same way every time..." and "...cartridges are cycled through firearms the same way...".[22] While this claim is true regarding the macro-mechanical process of firing a cartridge, it is not true with regard to the critical physical process parameters that influence the transfer of characteristics (striae and impression) on a microscopic level.

F. National Academy of Sciences and AFTE Practice[23]

16. In the Ballistic Imaging Report, the NAS stated:

> Forensic individualization sciences that lack actual data, which is most of them, have no choice but to either intuitively estimate those underlying probabilities and calculate the coincidental match probability from those subjective probabilities, or simply to assume the conclusion of a miniscule probability of a coincidental match (and in fact they do the latter).
>
> In the specific context of firearms and toolmark examination, derivation of an objective, statistical basis for rendering decisions is hampered by the fundamentally random nature of parts of the firing process. The exact same conditions—of ammunition, of wear and cleanliness of firearms parts, of burning of propellant particles and the resulting gas pressure, and so forth—do not necessarily apply for every shot from the same gun. Ultimately, as firearms identification is currently practiced, an examiner's assessment of the quality and quantity of resulting toolmarks and

[21] A cartridge is the entire assembly of projectile (bullet), case, propellant and primer.

[22] *U.S. v. Joseph Thomas*, D.C. Superior Ct., Cr. No. 2007-CF1-25845, Sept. 4, 2009, representations by AUSA Michael Ambrosino, T.tr. of affiant at 28-29.

[23] AFTE is the Association of Firearms and Toolmarks Examiners, a trade association that established guidelines for forensic practice that virtually all examiners follow.

> the decision of what does or does not constitute a match comes down to a subjective determination based on intuition and experience. By comparison, DNA analysis is practically unique among forensic science specialties as having a strong objective basis for determination and as being amenable to formal probability statements.
>
> [Two authors of a study] rank various forensic science subfields on a continuum of relative subjectivity. On the low end of that scale is DNA analysis, along with serology (blood type determination) and drug and narcotic identification. They identify firearms and toolmark identification as having relatively high subjectivity, on par with fiber identification. They identify blood spatter interpretation, voiceprint analysis, and bite-marks as a group of forensic science specialties just slightly more subjective than toolmark identification, and handwriting and hair identification as a cluster slightly more subjective yet.[24]

17. The nation's most prestigious voice of the scientific community and authority in matters of science, the National Academy of Sciences (NAS), concluded that the basic premises of toolmark identifications had not been scientifically established. Two separate NAS committees have rejected the notion that individualization in firearms/toolmarks practice has gained acceptance in the true scientific community. In one report, the NAS concluded, "the needs for research are extensive" and "[a] significant amount of research would be needed to scientifically determine the degree to which firearms-related toolmarks are unique or even to

[24] National Research Council, National Academy of Sciences, report on "Ballistic Imaging," National Academies Press (2008), at 55, available online at 82, available online at: http://www.nap.edu/openbook.php?record_id=12162&page=82.

quantitatively characterize the probability of uniqueness." "Very early in its work the committee found that this question [whether toolmarks are unique] cannot now be definitively answered."[25] The final report of the second NAS Committee observed that "the scientific knowledge base for firearms and toolmarks analysis is fairly limited", and,

> "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."[26] [author's emphasis]

18. Notwithstanding that the premise of uniqueness has not been scientifically established, the ability to differentiate between class, subclass, and individual characteristics is *critical* to support claims of specific source attribution (individualization).

G. Uniqueness & Misleading Expressions of Certainty

[25] *Ballistic Imaging* (National Academies Press 2008), at 3, available at http://books.nap.edu/catalog/12162.html.

[26] *Strengthening Forensic Science in the U.S.-A Path Forward* (National Academies Press, 2009), at 154, available online at: http://www.nap.edu/openbook.php?record_id=12589&page=154.

19. The NAS/NRC 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." Thus, "[c]onclusions drawn in firearms identification should not be made to imply the presence of a firm statistical basis when none has been demonstrated."[27]

20. Claims of individualization (specific source attribution) are inherently probabilistic. The NAS/NRC also noted that "[i]n most forensic science disciplines, no studies have been conducted of large populations to establish the uniqueness of marks or features. Yet, despite the lack of a statistical foundation, examiners make probabilistic claims based on their experience. A statistical framework that allows quantification of these claims is greatly needed."[28] Claims of specific source attribution imply a probability of 100 percent. There is no scientific basis for such claims.

A scientifically acceptable reporting, at this stage of firearms/toolmarks practice development, would be similar to that adopted by The Centre of Forensic Sciences in ending use of the term 'match' in reporting DNA results and testimonies[29]: that 'source A' cannot be excluded as the source of a particular fired

[27] See "*Ballistic Imaging*" (National Academies Press, 2008), p. 3 and 82, at fn. 10, *supra*.

[28] *Strengthening Forensic Science in the United States: A Path Forward*, National Research Council (2009), ISBN 0-309-13131-6, at 6:189.

[29] Memorandum from R.J. Prime, Director of The Centre of Forensic Sciences for the province of Ontario to Crown attorneys for the province, cited by Koehler, J.J., Saks, M.J.,

bullet or cartridge case. This is both accurate and, unlike a word like "match," does not have inherent meaning that is at odds with its evidentiary value. Further, "cannot exclude" can be scaled up and down based on efforts by the individual examiner, and by advances in the field. For example, right now it is generally accepted that firearms examiners can narrow the "pool of tools" that cannot be excluded down to those that share class characteristics (*e.g.* direction of twist and number of rifling grooves).[30]

H. Differentiating 'Individual' From Subclass Characteristics & Likelihood of Adventitious Hits

21. As the firearms examination community has started collecting and storing images of bullet and casing markings, there is evidence that confidence in the premise of uniqueness necessary for source attributions is being undermined by the availability of more data. It is logical that as a database sample size increases, the likelihood of an adventitious 'cold hit' increases, as well, when comparing unknown or questioned specimens with 'knowns' of a sample from the population. The phenomenon has been demonstrated in a study of the Integrated Ballistic Identification System (IBIS), which "...is used successfully with numerous regional..." open case file" databases...[and] performs automated comparisons

"Individualization Claims in Forensic Science: Still Unwarranted," *Brooklyn LR* 75:4 at pre-publication page 4, actual page unknown at this time.

[30] See *Strengthening Forensic Science in the United States: A Path Forward* (NRC) at 154.

between bullets and cartridge cases from different crime scenes and is the cornerstone of the National Integrated Ballistics Information Network (NIBIN) deployed by BATF."[31] This study noted that "...the situation [correlation, or rankings, of firearms considered likely candidates] worsens as the number of firearms in the database is increased"[32] and "...increasing the database size, the ranking of a cartridge case decreases substantially."[33] Likewise, a federal Alcohol, Tobacco, and Firearms toolmark examiner noted that "[a]s the [computer] database grew within a particular caliber, 9mm for instance, there were a number of known non-match test-fires from different firearms that were coming up near the top of the candidate list. When retrieving these known nonmatches on the comparison screen, there were numerous two dimensional similarities."[34] These striking similarities persisted even when the examiner looked at the bullets themselves. "When using a comparison microscope, these similarities are still present and it is difficult to eliminate comparisons even though we know they are from different firearms." *Id.* The phenomenon is not bullet-specific; it assuredly encompasses all firearm

[31] *See* Review: AB1717 report, "Technical Evaluation: Feasibility of a Ballistics Imaging Database for All New Handgun Sales," Dr. Jan De Kinder, Ballistics Section Head, National Institute for Forensic Science, Department of Justice, Vilvoordsesteenweg 98-100, B-1120 Brussels, Belgium.

[32] [De Kinder, *ibid* at 3]

[33] [De Kinder, *ibid* at 21]

[34] Joseph Masson, "Confidence Level Variations in Firearms," *AFTE Journal* 29(1) (Winter 1997).

component comparisons and is logically not restricted to purely automated comparisons.

I. Unfounded Assumption, Uniqueness & 'Individual' Characteristics

22. Firearms/toolmarks examiners' repetitive assertions of various forms of uniqueness, such as "unique signature", "particular weapon", and "no other weapon in the world", *inter alia*, are unfounded assertions. They are unfounded because, as previously discussed, the assumption of uniqueness has not been scientifically established and constitutes nothing more than subjective belief (speculation).[35] They are also misleading because the tools and dies involved in many fabrication processes involving primarily compressive and shear stresses are not sufficiently volatile over time as to change so quickly due to wear that most toolmarks transferred to firearms components are "individual." In reality, the overwhelming majority of toolmarks imparted in various production processes are subclass in nature, not "individual" characteristics. It is paradoxical because it would appear that the position of practitioners is that fabricating tools (many of which use tungsten carbide inserts) change so quickly as to leave "individual" toolmarks on each work piece fabricated, but that the component surfaces of barrels, breech faces, firing pins,

[35] See para. 5, above, and footnotes 5 and 6.

ejectors, and extractors (significantly more vulnerable to wear than tungsten carbide and most tool steels) virtually never change. That position is irrational.

23. Characteristics claimed as "individual" and observed on a cartridge case or bullet, the basis by which toolmarks examiners claim specific source attributions (individualizations), are considered to derive from any of several sources: during manufacturing, subsequent materials handling/processing, and/or during service. Even assuming that discernible extraneous ("individual") characteristics are introduced in the fabrication process, it is difficult to understand, even as a former plant metallurgist, how a forensic examiner far removed from the production process can reliably assess the difference between "individual" characteristics and subclass characteristics imparted during production for the majority of metallurgical processes available. Without personal knowledge of the individual and subclass characteristics produced by a particular manufacturing run, an examiner cannot necessarily tell the two apart, for most forming processes. Except for certain processes, such knowledge must be specific to a particular production run and/or even to aftermarket events. While some examiners have a general knowledge of how firearms are produced, such general knowledge does not provide any information in a significant number of circumstances about whether a particular mark(s) on a bullet

or casing is individual or subclass in nature.[36] As a plant metallurgist, I frequently observed that some of the characteristics imparted by a die and/or during production were intermittent over various runs, and even during a single work piece run in production, such that even if a firearm does not share particular subclass, or what would likely be interpreted as individual, characteristics with a consecutively manufactured firearm, it may share the characteristics with earlier or later work pieces (firearms components in this case) manufactured with the same tooling.

J. Subjectivity of Forensic Firearms/Toolmarks Practice

24. Firearms and toolmark examiners do not have objective criteria for declaring a match, a fact that the Association of Firearms and Toolmark Examiners (AFTE) organization and toolmarks examiner community concede. The focus of a firearms/toolmark examiner is generally on finding *similarities* and dismissing or rationalizing non-matching (dissimilar) characteristics (generally lines) as irrelevant, without compelling objective evidence or scientific explanation to support rejection, in effect selecting the data they wish to use to support identification. They do not employ the 'single dissimilarity exclusion rule' employed

[36] Some class characteristics are similarly difficult to distinguish from subclass or individual characteristics. However, many class characteristics are reliably identifiable, such as the caliber of a bullet or the direction and number of lands and grooves. These are useful pieces of data and can dramatically narrow the range of possible firearms that could have been used to fire a particular bullet or casing. However, these characteristics cannot be used to support the absolute individualization identification claims made by firearm examiners.

in other forensic areas, such as DNA, fingerprints, and even the now-defunct comparative bullet lead analysis (CBLA), where a single dissimilarity required exclusion ("non-ident").[37] The quality of both agreements and disagreements can be difficult to assess, particularly given that the characteristics used for comparison are a generally low combination (3 to 5 in many cases) of non-unique geometric form (lines). Firearms examiners generally do not make exclusions based on dissimilarity of individual characteristics within a field of view under the theory that bullets or casings fired from the same gun may pick up a number of dissimilar individual characteristics. It should be noted that, according to one study, the toolmark examiner typically encountered 15-20 percent matching striations between bullets fired from different firearms of the same manufacturer and type, and 36-38 percent on bullets fired from the same firearm.[38] A more recent work indicates that "...up to 25% of the striae in a non-match and more than 75% of the striae in a match will show concordance."[39] Inasmuch as firearm examinations are largely subjective in nature, each examiner must decide whether the non-matching characteristics viewed should preclude declaration as a match. As noted by one scholar of forensic science,

[37] Even with its numerous flaws, the forensic practice of CBLA had that aspect correct: if any single analyte in one sample was considered dissimilar in quantitative presence to that in another sample, an exclusion was declared.

[38] *See* Biasotti, "A Statistical Study of the Individual Characteristics of Fired Bullets," 4:1 *J.For.Sci.* 34, 34-50 (1959).

[39] *See* Heard, *Handbook of Firearms & Ballistics: Examining and Interpreting Forensic Evidence* (1997).

"[disagreements among toolmarks examiners] stem from one examiner ascribing too much significance to a small amount of matching striae and not appreciating that such agreement is achievable in known non-match comparisons."[40] Notwithstanding the number of AFTE studies, and even if control samples acquired contemporaneously to fabrication were made available to each examiner for a specific examination, inferences of specific source attribution (individualization), would not be generally accepted among materials scientists and forensic scholars given the lack of objective criteria for calling a match.

25. As critical as the skill is in discerning between subclass and individual characteristics, there is no articulated technique purporting to guide examiners in that regard. To rationalize the absence of articulated protocol, literature, and research for such a purported skill, practitioners repeatedly claim that the skill derives from 'training and experience' and that it cannot be explained, hence no articles in the public domain, peer reviewed or otherwise, articulating how to discern purported 'individual' characteristics from subclass characteristics. Such an explanation raises the question as to how toolmark trainers communicate behind closed doors with trainees to recognize the difference between subclass and individual characteristics if instructors cannot articulate such differences in published articles. This is

[40] Faigman, D.L., Saks, M.J., *et al*., *Modern Scientific Evidence: Forensics*, 5:10 at 426: Thomson-West (2008), ISBN 978-0-314-18415-3.

particularly problematic given the numerous acknowledgements by experienced examiners in the literature that "sub-class characteristics can be easily mistaken for individual characteristics"[41] and "…features that produce markings on bullets which may be mistaken as individualizing marks when in fact they are really a more restrictive form of class characteristics."[42]

26. In my reviews of underlying benchnotes and/or worksheets of firearms/toolmarks examiners, I most frequently see no acknowledgement, discussion, or reference to, subclass characteristics or "subclass carryover", considered to be the most critical consideration in eliminating the contribution of manufacturing characteristics ("subclass" characteristics) that are virtually analytically indistinguishable and repeat over hundreds, if not thousands, of firearm components but, rather, observe a direct leap from class characteristics to presumed "individual" characteristics. For example, in the Mississippi Crime Laboratory report provided in this case, I see no discussion, reference, or acknowledgement as to the existence of, subclass characteristics, or implicit indication as to how the examiner differentiated between subclass and purportedly individual characteristics. Although there were no underlying laboratory examiner benchnotes or photographs

[41] Sutton, Gerald, "Assignments and Exercises – Advanced Comparative Macroscopy," training materials in Australia on the technique of CMS (consecutively matching striae).

[42] Moran, Bruce, "Firearms Examiner Expert Witness Testimony: The Forensic Firearms Identification Process Including Criteria for Identification and Distance Determination," *32(3) AFTE Journal*, 231 (2000), at 239.

provided for review, this circumstance is consistent with virtually every other case I've reviewed with firearms/toolmarks benchnotes and testimony, suggesting an "all or nothing" approach, where the examiner presumes that the fabrication process left no subclass characteristics whatsoever and that all the characteristics used for comparison are "individual" characteristics. This presumption is also consistent with the history of firearms/toolmarks examination methodology: in the many decades of its existence, the firearms identification community did not even acknowledge the existence of manufacturing characteristics carrying over to product use until 1989, when it was incorporated into the AFTE guidelines. However, that presumption (of only class and purportedly 'individual' characteristics) remains in many, if not most, firearms identification practice today. It is metallurgically nonsensical to assume that characteristics of manufacture never exist in the final product available to consumers. Such a leap of faith is not unexpected in view of the fervent belief by practitioners in the unproven premise of uniqueness. Subclass carryover is particularly critical to probative value because of the alarming inhomogeneity of product distribution (in other words, geographical concentrations of indistinguishable product) known to exist, for example, in retail marketing of bullets studied by researchers.[43]

[43] See Cole and Tobin, *infra*, at fn. 49 and 50.

27. In fact, barrels produced in the manner described by the FBI firearms identification examiner (by broaching) almost always DO result in subclass carryover (manufacturing characteristics imparted sometimes to relatively large production lots). However, the FBI firearms examiner in this case testified that he had "no idea" as to the identity of the firearm manufacturer, which means he had no idea how the barrel was manufactured.[44] This was due, in part, to the fact that the murder weapon in this case was never found. He further testified that production of the .380 caliber firearm has spanned "many years."[45] There are varying metallurgical methods of producing firearms barrels; all firearms manufacturers do not use broaching to form the internal surface of their barrels. In fact, even the same manufacturer may use different production techniques at different plant locations, and/or even over time. In part because the manufacturer of the gun at issue in this case was not known, the method of barrel manufacture described by Mr. Lewoczko may well be irrelevant. It is [now] well-known in the firearms identification community that it is absolutely essential for a firearms identification examiner to know how a *specific* firearm involved in a case was manufactured before rendering an opinion as to individualization such as was made in the case *sub judice*. A noted toolmark author observed, "The difficulty of addressing subclass characteristics is

[44] T.tr. of John Lewoczko, 1094 at 1-8.
[45] *Ibid* at 12-13.

not in debate."[46] The necessity of determining the manner in which a gun was manufactured is such common knowledge within the firearm identification community that some crime laboratories do not allow firearms examiners to opine an individualization (specific source attribution) unless the firearm to which bullets or cartridge cases are attributed has been recovered and made available for examination.

28. A subtle, and easily overlooked, consideration rendering the practice of toolmark associations even more subjective than is already immediately apparent is the issue of line (striae) quality to which I previously alluded. Line quality is quite significant but is also unquantifiable and inherently subjective.[47] Because of the lack of scientifically acceptable parameters and descriptors to describe 'lines,' toolmark examiners frequently resort to ascribing nebulous and unquantifiable terms in frustrated efforts to give lines some 'character.' In one trial of another defendant, the toolmarks examiner described how he matched lines: "Just one or two fine lines is never going to make it, but if they have some character to them, there is some design

[46] Nichols, Ronald G., "Defending the Scientific Foundations of the Firearms and Toolmark Identification Discipline: Responding to Recent Challenges," *J.For.Sci*, May 2007, Vol. 52 No. 3, at. 587.

[47] This is true notwithstanding efforts by the AFTE community to reduce subjectivity in toolmark associations by introducing an element of quantifiability. For example, a methodology known as consecutive matching striae (CMS) promotes line counting and additionally requires that the 'matching' lines be consecutively occurring. However, a claim that six lines matched is deceptive in that it is perceived as specific, objective, unambiguous and inarguable, but whether each line is of sufficient quality to be included in the count of matching striae remains a subjective determination.

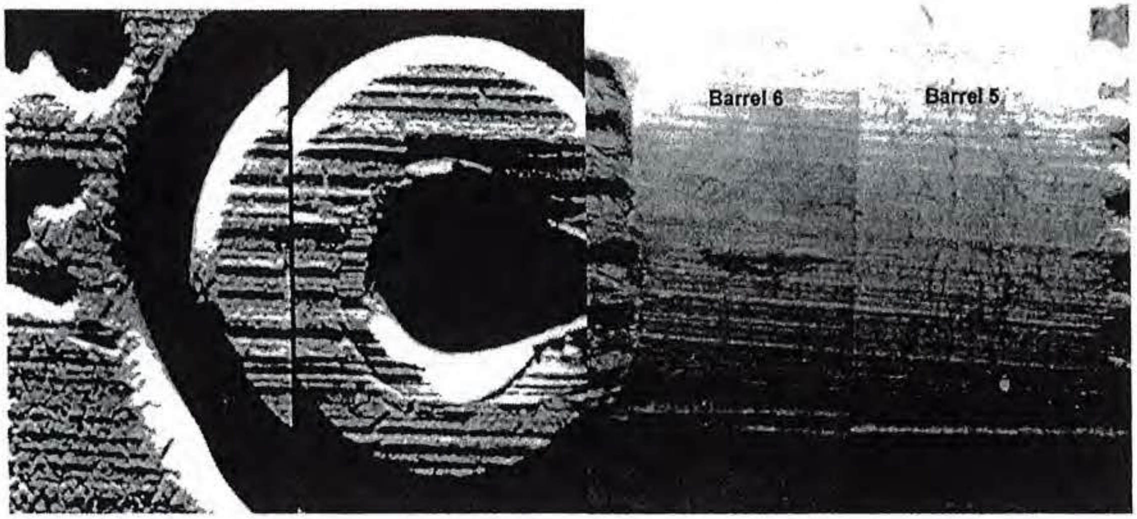
to them, and there are no significant differences between those two areas, then - - [sic]."[48] From the perspective of a metallurgist/materials scientist, representations that a line can have 'character' or 'design' are nonsensical. They are subjective descriptors with meaning only to the observer. They are not quantifiable, reproducible, data that can be conveyed for peer review, nor are they falsifiable, a critical element of the scientific method.

29. To demonstrate the unreasonableness of drawing on subjective recollection of samples presented in temporally remote settings, the following cartridge case and bullet comparisons, respectively, were presented in the AFTE literature as having been fired from *different* firearms which were contemporaneously fabricated.[49] The cartridge cases (left photo) and bullets (right photo) are virtually indistinguishable within each pair of photos and quite vulnerable to false positive associations if not presented for forensic examinations contemporaneously and/or the characteristics were the only ones used for identification (the most common occurrence). In my experience and opinion, it is quite likely they would be declared to be matches under either classical pattern-

[48] Testimony of firearms/toolmarks examiner Jon Kokanovich, Dec. 3, 1992, in *re State of Arizona v. Anthony Spears*, Maricopa County Superior Court Case CR92-90457, T.tr. at 855.

[49] *Left*: Rivera, Gene C., "Subclass Characteristics in Smith & Wesson SW40VE Sigma Pistols", *AFTE J.* 39(3), Summer 2007, at 247. *Right*: Tulleners & Hamiel, "Sub Class Characteristics of Sequentially Rifled 38 Special S&W Revolver Barrels", *AFTE J.* 31(2), Spring 1999 at 118.

matching (likely used in this case) or consecutively matching striae (CMS) methodology (generally on the West coast of the U.S.):



Overwhelming concordance of striae between cartridge cases (left pair with split screen image) and bullets (right pair in split screen image) fired in/from different consecutively manufactured firearms. (See Footnote 49)

Similar to the basis for the eventual demise of the forensic practice of comparative bullet lead analysis, the possibility and number of similarly manufactured firearms distributed in a local or regional community that could easily be confused, most particularly when not available for direct examination and comparison, unquestionably affects probative value of a claimed "match." This is the principal reason why some crime laboratories do not allow examiners to opine "same firearm" based on comparisons of recovered bullets or cartridge cases absent recovery of a questioned firearm for direct comparisons. Thus, in some jurisdictions, Mr. Lewoczko's opinion would be systemically barred.

K. Geographic Distribution of Highest Likelihood Coincidental Match Firearms

30. Particularly given that the assumption of uniqueness has not been scientifically established and that individualization is considered a fallacy in the scholarly scientific and forensic communities, a major determinant of possible probative value of even acceptably founded evidentiary product associations is the issue of product density and distribution. Firearms products are likely not uniformly dispersed throughout the U.S. Instead, there may be some clustering effects similar to those found by bullet lead researchers, where groups of firearms produced at the same time and in the same manufacturing process cluster in one particular area or region.[50] In assessing probative value of the significance of matching characteristics on a bullet or casing, one should take into account the density and distribution patterns of the particular type of firearm and, indeed, even the prevalence of other similar caliber firearm types, in a particular region. Firearms/toolmarks examiners have apparently never determined, or even attempted to test, claimed probative value of purported source attributions. Exhaustive literature search reveals no studies of the density and distribution patterns of firearms to assess probative value. This is

[50] *See* "A Retail Sampling Approach to Assess Impact of Geographic Concentrations on Probative Value of Comparative Bullet Lead Analysis," S.A. Cole, W.A. Tobin, L. Burgess, H. Stern, *Law, Probability and Risk*, Vol. 4, No. 4 (2005), Oxford University Press (probabilities of 1 were found in some geographic areas for some product lines, meaning that consumers had no choice but to purchase the same packing coded (composition) bullets even if they wanted others).

assuredly attributable to their intuitive belief in the unvalidated premise upon which they rely - - that each firearm exhibits and transfers 'unique' toolmarks. It should be noted that proponents of comparative bullet lead analysis (CBLA) maintained the same position for almost 40 years with regard to the similar underlying premise of uniqueness, and also of probative value, until relatively recent research proved CBLA invalid, misleading as proffered, and without evidentiary value (forensically meaningless).[51]

31. Realizing that the courtroom is not a laboratory, even the "test of time," also known as "implicit testing," as suggested by the lengthy admissibility of toolmarks testimony and conclusions, is not a valid measure of practice validity or rate of error because of the absence of effective feedback loop for expert witnesses testifying to exclusive source attributions. The most recent forensic committee of the NAS has observed,

> For years in the forensic science community, the dominant argument against regulating experts was that every time a forensic scientist steps into a courtroom, his work is vigorously peer reviewed and scrutinized by opposing counsel. A forensic scientist might occasionally make an error in the crime laboratory, but the crucible of courtroom cross-

[51] *See* "Comparative Bullet Lead Evidence (CBLA): Valid Evidence or *Ipse Dixit*?," E.J. Imwinkelried and W.A. Tobin, *Okla. City Univ. LR*, Vol. 28 No. 1 (2003). *Cf.*, "A Retail Sampling Approach to Assess Impact of Geographic Concentrations on Probative Value of Comparative Bullet Lead Analysis," S.A. Cole, W.A. Tobin, L. Burgess, H. Stern, *Law, Probability & Risk*, Vol. 4, No. 4 (2005), Oxford University Press, and FBI Press Release (where FBI concedes lack of probative value) dated Sept. 1, 2005, available at http://www.fbi.gov/pressrel/pressrel05/bullet_lead_analysis.htm.

> examination would expose it at trial. This "crucible," however, turned out to be utterly ineffective...
>
> Unlike the extremely well-litigated civil challenges, the criminal defendant's challenge is usually perfunctory. Even when the most vulnerable forensic sciences—hair microscopy, bite marks, and handwriting—are attacked, the courts routinely affirm admissibility citing earlier decisions rather than facts established at a hearing. Defense lawyers generally fail to build a challenge with appropriate witnesses and new data. Thus, even if inclined to mount a *Daubert* challenge, they lack the requisite knowledge and skills, as well as the funds, to succeed.[52]

32. Source attributions such as the one used by the firearms examiner in this case are without scientific foundation. Inferences, implications and assertions of "to the exclusion of all others," "no other weapon in the world", "it's like your fingerprints are to you,"[53] "these bullets were all fired from one barrel,"[54], and repeated claims of "unique",[55] and "individual"[56], inherently imply a statistical basis (and a high degree of certainty) that scientists do not accept and do not believe has been established. In the FBI examiner's testimony in this case, such references were redundantly pervasive: 22 such references over only 9 pages of substantive testimony.[57] An earlier NAS report concluded that, "Conclusions drawn in firearms

[52] *Strengthening Forensic Science in the United States: A Path Forward*, National Research Council, National Academy of Science (2009), at 107, available online at: http://www.nap.edu/openbook.php?record_id=12589&page=106

[53] T.tr. of FBI Examiner John Lewoczko, 1092 at 15.

[54] *Ibid* at 16.

[55] *Ibid,* various, but *e.g.* 1089 at 13.

[56] *Ibid*, various, but *e.g.* 1090 at 19.

[57] The 22 characterizations used by Lewoczko are inherently probabilistic (implying probability of 1 and, thus, a certainty) and, accordingly, they are without scientific foundation. The characterizations appear in the transcript as follows:

identification should not be made to imply the presence of a firm statistical basis when none has been demonstrated."[58] In particular, the NAS report on Ballistic Imaging was concerned about testimony cast "in bold absolutes" such as that a match can be made to the exclusion of all other firearms in the world: "Such comments cloak an inherently subjective assessment of a match with an extreme probability statement that has no firm grounding and unrealistically implies an error rate of zero."[59] From a forensic materials science perspective, the analogical reference to fingerprints is particularly egregiously misleading because fingerprints derive characteristics for forensic comparisons as a result of generally random biological

1088 at 10: "same firearm"
1089 at 13: "unique"
1089 at 13: "individual"
1090 at 4: "unique"
1090 at 8: "unique"
1090 at 19: "individual"
1090 at 22: "same firearm"
1091 at 4: "individual"
1091 at 14: "same firearm"
1091 at 16: "exclusion of every other firearm in the world"
1092 at 3: "same firearm"
1092 at 5: "individual"
1092 at 17: "same barrel"
1092at 10: "individual"
1092 at 11: "exact same"
1092 at 13: "exclusion of every other firearm"
1092 at 16: "one barrel"
1095 at 13: "individual"
1095 at 13: "unique"
1095 at 16: "individual"
1096 at 3: "same…"
1096 at 4: "one barrel"

[58] National Research Council, National Academy of Sciences, "Ballistic Imaging," 82 (2008).

[59] "The NAS Report And Its Implications for Criminal Litigation," Paul C. Giannelli, *Jurimetrics*, April 22, 2009 [May 6, 2009], citing NRC/NAS Report on Ballistic Imaging 82 (2008).

processes. Plant metallurgists generally take great pains to insure that their processes are anything BUT random, to insure continuity of production, maximum tool/die life, and suitability-for-service quality. Thus it is expected, and my experience as a plant metallurgist confirms, that there would be a very high degree of characteristic continuity within production runs.

L. Known Misattributions (Type I Errors: False Positives) & Error Rates

33. There have been numerous indications of disagreements, misidentifications, and various rates of error exceeding the oft-quoted "0.1 percent," "zero," "0 to 1%", "1-2%", or "near zero," rates of error. It has been indicated in the field literature that disagreements typically arise from an examiner[s] ascribing too much significance to small amounts of matching striae that are achievable in known non-matches.[60] This observation is not surprising given that up to 51 percent matching lines have been found in known non-matches, that examiners have differed by 39 "matching" lines, and other similar findings.[61] Further, it is known that not every manufactured tool is unique.[62] With regard to error rates, scholars have noted

[60] For an example, see Faigman, D, Kaye, D., Saks, M., *Modern Scientific Evidence: The Law and Science of Expert Testimony* (2002),

[61] "Criteria for Identification of Toolmarks, Part II: Supporting the Conclusion," Miller and Neel, *AFTE Journal*, Winter 2004, Vol.36 No.1.

[62] *Faigman, et al., supra*, at 500-501.

that with modern statistics, forensic examiner decision-making could, but to date has not been subjected to quantitative analysis.[63]

34. The forensic community does not engage in error detection. Most errors are discovered fortuitously. In spite of claims by the AFTE community that errors are rare, there are numerous references discussing the existence and frequencies of examiner error in the firearms/toolmarks literature. Several are by individuals formerly in capacities as crime lab and/or heads of firearms/toolmarks units, both in positions of being requested to adjudicate or arbitrate differences of opinion in the field of firearms/toolmarks. One states that there have been an "appalling number of misidentifications" in the firearms ID field, and discusses one American Association of Forensic Scientists (AAFS) funded study by the Law Enforcement Assistance Administration (LEAA) that found 24 percent of labs turning in unacceptable results.[64] The second, a former Unit Chief of the FBI Laboratory's Firearms/Toolmarks Unit, speaking in an AFTE training seminar, indicated that "most of us know someone who has committed serious error" and that examiners "might not be allowed to forget such error if it becomes public knowledge."[65] He proceeded to describe a false positive identification on a 1911A1 .45 caliber

[63] *Ibid*, at 508, *inter alia*.

[64] Bradford, Lowell, "Forensic Firearms Identification: Competence or Incompetence?," *AFTE Journal*, Vol. 11 No. 2 (April 1979).

[65] Hodge, Evan, "Guarding Against Error," 20(3) *AFTE Journal* (July 1988).

semiautomatic and mentions that although "[his] FBI" was not the only entity that did referee work (reviewing cases), it did enough to know that the described false positive case was "only one of many we have seen over the years."[66]

35. One study has indicated that 9.1 percent of firearms examiners responding to a proficiency test were "clearly in error" and noted that even more examiners gave unacceptable results.[67]

36. An account of one of the false positive identifications against a law enforcement officer serves as an overview summary of the problematic nature of the subjective forensic practice of firearms/toolmark identifications. In the prosecution of a sheriff's deputy, similarities were observed between marks on casings and bullets by a Los Angeles Police Department forensic examiner claiming a "match" but, according to independent experts, undue significance was attributed to those similarities. The marks were eventually determined to be coincidental and insufficient to support an identification (and, in fact, reportedly showed an exclusion). One of the independent examiners noted that, "It was clearly an exclusion. It was not an identification at all. It was flat-out error on the part of LAPD."[68] According to noted firearms/toolmarks authority John Murdock,

[66] *Ibid.*

[67] Jonakait, Randolph N., "Forensic Science: The Need for Regulation," *4 Harvard Journal of Law & Technology*," 109 & n.8 (1991).

[68] "Review of LAPD Ballistics Unit Set After Botched Test in Murder Case: Charges Dropped Against LA Sheriff's Deputy," *Law Enforcement News*, Vol. XV, No. 295 (June 30, 1989) at 7.

"Firearms identification is an area that is somewhat problematic in forensic science, because the determinations that are made are mostly subjective in nature and they're based on the experience of the examiner. Let's face it, an examiner can be around for a number of years and not have the right kinds of experience."[69]

37. A relatively recent audit of the Detroit Crime Laboratory in Detroit, Michigan, reportedly revealed a shocking 10 percent rate of error attributable to human error, not malfeasance. "Firearms work at the city police lab…first was halted in the spring [2008]…" after suspicion of an unacceptable rate of error in firearms/toolmark examinations. In one case, the lab had reportedly determined that 42 shell casings from a May 2007 shooting were fired by the same weapon. State police later determined that two different weapons were used.[70]

38. Other significant misattributions have been fortuitously discovered. In *Trotter v. Missouri*, a police officer was killed at the scene of a shooting. Investigators originally believed that the officer was shot with his own weapon, but the weapon was not found at the scene. A suspect involved in a completely different criminal matter was arrested and a firearms/toolmarks examiner "matched" the suspect's weapon to the bullet recovered from the deceased officer. Sometime later, the deceased officer's weapon was eventually found and it was confirmed to have

[69] *Ibid.*

[70] "Error Prone Detroit Crime Lab Shut Down," *USA Today*, 9/25/2008, available online at: http://www.usatoday.com/news/nation/2008-09-25-crime-lab_N.htm, *inter alia*.

been the weapon used to kill the officer, as admitted by the original firearms/toolmarks examiner.[71] Yet another misattribution was revealed in the matter of *Williams v. Quarterman*, where a firearms/toolmarks examiner testified *to an absolute certainty* that a bullet was fired from a certain .25 cal. pistol. It was eventually determined to have been fired from a .22 caliber pistol owned by another individual.[72]

39. As discussed in my latest paper, coauthored with a nationally known tribology research scientist at Oak Ridge National Laboratory, for numerous reasons, the forensic practice of firearms identification is not a science and is virtually entirely subjective. There exist no comprehensive or meaningful studies validating the ability of a firearms identification examiner to reliably opine source attributions.[73] Additionally, it has already been determined by a variety of empirical studies and incidents, that rates of firearms identification practice error significantly exceed the often claimed "1-2%" used by the State to claim that "there exists a 98-99% certainty that the bullets were properly identified" in this case.[74] There is no scientific foundation for such a claim. Ironically, for decades, firearms identification experts

[71] *Trotter v. Missouri*, 736 S.W.2d 536.

[72] *Williams v. Quarterman*, 551 F.3d 352 (5th Cir. 2008).

[73] Tobin, W.A. and Blau, P.J., "Hypothesis Testing of the Critical Underlying Premise of Discernible Uniqueness in Firearms/Toolmarks Forensic Practice", 53 *Jurimetrics J.*, 121-142 (Winter 2013). The paper is attached as Exhibit E-2.

[74] State's "Motion To Supplemental To Motion To Stay Execution…", page 9, line 17.

claimed "0% error" (infallibility) in opining specific source attributions until true (mainstream) scientists began calling such absurd claims into question.

40. In summary, the forensic practice of firearms/toolmarks associations lacks the rigor of science and should not be permitted to render inferences of specific source attribution (individualization), unfounded expressions of certainty of any kind, or other conclusions implying an aura of precision generally associated with scientific endeavor, without comprehensive or meaningful scientific foundation. At the time of Manning's trial, and even currently, the strongest opinion that is scientifically defensible is that, *in the examiner's opinion*, [either] the characteristics exhibited by the questioned bullets were *consistent* with having been fired from the same weapon [or], alternatively, that the possibility that the questioned bullets were fired from the same firearm could not be eliminated.

Further Affiant sayeth not.

William A. Tobin
William A. Tobin

Subscribed and sworn to before me
this ________ day of May, 2013.

____

Notary Public, State of ____________

My commission expires: ______________

Curriculum Vitae of

William A. Tobin

-- Educational --

Bachelor of Science, Metallurgy, Case Institute of Technology
Master of Arts, Special Studies, George Washington University
Graduate studies, Materials Science & Engineering, University of Virginia

Additional Courses & Symposia
Physical Metallurgy, Ohio State University
Shaping, Forming of Metals, Ohio State University
Engineering Metallurgy, Ohio State University
Principles of Failure Analysis, American Society for Metals (ASM)
Fractography: Practical Applications in Failure Analysis (ASM)
Metallographic Interpretation (ASM)
Energy Dispersive X-ray Fluorescence, Kevex Corporation
Statistics I, Northern Virginia Community College
Statistics II, Northern Virginia Community College
Detection and Recovery of Human Remains, FSRTC
Calculus I (refresher), Northern Virginia Community College
Calculus II (refresher), Northern Virginia Community College
Applied Statistics for Engineers and Physical Scientists, Va. Commonwealth Univ.
Structure and Properties of Materials, University of Virginia
Fastener Characterization by Mechanical & Metallographic Methods
Manufacturing Processes & Materials, University of Virginia
Applied Electrochemistry, University of Virginia
Explosion Effects & Structural Design for Blast
Metallurgy of Ductile Iron, American Foundry Society

-- Professional Experience --

Battelle Memorial Institute, Research Metallurgist
Chase Brass and Copper Company, Plant Metallurgist
National Aeronautics and Space Administration, Research Metallurgist
Monarch Aluminum Company, Manufacturing/Production Process Control
U.S. Marine Corps, Platoon Commander, Republic of South Vietnam
Federal Bureau of Investigation, Supervisory Special Agent
Manager of forensic metallurgy operations, FBI Laboratory
Forensic Engineering International, Principal

-- Court Appearances and Depositions --

Testified as an expert witness in 302 local, state, federal criminal & civil matters, in 46 states, D.C., P.R. (excl. Congressional testimonies & grand juries).

Exhibit B

-- *Commendations* --

- ♦ Bronze Star with Combat 'V', U.S. Marine Corps
- ♦ 2 Crosses of Gallantry, Republic of South Vietnam
- ♦ 20 additional military combat decorations

Numerous letters of commendation, including:

- Personal commendation from U. S. Attorney General William French Smith
- Three commendations with cash awards, from FBI Director William H. Webster
- Two commendations and cash award from FBI Director William S. Sessions

-- *Professional Affiliations* --

National Association of Corrosion Engineers (NACE, now AMPP)
Statistical & Applied Mathematical Sciences Institute (SAMSI)
American Society for Testing and Materials (ASTM)
American Society for Metals, International (ASM)
The Minerals, Metals & Materials Society (TMS)
National Fire Protection Association (NFPA)
Society for Experimental Mechanics (SEM)
International Metallographic Society (IMS)
American Foundry Society (AFS)
Failure Analysis Society (FAS)
1st Marine Division Association

-- *Literary Acknowledgments / References / Media* --

And The Sea Will Tell, Vincent Bugliosi, Ballantine Books, 1992; former prosecutor of Charles Manson and author of *Helter Skelter*.

Bones, Dr. Douglas Ubelaker (Smithsonian Institution) and Henry Scammell; Harper Collins Publishers, 1992, New York, NY.

Hard Evidence, David Fisher, Simon & Schuster, 1995; author of bestsellers *Gracie With George Burns*, *What's What*, *Killer*, and *The Umpire Strikes Back*.

"60 Minutes", CBS televised interview November 18, 2007; re-aired Sept. 14, 2008

-- *Other* --

Referee for *Fire Technology*, NFPA

Editorial Advisor, *The Forensic Examiner*, ACFEI

Requested by UNSCOM to serve as U.N. Weapons Inspector, Iraq (1998)

Editorial Reviewer, National Research Council, National Academy of Sciences

-- Publications --

(1) **Evidentiary Comparison of Plastic Materials and Products Based Upon Fabrication Characteristics** (Toolmarks), F.S. DeRonja and W.A. Tobin, *Proceedings of the International Symposium on the Analysis and Identification of Polymers*, July 31 to Aug. 2, 1984, FBI Academy, Quantico, Virginia.

(2) **Collapsed Springs in Arson Investigation: A Critical Metallurgical Evaluation**, W.A. Tobin & K.L. Monson, *Fire Technology*, Volume 25, Number 4 (November 1989), National Fire Protection Association.

(3) **Arson Investigations**, W.A. Tobin, *Law Enforcement Bulletin* ('Focus' feature), February 1990, Federal Bureau of Investigation.

(4) **What Collapsed Springs Really Tell Arson Investigators**, W.A. Tobin, *Fire Journal*, Volume 84, No. 2 (March/April 1990), National Fire Protection Association.

(5) **What Collapsed Springs Really Tell Arson Investigators**, W.A. Tobin; course instructional material, *Fire/Arson Investigation Resident Course*, October 1994, U.S. Fire Administration, National Fire Academy; requested and reprinted with permission.

(6) **Noninvasive Evaluation of Vehicular Lampbulbs**, W.A. Tobin, *Crime Laboratory Digest*, Volume 21, Number 1 (January 1994), Federal Bureau of Investigation.

(7) **Noninvasive Evaluation of Vehicular Lampbulbs**, W.A. Tobin, *Forensic News*, April-June 1994, Arizona Identification Council, Division of the International Association for Identification; reprinted with permission.

(8) **FBI Investigates Aircraft Corrosion** (submitted as "Aircraft Corrosion in Law Enforcement"), W.A. Tobin, *Materials Performance*, Volume 33, Number 6 (June 1994), National Association of Corrosion Engineers (NACE).

(9) **Inferring Duration of Exposure to a Hostile Environment Based on Measurement of Corrosion Product Thickness**, W. A. Tobin, *The Customs Laboratory Bulletin*, Volume 7, Number 1 (1995), U.S. Customs Service, S. M. Dyszel, Ed., Washington, D.C.

(10) **A Metallurgical Review of the Interpretation of Bullet Lead Compositional Analysis**, E. Randich, W. Duerfeldt, W. McClendon, W. Tobin, *Forensic Science International*, Volume 127, Issue 3 (September 2002), pp.174-191, Elsevier Science Publishing.

(11) **How Probative is Comparative Bullet Lead Analysis?**, W. A. Tobin, W. Duerfeldt, *Criminal Justice*, Volume 17, Number 3 (Fall 2002), pp.26-34, American Bar Association.

(12) **Comparative Bullet Lead Evidence (CBLA): Valid Evidence or *Ipse Dixit*?**, E. J. Imwinkelried and W. A. Tobin, *Oklahoma City University Law Review*, Vol. 28 No. 1 (2003), pp.43-72.

(13) **Comparative Bullet Lead Analysis: A Case Study in Flawed Forensics**, Tobin, W.A., *The Champion*, July 2004, pp.12-22, National Association of Criminal Defense Lawyers.

(14) **A Retail Sampling Approach to Assess Impact of Geographic Concentrations on Probative Value of Comparative Bullet Lead Analysis,** S.A. Cole, W.A. Tobin, L. Burgess, H. Stern, *Law, Probability & Risk*, Vol. 4, No. 4 (2005), Oxford University Press.

(15) **Evaluating and Challenging Forensic Identification Evidence,** W.A. Tobin, W.C. Thompson, *The Champion*, July 2006, pp. 12-21, National Association of Criminal Defense Lawyers.

(16) **Expert Opinion: Evidentiary Value**, Chapter 8: “Evaluating and Challenging Forensic Identification Evidence”, W.A. Tobin, W.C. Thompson, reprinted with permission, pp. 137-160; The Icfai University Press, Hyderabad, India (2007).

(17) **Chemical and Forensic Analysis of JFK Assassination Bullet Lots: Is A Second Shooter Possible?,** C. Spiegelman, W. A. Tobin, William D. James, Simon J. Sheather, Stuart Wexler, D. Max Roundhill, *The Annals of Applied Statistics*, Vol. 1 No. 2, 287-301 (2007); Institute of Mathematical Statistics. http://dx.doi.org/10.1214/07-AOAS119 or http://arxiv.org/abs/0712.2150. Winner of "2008 Statistics in Chemistry Award" from American Statistical Association with cash award.

(18) **Analysis of Experiments in Forensic Firearms/Toolmarks Practice Offered As Support for Low Rates of Practice Error & Claims of Inferential Certainty**, C. Spiegelman, W. A. Tobin, *Law, Probability and Risk*, (2013) 12 (2), 115-133, doi:10.1093/lpr/mgs028, first published online at http://lpr.oxfordjournals.org/cgi/content/full/mgs028?ijkey=ebC7b3Y008FdKhU&keytype=ref on October 1, 2012.

(19) **Hypothesis Testing of the Critical Underlying Premise of Discernible Uniqueness in Firearms-Toolmarks Forensic Practice**, W. A. Tobin, P. J. Blau, 53 *Jurimetrics* 121-146 (Winter 2013), available at: http://www.ssrn.com/author=1521077.

(20) **Absence of Statistical and Scientific Ethos: The Common Denominator in Deficient Forensic Practices,** W. A. Tobin, H. D. Sheets, C. Spiegelman, *Statistics and Public Policy*, 4:1, 1-11 (2017), DOI: 10.1080/2330443X.2016.1270175, available at http://dx.doi.org/10.1080/2330443X.2016.1270175. Biannual 'Editors' Choice' selection, with cash award, of the American Statistical Association (ASA) at http://explore.tandfonline.com/content/pgas/asa-editors-choice?utm_medium=email&utm_source=EmailStudio&utm_campaign=JMH01962_2544310.

-- Noteworthy / Sample Cases --

TWA 800 Aircraft Disaster: Mid-air explosion of flight TWA 800 enroute from New York's Kennedy Airport to Paris, France, on July 17, 1996.

Mid-air Breakup of Missouri Air National Guard F-15C: Crash of F-15C from longeron fatigue failure resulting in nationwide grounding of all F-15A/B/C/D aircraft.

Sago Coal Mine Disaster*:* Complex materials interaction issues relating to methane gas explosion, Sago, W.V., January 2, 2006. Thirteen trapped miners; one survivor.

U.S. v. [Blackwater Worldwide personnel]**;** Incident involving Blackwater Personal Security Detail (PSD) in September 2007 escorting convoy of U.S. State Department vehicles en route to meeting in western Baghdad with USAID officials, resulting in 17 Iraqi civilian fatalities in Nisour Square, Baghdad.

U.S. v. Aafia Siddiqui; Trial of Dr. Aafia Siddiqui for attempted murder with M4 rifle; trial in NYC, NY (from Guantanamo). Terminal ballistics and GSR (gunshot residue) issues from shooting reconstruction: evaluation of wall damage claimed to be bullet holes from high velocity impact of M855 (SS109) projectiles (bullets).

Olympic Park Bombing: Pipe bomb explosion at Centennial Park, Atlanta, GA, during 1996 Olympics.

Charles Stuart: National notoriety and local racial strife in Massachusetts resulting from incident where Stuart and his pregnant wife were shot in their vehicle; Stuart called "911" from his vehicle while wounded. Notoriety resulted in TV movie "Good Night, Sweet Wife" (CBS) and several books.

U.S. v. Walter Leroy Moody: Defendant sentenced to 7 life terms plus 400 years for mailing package bombs that killed U.S. Appellate Court Judge Robert S. Vance and civil rights attorney Edward Robinson.

USS Iowa: Explosion aboard ship that killed numerous sailors during training operation.

Susan B. Anthony silver dollar recovery. Developed technique adopted by U.S. Mint to recover thousands of mis-minted silver dollars embedded in Lucite for collectors.

U.S. v. Joseph Earl Meling; Product tampering of SUDAFED capsules; defendant convicted of contaminating capsules with sodium cyanide to murder his wife, causing the deaths of several consumers purchasing SUDAFED from store shelves.

Girl Scout Cookie Tampering; Nationwide alert for contaminated Girl Scout cookies.

Train Derailment, Panama City, FL; 129 car derailment releasing chlorine gas causing deaths of 8 people. Incident featured in *Newsweek* and numerous other news periodicals.

Wilberg Coal Mine Explosion, Orangeville, UT; coal mine explosion of such severity that it took approximately two years to recover bodies of 27 miners who died in the mine.

Scaffold Collapse, Willow Island, WV; Wire rope failure that caused collapse of scaffold used in construction of nuclear facility, resulting in 51 deaths, many from same family.

U.S. v Buck Walker & U.S. v. Stephanie Stearns; "Hippie" couple charged with murders of Malcolm ("Mac") & Eleanor ("Muff") Graham on Palmyra Island in the South Seas. Skull found by beachcomber on deserted beach in the South Seas 12 years later, depicted on the cover of *And The Sea Will Tell* by Vincent Bugliosi (author of *Helter Skelter* and prosecutor of Charles Manson); subject of popular TV movie "And The Sea Will Tell" (CBS) aired numerous times.

Lt. Colonel William Higgins, Commander of U. N. Forces, kidnapped and killed, Beirut, Lebanon.

Achille Lauro Cruise Ship; Terrorism aboard cruise ship.

Judge Alcee Hastings; impeached Federal judge accused of misconduct and obstruction of justice. Tensile testing and failure analysis of purse strap carried by Judge Hastings.

Train Derailment, Mobile, AL; Derailment of the "Sunset Limited," worst rail disaster in U.S. history, resulting in the deaths of 47 passengers on September 23, 1993.

Environmental Disaster; Oil spill, Escambron Beach, San Juan, Puerto Rico, January 7, 1994, involving motorized vessel (M/V) Emily S. (tug) and barge Morris J. Berman, with 662,000 gallons of #6 fuel oil.

UNABOM; Sixteen package bombs sent to/opened by various technical personnel.

Oklahoma City, OK; Bombing of Murrah Federal Building on April 19, 1995.

Patent Infringement Litigation: *Brunswick v. U.S. Army*; materials design of radar-scattering camouflage netting used by U.S. Army in Kuwait-Iraq conflict. Devised unique testing technique to determine spatial relationship of critical component fibers for U.S. Department of Justice.

Auto Accident Due To Roadway Debris; Tragic automobile accident caused by 50-lb. steel plate falling from commercial truck under tow, nearly decapitating victim driver in vehicle behind tow, September 2000, causing massive I-95 traffic stoppage. Cause: poor maintenance and defective weldment on battery compartment of truck under tow.

Dogwood Elementary School Fire, Reston VA; Elementary school fire resulting in total destruction of school ($17 million loss), November 2000. Unsolved by fire investigators for many months. Forensic metallurgical assistance provided to Fairfax County Fire & Rescue; cause of fire determined to be defective ceiling-hung clock.

COLLAPSE OF BUCKET TRUCK BOOM ARM; bucket truck boom arm, used to trim and clear tree limbs from vicinity of electrical power lines in Warrenton, Virginia, collapsed during use, November 2000. Failure attributable to defective manufacturing technique (weldment).

BICYCLE FATALITY; Moped conversion bike, with caliper hand brakes, became uncontrollable when brakes were applied, causing rider fatality from ejection over handlebars. Loss of control attributable to improper bicycle modification.

VEHICULAR FATALITY; Driver stopped on Interstate 95 with mechanical problems killed by commercial truck while awaiting roadside assistance. Metallurgical examinations confirmed that disabled vehicle's lights, including emergency flashers, were incandescent and visible at time of truck impact.

CORROSION: Premature condenser tubing failures. Internationally-known construction contractor experienced through-wall corrosion of stainless steel condenser tubing within one year of construction for utility client in Colombia, South America. Three metallurgical entities disagreed as to cause but all concluded microbiologically induced corrosion (MIC) involved either as proximate or related cause. Indisputable determination of cause: improper heat treatment of tubing; MIC not involved.

CORROSION: Determination of cause and fault for metal building corrosion of roof that had been installed one year earlier; Wilmington, NC.

CORROSION: Determination of cause & fault for metal building roof corrosion, Annandale, VA.

CORROSION: Determination of cause & fault for multi-million-dollar power generating trailers for large-scale emergency power, Wheeling, IL.

CORROSION**:** Determination of cause & fault for multi-million-dollar firing range automatic target system failures installed one year earlier for FBI Quantico training facility.

CORROSION**:** Authentication of disputed origin of historical artifact: “six-shooter” firearm (pistol) owned by leader of last famous outlaw gang in the U.S.: the ‘Bob Dalton Gang’.

CORROSION**:** Numerous cases of corrosion in marine and industrial environments, including environmental disaster at Escambron Beach, Puerto Rico, mentioned above, where corrosion played integral role in sequence of disastrous events.

CORROSION**:** Determination of cause & fault for pervasive corrosion of expensive paper testing laboratory instruments for well-known paper manufacturer in Richmond, Virginia.

EXPLOSION failure of chamber used for demilitarization processes at Army Research Laboratory (ARL), Aberdeen Proving Grounds (APG), Aberdeen, MD.

TRAJECTORY ANALYSIS: Terminal ballistics shooting reconstruction analysis of military engagement in jungle/rain forest by Philippine Army near Kananga, Leyte, P.R., November 15, 2010, where over 245 military bullets (5.56 mm M193 in M-16 rifles) were expended. Victims of mistaken identity were unarmed and engaged in scientific research, including renowned botanist, Dr. Leonardo Co.

YACHT FASTENER FAILURE**:** Fatigue failure of threaded fastener aboard “Destiny” yacht attributable to deficient design, manufacturing, construction, installation practice.

MISCELLANEOUS: Work ladders; hunter’s tree stand; wire rope & cables; fire sprinkler system corrosion; foundry & casting matters; obliterated serial number & identification marking restorations; oil drilling equipment; fasteners (nails, screws, staples, bolts, nuts, *etc*.); missile guidance system components (radar waveguides); aircraft, boat and ship corrosion; aviation components; false claims act; fraud against the government; automobile accidents & components (fractures, failures, speedometer, headlights, taillights, *etc*.); timing mechanisms (clocks, watches, *etc*.); manufacturing processes; statistical process control; metal building corrosion; mine disasters; transport disasters (maritime, aviation, rail); quality control; standards & specifications; welding; fires & explosions; M4 launch and penetration mechanics with M855 (SS109); terminal ballistics; gunshot residue (GSR); bullets; firearms; toolmarks; body armor; apple brandy alembic in distilling process; operational failures/explosions of firearms during user use (*aka*, ‘kaBooms’).

Various cases featured on "America's Most Wanted", "Unsolved Mysteries", "60 Minutes", "20/20", "Dateline", "Primetime", "Eye to Eye", "48 Hours", “Forensic Files”, “FBI Files”, “The Discovery Channel”, “The Learning Channel”, CNN, Canadian Broadcasting Corp.(CBC), British Broadcasting Company (BBC), and National Geographic Channel.

-- Speaking Engagements --

ASM, COMS (Central Ohio Metallographic Society), Columbus, OH
Ohio State University
Welding & Testing Technology 8th Annual National Conference (31 professional societies, Knoxville, TN)
ASM, Philadelphia, PA (Liberty Bell Chapter)
MTI (Metal Treating Institute), Secaucus, NJ
ASM, Hartford, CT
ASM, Bethlehem, PA
ASM, New Haven, CT
ASM, Nashua, NH
ASM, York, PA
ASM, Charlotte, NC
ASM, Cincinnati, OH
AWS (American Welding Society), York, PA
University of Pittsburgh
ASNT (American Society for Nondestructive Testing), ASM, Hampton, VA
AWS, ASM, Houston, TX
ASM, Peoria, IL
AWS, Los Angeles, CA
AWS, Baltimore, MD
AWS, Hampton, VA
ASM, Baltimore, MD
ASM, Washington, DC
ASM, Johnson City, TN
ASM, South Bend, IN (Notre Dame Chapter)
AWS, Houston, TX
AIME (American Institute of Mechanical Engineers), AWS, ASM, Beaumont, TX
SCTE (Society of Carbide and Tool Engineers), ASM, Philadelphia, PA
ASM, Portland, OR
ASM, Greensboro, NC
ASQC (American Society for Quality Control), ASM, AIME, Worcester, MA
Metal Treating Institute International Convention, Washington, DC
ASM, Baton Rouge, LA
Florida International Arson Seminar, 46th Annual, Orlando, FL
AWMI (Association of Women in the Metal Industries), Marlboro, MA
AWS, Washington, DC
SAMPE (Society for Advancement of Materials and Processing Engineers), SCTE, ASM, San Diego, CA
Florida International University
ASM, AWS, Miami, FL
MFPG (Mechanical Failures Prevention Group), 45th Session Symposium
AICE (American Institute of Carbide Engineers), ASM, AIME, Kansas City, MO
ASM, Grand Rapids, MI
ASM, Battle Creek, MI
ASM, AWS, ASNT, Rahwah, NJ
ASM, Oak Ridge, TN
ASM, South Bend, IN (Notre Dame Chapter)
Roger Williams College
ASME, ASM, East Providence, RI
ASM, Bethlehem, PA (Lehigh Valley Chapter)
COMS, ASM, ASNT, Ohio State University, Columbus, OH
ASM, AES (American Electroplaters Society), ASQC, Springfield, MA
ASM International, Montreal, Quebec, Canada
TMS (The Metallurgical Society), New Haven, CT
AWS, ASM, Beaumont, TX
AWS, ASM, Houston, TX
AWS, Tampa, FL

Case Alumni Association, Washington, DC
National Thermal Spray Convention (NTSC) '93, Anaheim, CA
26th Annual IMS Symposium, Charleston, SC
ASM, Dayton, OH
ASM, Central Carolinas Chapter, Raleigh, NC
SWE, ASM, Peoria, IL
AWMI, Cleveland, OH
AFS (American Foundrymen's Society), ASM, Saginaw, MI
U. S. Attorney's Office, Dept. of Justice, San Diego, CA
AWS, San Diego, CA
ASM, Milwaukee, WI
AWMI, Baltimore, MD
AWS, Tysons Corner, VA
ASM, Indianapolis & Muncie, IN
National Engineers Week, Akron, OH: AIIA, ASM, ASCE, ASDPE, ASME, ASHE, IEEE, SME, NAWIC, ASQC, ASHRAE, AIChE, ACESS, Univ. of Akron, Kent State Univ.
ASM, MIT Faculty Club, Cambridge, MA
AWMI, Dallas, TX
ASM, Baltimore, MD
University of Virginia (graduate seminar)
AWMI, Minneapolis, MN
AWMI, St. Louis, MO
Oklahoma City University School of Law
Florida Assoc. of Criminal Defense Lawyers (FACDL), Palm Beach, FL
Wisconsin Assoc. of Criminal Defense Lawyers (WACDL), Madison, WI
American University, Washington School of Law, Washington, D.C. (guest lecturer)
CLE: "Life In The Balance" Seminar, NLADA, Memphis, TN
CLE: North Carolina Association of Trial Lawyers (NCATL), Raleigh, NC.
Joint Statistics Meeting (JSM 2004), Toronto, Canada
CLE: NLADA Conference, Washington, DC.
CLE: NACDL Midwinter Meeting & Seminar, New Orleans, LA
CLE: NCATL Conference, Sunset Beach, NC.
CLE: CPD, Copper Mountain, CO
CLE: TCDLA, Dallas, TX (co-director)
Georgetown University School of Law, Washington, D.C. (guest lecturer)
CLE: DCACDL, Washington, D.C.
CLE (judges only): "Science in the Courtroom", Judicial Institute of Maryland, Annapolis, MD
CLE: TCDLA, Houston, TX (co-director)
CLE: NCAJ, Raleigh, N.C.
CLE: WISBA, Milwaukee, WI
Northwestern University School of Law (guest lecturer for Prof. Jonathan J. Koehler)
CLE (invitation only): Cardozo Law School, NYC, NY
Metropolitan Public Defender's Office, Louisville, KY
CLE, TCDLA, Houston, TX (Oct. 2016)
CLE, NACDL, Las Vegas, NV (May 2017)
FAPI (Feb. 2021)

Contact Information:

Forensic Engineering International
2708 Little Gunstock Road, Lake Anna, VA 23024
(804) 448-3955 voice (540) 903-0423 mobile
e-mail: wt.matsci@gmail.com
Website: forensicengineersintl.com

REPORT TO THE PRESIDENT

Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods

Executive Office of the President
President's Council of Advisors on
Science and Technology

September 2016

Exhibit C

REPORT TO THE PRESIDENT
Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods

Executive Office of the President
President's Council of Advisors on
Science and Technology

September 2016

About the President’s Council of Advisors on Science and Technology

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

For more information about PCAST, see www.whitehouse.gov/ostp/pcast.

The President's Council of Advisors on Science and Technology

Co-Chairs

John P. Holdren
Assistant to the President for Science and Technology
Director, Office of Science and Technology Policy

Eric S. Lander
President
Broad Institute of Harvard and MIT

Vice Chairs

William Press
Raymer Professor in Computer Science and Integrative Biology
University of Texas at Austin

Maxine Savitz
Honeywell (ret.)

Members

Wanda M. Austin
President and CEO
The Aerospace Corporation

Rosina Bierbaum
Professor, School of Natural Resources and Environment, University of Michigan
Roy F. Westin Chair in Natural Economics, School of Public Policy, University of Maryland

Christine Cassel
Planning Dean
Kaiser Permanente School of Medicine

Christopher Chyba
Professor, Astrophysical Sciences and International Affairs
Princeton University

S. James Gates, Jr.
John S. Toll Professor of Physics
Director, Center for String and Particle Theory
University of Maryland, College Park

Mark Gorenberg
Managing Member
Zetta Venture Partners

Susan L. Graham
Pehong Chen Distinguished Professor Emerita in Electrical Engineering and Computer Science
University of California, Berkeley

Michael McQuade
Senior Vice President for Science and Technology
United Technologies Corporation

Chad Mirkin
George B. Rathmann Professor of Chemistry
Director, International Institute for Nanotechnology
Northwestern University

Mario Molina
Distinguished Professor, Chemistry and Biochemistry
University of California, San Diego
Professor, Center for Atmospheric Sciences
Scripps Institution of Oceanography

Craig Mundie
President
Mundie Associates

Ed Penhoet
Director
Alta Partners
Professor Emeritus, Biochemistry and Public Health
University of California, Berkeley

Barbara Schaal
Dean of the Faculty of Arts and Sciences
Mary-Dell Chilton Distinguished Professor of Biology
Washington University of St. Louis

Eric Schmidt
Executive Chairman
Alphabet, Inc.

Daniel Schrag
Sturgis Hooper Professor of Geology
Professor, Environmental Science and Engineering
Director, Harvard University Center for Environment
Harvard University

Staff

Ashley Predith
Executive Director

Jennifer L. Michael
Program Support Specialist

Diana E. Pankevich
AAAS Science & Technology Policy Fellow

PCAST Working Group

Working Group members participated in the preparation of this report. The full membership of PCAST reviewed and approved it.

Working Group

Eric S. Lander (Working Group Chair)
President
Broad Institute of Harvard and MIT

S. James Gates, Jr.
John S. Toll Professor of Physics
Director, Center for String and Particle Theory
University of Maryland, College Park

Susan L. Graham
Pehong Chen Distinguished Professor Emerita in Electrical Engineering and Computer Science
University of California, Berkeley

Michael McQuade
Senior Vice President for Science and Technology
United Technologies Corporation

William Press
Raymer Professor in Computer Science and Integrative Biology
University of Texas at Austin

Daniel Schrag
Sturgis Hooper Professor of Geology
Professor, Environmental Science and Engineering
Director, Harvard University Center for Environment
Harvard University

Staff

Diana E. Pankevich
AAAS Science & Technology Policy Fellow

Kristen Zarrelli
Advisor, Public Policy & Special Projects
Broad Institute of Harvard and MIT

Writer

Tania Simoncelli
Senior Advisor to the Director
Broad Institute of Harvard and MIT

Senior Advisors

PCAST consulted with a panel of legal experts to provide guidance on factual matters relating to the interaction between science and the law. PCAST also sought guidance and input from two statisticians, who have expertise in this domain. Senior advisors were given an opportunity to review early drafts to ensure factual accuracy. PCAST expresses its gratitude to those listed here. Their willingness to engage with PCAST on specific points does not imply endorsement of the views expressed in this report. Responsibility for the opinions, findings, and recommendations in this report and for any errors of fact or interpretation rests solely with PCAST.

Senior Advisor Co-Chairs

The Honorable Harry T. Edwards
Judge
United States Court of Appeals
District of Columbia Circuit

Jennifer L. Mnookin
Dean, David G. Price and Dallas P. Price
Professor of Law
University of California Los Angeles Law

Senior Advisors

The Honorable James E. Boasberg
District Judge
United States District Court
District of Columbia

The Honorable Andre M. Davis
Senior Judge
United States Court of Appeals
Fourth Circuit

David L. Faigman
Acting Chancellor & Dean
University of California Hastings College of the Law

Stephen Fienberg
Maurice Falk University Professor of Statistics and Social Science (Emeritus)
Carnegie Mellon University

The Honorable Pamela Harris
Judge
United States Court of Appeals
Fourth Circuit

Karen Kafadar
Commonwealth Professor and Chair
Department of Statistics
University of Virginia

The Honorable Alex Kozinski
Judge
United States Court of Appeals
Ninth Circuit

The Honorable Cornelia T.L. Pillard
Judge
United States Court of Appeals
District of Columbia Circuit

The Honorable Charles Fried
Beneficial Professor of Law
Harvard Law School
Harvard University

The Honorable Nancy Gertner
Senior Lecturer on Law
Harvard Law School
Harvard University

The Honorable Jed S. Rakoff
District Judge
United States District Court
Southern District of New York

The Honorable Patti B. Saris
Chief Judge
United States District Court
District of Massachusetts

EXECUTIVE OFFICE OF THE PRESIDENT
PRESIDENT'S COUNCIL OF ADVISORS ON SCIENCE AND TECHNOLOGY
WASHINGTON, D.C. 20502

President Barack Obama
The White House
Washington, DC 20502

Dear Mr. President:

We are pleased to send you this PCAST report on *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. The study that led to the report was a response to your question to PCAST, in 2015, as to whether there are additional steps on the scientific side, beyond those already taken by the Administration in the aftermath of the highly critical 2009 National Research Council report on the state of the forensic sciences, that could help ensure the validity of forensic evidence used in the Nation's legal system.

PCAST concluded that there are two important gaps: (1) the need for clarity about the scientific standards for the validity and reliability of forensic methods and (2) the need to evaluate specific forensic methods to determine whether they have been scientifically established to be valid and reliable. Our study aimed to help close these gaps for a number of forensic "feature-comparison" methods—specifically, methods for comparing DNA samples, bitemarks, latent fingerprints, firearm marks, footwear, and hair.

Our study, which included an extensive literature review, was also informed by inputs from forensic researchers at the Federal Bureau of Investigation Laboratory and the National Institute of Standards and Technology as well as from many other forensic scientists and practitioners, judges, prosecutors, defense attorneys, academic researchers, criminal-justice-reform advocates, and representatives of Federal agencies. The findings and recommendations conveyed in this report, of course, are PCAST's alone.

Our report reviews previous studies relating to forensic practice and Federal actions currently underway to strengthen forensic science; discusses the role of scientific validity within the legal system; explains the criteria by which the scientific validity of feature-comparison forensic methods can be judged; and applies those criteria to the selected feature-comparison methods.

Based on our findings concerning the “foundational validity” of the indicated methods as well as their “validity as applied” in practice in the courts, we offer recommendations on actions that could be taken by the National Institute of Standards and Technology, the Office of Science and Technology Policy, and the Federal Bureau of Investigation Laboratory to strengthen the scientific underpinnings of the forensic disciplines, as well as on actions that could be taken by the Attorney General and the judiciary to promote the more rigorous use of these disciplines in the courtroom.

Sincerely,

John P. Holdren

Eric S Lander

John P. Holdren
Co-Chair

Eric S. Lander
Co-Chair

Table of Contents

Executive Summary

“Forensic science” has been defined as the application of scientific or technical practices to the recognition, collection, analysis, and interpretation of evidence for criminal and civil law or regulatory issues. Developments over the past two decades—including the exoneration of defendants who had been wrongfully convicted based in part on forensic-science evidence, a variety of studies of the scientific underpinnings of the forensic disciplines, reviews of expert testimony based on forensic findings, and scandals in state crime laboratories—have called increasing attention to the question of the validity and reliability of some important forms of forensic evidence and of testimony based upon them.[1]

A multi-year, Congressionally-mandated study of this issue released in 2009 by the National Research Council[2] (*Strengthening Forensic Science in the United States: A Path Forward*) was particularly critical of weaknesses in the scientific underpinnings of a number of the forensic disciplines routinely used in the criminal justice system. That report led to extensive discussion, inside and outside the Federal government, of a path forward, and ultimately to the establishment of two groups: the National Commission on Forensic Science hosted by the Department of Justice and the Organization for Scientific Area Committees for Forensic Science at the National Institute of Standards and Technology.

When President Obama asked the President’s Council of Advisors on Science and Technology (PCAST) in 2015 to consider whether there are additional steps that could usefully be taken on the scientific side to strengthen the forensic-science disciplines and ensure the validity of forensic evidence used in the Nation’s legal system, PCAST concluded that there are two important gaps: (1) the need for clarity about the scientific standards for the validity and reliability of forensic methods and (2) the need to evaluate specific forensic methods to determine whether they have been scientifically established to be valid and reliable.

This report aims to help close these gaps for the case of forensic “feature-comparison” methods—that is, methods that attempt to determine whether an evidentiary sample (e.g., from a crime scene) is or is not associated with a potential “source” sample (e.g., from a suspect), based on the presence of similar patterns, impressions, or other features in the sample and the source. Examples of such methods include the analysis of DNA, hair, latent fingerprints, firearms and spent ammunition, toolmarks and bitemarks, shoeprints and tire tracks, and handwriting.

[1] Citations to literature in support of points made in the Executive Summary are found in the main body of the report.

[2] The National Research Council is the study-conducting arm of the National Academies of Science, Engineering, and Medicine.

In the course of its study, PCAST compiled and reviewed a set of more than 2,000 papers from various sources—including bibliographies prepared by the Subcommittee on Forensic Science of the National Science and Technology Council and the relevant Working Groups organized by the National Institute of Standards and Technology (NIST); submissions in response to PCAST's request for information from the forensic-science stakeholder community; and PCAST's own literature searches.

To educate itself on factual matters relating to the interaction between science and the law, PCAST consulted with a panel of Senior Advisors comprising nine current or former Federal judges, a former U.S. Solicitor General, a former state Supreme Court justice, two law-school deans, and two distinguished statisticians who have expertise in this domain. Additional input was obtained from the Federal Bureau of Investigation (FBI) Laboratory and individual scientists at NIST, as well as from many other forensic scientists and practitioners, judges, prosecutors, defense attorneys, academic researchers, criminal-justice-reform advocates, and representatives of Federal agencies. The willingness of these groups and individuals to engage with PCAST does not imply endorsement of the views expressed in the report. The findings and recommendations conveyed in this report are the responsibility of PCAST alone.

The resulting report—summarized here without the extensive technical elaborations and dense citations in the main text that follows—begins with a review of previous studies relating to forensic practice and Federal actions currently underway to strengthen forensic science; discusses the role of scientific validity within the legal system; explains the criteria by which the scientific validity of forensic feature-comparison methods can be judged; applies those criteria to six such methods in detail and reviews an evaluation by others of a seventh method; and offers recommendations on Federal actions that could be taken to strengthen forensic science and promote its more rigorous use in the courtroom.

We believe the findings and recommendations will be of use both to the judiciary and to those working to strengthen forensic science.

Previous Work on Scientific Validity of Forensic-Science Disciplines

Ironically, it was the emergence and maturation of a *new* forensic science, DNA analysis, in the 1990s that first led to serious questioning of the validity of many of the traditional forensic disciplines. When DNA evidence was first introduced in the courts, beginning in the late 1980s, it was initially hailed as infallible; but the methods used in early cases turned out to be unreliable: testing labs lacked validated and consistently-applied procedures for defining DNA patterns from samples, for declaring whether two patterns matched within a given tolerance, and for determining the probability of such matches arising by chance in the population. When, as a result, DNA evidence was declared inadmissible in a 1989 case in New York, scientists engaged in DNA analysis in both forensic and non-forensic applications came together to promote the development of reliable principles and methods that have enabled DNA analysis of single-source samples to become the "gold standard" of forensic science for both investigation and prosecution.

Once DNA analysis became a reliable methodology, the power of the technology—including its ability to analyze small samples and to distinguish between individuals—made it possible not only to identify and convict true perpetrators but also to clear wrongly accused suspects before prosecution and to re-examine a number of past convictions. Reviews by the National Institute of Justice and others have found that DNA testing during the course of investigations has cleared tens of thousands of suspects and that DNA-based re-examination of past cases has led so far to the exonerations of 342 defendants. Independent reviews of these cases have revealed that many relied in part on faulty expert testimony from forensic scientists who had told juries incorrectly that similar features in a pair of samples taken from a suspect and from a crime scene (hair, bullets, bitemarks, tire or shoe treads, or other items) implicated defendants in a crime with a high degree of certainty.

The questions that DNA analysis had raised about the scientific validity of traditional forensic disciplines and testimony based on them led, naturally, to increased efforts to test empirically the reliability of the methods that those disciplines employed. Relevant studies that followed included:

- a 2002 FBI re-examination of microscopic hair comparisons the agency's scientists had performed in criminal cases, in which DNA testing revealed that 11 percent of hair samples found to match microscopically actually came from different individuals;

- a 2004 National Research Council report, commissioned by the FBI, on bullet-lead evidence, which found that there was insufficient research and data to support drawing a definitive connection between two bullets based on compositional similarity of the lead they contain;

- a 2005 report of an international committee established by the FBI to review the use of latent fingerprint evidence in the case of a terrorist bombing in Spain, in which the committee found that "confirmation bias"—the inclination to confirm a suspicion based on other grounds—contributed to a misidentification and improper detention; and

- studies reported in 2009 and 2010 on bitemark evidence, which found that current procedures for comparing bitemarks are unable to reliably exclude or include a suspect as a potential biter.

Beyond these kinds of shortfalls with respect to "reliable methods" in forensic feature-comparison disciplines, reviews have found that expert witnesses have often overstated the probative value of their evidence, going far beyond what the relevant science can justify. Examiners have sometimes testified, for example, that their conclusions are "100 percent certain;" or have "zero," "essentially zero," or "negligible," error rate. As many reviews—including the highly regarded 2009 National Research Council study—have noted, however, such statements are not scientifically defensible: all laboratory tests and feature-comparison analyses have non-zero error rates.

Starting in 2012, the Department of Justice (DOJ) and FBI undertook an unprecedented review of testimony in more than 3,000 criminal cases involving microscopic hair analysis. Their initial results, released in 2015, showed that FBI examiners had provided scientifically invalid testimony in more than 95 percent of cases where that testimony was used to inculpate a defendant at trial. In March 2016, the Department of Justice announced its intention to expand to additional forensic-science methods its review of forensic testimony by the FBI Laboratory in closed criminal cases. This review will help assess the extent to which similar testimonial overstatement has occurred in other forensic disciplines.

The 2009 National Research Council report was the most comprehensive review to date of the forensic sciences in this country. The report made clear that some types of problems, irregularities, and miscarriages of justice cannot simply be attributed to a handful of rogue analysts or underperforming laboratories, but are systemic and pervasive—the result of factors including a high degree of fragmentation (including disparate and often inadequate training and educational requirements, resources, and capacities of laboratories), a lack of standardization of the disciplines, insufficient high-quality research and education, and a dearth of peer-reviewed studies establishing the scientific basis and validity of many routinely used forensic methods.

The 2009 report found that shortcomings in the forensic sciences were especially prevalent among the feature-comparison disciplines, many of which, the report said, lacked well-defined systems for determining error rates and had not done studies to establish the uniqueness or relative rarity or commonality of the particular marks or features examined. In addition, proficiency testing, where it had been conducted, showed instances of poor performance by specific examiners. In short, the report concluded that "much forensic evidence—including, for example, bitemarks and firearm and toolmark identifications—is introduced in criminal trials without any meaningful scientific validation, determination of error rates, or reliability testing to explain the limits of the discipline."

The Legal Context

Historically, forensic science has been used primarily in two phases of the criminal-justice process: (1) *investigation*, which seeks to identify the likely perpetrator of a crime, and (2) *prosecution*, which seeks to prove the guilt of a defendant beyond a reasonable doubt. In recent years, forensic science—particularly DNA analysis—has also come into wide use for challenging past convictions.

Importantly, the investigative and prosecutorial phases involve different standards for the use of forensic science and other investigative tools. In investigations, insights and information may come from both well-established science and exploratory approaches. In the prosecution phase, forensic science must satisfy a higher standard. Specifically, the Federal Rules of Evidence (Rule 702(c,d)) require that expert testimony be based, among other things, on "reliable principles and methods" that have been "reliably applied" to the facts of the case. And, the Supreme Court has stated that judges must determine "whether the reasoning or methodology underlying the testimony is scientifically valid."

This is where legal standards and scientific standards intersect. Judges' decisions about the admissibility of scientific evidence rest solely on *legal* standards; they are exclusively the province of the courts and PCAST does not opine on them. But, these decisions require making determinations about scientific validity. It is the proper province of the scientific community to provide guidance concerning scientific standards for scientific validity, and it is on those *scientific* standards that PCAST focuses here.

We distinguish here between two types of scientific validity: foundational validity and validity as applied.

(1) *Foundational validity* for a forensic-science method requires that it be shown, based on empirical studies, to be *repeatable*, *reproducible*, and *accurate*, at levels that have been measured and are appropriate to the intended application. Foundational validity, then, means that a method can, *in*

principle, be reliable. It is the *scientific* concept we mean to correspond to the *legal* requirement, in Rule 702(c), of "reliable principles and methods."

(2) *Validity as applied* means that the method has been reliably applied *in practice.* It is the *scientific* concept we mean to correspond to the *legal* requirement, in Rule 702(d), that an expert "has reliably applied the principles and methods to the facts of the case."

Scientific Criteria for Validity and Reliability of Forensic Feature-Comparison Methods

Chapter 4 of the main report provides a detailed description of the scientific criteria for establishing the foundationally validity and reliability of forensic feature-comparison methods, including both objective and subjective methods.[3]

Subjective methods require particularly careful scrutiny because their heavy reliance on human judgment means they are especially vulnerable to human error, inconsistency across examiners, and cognitive bias. In the forensic feature-comparison disciplines, cognitive bias includes the phenomena that, in certain settings, humans may tend naturally to focus on similarities between samples and discount differences and may also be influenced by extraneous information and external pressures about a case.

The essential points of foundational validity include the following:

(1) Foundational validity requires that a method has been subjected to *empirical* testing by multiple groups, under conditions appropriate to its intended use. The studies must (a) demonstrate that the method is repeatable and reproducible and (b) provide valid estimates of the method's accuracy (that is, how often the method reaches an incorrect conclusion) that indicate the method is appropriate to the intended application.

(2) For objective methods, the foundational validity of the method can be established by studying measuring the accuracy, reproducibility, and consistency of each of its individual steps.

(3) For subjective feature-comparison methods, because the individual steps are not objectively specified, the method must be evaluated as if it were a "black box" in the examiner's head. Evaluations of validity and reliability must therefore be based on "black-box studies," in which many examiners render

[3] Feature-comparison methods may be classified as either objective or subjective. By objective feature-comparison methods, we mean methods consisting of procedures that are each defined with enough standardized and quantifiable detail that they can be performed by either an automated system or human examiners exercising little or no judgment. By subjective methods, we mean methods including key procedures that involve significant human judgment—for example, about which features to select within a pattern or how to determine whether the features are sufficiently similar to be called a probable match.

decisions about many independent tests (typically, involving "questioned" samples and one or more "known" samples) and the error rates are determined.

(4) 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.

Once a method has been established as foundationally valid based on appropriate empirical studies, claims about the method's accuracy and the probative value of proposed identifications, in order to be valid, must be based on such empirical studies. *Statements claiming or implying greater certainty than demonstrated by empirical evidence are scientifically invalid*. Forensic examiners should therefore report findings of a proposed identification with clarity and restraint, explaining in each case that the fact that two samples satisfy a method's criteria for a proposed match does not mean that the samples are from the same source. For example, if the false positive rate of a method has been found to be 1 in 50, experts should not imply that the method is able to produce results at a higher accuracy.

To meet the scientific criteria for validity as applied, two tests must be met:

(1) The forensic examiner must have been shown to be *capable* of reliably applying the method and must *actually* have done so. Demonstrating that an expert is *capable* of reliably applying the method is crucial—especially for subjective methods, in which human judgment plays a central role. From a scientific standpoint, the ability to apply a method reliably can be demonstrated only through empirical testing that measures how often the expert reaches the correct answer. Determining whether an examiner has *actually* reliably applied the method requires that the procedures actually used in the case, the results obtained, and the laboratory notes be made available for scientific review by others.

(2) The practitioner's assertions about the probative value of proposed identifications must be scientifically valid. The expert should report the overall false-positive rate and sensitivity for the method established in the studies of foundational validity and should demonstrate that the samples used in the foundational studies are relevant to the facts of the case. Where applicable, the expert should report the probative value of the observed match based on the specific features observed in the case. And the expert should not make claims or implications that go beyond the empirical evidence and the applications of valid statistical principles to that evidence.

We note, finally, that neither experience, nor judgment, nor good professional practices (such as certification programs and accreditation programs, standardized protocols, proficiency testing, and codes of ethics) can substitute for actual evidence of foundational validity and reliability. The frequency with which a particular pattern or set of features will be observed in different samples, which is an essential element in drawing conclusions, is not a matter of "judgment." It is an empirical matter for which only empirical evidence is relevant. Similarly, an expert's expression of *confidence* based on personal professional experience or expressions of *consensus* among practitioners about the accuracy of their field is no substitute for error rates estimated from relevant studies. For forensic feature-comparison methods, establishing foundational validity based on empirical evidence is thus a *sine qua non.* Nothing can substitute for it.

Evaluation of Scientific Validity for Seven Feature-Comparison Methods

For this study, PCAST applied the criteria discussed above to six forensic feature-comparison methods: (1) DNA analysis of single-source and simple-mixture samples, (2) DNA analysis of complex-mixture samples, (3) bitemarks, (4) latent fingerprints, (5) firearms identification, and (6) footwear analysis. For each method, Chapter 5 of the main report provides a brief overview of the methodology, discusses background information and studies, provides an evaluation on scientific validity, and offers suggestions on a path forward. For a seventh feature-comparison method—hair analysis—we do not undertake a full evaluation of scientific validity, but review supporting material recently released for comment by the Department of Justice. This Executive Summary provides only a brief summary of some key findings concerning these seven methods.

DNA Analysis of Single-Source and Simple-Mixture Samples

The vast majority of DNA analysis currently involves samples from a single individual or from a simple mixture of two individuals (such as from a rape kit). DNA analysis in such cases is an objective method in which the laboratory protocols are precisely defined and the interpretation involves little or no human judgment.

To evaluate the foundational validity of an objective method, one can examine the reliability of each of the individual steps rather than having to rely on black-box studies. In the case of DNA analysis of single-source and simple-mixture samples, each of the steps has been found to be “repeatable, reproducible, and accurate” with levels that have been measured and are “appropriate to the intended application” (to quote the requirement for foundational validity as stated above), and the probability of a match arising by chance in the population by chance can be estimated directly from appropriate genetic databases and is extremely low.

Concerning validity as applied, DNA analysis, like all forensic analyses, is not infallible in practice. Errors can and do occur. Although the probability that two samples from different sources have the same DNA profile is tiny, the chance of human error is much higher. Such errors may stem from sample mix-ups, contamination, incorrect interpretation, and errors in reporting.

To minimize human error, the FBI requires, as a condition of participating in the National DNA Index System, that laboratories follow the FBI’s Quality Assurance Standards. These require that the examiner run a series of controls to check for possible contamination and ensure that the PCR process ran properly. The Standards also requires semi-annual proficiency testing of all analysts who perform DNA testing for criminal cases. We find, though, that there is a need to improve proficiency testing.

DNA Analysis of Complex-Mixture Samples

Some investigations involve DNA analysis of complex mixtures of biological samples from multiple unknown individuals in unknown proportions. (Such samples arise, for example, from mixed blood stains, and increasingly from multiple individual touching a surface.) The fundamental difference between DNA analysis of complex-mixture samples and DNA analysis of single-source and simple mixtures lies not in the laboratory processing, but in the interpretation of the resulting DNA profile.

DNA analysis of complex mixtures is inherently difficult. Such samples result in a DNA profile that superimposes multiple individual DNA profiles. Interpreting a mixed profile is different from and more challenging than interpreting a simple profile, for many reasons. It is often impossible to tell with certainty which genetic variants are present in the mixture or how many separate individuals contributed to the mixture, let alone accurately to infer the DNA profile of each one.

The questions an examiner must ask, then, are, "Could a suspect's DNA profile be present *within* the mixture profile? And, what is the probability that such an observation might occur by chance?" Because many different DNA profiles may fit within some mixture profiles, the probability that a suspect "cannot be excluded" as a possible contributor to complex mixture may be *much higher* (in some cases, millions of times higher) than the probabilities encountered for single-source DNA profiles.

Initial approaches to the interpretation of complex mixtures relied on subjective judgment by examiners and simplified calculations. This approach is problematic because subjective choices made by examiners can dramatically affect the answer and the estimated probative value—introducing significant risk of both analytical error and confirmation bias. PCAST finds that subjective analysis of complex DNA mixtures has not been established to be foundationally valid and is not a reliable methodology.

Given the problems with subjective interpretation of complex DNA mixtures, a number of groups launched efforts to develop computer programs that apply various algorithms to interpret complex mixtures in an objective manner. The programs clearly represent a major improvement over purely subjective interpretation. They still require scientific scrutiny, however, to determine (1) whether the methods are scientifically valid, including defining the limitations on their reliability (that is, the circumstances in which they may yield unreliable results) and (2) whether the software correctly implements the methods.

PCAST finds that, at present, studies have established the foundational validity of some objective methods under limited circumstances (specifically, a three-person mixture in which the minor contributor constitutes at least 20 percent of the intact DNA in the mixture) but that substantially more evidence is needed to establish foundational validity across broader settings.

Bitemark Analysis

Bitemark analysis typically involves examining marks left on a victim or an object at the crime scene and comparing those marks with dental impressions taken from a suspect. Bitemark comparison is based on the premises that (1) dental characteristics, particularly the arrangement of the front teeth, differ substantially among people and (2) skin (or some other marked surface at a crime scene) can reliably capture these distinctive features. Bitemark analysis begins with an examiner deciding whether an injury is a mark caused by human teeth. If so, the examiner creates photographs or impressions of the questioned bitemark and of the suspect's dentition; compares the bitemark and the dentition; and determines if the dentition (1) cannot be excluded as having made the bitemark, (2) can be excluded as having made the bitemark, or (3) is inconclusive.

Bitemark analysis is a subjective method. Current protocols do not provide well-defined standards concerning the identification of features or the degree of similarity that must be identified to support a reliable conclusion that the mark could have or could not have been created by the dentition in question. Conclusions about all these matters are left to the examiner's judgment.

As noted above, the foundational validity of a subjective method can only be established through multiple, appropriately designed black-box studies. Few studies—and no appropriate black-box studies—have been undertaken to study the ability of examiners to accurately identify the source of a bitemark. In these studies, the observed false-positive rates were very high—typically above ten percent and sometimes far above. Moreover, several of these studies employed inappropriate closed-set designs that are likely to *under*estimate the true false positive rate. Indeed, available scientific evidence strongly suggests that examiners not only cannot identify the source of bitemark with reasonable accuracy, they cannot even consistently agree on whether an injury *is* a human bitemark. For these reasons, PCAST finds that bitemark analysis is far from meeting the scientific standards for foundational validity.

We note that some practitioners have expressed concern that the exclusion of bitemarks in court could hamper efforts to convict defendants in some cases. If so, the correct solution, from a scientific perspective, would not be to admit expert testimony based on invalid and unreliable methods but rather to attempt to develop scientifically valid methods. But, PCAST considers the prospects of developing bitemark analysis into a scientifically valid method to be low. We advise against devoting significant resources to such efforts.

Latent Fingerprint Analysis

Latent fingerprint analysis typically involves comparing (1) a "latent print" (a complete or partial friction-ridge impression from an unknown subject) that has been developed or observed on an item with (2) one or more "known prints" (fingerprints deliberately collected under a controlled setting from known subjects; also referred to as "ten prints"), to assess whether the two may have originated from the same source. It may also involve comparing latent prints with one another. An examiner might be called upon to (1) compare a latent print to the fingerprints of a known suspect who has been identified by other means ("identified suspect") or (2) search a large database of fingerprints to identify a suspect ("database search").

Latent fingerprint analysis was first proposed for use in criminal identification in the 1800s and has been used for more than a century. The method was long hailed as infallible, despite the lack of appropriate empirical studies to assess its error rate. In response to criticism on this point in the 2009 National Research Council report, those working in the field of latent fingerprint analysis recognized the need to perform empirical studies to assess foundational validity and measure reliability and have made progress in doing so. Much credit goes to the FBI Laboratory, which has led the way in performing black-box studies to assess validity and estimate reliability, as well as so-called "white-box" studies to understand the factors that affect examiners' decisions. PCAST applauds the FBI Laboratory's efforts. There are also nascent efforts to begin to move the field from a purely subjective method toward an objective method—although there is still a considerable way to go to achieve this important goal.

PCAST finds that latent fingerprint analysis is a foundationally valid subjective methodology—albeit with a false positive rate that is substantial and is likely to be higher than expected by many jurors based on longstanding claims about the infallibility of fingerprint analysis. The false-positive rate could be as high as 1 error in 306 cases based on the FBI study and 1 error in 18 cases based on a study by another crime laboratory.[4] In reporting results of latent-fingerprint examination, it is important to state the false-positive rates based on properly designed validation studies

With respect to validity as applied, there are, however, a number of open issues, notably:

(1) *Confirmation bias.* Work by FBI scientists has shown that examiners often alter the features that they initially mark in a latent print based on comparison with an apparently matching exemplar. Such circular reasoning introduces a serious risk of confirmation bias. Examiners should be required to complete and document their analysis of a latent fingerprint *before* looking at any known fingerprint and should separately document any additional data used during their comparison and evaluation.

(2) *Contextual bias*. Work by academic scholars has shown that examiners' judgments can be influenced by irrelevant information about the facts of a case. Efforts should be made to ensure that examiners are not exposed to potentially biasing information.

(3) *Proficiency testing*. Proficiency testing is essential for assessing an examiner's capability and performance in making accurate judgments. As discussed elsewhere in this report, proficiency testing needs to be improved by making it more rigorous, by incorporating it systematically within the flow of casework, and by disclosing tests for evaluation by the scientific community.

Scientific validity as applied, then, requires that an expert: (1) has undergone relevant proficiency testing to test his or her accuracy and reports the results of the proficiency testing; (2) discloses whether he or she documented the features in the latent print in writing before comparing it to the known print; (3) provides a written analysis explaining the selection and comparison of the features; (4) discloses whether, when performing the examination, he or she was aware of any other facts of the case that might influence the conclusion; and (5) verifies that the latent print in the case at hand is similar in quality to the range of latent prints considered in the foundational studies.

Concerning the path forward, continuing efforts are needed to improve the state of latent-print analysis—and these efforts will pay clear dividends for the criminal justice system. One direction is to continue to improve latent print analysis as a subjective method. There is a need for additional empirical studies to estimate error rates for latent prints of varying quality and completeness, using well-defined measures.

A second—and more important—direction is to convert latent-print analysis from a subjective method to an objective method. The past decade has seen extraordinary advances in automated image analysis based on machine learning and other approaches—leading to dramatic improvements in such tasks as face recognition and the interpretation of medical images. This progress holds promise of making fully automated latent

[4] The main report discusses the appropriate calculations of error rates, including best estimates (which are 1 in 604 and 1 in 24, respectively, for the two studies cited) and confidence bounds (stated above). It also discusses issues with specific studies, including problems with studies that may contribute to differences in rates (as in the two studies cited).

fingerprint analysis possible in the near future. There have already been initial steps in this direction, both in academia and industry.

The most important resource to propel the development of objective methods would be the creation of huge databases containing known prints, each with many corresponding ”simulated” latent prints of varying qualities and completeness, which would be made available to scientifically-trained researchers in academia and industry. The simulated latent prints could be created by “morphing” the known prints, based on transformations derived from collections of actual latent print-record print pairs.

Firearms Analysis

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. 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. 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 markings that arise during firing from a particular gun.

Firearms analysts have long stated that their discipline has near-perfect accuracy; however, the 2009 National Research Council study of all the forensic disciplines concluded about firearms analysis 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.

Our own extensive review of the relevant literature prior to 2009 is consistent with the National Research Council’s conclusion. We find that many of these earlier studies were inappropriately designed to assess foundational validity and estimate reliability. Indeed, there is internal evidence among the studies themselves indicating that many previous studies underestimated the false positive rate by at least 100-fold.

We identified one notable advance since 2009: the completion of the first appropriately designed black-box study of firearms. The work was commissioned and funded by the Defense Department’s Forensic Science Center and was conducted by an independent testing lab (the Ames Laboratory, a Department of Energy national laboratory affiliated with Iowa State University). The false-positive rate was estimated at 1 in 66, with a confidence bound indicating that the rate could be as high as 1 in 46. While the study is available as a report to the Federal government, it has not been published in a scientific journal.

The scientific criteria for foundational validity require that there be more than one such study, to demonstrate reproducibility, and that studies should ideally be published in the peer-reviewed scientific literature. Accordingly, the current evidence still falls short of the scientific criteria for foundational validity.

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 the one appropriately designed black-box study. Claims of higher accuracy are not scientifically justified at present.

Validity as applied would also require, from a scientific standpoint, that an expert testifying on firearms analysis (1) has undergone rigorous proficiency testing on a large number of test problems to measure his or her accuracy 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.

Concerning the path forward, with firearms analysis as with latent fingerprint analysis, two directions are available for strengthening the scientific underpinnings of the discipline. The first is to improve firearms analysis as a subjective method, which would require additional black-box studies to assess scientific validity and reliability and more rigorous proficiency testing of examiners, using problems that are appropriately challenging and publically disclosed after the test.

The second direction, as with latent print analysis, is 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. The same tremendous progress over the past decade in image analysis that gives us reason to expect early achievement of fully automated latent print analysis is cause for optimism that fully automated firearms analysis may be possible in the near future. Efforts in this direction are currently hampered, however, 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 the needed 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 images—could attract significant interest from academia and industry.

Footwear Analysis

Footwear analysis is a process that typically involves comparing a known object, such as a shoe, to a complete or partial impression found at a crime scene, to assess whether the object is likely to be the source of the impression. The process proceeds in a stepwise manner, beginning with a comparison of “class characteristics” (such as design, physical size, and general wear) and then moving to “identifying characteristics” or “randomly acquired characteristics” (such as marks on a shoe caused by cuts, nicks, and gouges in the course of use).

PCAST has not addressed the question of whether examiners can reliably determine class characteristics—for example, whether a particular shoeprint was made by a size 12 shoe of a particular make. While it is important that studies be undertaken to estimate the reliability of footwear analysis aimed at determining class characteristics, PCAST chose not to focus on this aspect of footwear examination because it is not *inherently* a challenging measurement problem to determine class characteristics, to estimate the frequency of shoes having a particular class characteristic, or (for jurors) to understand the nature of the features in question.

Instead, PCAST focused on the reliability of conclusions that an impression was likely to have come from a *specific* piece of footwear. This is a much harder problem because it requires knowing how accurately examiners can identify specific features shared between a shoe and an impression, how often they fail to identify features that would distinguish them, and what probative value should be ascribed to a particular "randomly acquired characteristic."

PCAST finds that there are no appropriate black-box studies to support the foundational validity of footwear analysis to associate shoeprints with particular shoes based on specific identifying marks. Such associations are unsupported by any meaningful evidence or estimates of their accuracy and thus are not scientifically valid.

Hair Analysis

Forensic hair analysis is a process by which examiners compare microscopic features of hair to determine whether a particular person may be the source of a questioned hair. As PCAST was completing this report, the Department of Justice released for comment proposed guidelines concerning testimony on hair examination, including a supporting document addressing the validity and reliability of the discipline. While PCAST has not performed the sort of in-depth evaluation for the hair-analysis discipline that we did for other feature-comparison disciplines discussed here, we undertook a review of the DOJ's supporting document in order to shed further light on the standards for conducting a scientific evaluation of a forensic feature-comparison discipline.

The document states that "microscopic hair comparison has been demonstrated to be a valid and reliable scientific methodology," while noting that "microscopic hair comparisons alone cannot lead to personal identification and it is crucial that this limitation be conveyed both in the written report and in testimony." In support of its conclusion that hair examination is valid and reliable, however, the document discusses only a handful of studies of human hair comparison, from the 1970s and 1980s. The supporting documents fail to note that subsequent studies found substantial flaws in the methodology and results of the key papers. PCAST's own review of the cited papers finds that these studies do not establish the foundational validity and reliability of hair analysis.

The DOJ's supporting document also cites a 2002 FBI study that used mitochondrial DNA analysis to re-examine 170 samples from previous cases in which the FBI Laboratory had performed microscopic hair examination. But that study's key conclusion does *not* support the conclusion that hair analysis is a "valid and reliable scientific methodology." The FBI authors actually found that, in 9 of 80 cases (11 percent) the FBI Laboratory had found the hairs to be microscopically indistinguishable, the DNA analysis showed that the hairs actually came from *different* individuals.

These shortcomings illustrate both the difficulty of these scientific evaluations and the reason they are best carried out by a science-based agency that is not itself involved in the application of forensic science within the legal system. They also underscore why it is important that *quantitative* information about the reliability of methods (e.g., the frequency of false associations in hair analysis) be stated clearly in expert testimony.

Closing Observations on the Seven Evaluations

Although we have undertaken detailed evaluations of only six specific methods—and a review of an evaluation by others of a seventh—our approach could be applied to assess the foundational validity and validity as applied of any forensic feature-comparison method, including traditional forensic disciplines as well as methods yet to be developed (such as microbiome analysis or internet-browsing patterns).

We note, finally, that the evaluation of scientific validity is necessarily based on the available scientific evidence at a point in time. Some methods that have not been shown to be foundationally valid may ultimately be found to be reliable, although significant modifications to the methods may be required to achieve this goal. Other methods may not be salvageable, as was the case with compositional bullet lead analysis and is likely the case with bitemarks. Still others may be subsumed by different but more reliable methods, much as DNA analysis has replaced other methods in some instances.

Recommendations to NIST and OSTP

Recommendation 1. Assessment of foundational validity

It is important that scientific evaluations of the foundational validity be conducted, on an ongoing basis, to assess the foundational validity of current and newly developed forensic feature-comparison technologies. To ensure the scientific judgments are unbiased and independent, such evaluations should be conducted by an agency which has no stake in the outcome.

(A) The National Institute of Standards and Technology (NIST) should perform such evaluations and should issue an annual public report evaluating the foundational validity of key forensic feature-comparison methods.

> (i) The evaluations should (a) assess whether each method reviewed has been adequately defined and whether its foundational validity has been adequately established and its level of accuracy estimated based on empirical evidence; (b) be based on studies published in the scientific literature by the laboratories and agencies in the U.S. and in other countries, as well as any work conducted by NIST’s own staff and grantees; (c) as a minimum, produce assessments along the lines of those in this report, updated as appropriate; and (d) be conducted under the auspices of NIST, with additional expertise as deemed necessary from experts outside forensic science.
>
> (ii) NIST should establish an advisory committee of experimental and statistical scientists from outside the forensic science community to provide advice concerning the evaluations and to ensure that they are rigorous and independent. The members of the advisory committee should be selected jointly by NIST and the Office of Science and Technology Policy.

(iii) NIST should prioritize forensic feature-comparison methods that are most in need of evaluation, including those currently in use and in late-stage development, based on input from the Department of Justice and the scientific community.

(iv) Where NIST assesses that a method has been established as foundationally valid, it should (a) indicate appropriate estimates of error rates based on foundational studies and (b) identify any issues relevant to validity as applied.

(v) Where NIST assesses that a method has not been established as foundationally valid, it should suggest what steps, if any, could be taken to establish the method's validity.

(vi) NIST should not have regulatory responsibilities with respect to forensic science.

(vii) NIST should encourage one or more leading scientific journals outside the forensic community to develop mechanisms to promote the rigorous peer review and publication of papers addressing the foundational validity of forensic feature-comparison methods.

(B) The President should request and Congress should provide increased appropriations to NIST of (a) $4 million to support the evaluation activities described above and (b) $10 million to support increased research activities in forensic science, including on complex DNA mixtures, latent fingerprints, voice/speaker recognition, and face/iris biometrics.

Recommendation 2. Development of objective methods for DNA analysis of complex mixture samples, latent fingerprint analysis, and firearms analysis

The National Institute of Standards and Technology (NIST) should take a leadership role in transforming three important feature-comparison methods that are currently subjective—latent fingerprint analysis, firearms analysis, and, under some circumstances, DNA analysis of complex mixtures—into objective methods.

(A) NIST should coordinate these efforts with the Federal Bureau of Investigation Laboratory, the Defense Forensic Science Center, the National Institute of Justice, and other relevant agencies.

(B) These efforts should include (i) the creation and dissemination of large datasets and test materials to support the development and testing of methods by both companies and academic researchers, (ii) grant and contract support, and (iii) sponsoring processes, such as prize competitions, to evaluate methods.

Recommendation 3. Improving the Organization for Scientific Area Committees Process

(A) The National Institute of Standards and Technology (NIST) should improve the Organization for Scientific Area Committees (OSAC), which was established to develop and promulgate standards and guidelines to improve best practices in the forensic science community.

(i) NIST should establish a Metrology Resource Committee, composed of metrologists, statisticians, and other scientists from outside the forensic-science community. A representative of the Metrology Resource

Committee should serve on each of the Scientific Area Committees (SACs) to provide direct guidance on the application of measurement and statistical principles to the developing documentary standards.

(ii) The Metrology Resource Committee, as a whole, should review and publically approve or disapprove all standards proposed by the Scientific Area Committees before they are transmitted to the Forensic Science Standards Board.

(B) NIST should ensure that the content of OSAC-registered standards and guidelines are freely available to any party that may desire them in connection with a legal case or for evaluation and research, including by aligning with the policies related to reasonable availability of standards in the Office of Management and Budget Circular A-119, Federal Participation in the Development and Use of Voluntary Consensus Standards and Conformity Assessment Activities and the Office of the Federal Register, IBR (incorporation by reference) Handbook.

Recommendation 4. R&D strategy for forensic science

(A) The Office of Science and Technology Policy (OSTP) should coordinate the creation of a national forensic science research and development strategy. The strategy should address plans and funding needs for:

(i) major expansion and strengthening of the academic research community working on forensic sciences, including substantially increased funding for both research and training;

(ii) studies of foundational validity of forensic feature-comparison methods;

(iii) improvement of current forensic methods, including converting subjective methods into objective methods, and development of new forensic methods;

(iv) development of forensic feature databases, with adequate privacy protections, that can be used in research;

(v) bridging the gap between research scientists and forensic practitioners; and

(vi) oversight and regular review of forensic-science research.

(B) In preparing the strategy, OSTP should seek input from appropriate Federal agencies, including especially the Department of Justice, Department of Defense, National Science Foundation, and National Institute of Standards and Technology; Federal and State forensic science practitioners; forensic science and non-forensic science researchers; and other stakeholders.

Recommendation to the FBI Laboratory

Recommendation 5. Expanded forensic-science agenda at the Federal Bureau of Investigation Laboratory

(A) ***Research programs.*** **The Federal Bureau of Investigation (FBI) Laboratory should undertake a vigorous research program to improve forensic science, building on its recent important work on latent fingerprint analysis.** The program should include:

(i) conducting studies on the reliability of feature-comparison methods, in conjunction with independent third parties without a stake in the outcome;

(ii) developing new approaches to improve reliability of feature-comparison methods;

(iii) expanding collaborative programs with external scientists; and

(iv) ensuring that external scientists have appropriate access to datasets and sample collections, so that they can carry out independent studies.

(B) ***Black-box studies.*** **Drawing on its expertise in forensic science research, the FBI Laboratory should assist in the design and execution of additional empirical 'black-box' studies for subjective methods, including for latent fingerprint analysis and firearms analysis.** These studies should be conducted by or in conjunction with independent third parties with no stake in the outcome.

(C) ***Development of objective methods.*** **The FBI Laboratory should work with the National Institute of Standards and Technology to transform three important feature-comparison methods that are currently subjective—latent fingerprint analysis, firearm analysis, and, under some circumstances, DNA analysis of complex mixtures—into objective methods.** These efforts should include (i) the creation and dissemination of large datasets to support the development and testing of methods by both companies and academic researchers, (ii) grant and contract support, and (iii) sponsoring prize competitions to evaluate methods.

(D) ***Proficiency testing.*** **The FBI Laboratory, should promote increased rigor in proficiency testing by (i) within the next four years, instituting routine blind proficiency testing within the flow of casework in its own laboratory, (ii) assisting other Federal, State, and local laboratories in doing so as well, and (iii) encouraging routine access to and evaluation of the tests used in commercial proficiency testing.**

(E) ***Latent fingerprint analysis.*** **The FBI Laboratory should vigorously promote the adoption, by all laboratories that perform latent fingerprint analysis, of rules requiring a "linear Analysis, Comparison, Evaluation" process—whereby examiners must complete and document their analysis of a latent fingerprint** ***before*** **looking at any known fingerprint and should separately document any additional data used during comparison and evaluation.**

(F) ***Transparency concerning quality issues in casework.*** **The FBI Laboratory, as well as other Federal forensic laboratories, should regularly and publicly report quality issues in casework (in a manner similar to the practices employed by the Netherlands Forensic Institute, described in Chapter 5), as a means to improve quality and promote transparency.**

(G) ***Budget.*** **The President should request and Congress should provide increased appropriations to the FBI to restore the FBI Laboratory's budget for forensic science research activities from its current level to $30 million and should evaluate the need for increased funding for other forensic-science research activities in the Department of Justice.**

Recommendations to the Attorney General

Recommendation 6. Use of feature-comparison methods in Federal prosecutions

(A) The Attorney General should direct attorneys appearing on behalf of the Department of Justice (DOJ) to ensure expert testimony in court about forensic feature-comparison methods meets the scientific standards for scientific validity.

While pretrial investigations may draw on a wider range of methods, expert testimony in court about forensic feature-comparison methods in criminal cases—which can be highly influential and has led to many wrongful convictions—must meet a higher standard. In particular, attorneys appearing on behalf of the DOJ should ensure that:

> (i) the forensic feature-comparison methods upon which testimony is based have been established to be foundationally valid with a level of accuracy suitable to their intended application, as shown by appropriate empirical studies and consistency with evaluations by the National Institute of Standards and Technology (NIST), where available; and
>
> (ii) the testimony is scientifically valid, with the expert's statements concerning the accuracy of methods and the probative value of proposed identifications being constrained by the empirically supported evidence and not implying a higher degree of certainty.

(B) DOJ should undertake an initial review, with assistance from NIST, of subjective feature-comparison methods used by DOJ to identify which methods (beyond those reviewed in this report) lack appropriate black-box studies necessary to assess foundational validity. Because such subjective methods are presumptively not established to be foundationally valid, DOJ should evaluate whether it is appropriate to present in court conclusions based on such methods.

(C) Where relevant methods have not yet been established to be foundationally valid, DOJ should encourage and provide support for appropriate black-box studies to assess foundational validity and measure reliability. The design and execution of these studies should be conducted by or in conjunction with independent third parties with no stake in the outcome.

Recommendation 7. Department of Justice guidelines on expert testimony

(A) The Attorney General should revise and reissue for public comment the Department of Justice's (DOJ) proposed "Uniform Language for Testimony and Reports" and supporting documents to bring them into alignment with scientific standards for scientific validity.

(B) The Attorney General should issue instructions directing that:

(i) Where empirical studies and/or statistical models exist to shed light on the accuracy of a forensic feature-comparison method, an examiner should provide quantitative information about error rates, in accordance with guidelines to be established by DOJ and the National Institute of Standards and Technology, based on advice from the scientific community.

(ii) Where there are not adequate empirical studies and/or statistical models to provide meaningful information about the accuracy of a forensic feature-comparison method, DOJ attorneys and examiners should not offer testimony based on the method. If it is necessary to provide testimony concerning the method, they should clearly acknowledge to courts the lack of such evidence.

(iii) In testimony, examiners should always state clearly that errors can and do occur, due both to similarities between features and to human mistakes in the laboratory.

Recommendation to the Judiciary

Recommendation 8. Scientific validity as a foundation for expert testimony

(A) When deciding the admissibility of expert testimony, Federal judges should take into account the appropriate scientific criteria for assessing scientific validity including:

(i) foundational validity, with respect to the requirement under Rule 702(c) that testimony is the product of reliable principles and methods; and

(ii) validity as applied, with respect to requirement under Rule 702(d) that an expert has reliably applied the principles and methods to the facts of the case.

These scientific criteria are described in Finding 1.

(B) Federal judges, when permitting an expert to testify about a foundationally valid feature-comparison method, should ensure that testimony about the accuracy of the method and the probative value of proposed identifications is scientifically valid in that it is limited to what the empirical evidence supports. Statements suggesting or implying greater certainty are not scientifically valid and should not be permitted. In particular, courts should never permit scientifically indefensible claims such as: "zero," "vanishingly small," "essentially zero," "negligible," "minimal," or "microscopic" error rates; "100 percent certainty" or proof "to a reasonable degree of scientific certainty;" identification "to the exclusion of all other sources;" or a chance of error so remote as to be a "practical impossibility."

(C) To assist judges, the Judicial Conference of the United States, through its Standing Advisory Committee on the Federal Rules of Evidence, should prepare, with advice from the scientific community, a best practices manual and an Advisory Committee note, providing guidance to Federal judges concerning the admissibility under Rule 702 of expert testimony based on forensic feature-comparison methods.

(D) To assist judges, the Federal Judicial Center should develop programs concerning the scientific criteria for scientific validity of forensic feature-comparison methods.

1. Introduction

"Forensic science" has been defined as the application of scientific or technical practices to the recognition, collection, analysis, and interpretation of evidence for criminal and civil law or regulatory issues.[5] The forensic sciences encompass a broad range of disciplines, each with its own set of technologies and practices. The National Institute of Justice (NIJ) divides those disciplines into twelve categories: general toxicology; firearms and toolmarks; questioned documents; trace evidence (such as hair and fiber analysis); controlled substances; biological/serology screening (including DNA analysis); fire debris/arson analysis; impression evidence; blood pattern evidence; crime scene investigation; medicolegal death investigation; and digital evidence.[6] In the years ahead, science and technology will likely offer additional powerful tools for the forensic domain—perhaps the ability to compare populations of bacteria in the gut or patterns of search on the Internet.

Historically, forensic science has been used primarily in two phases of the criminal-justice process: (1) *investigation*, which seeks to identify the likely perpetrator of a crime, and (2) *prosecution*, which seeks to prove the guilt of a defendant beyond a reasonable doubt. (In recent years, forensic science—particularly DNA analysis—has also come into wide use for challenging past convictions.) Importantly, the investigative and prosecutorial phases involve different standards for the use of forensic science and other investigative tools. In investigations, insights and information may come from both well-established science and exploratory approaches.[7] In the prosecution phase, forensic science must satisfy a higher standard. Specifically, the Federal Rules of Evidence require that expert testimony be based, among other things, on "reliable principles and methods" that have been "reliably applied" to the facts of the case.[8] And, the Supreme Court has stated that judges must determine "whether the reasoning or methodology underlying the testimony is scientifically valid."[9]

This is where legal standards and scientific standards intersect. Judges' decisions about the admissibility of scientific evidence rest solely on *legal* standards; they are exclusively the province of the courts. But, the overarching subject of the judges' inquiry is scientific validity.[10] It is the proper province of the scientific community to provide guidance concerning scientific standards for scientific validity.[11]

[5] Definition of "forensic science" as provided by the National Commission on Forensic Science in its Views Document, "Defining forensic science and related terms." Adopted April 30-May 1, 2015. www.justice.gov/ncfs/file/786571/download.

[6] See: National Institute of Justice. *Status and Needs of Forensic Science Service Providers: A Report to Congress.* 2006. www.ojp.usdoj.gov/nij/pubs-sum/213420.htm.

[7] While investigative methods need not meet the standards of reliability required under the Federal Rules of Evidence, they should be based in sound scientific principles and practices so as to avoid false accusations.

[8] Fed. R. Evid. 702.

[9] *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993) at 592.

[10] *Daubert,* at 594.

[11] In this report, PCAST addresses solely the *scientific* standards for scientific validity and reliability. We do not offer opinions concerning *legal* standards.

A focus on the scientific side of this intersection is timely because it has become increasingly clear in recent years that lack of rigor in the assessment of the scientific validity of forensic evidence is not just a hypothetical problem but a real and significant weakness in the judicial system. As recounted in Chapter 2, reviews by competent bodies of the scientific underpinnings of forensic disciplines and the use in courtrooms of evidence based on those disciplines have revealed a dismaying frequency of instances of use of forensic evidence that do not pass an objective test of scientific validity.

The most comprehensive such review to date was conducted by a National Research Council (NRC) committee co-chaired by Judge Harry Edwards of the U.S. Court of Appeals for the District of Columbia Circuit and Constantine Gatsonis, Director of the Center for Statistical Sciences at Brown University. Mandated by Congress in an appropriations bill signed into law in late 2005, the study launched in the fall of 2006 and the committee released its report in February 2009.[12]

The 2009 NRC report described a disturbing pattern of deficiencies common to many of the forensic methods routinely used in the criminal justice system, most importantly a lack of rigorous and appropriate studies establishing their scientific validity, concluding that "much forensic evidence—including, for example, bitemarks and firearm and toolmark identifications—is introduced in criminal trials without any meaningful scientific validation, determination of error rates, or reliability testing to explain the limits of the discipline."[13]

In 2013, after prolonged discussion of the NRC report's findings and recommendations inside and outside the Federal government, the Department of Justice (DOJ)—in collaboration with the National Institute of Standards and Technology (NIST)—established the National Commission on Forensic Science (NCFS) as a Federal advisory body charged with providing forensic-science guidance and policy recommendations to the Attorney General. Co-chaired by the Deputy Attorney General and the Director of NIST, the NCFS's 32 members include eight academic scientists and five other science Ph.D.s; the other members include judges, attorneys, and forensic practitioners. To strengthen forensic science more generally, in 2014 NIST established the Organization for Scientific Area Committees for Forensic Science (OSAC) to "coordinate development of standards and guidelines…to improve quality and consistency of work in the forensic science community."[14]

In September 2015, President Obama asked his Council of Advisors on Science and Technology (PCAST) to explore, in light of the work being done by the NCSF and OSAC, what additional efforts could contribute to strengthening the forensic-science disciplines and ensuring the scientific reliability of forensic evidence used in the Nation's legal system. After review of the ongoing activities and the relevant scientific and legal literatures—including particularly the scientific and legal assessments in the 2009 NRC report—PCAST concluded that there are two important gaps: (1) the need for clarity on the scientific meaning of "reliable principles and methods" and "scientific validity" in the context of certain forensic disciplines, and (2) the need to evaluate

[12] National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (2009).

[13] Ibid., 107-8.

[14] See: www.nist.gov/forensics/organization-scientific-area-committees-forensic-science.

specific forensic methods to determine whether they have been scientifically established to be valid and reliable.

Within the broad span of forensic disciplines, we chose to narrow our focus to techniques that we refer to here as forensic "feature-comparison" methods (see Box 1).[15] While one motivation for this narrowing was to make our task tractable within the limits of available time and resources, we chose this particular class of methods because: (1) they are commonly used in criminal cases; (2) they have attracted a high degree of concern with respect to validity (e.g., the 2009 NRC report); and (3) they all belong to the same broad scientific discipline, *metrology*, which is "the science of measurement and its application," in this case to measuring and comparing features.[16]

BOX 1. Forensic feature-comparison methods

PCAST uses the term "forensic feature-comparison methods" to refer to the wide variety of methods that aim to determine whether an evidentiary sample (e.g., from a crime scene) is or is not associated with a potential source sample (e.g., from a suspect) based on the presence of similar patterns, impressions, features, or characteristics in the sample and the source. Examples include the analyses of DNA, hair, latent fingerprints, firearms and spent ammunition, tool and toolmarks, shoeprints and tire tracks, bitemarks, and handwriting.

PCAST began this study by forming a working group of six of its members to gather information for consideration.[17] To educate itself about factual matters relating to the interaction between science and law, PCAST consulted with a panel of Senior Advisors (listed in the front matter) comprising nine current or former Federal judges, one former U.S. Solicitor General and State supreme court justice, two law school deans, and two statisticians, who have expertise in this domain. PCAST also sought input from a diverse group of additional experts and stakeholders, including forensic scientists and practitioners, judges, prosecutors, defense attorneys, criminal justice reform advocates, statisticians, academic researchers, and Federal agency representatives (see Appendix B). Input was gathered through multiple in-person meetings and conference calls, including a session

[15] PCAST notes that there are issues related to the scientific validity of other types of forensic evidence that are beyond the scope of this report but require urgent attention—including notably arson science and abusive head trauma commonly referred to as "Shaken Baby Syndrome." In addition, a major area not addressed in this report is scientific methods for assessing causation—for example, whether exposure to substance was likely to have caused harm to an individual.

[16] *International Vocabulary of Metrology – Basic and General Concepts and Associated Terms* (VIM 3rd edition) JCGM 200 (2012).

[17] Two of the members have been involved with forensic science. PCAST Co-chair Eric Lander has served in various scientific roles (expert witness in *People v. Castro* 545 N.Y.S.2d 985 (Sup. Ct. 1989), a seminal case on the quality of DNA analysis discussed on p. 25; court's witness in *U.S. v. Yee,* 134 F.R.D. 161 in 1991; member of the NRC panel on forensic DNA analysis in 1992; scientific co-author with a forensic scientist from the FBI Laboratory in 1994; and a member of the Board of Directors of the Innocence Project from 2004 to the present). All of these roles have been unremunerated. PCAST member S. James Gates, Jr. has been a member, since its inception, of the National Commission on Forensic Science.

at a meeting of PCAST on January 15, 2016. PCAST also took the unusual step of initiating an online, open solicitation to broaden input, in particular from the forensic-science practitioner community; more than 70 responses were received.[18]

PCAST also shared a draft of this report with NIST and DOJ, which provided detailed and helpful comments that were carefully considered in revising the report.

PCAST expresses its gratitude to all those who shared their views. Their willingness to engage with PCAST does not imply endorsement of the views expressed in the report. Responsibility for the opinions, findings and recommendations expressed in this report and for any errors of fact or interpretation rests solely with PCAST.

The remainder of our report is organized as follows.

- Chapter 2 provides a brief overview of the findings of other studies relating to forensic practice and testimony based on it, and it reviews, as well, Federal actions currently underway to strengthen forensic science.

- Chapter 3 briefly reviews the role of scientific validity within the legal system. It describes the important distinction between legal standards and scientific standards.

- Chapter 4 then describes the scientific standards for "reliable principles and methods" and "scientific validity" as they apply to forensic feature-comparison methods and offers clear criteria that could be readily applied by courts.

- Chapter 5 illustrates the application of the indicated criteria by using them to evaluate the scientific validity of six important "feature-comparison" methods: DNA analysis of single-source and simple-mixture samples, DNA analysis of complex mixtures, bitemark analysis, latent fingerprint analysis, firearms analysis, and footwear analysis. We also discuss an evaluation by others of a seventh method, hair analysis.

- In Chapters 6–9, we offer recommendations, based on the findings of Chapters 4–5, concerning Federal actions that could be taken to strengthen forensic science and promote its more rigorous use in the courtroom.

[18] See: www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_request_for_information.pdf.

2. Previous Work on Validity of Forensic-Science Methods

Developments over the past two decades—including the exoneration of defendants who had been wrongfully convicted based in part on forensic-science evidence, a variety of studies of the scientific underpinnings of the forensic disciplines, reviews of expert testimony based on forensic findings, and scandals in state crime laboratories—have called increasing attention to the question of the validity and reliability of some important forensic methods evidence and testimony based upon them. (For definitions of key terms such as scientific validity and reliability, see Box 1 on page 47-8.)

In this chapter, we briefly review this history to inform our assessment of the current state of forensic science methods and their validity and the path forward.[19]

2.1 DNA Evidence and Wrongful Convictions

Ironically, it was the emergence and maturation of a new forensic science, DNA analysis, that first led to serious questioning of the validity of many of the traditional forensic disciplines. When defendants convicted with the help of forensic evidence from those traditional disciplines began to be exonerated on the basis of persuasive DNA comparisons deeper inquiry into scientific validity began. How this came to pass provides useful context for our inquiry here.

When DNA evidence was first introduced in the courts, beginning in the late 1980s, it was initially hailed as infallible. But the methods used in early cases turned out to be unreliable: testing labs lacked validated and consistently-applied procedures for defining DNA patterns from samples, for declaring whether two patterns matched within a given tolerance, and for determining the probability of such matches arising by chance in the population.[20]

When DNA evidence was declared inadmissible in *People v. Castro*, a New York case in 1989, scientists—including at the U.S. National Academy of Sciences and the Federal Bureau of Investigation (FBI)—came together

[19] In producing this summary we relied particularly on the National Research Council 2009 report, *Strengthening Forensic Science in the United States: A Path Forward* and the National Academies of Sciences, Engineering, and Medicine 2015 report, *Support for Forensic Science Research: Improving the Scientific Role of the National Institute of Justice.*

[20] See: Lander, E.S. "DNA fingerprinting on trial." *Nature,* Vol. 339 (1989): 501-5; Lander, E.S., and B. Budowle. "DNA fingerprinting dispute laid to rest." *Nature,* Vol. 371 (1994): 735-8; Kaye, D.H. "DNA Evidence: Probability, Population Genetics, and the Courts." *Harv. J. L. & Tech*, Vol. 7 (1993): 101-72; Roberts, L. "Fight erupts over DNA fingerprinting." *Science*, Vol. 254 (1991): 1721-3; Thompson, W.C., and S. Ford. "Is DNA fingerprinting ready for the courts?" *New Scientist,* Vol. 125 (1990): 38-43; Neufeld, P.J., and N. Colman. "When science takes the witness stand." *Scientific American*, Vol. 262 (1991): 46-53.

to promote the development of reliable principles and methods that have enabled DNA analysis of single-source samples to become the "gold standard" of forensic science for both investigation and prosecution.[21]

Both the initial recognition of serious problems and the subsequent development of reliable procedures were aided by the existence of a robust community of molecular biologists who used DNA analysis in non-forensic applications, such as in biomedical and agricultural sciences. They were also aided by judges who recognized that this powerful forensic method should only be admitted as courtroom evidence once its reliability was properly established.

Once DNA analysis became a reliable methodology, the power of the technology—including its ability to analyze small samples and to distinguish between individuals—made it possible not only to identify and convict true perpetrators but also to clear mistakenly accused suspects before prosecution and to re-examine a number of past convictions. Reviews by the National Institute of Justice (NIJ)[22] and others have found that DNA testing during the course of investigations has cleared tens of thousands of suspects. DNA-based re-examination of past cases, moreover, has led so far to the exonerations of 342 defendants, including 20 who had been sentenced to death, and to the identification of 147 real perpetrators.[23]

Independent reviews of these cases have revealed that many relied in part on faulty expert testimony from forensic scientists who had told juries that similar features in a pair of samples taken from a suspect and from a crime scene (e.g., hair, bullets, bitemarks, tire or shoe treads, or other items) implicated defendants in a crime with a high degree of certainty.[24] According to the reviews, these errors were not simply a matter of individual examiners testifying to conclusions that turned out to be incorrect; rather, they reflected a systemic problem—the testimony was based on methods and included claims of accuracy that were cloaked in purported scientific respectability but actually had never been subjected to meaningful scientific scrutiny.[25]

[21] *People v. Castro* 545 N.Y.S.2d 985 (Sup. Ct. 1989). The case, in which a janitor was charged with the murder of a woman in the Bronx, was among the first criminal cases involving DNA analysis in the United States. The court held a 15-week-long pretrial hearing about the admissibility of the DNA evidence. By the end of the hearing, the independent experts for both the defense and prosecution unanimously agreed that the DNA evidence presented was not scientifically reliable—and the judge ruled the evidence inadmissible. See: Lander, E.S. "DNA fingerprinting on trial." *Nature,* Vol. 339 (1989): 501-5. These events eventually led to two NRC reports on forensic DNA analysis, in 1992 and 1996, and to the founding of the Innocence Project (www.innocenceproject.org).

[22] DNA testing has excluded 20-25 percent of initial suspects in sexual assault cases. U.S Department of Justice, Office of Justice Programs, National Institute of Justice. *Convicted by Juries, Exonerated by Science: Case Studies in the Use of DNA Evidence to Establish Innocence after Trial,* (1996): xxviii.

[23] Innocence Project, "DNA Exonerations in the United States." See: www.innocenceproject.org/dna-exonerations-in-the-united-states.

[24] For example, see: Gross, S.R., and M. Shaffer. "Exonerations in the United States, 1989-2012." National Registry of Exonerations, (2012) available at: www.law.umich.edu/special/exoneration/Documents/exonerations_us_1989_2012_full_report.pdf. See also: Saks, M.J., and J.J. Koehler. "The coming paradigm shift in forensic identification science." *Science,* Vol. 309, No. 5736 (2005): 892-5.

[25] Garrett, B.L., and P.J. Neufeld. "Invalid forensic science testimony and wrongful convictions." *Virginia Law Review*, Vol. 91, No. 1 (2009): 1-97; National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (2009): 42-3.

2.2 Studies of Specific Forensic-Science Methods and Laboratory Practices

The questions that DNA analysis had raised about the scientific validity of traditional forensic disciplines and testimony based on them led, naturally, to increased efforts to test empirically the reliability of the methods that those disciplines employed. Scrutiny was directed, similarly, to the practices by which forensic evidence is collected, stored, and analyzed in crime laboratories around the country. The FBI Laboratory, widely regarded as one of the best in the country, played an important role in the latter investigations, re-assessing its own practices as well as those of others. In what follows we summarize some of the key findings of the studies of methods and practices that ensued in the case of the "comparison" disciplines that are the focus in this report.

Bullet Lead Examination

From the 1960s until 2005, the FBI used compositional analysis of bullet lead as a forensic tool of analysis to identify the source of bullets. Yet, an NRC report commissioned by the FBI and released in 2004 challenged the foundational validity of identifications based on the discipline. The technique involved comparing the quantity of various elements in bullets found at a crime scene with that of unused bullets to determine whether the bullets came from the same box of ammunition. The 2004 NRC report found that there is no scientific basis for making such a determination.[26] While the method for determining the concentrations of different elements within a bullet was found to be reliable, the report found there was insufficient research and data to support drawing a connection, based on compositional similarity between a particular bullet and a given batch of ammunition, which is usually the relevant question in a criminal case.[27] In 2005, the FBI announced that it would discontinue the practice of bullet lead examinations, noting that while it "firmly supports the scientific foundation of bullet lead analysis," the manufacturing and distribution of bullets was too variable to make the matching reliable.[28]

[26] National Research Council. *Forensic Analysis: Weighing Bullet Lead Evidence.* The National Academies Press. Washington DC. (2004). Lead bullet examination, also known as Compositional Analysis of Bullet Lead (CABL), involves comparing the elemental composition of bullets found at a crime scene with unused cartridges in the possession of a suspect. This technique assumes that (1) the molten source used to produce a single "lot" of bullets has a uniform composition throughout, (2) no two molten sources have the same composition, and (3) bullets with different compositions are not mixed during the manufacturing or shipping processes. However, in practice, this is not the case. The 2004 NRC report found that compositionally indistinguishable volumes of lead could produce small lots of bullets—on the order of 12,000 bullets—or large lots—with more than 35 million bullets. The report also found no assurance that indistinguishable volumes of lead could not occur at different times and places. Neither scientists nor bullet manufacturers are able to definitively attest to the significance of an association made between bullets in the course of a bullet lead examination. The most that one can say is that bullets that are indistinguishable by CABL *could* have come from the same source.

[27] Faigman, D.L., Cheng, E.K., Mnookin, J.L., Murphy, E.E., Sander, J., and C. Slobogin (Eds.) *Modern Scientific Evidence: The Law and Science of Expert Testimony, 2015-2016 ed.* Thomson/West Publishing (2016).

[28] Federal Bureau of Investigation. *FBI Laboratory Announces Discontinuation of Bullet Lead Examinations.* (September 1, 2005, press release). www.fbi.gov/news/pressrel/press-releases/fbi-laboratory-announces-discontinuation-of-bullet-lead-examinations (accessed May 6, 2016).

Latent Fingerprints

In 2005, an international committee established by the FBI released a report concerning flaws in the FBI's practices for fingerprint identification that had led to a prominent misidentification. Based almost entirely on a latent fingerprint recovered from the 2004 bombing of the Madrid commuter train system, the FBI erroneously detained an American in Portland, Oregon and held him for two weeks as a material witness.[29] An FBI examiner concluded the fingerprints matched with "100 percent certainty," although Spanish authorities were unable to confirm the match.[30] The review committee concluded that the FBI's misidentification had occurred primarily as a result of "confirmation bias."[31] Similarly, a report by the DOJ's Office of the Inspector General highlighted "reverse reasoning" from the known print to the latent image that led to an exaggerated focus on apparent similarities and inadequate attention to differences between the images.[32]

Hair Analysis

In 2002, FBI scientists used mitochondrial DNA sequencing to re-examine 170 microscopic hair comparisons that the agency's scientists had performed in criminal cases. The DNA analysis showed that, in 11 percent of cases in which the FBI examiners had found the hair samples to match microscopically, DNA testing of the samples revealed they actually came from different individuals.[33] These false associations may not have been the result of a failure of the examiner to perform the analysis correctly; instead, the characteristics could have just happened to have been shared by chance. The study showed that the power of microscopic hair comparison to distinguish between samples from different sources was much lower than previously assumed. (For example, earlier studies suggested that the false positive rate for of hair analysis is in the range of 1 in 40,000.[34])

Bitemarks

A 2010 study of experimentally created bitemarks produced by known biters found that skin deformation distorts bitemarks so substantially and so variably that current procedures for comparing bitemarks are unable to reliably exclude or include a suspect as a potential biter. ("The data derived showed no correlation and was

[29] Stacey, R.B. "Report on the erroneous fingerprint individualization in the Madrid train bombing case." *Forensic Science Communications*, Vol. 7, No. 1 (2005).

[30] Application for Material Witness Order and Warrant Regarding Witness: Brandon Bieri Mayfield, *In re* Federal Grand Jury Proceedings 03-01, 337 F. Supp. 2d 1218 (D. Or. 2004) (No. 04-MC-9071).

[31] Specifically, similarities between the two prints, combined with the inherent pressure of working on an extremely high-profile case, influenced the initial examiner's judgment: ambiguous characteristics were interpreted as points of similarity and differences between the two prints were explained away. A second examiner, not shielded from the first examiner's conclusions, simply confirmed the first examiner's results. See: Stacey, R.B. "Report on the erroneous fingerprint individualization in the Madrid train bombing case." *Forensic Science Communications*, Vol. 7, No. 1 (2005).

[32] U.S. Department of Justice, Office of the Inspector General. "A review of the FBI's handling of the Brandon Mayfield case." (2006). oig.justice.special/s0601/final.pdf.

[33] Houck, M.M., and B. Budowle. "Correlation of microscopic and mitochondrial DNA hair comparisons." *Journal of Forensic Sciences*, Vol. 47, No. 5 (2002): 964-7.

[34] Gaudette, B. D., and E.S. Keeping. "An attempt at determining probabilities in human scalp hair comparisons." *Journal of Forensic Sciences,* Vol. 19 (1975): 599-606. This study was recently cited by DOJ to support the assertion that hair analysis is a valid and reliable scientific methodology. www.justice.gov/dag/file/877741/download. The topic of hair analysis is discussed in Chapter 5.

not reproducible, that is, the same dentition could not create a measurable impression that was consistent in all of the parameters in any of the test circumstances.[35]) A recent study by the American Board of Forensic Odontology also showed a disturbing lack of consistency in the way that forensic odontologists go about analyzing bitemarks, including even on deciding whether there was sufficient evidence to determine whether a photographed bitemark was a human bitemark.[36] In February 2016, following a six-month investigation, the Texas Forensic Science Commission unanimously recommended a moratorium on the use of bitemark identifications in criminal trials, concluding that the validity of the technique has not been scientifically established. [37]

These examples illustrate how several forensic feature-comparison methods that have been in wide use have nonetheless not been subjected to meaningful tests of scientific validity or measures of reliability.

2.3 Testimony Concerning Forensic Evidence

Reviews of trial transcripts have found that expert witnesses have often overstated the probative value of their evidence, going far beyond what the relevant science can justify. For example, some examiners have testified:

- that their conclusions are "100 percent certain;" have "zero," "essentially zero," vanishingly small," "negligible," "minimal," or "microscopic" error rate; or have a chance of error so remote as to be a "practical impossibility."[38] As many reviews have noted, however, such statements are not scientifically defensible. All laboratory tests and feature-comparison analyses have non-zero error rates, even if an

[35] Bush, M.A., Cooper, H.I., and R.B. Dorion. "Inquiry into the scientific basis for bitemark profiling and arbitrary distortion compensation." *Journal of Forensic Sciences*, Vol. 55, No. 4 (2010): 976-83. See also
Bush, M.A., Miller, R.G., Bush, P.J., and R.B. Dorion. "Biomechanical factors in human dermal bitemarks in a cadaver model." *Journal of Forensic Sciences,* Vol. 54, No. 1 (2009): 167-76.

[36] Balko, R. "A bite mark matching advocacy group just conducted a study that discredits bite mark evidence." *Washington Post,* April 8, 2015. www.washingtonpost.com/news/the-watch/wp/2015/04/08/a-bite-mark-matching-advocacy-group-just-conducted-a-study-that-discredits-bite-mark-evidence.; Adam J. Freeman & Iain A. Pretty, Construct Validity of Bitemark Assessments Using the ABO Bitemark Decision Tree, American Academy of Forensic Sciences, Annual Meeting, Odontology Section, G14, February 2015 (data made available by the authors upon request).

[37] Texas Forensic Science Commission. "Forensic bitemark comparison complaint filed by National Innocence Project on behalf of Steven Mark Chaney – Final Report." (2016). www.fsc.texas.gov/sites/default/files/FinalBiteMarkReport.pdf.

[38] Thompson, W.C., Taroni, F., and C.G.G. Aitken. "How the Probability of a False Positive Affects the Value of DNA Evidence." *J Forensic Sci,* Vol. 48, No. 1 (2003): 1-8; Thompson, W.C. "The Myth of Infallibility," In Sheldon Krimsky & Jeremy Gruber (Eds.) *Genetic Explanations: Sense and Nonsense*, Harvard University Press (2013); Cole, S.A. "More than zero: Accounting for error in latent fingerprint identification." *Journal of Criminal Law and Criminology,* Vol. 95, No.3 (2005): 985-1078; and Koehler, J.J. "Forensics or fauxrensics? Ascertaining accuracy in the forensic sciences." papers.ssrn.com/sol3/papers.cfm?abstract_id=2773255 (accessed June 28, 2016).

examiner received a perfect score on a particular performance test involving a limited number of samples.[39] Even highly automated tests do not have a zero error rate.[40,41]

- that they can "individualize" evidence—for example, using markings on a bullet to attribute it to a specific weapon "to the exclusion of every other firearm in the world"—an assertion that is not supportable by the relevant science.[42]
- that a result is true "to a reasonable degree of scientific certainty." This phrase has no generally accepted meaning in science and is open to widely differing interpretations by different scientists.[43] Moreover, the statement may be taken as implying certainty.

DOJ Review of Testimony on Hair Analysis

In 2012, the DOJ and FBI announced that they would initiate a formal review of testimony in more than 3,000 criminal cases involving microscopic hair analysis. Initial results of this unprecedented review, conducted in consultation with the Innocence Project and the National Association of Criminal Defense Lawyers, found that FBI examiners had provided scientifically invalid testimony in more than 95 percent of cases where examiner-provided testimony was used to inculpate a defendant at trial. These problems were systemic: 26 of the 28 FBI hair examiners who testified in the 328 cases provided scientifically invalid testimony.[44,45]

[39] Cole, S.A. "More than zero: Accounting for error in latent fingerprint identification." *Journal of Criminal Law and Criminology,* Vol. 95, No.3 (2005): 985-1078 and Koehler, J.J. "Forensics or fauxrensics? Ascertaining accuracy in the forensic sciences." papers.ssrn.com/sol3/papers.cfm?abstract_id=2773255 (accessed June 28, 2016).

[40] Thompson, W.C., Franco, T., and C.G.G. Aitken. "How the probability of a false positive affects the value of DNA evidence." *Journal of Forensic Science,* Vol. 48, No. 1 (2003): 1-8.

[41] False positive results can arise from two sources: (1) similarity between two features that occur by chance and (2) human/technical failures. See discussion in Chapter 4, p. 50-1.

[42] See: National Research Council. *Ballistic Imaging.* The National Academies Press. Washington DC. 2008 and Saks, M. J., and J.J. Koehler. "The individualization fallacy in forensic science evidence." Forensic Science Evidence." *Vanderbilt Law Review,* Vol. 61, No. 1 (2008): 199-218.

[43] National Commission on Forensic Science, "Recommendations to the Attorney General Regarding Use of the Term 'Reasonable Scientific Certainty'," Approved March 22, 2016, available at: www.justice.gov/ncfs/file/839726/download. The NCSF states that "forensic discipline conclusions are often testified to as being held 'to a reasonable degree of scientific certainty' or 'to a reasonable degree of [discipline] certainty.' These terms have no scientific meaning and may mislead factfinders about the level of objectivity involved in the analysis, its scientific reliability and limitations, and the ability of the analysis to reach a conclusion."

[44] Federal Bureau of Investigation. *FBI Testimony on Microscopic Hair Analysis Contained Errors in at Least 90 Percent of Cases in Ongoing Review,* (April 20, 2015, press release). www.fbi.gov/news/pressrel/press-releases/fbi-testimony-on-microscopic-hair-analysis-contained-errors-in-at-least-90-percent-of-cases-in-ongoing-review.

[45] The erroneous statements fell into three categories, in which the examiner: (1) stated or implied that evidentiary hair could be associated with a specific individual to the exclusion of all others; (2) assigned to the positive association a statistical weight or a probability that the evidentiary hair originated from a particular source; or (3) cited the number of cases worked in the lab and the number of successful matches to support a conclusion that an evidentiary hair belonged to a specific individual. Reimer, N.L. "The hair microscopy review project: An historic breakthrough for law enforcement and a daunting challenge for the defense bar." *The Champion*, (July 2013): 16. www.nacdl.org/champion.aspx?id=29488.

The importance of the FBI's hair analysis review was illustrated by the decision in January 2016 by Massachusetts Superior Court Judge Robert Kane to vacate the conviction of George Perrot, based in part on the FBI's acknowledgment of errors in hair analysis.[46]

Expanded DOJ Review

In March 2016, DOJ announced its intention to expand its review of forensic testimony by the FBI Laboratory in closed criminal cases to additional forensic science methods. The review will provide the opportunity to assess the extent to which similar testimonial overstatement has occurred in other disciplines.[47] DOJ plans to lay out a framework for auditing samples of testimony that came from FBI units handling additional kinds of feature-based evidence, such as tracing the impressions that guns leave on bullets, shoe treads, fibers, soil and other crime-scene evidence.

2.4 Cognitive Bias

In addition to the issues previously described, scientists have studied a subtler but equally important problem that affects the reliability of conclusions in many fields, including forensic science: cognitive bias. Cognitive bias refers to ways in which human perceptions and judgments can be shaped by factors other than those relevant to the decision at hand. It includes "contextual bias," where individuals are influenced by irrelevant background information; "confirmation bias," where individuals interpret information, or look for new evidence, in a way that conforms to their pre-existing beliefs or assumptions; and "avoidance of cognitive dissonance," where individuals are reluctant to accept new information that is inconsistent with their tentative conclusion. The biomedical science community, for example, goes to great lengths to minimize cognitive bias by employing strict protocols, such as double-blinding in clinical trials.

Studies have demonstrated that cognitive bias may be a serious issue in forensic science. For example, a study by Itiel Dror and colleagues demonstrated that the judgment of latent fingerprint examiners can be influenced by knowledge about other forensic examiners' decisions (a form of confirmation bias).[48] These studies are discussed in more detail in Section 5.4. Similar studies have replicated these findings in other forensic domains, including DNA mixture interpretation, microscopic hair analysis, and fire investigation.[49,50]

[46] *Commonwealth v. Perrot,* No. 85-5415, 2016 WL 380123 (Mass. Super. Man. 26, 2016).

[47] See: www.justice.gov/dag/file/870671/download.

[48] Dror, I.E., Charlton, D., and A.E. Peron. "Contextual information renders experts vulnerable to making erroneous identifications." *Forensic Science International*, Vol. 156 (2006): 74-8.

[49] See, for example: Dror, I.E., and G. Hampikian. "Subjectivity and bias in forensic DNA mixture interpretation." *Science & Justice,* Vol. 51, No. 4 (2011): 204-8; Miller, L.S. "Procedural bias in forensic examinations of human hair." *Law and Human Behavior,* Vol. 11 (1987): 157; and Bieber, P. "Fire investigation and cognitive bias." *Wiley Encyclopedia of Forensic Science,* 2014, available through onlinelibrary.wiley.com/doi/10.1002/9780470061589.fsa1119/abstract.

[50] See, generally, Dror, I.E. "A hierarchy of expert performance." *Journal of Applied Research in Memory and Cognition*, Vol. 5 (2016): 121-127.

Several strategies have been proposed for mitigating cognitive bias in forensic laboratories, including managing the flow of information in a crime laboratory to minimize exposure of the forensic analyst to irrelevant contextual information (such as confessions or eyewitness identification) and ensuring that examiners work in a linear fashion, documenting their finding about evidence from crime science *before* performing comparisons with samples from a suspect.[51]

2.5 State of Forensic Science

The 2009 NRC study concluded that many of these difficulties with forensic science may stem from the historical reality that many methods were devised as rough heuristics to aid criminal investigations and were not grounded in the validation practices of scientific research.[52] Although many forensic laboratories do now require newly-hired forensic science practitioners to have an undergraduate science degree, many practitioners in forensic laboratories do not have advanced degrees in a scientific discipline.[53] In addition, until 2015, there were no Ph.D. programs specific to forensic science in the United States (although such programs exist in Europe).[54] There has been very limited funding for forensic science research, especially to study the validity or reliability of these disciplines. Serious peer-reviewed forensic science journals focused on feature-comparison fields remain quite limited.

As the 2009 NRC study and others have noted, fundamentally, the forensic sciences do not yet have a well-developed "research culture." [55] Importantly, a research culture includes the principles that (1) methods must be presumed to be unreliable until their foundational validity has been established based on empirical evidence and (2) even then, scientific questioning and review of methods must continue on an ongoing basis. Notably, some forensic practitioners espouse the notion that extensive "experience" in casework can substitute for empirical studies of scientific validity.[56] Casework is not scientifically valid research, and experience alone

[51] Kassin, S.M., Dror, I.E., and J. Kakucka. "The forensic confirmation bias: Problems, perspectives, and proposed solutions." *Journal of Applied Research in Memory and Cognition*, Vol. 2, No. 1 (2013): 42-52. See also: Krane, D.E., Ford, S., Gilder, J., Iman, K., Jamieson, A., Taylor, M.S., and W.C. Thompson. "Sequential unmasking: A means of minimizing observer effects in forensic DNA interpretation." *Journal of Forensic Sciences*, Vol. 53, No. 4 (July 2008): 1006-7.

[52] National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (2009): 128.

[53] National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (2009): 223-230. See also: Cooney, L. "Latent Print Training to Competency: Is it Time for a Universal Training Program?" *Journal of Forensic Identification*, Vol. 60 (2010): 223–58. ("The areas where there was no consensus included degree requirements (almost a 50/50 split between agencies that required a four-year degree or higher versus those agencies that required less than a four-year degree or no degree at all.")

[54] National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (2009): 223. While there are several Ph.D. programs in criminal justice, forensic psychology, forensic anthropology or programs in chemistry or related disciplines that offer a concentration in forensic science, only Sam Houston State University College of Criminal Justice offers a doctoral program in "forensic science." See: www.shsu.edu/programs/doctorate-of-philosophy-in-forensic-science.

[55] Mnookin, J.L., Cole, S.A., Dror, I.E., Fisher, B.A.J., Houck, M.M., Inman, K., Kaye, D.H., Koehler, J.J., Langenburg, G., Risinger, D.M., Rudin, N., Siegel, J., and D.A. Stoney. "The need for a research culture in the forensic sciences." *UCLA Law Review,* Vol. 725 (2011): 754-8.

[56] See Section 4.7.

cannot establish scientific validity. In particular, one cannot reliably estimate error rates from casework because one typically does not have independent knowledge of the "ground truth" or "right answer." [57]

Beyond the foundational issue of scientific validity, most feature-comparison fields historically gave insufficient attention to the importance of blinding practitioners to potentially biasing information; developing objective measures of assessment and interpretation; paying careful attention to error rates and their measurement; and developing objective assessments of the meaning of an association between a sample and its potential source.[58]

The 2009 NRC report stimulated some in the forensic science community to recognize these flaws. Some forensic scientists have embraced the need to place forensics on a solid scientific foundation and have undertaken initial efforts to do so.[59]

2.6 State of Forensic Practice

Investigations of forensic practice have likewise unearthed problems stemming from the lack of a strong "quality culture." Specifically, dozens of investigations of crime laboratories—primarily at the state and local level—have revealed repeated failures concerning the handling and processing of evidence and incorrect interpretation of forensic analysis results.[60]

Various commentators have pointed out a fundamental issue that may underlie these serious problems: the fact that nearly all crime laboratories are closely tied to the prosecution in criminal cases. This structure undermines

[57] See Section 4.7.

[58] National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (2009): 8, 124, 184-5, 188-91. See also Koppl, R., and D. Krane. "Minimizing and leveraging bias in forensic science." In Robertson C.T., and A.S. Kesselheim (Eds.) *Blinding as a solution to bias: Strengthening biomedical science, forensic science, and law*. Atlanta, GA: Elsevier (2016).

[59] See Section 4.8.

[60] A few examples of such investigations include: (1) a 2-year independent investigation of the Houston Police Department's crime lab that resulted in the review of 3,500 cases (Final Report of the Independent Investigator for the Houston Police Department Crime Laboratory and Property Room, prepared by Michael R. Bromwich, June 13, 2007 (www.hpdlabinvestigation.org/reports/070613report.pdf); (2) the investigation and closure of the Detroit Police Crime Lab's firearms unit following the discovery of evidence contamination and failure to properly maintain testing equipment (see Bunkley, N. "Detroit police lab is closed after audit finds serious errors in many cases." *New York Times*, September 25, 2008, www.nytimes.com/2008/09/26/us/26detroit.html?_r=0); (3) a 2010 investigation of North Carolina's State Bureau of Investigation crime laboratory that found that agents consistently withheld exculpatory evidence or distorted evidence in more than 230 cases over a 16 year period (see Swecker, C., and M. Wolf, "An Independent Review of the SBI Forensic Laboratory" images.bimedia.net/documents/SBI+Report.pdf); and (4) a 2013 review of the New York City medical examiner's office handling of DNA evidence in more than 800 rape cases (see State of New York, Office of the Inspector General. December 2013, www.ig.ny.gov/sites/default/files/pdfs/OCMEFinalReport.pdf). One analysis estimated that at least fifty major laboratories reported fraud by analysts, evidence destruction, failed proficiency tests, misrepresenting findings in testimony, or tampering with drugs between 2005 and 2011. Twenty-eight of these labs were nationally accredited. Memorandum from Marvin Schechter to New York State Commission on Forensic Science (March 25, 2011): 243-4 (see www.americanbar.org/content/dam/aba/administrative/legal_aid_indigent_defendants/ls_sclaid_def_train_memo_schechter.authcheckdam.pdf).

the greater objectivity typically found in testing laboratories in other fields and creates situations where personnel may make errors due to subtle cognitive bias or overt pressure.[61]

The 2009 NRC report recommended that all public forensic laboratories and facilities be removed from the administrative control of law enforcement agencies or prosecutors' offices.[62] For example, Houston—after disbanding its crime laboratory twice in three years—followed this recommendation and, despite significant political pushback, succeeded in transitioning the laboratory into an independent forensic science center.[63]

2.7 National Research Council Report

The 2009 NRC report, *Strengthening Forensic Science in the United States: A Path Forward,* was the most comprehensive review to date of the forensic sciences in the United States. The report made clear that the types of problems, irregularities, and miscarriages of justice outlined in this report cannot simply be attributed to a handful of rogue analysts or underperforming laboratories. Instead, the report found the problems plaguing the forensic science community are systemic and pervasive—the result of factors including a high degree of fragmentation (including disparate and often inadequate training and educational requirements, resources, and capacities of laboratories); a lack of standardization of the disciplines, insufficient high-quality research and education; and a dearth of peer-reviewed studies establishing the scientific basis and validity of many routinely used forensic methods.

Shortcomings in the forensic sciences were especially prevalent among the feature-comparison disciplines. The 2009 NRC report found that many of these disciplines lacked well-defined systems for determining error rates and had not done studies to establish the uniqueness or relative rarity or commonality of the particular marks or features examined. In addition, proficiency testing, where it had been conducted, showed instances of poor performance by specific examiners. In short, the report concluded that "much forensic evidence—including, for example, bitemarks and firearm and toolmark identifications—is introduced in criminal trials without any

[61] The 2009 NRC Report (pp. 24-5) states, "The best science is conducted in a scientific setting as opposed to a law enforcement setting. Because forensic scientists often are driven in their work by a need to answer a particular question related to the issues of a particular case, they sometimes face pressure to sacrifice appropriate methodology for the sake of expediency." See also: Giannelli, P.G. "Independent crime laboratories: The problem of motivational and cognitive bias." *Utah Law Review,* (2010): 247-66 and Thompson, S.G. *Cops in Lab Coats: Curbing Wrongful Convictions through Independent Forensic Laboratories.* Carolina Academic Press (2015).

[62] National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (2009): Recommendation 4, p. 24.

[63] The Houston Forensic Science Center opened in April 2014, replacing the former Houston Police Department Crime Laboratory. The Center operates as a "local government corporation" with its own directors, officers, and employees. The structure was intentionally designed to insulate the Center from undue influence by police, prosecutors, elected officials, or special interest groups. See: Thompson, S.G. *Cops in Lab Coats: Curbing Wrongful Convictions through Independent Forensic Laboratories.* Carolina Academic Press (2015): 214.

meaningful scientific validation, determination of error rates, or reliability testing to explain the limits of the discipline."[64]

The 2009 NRC report found that the problems plaguing the forensic sciences were so severe that they could only be addressed by "a national commitment to overhaul the current structure that supports the forensic science community in this country."[65] Underlying the report's 13 core recommendations was a call for leadership at the highest levels of both Federal and State governments and the promotion and adoption of a long-term agenda to pull the forensic science enterprise up from its current weaknesses.

The 2009 NRC report called for studies to test whether various forensic methods are foundationally valid, including performing empirical tests of the accuracy of the results. It also called for the creation of a new, independent Federal agency to provide needed oversight of the forensic science system; standardization of terminology used in reporting and testifying about the results of forensic sciences; the removal of public forensic laboratories from the administrative control of law enforcement agencies; implementation of mandatory certification requirements for practitioners and mandatory accreditation programs for laboratories; research on human observer bias and sources of human error in forensic examinations; the development of tools for advancing measurement, validation, reliability, and proficiency testing in forensic science; and the strengthening and development of graduate and continuous education and training programs.

2.8 Recent Progress

In response to the 2009 NRC report, the Obama Administration initiated a series of reform efforts aimed at strengthening the forensic sciences, beginning with the creation in 2009 of a Subcommittee on Forensic Science of the National Science and Technology Council's Committee on Science that was charged with considering how best to achieve the goals of the NRC report. The resulting activities are described in some detail below.

National Commission on Forensic Science

In 2013, the DOJ and NIST, with support from the White House, signed a Memorandum of Understanding that outlined a framework for cooperation and collaboration between the two agencies in support of efforts to strengthen forensic science.

In 2013, DOJ established a National Commission on Forensic Science (NCFS), a Federal advisory committee reporting to the Attorney General. Co-chaired by the Deputy Attorney General and the Director of NIST, the NCFS's 32 members include seven academic scientists and five other science Ph.D.s; the other members include judges, attorneys and forensic practitioners. It is charged with providing policy recommendations to the Attorney General.[66] The NCFS issues formal recommendations to the Attorney General, as well as "views

[64] National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (2009): 107-8.

[65] National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (2009).

[66] See: www.justice.gov/ncfs.

documents" that reflect two-thirds majority view of NCFS but do not request specific action by the Attorney General. To date, the NCFS has issued ten recommendations concerning, among other things, accreditation of forensic laboratories and certification of forensic practitioners, advancing the interoperability of fingerprint information systems, development of root cause analysis protocols for forensic service providers, and enhancing communications among medical-examiner and coroner offices.[67] To date, the Attorney General has formally adopted the first set of recommendations on accreditation[68] and has directed the Department to begin to take steps toward addressing some of the other recommendations put forward to date.[69]

In 2014, NIST established the Organization of Scientific Area Committees (OSAC), a collaborative body of more than 600 volunteer members largely drawn from the forensic science community.[70] OSAC was established to support the development of voluntary standards and guidelines for consideration by the forensic practitioner community.[71] The structure consists of six Scientific Area Committees (SACs) and 25 subcommittees that work to develop standards, guidelines, and codes of practice for each of the forensic science disciplines and methodologies. [72] Three overarching resource committees provide guidance on questions of law, human factors, and quality assurance. All documents developed by the SACs are approved by a Forensic Science Standards Board (FSSB), a component of the OSAC structure, for listing on the OSAC Registry of Approved Standards. OSAC is not a Federal advisory committee.

Federal Funding Of Research

The Federal government has also taken steps to address one factor contributing to the problems with forensic science—the lack of a robust and rigorous scientific research community in many disciplines in forensic science. While there are multiple reasons for the absence of such a research community, one reason is that, unlike most scientific disciplines, there has been too little funding to attract and sustain a substantial cadre of excellent scientists focused on *fundamental* research in forensic science.

The National Science Foundation (NSF) has recently begun efforts to help address this foundational shortcoming of forensic science. In 2013, NSF signaled its interest in this area and encouraged researchers to submit research proposals addressing fundamental questions that might advance knowledge and education in the forensic

[67] For a full list of documents approved by NCFS, see www.justice.gov/ncfs/work-products-adopted-commission.

[68] Department of Justice. "Justice Department announces new accreditation policies to advance forensic science." (December 7, 2015, press release). www.justice.gov/opa/pr/justice-department-announces-new-accreditation-policies-advance-forensic-science.

[69] Memorandum from the Attorney General to Heads of Department Components Regarding Recommendations of the National Commission on Forensic Science, March 17, 2016. www.justice.gov/ncfs/file/841861/download.

[70] Members include forensic science practitioners and other experts who represent local, State, and Federal agencies; academia; and industry.

[71] For more information see: www.nist.gov/forensics/osac.cfm.

[72] The six Scientific Area Committees under OSAC are: Biology/DNA, Chemistry/Instrumental Analysis, Crime Scene/Death Investigation, Digital/Multimedia, and Physics/Pattern Interpretation (www.nist.gov/forensics/upload/OSAC-Block-Org-Chart-3-17-2015.pdf).

sciences.[73] As a result of an interagency process led by OSTP and NSF, in collaboration with the National Institute of Justice (NIJ), invited proposals for the creation of new, multi-disciplinary research centers for funding in 2014.[74] Based on our review of grant abstracts, PCAST estimates that NSF commits a total of approximately $4.5 million per year in support for extramural research projects on foundational forensic science.

NIST has also taken steps to address this issue by creating a new Forensic Science Center of Excellence, called the Center for Statistics and Applications in Forensic Evidence (CSAFE), that will focus its research efforts on improving the statistical foundation for latent prints, ballistics, tiremarks, handwriting, bloodstain patterns, toolmarks, pattern evidence analyses, and for computer and information systems, mobile devices, network traffic, social media, and GPS digital evidence analyses.[75] CSAFE is funded under a cooperative agreement with Iowa State University, to set up a center in partnership with investigators at Carnegie Mellon University, the University of Virginia, and the University of California, Irvine; the total support is $20 million over five years. PCAST estimates that NIST commits a total of approximately $5 million per year in support for extramural research projects on foundational forensic science, consisting of approximately $4 million to CSAFE and approximately $1 million to other projects.

NIJ has no budget allocated specifically for forensic science research. In order to support research activities, NIJ must draw from its base funding, funding from the Office of Justice Programs' assistance programs for research and statistics, or from the DNA backlog reduction programs.[76] Most of its research support is directed to applied research. Although it is difficult to classify NIJ's research projects, we estimate that NIJ commits a total of approximately $4 million per year to support extramural research projects on fundamental forensic science.[77]

Even with the recent increases, the total extramural funding for fundamental research in forensic science across NSF, NIST, and NIJ is thus likely to be in the range of only $13.5 million per year.

[73] See: Dear Colleague Letter: Forensic Science – Opportunity for Breakthroughs in Fundamental and Basic Research and Education. www.nsf.gov/pubs/2013/nsf13120/nsf13120.jsp.

[74] The centers NSF is proposing to create are Industry/University Cooperative Research Centers (I/UCRCs). I/UCRCs are collaborative by design and could be effective in helping to bridge the scientific and cultural gap between academic researchers who work in forensics-relevant fields of science and forensic practitioners. www.nsf.gov/pubs/2014/nsf14066/nsf14066.pdf.

[75] National Institute of Standards and Technology. "New NIST Center of Excellence to Improve Statistical Analysis of Forensic Evidence." (2015). www.nist.gov/forensics/center-excellence-forensic052615.cfm.

[76] National Academies of Sciences, Engineering, and Medicine. *Support for Forensic Science Research: Improving the Scientific Role of the National Institute of Justice.* The National Academies Press. Washington DC. (2015). According to the report, "Congressional appropriations to support NIJ's research programs declined during the early to mid-2000s and remain insufficient, especially in light of the growing challenges facing the forensic science community…With limited base funding, NIJ funds research and development from the appropriations for DNA backlog reduction programs and other assistance programs. These carved-out funds are essentially supporting NIJ's current forensic science portfolio, but there are pressures to limit the amount used for research from these programs. In the past 3 years, funding for these assistance programs has declined; therefore, funds available for research have also been reduced."

[77] U.S. Department of Justice, National Institute of Justice. "Report Forensic Science: Fiscal Year 2015 Funding for DNA Analysis, Capacity Enhancement and Other Forensic Activities." 2016.

The 2009 NRC report found that

> *Forensic science research is [overall] not well supported. . . . Relative to other areas of science, the forensic science disciplines have extremely limited opportunities for research funding. Although the FBI and NIJ have supported some research in the forensic science disciplines, the level of support has been well short of what is necessary for the forensic science community to establish strong links with a broad base of research universities and the national research community. Moreover, funding for academic research is limited . . . , which can inhibit the pursuit of more fundamental scientific questions essential to establishing the foundation of forensic science. Finally, the broader research community generally is not engaged in conducting research relevant to advancing the forensic science disciplines.*[78]

A 2015 NRC report, *Support for Forensic Science Research: Improving the Scientific Role of the National Institute of Justice*, found that the status of forensic science research funding has not improved much since the 2009 NRC report.[79]

In addition, the Defense Forensic Science Center has recently begun to support extramural research spanning the forensic science disciplines as part of its mission to provide specialized forensic and biometric research capabilities and support to the Department of Defense. Redesignated as DFSC in 2013, the Center was formerly the U.S. Army Criminal Investigation Laboratory, originally charged with supporting criminal investigations within the military but additionally tasked in 2007 with providing an "enduring expeditionary forensics capability," in response in part to the need to investigate and prosecute explosives attacks in Iraq and Afghanistan. While the bulk of DFSC support has traditionally supported research in DNA analysis and biochemistry, the Center has recently directed resources toward projects to address critical foundational gaps in other disciplines, including firearms and latent print analysis.

Notably, DFSC has helped stimulate research in the forensic science community. Discussions between DFSC and the American Society of Crime Lab Directors (ASCLD) led ASCLD to host a meeting in 2011 to identify research priorities for the forensic science community. DFSC agreed to fund two foundational studies to address the highest priority research needs identified by the Forensic Research Committee of ASCLD: the first independent "black-box" study on firearms analysis and a DNA mixture interpretation study (see Chapter 5). In FY 2015, DFSC allocated approximately $9.2 million to external forensic science research. Seventy-five percent of DFSC's funding supported projects with regard to DNA/biochemistry; 9 percent digital evidence; 8 percent non-DNA pattern evidence; and 8 percent chemistry.[80] As is the case for NIJ, there is no line item in DFSC's budget dedicated to forensic science research; DFSC instead must solicit funding from multiple sources within the Department of Defense to support this research.

[78] National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (2009): 78.

[79] National Academies of Sciences, Engineering, and Medicine. *Support for Forensic Science Research: Improving the Scientific Role of the National Institute of Justice***.** The National Academies Press. Washington DC. (2015): 15.

[80] Defense Forensic Science Center, Office of the Chief Scientist, Annual Research Portfolio Report, January 5, 2016.

A Critical Gap: Scientific Validity

The Administration has taken important and much needed initial steps by creating mechanisms to discuss policy, develop best practices for practitioners of specific methods, and support scientific research. At the same time, work to date has not addressed the 2009 NRC report's call to examine the fundamental scientific validity and reliability of many forensic methods used every day in courts. The remainder of our report focuses on that issue.

3. The Role of Scientific Validity in the Courts

The central focus of this report is the scientific validity of forensic-science evidence—more specifically, evidence from scientific methods for comparison of features (in, for example, DNA, latent fingerprints, bullet marks and other items). The reliability of methods for interpreting evidence is a fundamental consideration throughout science. Accordingly, every scientific field has a well-developed, domain-specific understanding of what scientific validity of methods entails.

The concept of scientific validity also plays an important role in the legal system. In particular, as noted in Chapter 1, the Federal Rules of Evidence require that expert testimony about forensic science must be the product of "reliable principles and methods" that have been "reliably applied . . . to the facts of the case."

This report explicates the scientific criteria for scientific validity in the case of forensic feature-comparison methods, for use both within the legal system and by those working to strengthen the scientific underpinnings of those disciplines. Before delving into that scientific explication, we provide in this chapter a very brief summary, aimed principally at scientists and lay readers, of the relevant legal background and terms, as well as the nature of this intersection between law and science.

3.1 Evolution of Admissibility Standards

Over the course of the 20th century, the legal system's approach for determining the admissibility of scientific evidence has evolved in response to advances in science. In 1923, in *Frye v. United States,*[81] the Court of Appeals for the District of Columbia considered the admissibility of testimony concerning results of a purported "lie detector," a systolic-blood- pressure deception test that was a precursor to the polygraph machine. After describing the device and its operation, the Court rejected the testimony, stating:

> *[W]hile courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs.*[82]

The court found that the systolic test had "not yet gained such standing and scientific recognition among physiological and psychological authorities," and was therefore inadmissible.

More than a half-century later, the Federal Rules of Evidence were enacted into law in 1975 to guide criminal and civil litigation in Federal courts. Rule 702, in its original form, stated that:

[81] *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923).

[82] Ibid., 1014.

> *If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise.*[83]

There was considerable debate among litigants, judges, and legal scholars as to whether the rule embraced the *Frye* standard or established a new standard.[84] In 1993, the United States Supreme Court sought to resolve these questions in its landmark ruling in *Daubert v. Merrell Dow Pharmaceuticals*. In interpreting Rule 702, the *Daubert* Court held that the Federal Rules of Evidence superseded *Frye* as the standard for admissibility of expert evidence in Federal courts. The Court rejected "general acceptance" as the standard for admissibility and instead held that the admissibility of scientific expert testimony depended on its scientific reliability.

Where *Frye* told judges to defer to the judgment of the relevant expert community, *Daubert* assigned trial court judges the role of "gatekeepers" charged with ensuring that expert testimony "rests on reliable foundation."[85]

The Court stated that "the trial judge must determine . . . whether the reasoning or methodology underlying the testimony is scientifically valid."[86] It identified five factors that a judge should, among others, ordinarily consider in evaluating the validity of an underlying methodology. These factors are: (1) whether the theory or technique can be (and has been) tested; (2) whether the theory or technique has been subjected to peer review and publication; (3) the known or potential rate of error of a particular scientific technique; (4) the existence and maintenance of standards controlling the technique's operation; and (5) a scientific technique's degree of acceptance within a relevant scientific community.

The *Daubert* court also noted that judges evaluating proffers of expert scientific testimony should be mindful of other applicable rules, including:

- Rule 403, which permits the exclusion of relevant evidence "if its probative value is substantially outweighed by the danger of unfair prejudice, confusion of the issues, or misleading the jury…" (noting that expert evidence can be "both powerful and quite misleading because of the difficulty in evaluating it."); and
- Rule 706, which allows the court at its discretion to procure the assistance of an expert of its own choosing.[87]

[83] Act of January 2, 1975, Pub. Law No. 93-595, 88 Stat. 1926 (1975). See: federalevidence.com/pdf/FRE_Amendments/1975_Orig_Enact/1975-Pub.L._93-595_FRE.pdf.

[84] See: Giannelli, P.C. "The admissibility of novel scientific evidence: Frye v. United States, a half-century later." *Columbus Law Review*, Vol. 80, No. 6 (1980); McCabe, J. "DNA fingerprinting: The failings of Frye," *Norther Illinois University Law Review*, Vol. 16 (1996): 455-82; and Page, M., Taylor, J., and M. Blenkin. "Forensic identification science evidence since Daubert: Part II—judicial reasoning in decisions to exclude forensic identification evidence on grounds of reliability." *Journal of Forensic Sciences*, Vol. 56, No. 4 (2011): 913-7.

[85] *Daubert*, at 597.

[86] *Daubert*, at 580. See also, FN9 ("In a case involving scientific evidence, *evidentiary reliability* will be based on *scientific validity*." [emphasis in original]).

[87] *Daubert*, at 595, citing Weinstein, 138 F.R.D., at 632.

Congress amended Rule 702 in 2000 to make it more precise, and made further stylistic changes in 2011. In its current form, Rule 702 imposes four requirements:

> *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;*
> *(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.*

An Advisory Committee's Note to Rule 702 also specified a number of reliability factors that supplement the five factors enumerated in Daubert. Among those factors is "whether the field of expertise claimed by the expert is known to reach reliable results."[88,89]

Many states have adopted rules of evidence that track key aspects of these federal rules. Such rules are now the law in over half of the states, while other states continue to follow the Frye standard or variations of it.[90]

3.2 Foundational Validity and Validity as Applied

As described in *Daubert*, the legal system envisions an important conversation between law and science:

> *"The [judge's] inquiry envisioned by Rule 702 is, we emphasize, a flexible one. Its overarching subject is the scientific validity—and thus the evidentiary relevance and reliability—of the principles that underlie a proposed submission."*[91]

[88] See: Fed. R. Evid. 702 Advisory Committee note (2000). The following factors may be relevant under Rule 702: whether the underlying research was conducted independently of litigation; whether the expert unjustifiably extrapolated from an accepted premise to an unfounded conclusion; whether the expert has adequately accounted for obvious alternative explanations; whether the expert was as careful as she would be in her professional work outside of paid litigation; and *whether the field of expertise claimed by the expert is known to reach reliable results* [emphasis added].

[89] This note has been pointed to as support for efforts to challenge entire fields of forensic science, including fingerprints and hair comparisons. See: Giannelli, P.C. "The Supreme Court's 'Criminal' *Daubert* Cases." *Seton Hall Law Review,* Vol. 33 (2003): 1096.

[90] Even under the *Frye* formulation, the views of scientists about the meaning of reliability are relevant. *Frye* requires that a scientific technique or method must "have general acceptance" in the relevant scientific community to be admissible. As a scientific matter, the relevant scientific community for assessing the reliability of feature-comparison sciences includes metrologists (including statisticians) as well as other physical and life scientists from disciplines on which the specific methods are based. Importantly, the community is not limited to forensic scientists who practice the specific method. For example, the *Frye* court evaluated whether the proffered lie detector had gained "standing and scientific recognition among physiological and psychological authorities," rather than among lie detector experts. *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923).

[91] *Daubert*, at 594

Legal and scientific considerations thus both play important roles.

(1) The admissibility of expert testimony depends on a threshold test of, among other things, whether it meets certain *legal* standards embodied in Rule 702. These decisions about admissibility are exclusively the province of the courts.

(2) Yet, as noted above, the overarching subject of the judge's inquiry under Rule 702 is "scientific validity." It is the proper province of the scientific community to provide guidance concerning *scientific* standards for scientific validity.

PCAST does not opine here on the legal standards, but seeks only to clarify the scientific standards that underlie them. For complete clarity about our intent, we have adopted specific terms to refer to the *scientific* standards for two key types of scientific validity, which we mean to correspond, as scientific standards, to the legal standards in Rule 702 (c,d)):

(1) by "foundational validity," we mean the *scientific* standard corresponding to the legal standard of evidence being based on "reliable principles and methods," and

(2) by "validity as applied," we mean the *scientific* standard corresponding to the legal standard of an expert having "reliably applied the principles and methods."

In the next chapter, we turn to discussing the scientific standards for these concepts. We close this chapter by noting that answering the question of scientific validity in the forensic disciplines is important not just for the courts but also because it sets quality standards that ripple out throughout these disciplines—affecting practice and defining necessary research.

4. Scientific Criteria for Validity and Reliability of Forensic Feature-Comparison Methods

In this report, PCAST has chosen to focus on defining the validity and reliability of one specific area within forensic science: forensic feature-comparison methods. We have done so because it is both possible and important to do so for this particular class of methods.

- It is *possible* because feature comparison is a common scientific activity, and science has clear standards for determining whether such methods are reliable. In particular, feature-comparison methods belong squarely to the discipline of metrology—the science of measurement and its application.[92,93]

- It is *important* because it has become apparent, over the past decade, that faulty forensic feature comparison has led to numerous miscarriages of justice.[94] It has also been revealed that the problems

[92] International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM 3rd edition) JCGM 200 (2012).

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

[94] DNA-based re-examination of past cases has led so far to the exonerations of 342 defendants, including 20 who had been sentenced to death, and to the identification of 147 real perpetrators. See: Innocence Project, "DNA Exonerations in the United States." www.innocenceproject.org/dna-exonerations-in-the-united-states. Reviews of these cases have revealed that roughly half relied in part on expert testimony that was based on methods that had not been subjected to meaningful scientific scrutiny or that included scientifically invalid claims of accuracy. See: Gross, S.R., and M. Shaffer. "Exonerations in the United States, 1989-2012." National Registry of Exonerations, (2012) available at: www.law.umich.edu/special/exoneration/Documents/exonerations_us_1989_2012_full_report.pdf; Garrett, B.L., and P.J. Neufeld. "Invalid forensic science testimony and wrongful convictions." *Virginia Law Review*, Vol. 91, No. 1 (2009): 1-97; National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (2009): 42-3. The nature of the issues is illustrated by specific examples described in the materials cited: Levon Brooks and Kennedy Brewer, each convicted of separate child murders in the 1990s almost entirely on the basis of bitemark analysis testimony, spent more than 13 years in prison before DNA testing identified the actual perpetrator, who confessed to both crimes; Santae Tribble, convicted of murder after an FBI analyst testified that hair from a stocking mask linked Tribble to the crime and "matched in all microscopic characteristics," spent more than 20 years in prison before DNA testing revealed that none of the 13 hairs belonged to Tribble and that one came from a dog; Jimmy Ray Bromgard of Montana served 15 years in prison for rape before DNA testing showed that hairs collected from the victim's bed and reported as a match to Bromgard's could not have come from him; Stephan Cowans, convicted of shooting a Boston police officer after two fingerprint experts testified that a thumbprint left by the perpetrator was "unique and

are not due simply to poor performance by a few practitioners, but rather to the fact that the reliability of many forensic feature-comparison methods has never been meaningfully evaluated.[95]

Compared to many types of expert testimony, testimony based on forensic feature-comparison methods poses unique dangers of misleading jurors for two reasons:

- The vast majority of jurors have no independent ability to interpret the probative value of results based on the detection, comparison, and frequency of scientific evidence. If matching halves of a ransom note were found at a crime scene and at a defendant's home, jurors could rely on their own experiences to assess how unlikely it is that two torn scraps would match if they were not in fact from a single original note. If a witness were to describe a perpetrator as "tall and bushy haired," jurors could make a reasonable judgment of how many people might match the description. But, if an expert witness were to say that, in two DNA samples, the third exon of the *DYNC1H1* gene is precisely 174 nucleotides in length, most jurors would have no way to know if they should be impressed by the coincidence; they would be completely dependent on expert statements garbed in the mantle of science. (As it happens, they should not be impressed by the preceding statement: At the DNA locus cited, more than 99.9 percent of people have a fragment of the indicated size.[96])

- The potential prejudicial impact is unusually high, because jurors are likely to overestimate the probative value of a "match" between samples. Indeed, the DOJ itself historically overestimated the probative value of matches in its longstanding contention, now acknowledged to be inappropriate, that latent fingerprint analysis was "infallible."[97] Similarly, a former head of the FBI's fingerprint unit testified that the FBI had "an error rate of one per every 11 million cases."[98] In an online experiment, researchers asked mock jurors to estimate the frequency that a qualified, experienced forensic scientist would mistakenly conclude that two samples of specified types came from the same person when they actually came from two different people. The mock jurors believed such errors are likely to occur about 1 in 5.5 million for fingerprint analysis comparison; 1 in 1 million for bitemark comparison; 1 in 1 million for hair comparison; and 1 in 100 thousand for handwriting comparison.[99] While precise error rates are not known for most of these techniques, all indications point to the actual error rates being orders of magnitude higher. For example, the FBI's own studies of latent fingerprint analysis point to error rates in the range of one in several hundred.[100] (Because the term "match" is likely to imply an

identical," spent more than 5 years in prison before DNA testing on multiple items of evidence excluded him as the perpetrator; and Steven Barnes of upstate New York served 20 years in prison for a rape and murder he did not commit after a criminalist testified that a photographic overlay of fabric from the victim's jeans and an imprint on Barnes' truck showed patterns that were "similar" and hairs collected from the truck were similar to the victim's hairs.

[95] See: Chapter 5.

[96] See: ExAC database: exac.broadinstitute.org/gene/ENSG00000197102.

[97] See: www.justice.gov/olp/file/861906/download.

[98] *U.S. v. Baines* 573 F.3d 979 (2009) at 984.

[99] Koehler, J.J. "Intuitive error rate estimates for the forensic sciences." (August 2, 2016). Available at papers.ssrn.com/sol3/papers.cfm?abstract_id=2817443 .

[100] See: Section 5.4.

inappropriately high probative value, a more neutral term should be used for an examiner's belief that two samples come from the same source. We suggest the term "*proposed* identification" to appropriately convey the examiner's conclusion, along with the possibility that it might be wrong. We will use this term throughout this report.)

This chapter lays out PCAST's conclusions concerning the scientific criteria for scientific validity. The conclusions are based on the fundamental principles of the "scientific method"—applicable throughout science—that valid scientific knowledge can *only* be gained through *empirical* testing of specific propositions.[101] PCAST's conclusions in the chapter might be briefly summarized as follows:

> *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.*

The chapter is organized as follows:

- The first section describes the distinction between two fundamentally different types of feature-comparison methods: objective methods and subjective methods.

- The next five sections discuss the scientific criteria for the two types of scientific validity: foundational validity and validity as applied.

- The final two sections discuss views held in the forensic community.

4.1 Feature-Comparison Methods: Objective and Subjective Methods

A forensic feature-comparison method is a procedure by which an examiner seeks to determine whether an evidentiary sample (e.g., from a crime scene) is or is not associated with a source sample (e.g., from a suspect)[102] based on similar features. The evidentiary sample might be DNA, hair, fingerprints, bitemarks, toolmarks, bullets, tire tracks, voiceprints, visual images, and so on. The source sample would be biological material or an item (tool, gun, shoe, or tire) associated with the suspect.

[101] For example, the Oxford Online Dictionary defines the scientific method as "a method or procedure that has characterized the natural sciences since the 17th century, consisting in systematic observation, measurement, and experimentation, and the formulation, testing, and modification of hypotheses." "Scientific method" *Oxford Dictionaries Online*. Oxford University Press (accessed on August 19, 2016).

[102] A "source sample" refers to a specific individual or object (e.g., a tire or gun).

Feature-comparison methods may be classified as either objective or subjective. By objective feature-comparison methods, we mean methods consisting of procedures that are each defined with enough standardized and quantifiable detail that they can be performed by either an automated system or human examiners exercising little or no judgment. By subjective methods, we mean methods including key procedures that involve significant human judgment—for example, about which features to select or how to determine whether the features are sufficiently similar to be called a proposed identification.

Objective methods are, in general, preferable to subjective methods. Analyses that depend on human judgment (rather than a quantitative measure of similarity) are obviously more susceptible to human error, bias, and performance variability across examiners.[103] In contrast, objective, quantified methods tend to yield greater accuracy, repeatability and reliability, including reducing variation in results among examiners. Subjective methods can evolve into or be replaced by objective methods.[104]

4.2 Foundational Validity: Requirement for Empirical Studies

For a metrological method to be scientifically valid and reliable, the procedures that comprise it must be shown, based on empirical studies, to be *repeatable*, *reproducible*, and *accurate*, at levels that have been measured and are appropriate to the intended application.[105,106]

BOX 2. Definition of key terms

By "repeatable," we mean that, with known probability, an examiner obtains the same result, when analyzing samples from the same sources.

By "reproducible," we mean that, with known probability, different examiners obtain the same result, when analyzing the same samples.

By "accurate," we mean that, with known probabilities, an examiner obtains correct results both (1) for samples from the same source (true positives) and (2) for samples from different sources (true negatives).

By "reliability," we mean repeatability, reproducibility, and accuracy.[107]

[103] Dror, I.E. "A hierarchy of expert performance." *Journal of Applied Research in Memory and Cognitio*n, Vol. 5 (2016): 121-127.

[104] For example, before the development of objective tests for intoxication, courts had to rely exclusively on the testimony of police officers and others who in turn relied on behavioral indications of drunkenness and the presence of alcohol on the breath. The development of objective chemical tests drove a change from subjective to objective standards.

[105] National Physical Laboratory. "A Beginner's Guide to Measurement." (2010) available at: www.npl.co.uk/upload/pdf/NPL-Beginners-Guide-to-Measurement.pdf; Pavese, F. "An Introduction to Data Modelling Principles in Metrology and Testing." in *Data Modeling for Metrology and Testing in Measurement Science*, Pavese, F. and A.B. Forbes (Eds.) Birkhäuser (2009).

[106] Feature-comparison methods that get the wrong answer too often have, by definition, low probative value. As discussed above, the prejudicial impact will thus likely to outweigh the probative value.

[107] We note that "reliability" also has a narrow meaning within the field of statistics referring to "consistency"—that is, the extent to which a method produces the same result, regardless of whether the result is accurate. This is not the sense in which "reliability" is used in this report, or in the law.

By “scientific validity,” we mean that a method has shown, based on empirical studies, to be reliable with levels of repeatability, reproducibility, and accuracy that are appropriate to the intended application.

By an “empirical study,” we mean test in which a method has been used to analyze a large number of independent sets of samples, similar in relevant aspects to those encountered in casework, in order to estimate the method’s repeatability, reproducibility, and accuracy.

By a “black-box study,” we mean an empirical study that assesses a subjective method by having examiners analyze samples and render opinions about the origin or similarity of samples.

The method need not be perfect, but it is clearly *essential* that its accuracy has been measured based on appropriate empirical testing and is high enough to be appropriate to the application. Without an appropriate estimate of its accuracy, a metrological method is useless—because one has no idea how to interpret its results. The importance of knowing a method’s accuracy was emphasized by the 2009 NRC report on forensic science and by a 2010 NRC report on biometric technologies.[108]

To meet the scientific criteria of foundational validity, two key elements are required:

(1) a reproducible and consistent procedure for (a) identifying features within evidence samples; (b) comparing the features in two samples; and (c) determining, based on the similarity between the features in two samples, whether the samples should be declared to be a proposed identification (“matching rule”).

(2) empirical measurements, from multiple independent studies, of (a) the method’s false positive rate—that is, the probability it declares a proposed identification between samples that actually come from *different* sources and (b) the method’s sensitivity—that is, probability that it declares a proposed identification between samples that actually come from the *same* source.

We discuss these elements in turn.

Reproducible and Consistent Procedures

For a method to be objective, *each* of the three steps (feature identification, feature comparison, and matching rule) should be precisely defined, reproducible and consistent. Forensic examiners should identify relevant features in the same way and obtain the same result. They should compare features in the same quantitative manner. To declare a proposed identification, they should calculate whether the features in an evidentiary sample and the features in a sample from a suspected source lie within a pre-specified measurement tolerance

[108] “Biometric recognition is an inherently probabilistic endeavor…Consequently, even when the technology and the system it is embedded in are behaving as designed, there is inevitable uncertainty and risk of error.” National Research Council, *“Biometric Recognition: Challenges and Opportunities.”* The National Academies Press. Washington DC. (2010): viii-ix.

(matching rule).[109] For an objective method, one can establish the foundational validity of each of the individual steps by measuring its accuracy, reproducibility, and consistency.

For subjective methods, procedures must still be carefully defined—but they involve substantial human judgment. For example, different examiners may recognize or focus on different features, may attach different importance to the same features, and may have different criteria for declaring proposed identifications. Because the procedures for feature identification, the matching rule, and frequency determinations about features are not objectively specified, the overall procedure must be treated as a kind of "black box" inside the examiner's head.

Subjective methods require careful scrutiny, more generally, their heavy reliance on human judgment means that they are especially vulnerable to human error, inconsistency across examiners, and cognitive bias. In the forensic feature-comparison disciplines, cognitive bias includes the phenomena that, in certain settings, humans (1) may tend naturally to focus on similarities between samples and discount differences and (2) may also be influenced by extraneous information and external pressures about a case.[110] (The latter issues are illustrated by the FBI's misidentification of a latent fingerprint in the Madrid training bombing, discussed on p.9.)

Since the black box in the examiner's head cannot be examined directly for its foundational basis in science, the foundational validity of subjective methods can be established *only* through empirical studies of examiner's performance to determine whether they can provide accurate answers; such studies are referred to as "black-box" studies (Box 2). In black-box studies, many examiners are presented with many independent comparison problems—typically, involving "questioned" samples and one or more "known" samples—and asked to declare whether the questioned samples came from the same source as one of the known samples.[111] The researchers then determine how often examiners reach erroneous conclusions.

[109] If a source is declared *not* to share the same features, it is "excluded" by the test. The matching rule should be chosen carefully. If the "matching rule" is chosen to be too strict, samples that actually come from the same source will be declared a non-match (false negative). If it is too lax, then the method will not have much discriminatory power because the random match probability will be too high (false positive).

[110] See, for example: Boroditsky, L. "Comparison and the development of knowledge." *Cognition*, Vol. 102 (2007): 118-128; Hassin, R. "Making features similar: comparison processes affect perception." *Psychonomic Bulletin & Review*, Vol. 8 (2001): 728–31; Medin, D.L., Goldstone, R.L., and D. Gentner. "Respects for similarity." *Psychological Review*, Vol. 100 (1993): 254–78; Tversky, A. "Features of similarity." *Psychological Review*, Vol. 84 (1977): 327–52; Kim, J., Novemsky, N., and R. Dhar. "Adding small differences can increase similarity and choice." *Psychological Science*, Vol. 24 (2012): 225–9; Larkey, L.B., and A.B. Markman. "Processes of similarity judgment." *Cognitive Science*, Vol. 29 (2005): 1061–76; Medin, D.L., Goldstone, R.L., and A.B. Markman. "Comparison and choice: Relations between similarity processes and decision processes." *Psychonomic Bulletin and Review*, Vol. 2 (1995): 1–19; Goldstone, R. L. "The role of similarity in categorization: Providing a groundwork." *Cognition*, Vol. 52 (1994): 125–57; Nosofsky, R. M. "Attention, similarity, and the identification-categorization relation." *Journal of Experimental Psychology, General*, Vol. 115 (1986): 39–57.

[111] Answers may be expressed in such terms as "match/no match/inconclusive" or "identification/exclusion/inconclusive."

As an excellent example, the FBI recently conducted a black-box study of latent fingerprint analysis, involving 169 examiners and 744 fingerprint pairs, and published the results of the study in a leading scientific journal.[112]

(Some forensic scientists have cautioned that too much attention to the subjective aspects of forensic methods—such as studies of cognitive bias and black-box studies—might distract from the goal of improving knowledge about the objective features of the forensic evidence and developing truly objective methods.[113] Others have noted that this is not currently a problem, because current efforts and funding to address the challenges associated with subjective forensic methods are very limited.[114])

Empirical Measurements of Accuracy

It is necessary to have appropriate empirical measurements of a method's false positive rate and the method's sensitivity. As explained in Appendix A, it is necessary to know these two measures to assess the probative value of a method.

The false positive rate is the probability that the method declares a proposed identification between samples that actually come from *different* sources. For example, a false positive rate of 5 percent means that two samples from *different* sources will (due to limitations of the method) be incorrectly declared to come from the same source 5 percent of the time. (The quantity equal to one minus the false positive rate—95 percent, in the example—is referred to as the specificity.)

The method's sensitivity is the probability that the method declares a proposed identification between samples that actually come from the *same* source. For example, a sensitivity of 90 percent means two samples from the same source will be declared to come from the same source 90 percent of the time, and declared to come from different sources 10 percent of the time. (The latter quantity is referred to as the false negative rate.)

The false positive rate is especially important because false positive results can lead directly to wrongful convictions.[115] In some circumstances, it may be possible to estimate a false positive rate related to specific features of the evidence in the case. (For example, the random match probability calculated in DNA analysis depends in part on the specific genotype seen in an evidentiary sample. The false positive rate for latent fingerprint analysis may depend on the quality of the latent print.) For other feature-comparison methods, it may be only possible to make an overall estimate of the average false positive rate across samples.

For objective methods, the false positive rate is composed of two distinguishable sources—coincidental matches (where samples from different sources nonetheless have *features* that fall within the tolerance of the objective matching rule) and human/technical failures (where samples have features that fall outside the matching rule, but where a proposed identification was nonetheless declared due to a human or technical failure). For

[112] Ulery, B.T., Hicklin, R.A., Buscaglia, J., and M.A. Roberts. "Accuracy and reliability of forensic latent fingerprint decisions." *Proceedings of the National Academy of Sciences*, Vol. 108, No. 19 (2011): 7733-8.

[113] Champod, C. "Research focused mainly on bias will paralyse forensic science." *Science & Justice*, Vol. 54 (2014): 107–9.

[114] Risinger, D.M., Thompson, W.C., Jamieson, A., Koppl, R., Kornfield, I., Krane, D., Mnookin, J.L., Rosenthal, R., Saks, M.J., and S.L. Zabell. "Regarding Champod, editorial: "Research focused mainly on bias will paralyse forensic science." *Science and Justice*, Vol. 54 (2014):508-9.

[115] See footnote 94, p. 44. Under some circumstances, false-negative results can contribute to wrongful convictions as well.

objective methods where the probability of coincidental match is very low (such as DNA analysis), the false positive rate in application in a given case will be dominated by the rate of human/technical failures—which may well be hundreds of times larger.

For subjective methods, both types of error—coincidental matches and human/technical failures—occur as well, but, without an objective "matching rule," the two sources cannot be distinguished. In establishing foundational validity, it is thus essential to perform black-box studies that empirically measure the overall error rate across many examiners. (See Box 3 concerning the word "error.")

BOX 3. The meanings of "error"

The term "error" has differing meanings in science and law, which can lead to confusion. In legal settings, the term "error" often implies fault—e.g., that a person has made a mistake that could have been avoided if he or she had properly followed correct procedures or a machine has given an erroneous result that could have been avoided it if had been properly calibrated. In science, the term "error" also includes the situation in which the procedure itself, when properly applied, does not yield the correct answer owing to chance occurrence.

When one applies a forensic feature-comparison method with the goal of assessing whether two samples did or did not come from the same source, coincidental matches and human/technical failures are both regarded, from a statistical point of view, as "errors" because both can lead to incorrect conclusions.

Studies designed to estimate a method's false positive rate and sensitivity are necessarily conducted using only a finite number of samples. As a consequence, they cannot provide "exact" values for these quantities (and should not claim to do so), but only "confidence intervals," whose bounds reflect, respectively, the range of values that are reasonably compatible with the results. When reporting a false positive rate to a jury, it is scientifically important to state the "upper 95 percent one-sided confidence bound" to reflect the fact that the actual false positive rate could reasonably be as high as this value.[116] (For more information, see Appendix A.)

Studies often categorize their results as being conclusive (e.g., identification or exclusion) or inconclusive (no determination made).[117] When reporting a false positive rate to a jury, it is scientifically important to calculate the rate based on the proportion of *conclusive* examinations, rather than just the proportion of all examinations. This is appropriate because evidence used against a defendant will typically be based on *conclusive*, rather than inconclusive, examinations. To illustrate the point, consider an extreme case in which a method had been

[116] The upper confidence bound properly incorporates the precision of the estimate based on the sample size. For example, if a study found no errors in 100 tests, it would be misleading to tell a jury that the error rate was 0 percent. In fact, if the tests are independent, the upper 95 percent confidence bound for the true error rate is 3.0 percent. Accordingly a jury should be told that the error rate could be as high as 3.0 percent (that is, 1 in 33). The true error rate could be higher, but with rather small probability (less than 5 percent). If the study were much smaller, the upper 95 percent confidence limit would be higher. For a study that found no errors in 10 tests, the upper 95 percent confidence bound is 26 percent—that is, the actual false positive rate could be roughly 1 in 4 (see Appendix A).

[117] See: Chapter 5.

tested 1000 times and found to yield 990 inconclusive results, 10 false positives, and no correct results. It would be misleading to report that the false positive rate was 1 percent (10/1000 examinations). Rather, one should report that 100 percent of the conclusive results were false positives (10/10 examinations).

Whereas exploratory scientific studies may take many forms, scientific *validation* studies—intended to assess the validity and reliability of a metrological method for a particular forensic feature-comparison application—must satisfy a number of criteria, which are described in Box 4.

BOX 4. Key criteria for validation studies to establish foundational validity

Scientific validation studies—intended to assess the validity and reliability of a metrological method for a particular forensic feature-comparison application—must satisfy a number of criteria.

(1) The studies must involve a sufficiently large number of examiners and must be based on sufficiently *large* collections of *known* and *representative* samples from *relevant* populations to reflect the range of features or combinations of features that will occur in the application. In particular, the sample collections should be:

(a) representative of the quality of evidentiary samples seen in real cases. (For example, if a method is to be used on distorted, partial, latent fingerprints, one must determine the *random match probability*—that is, the probability that the match occurred by chance—for distorted, partial, latent fingerprints; the random match probability for full scanned fingerprints, or even very high quality latent prints would not be relevant.)

(b) chosen from populations relevant to real cases. For example, for features in biological samples, the false positive rate should be determined for the overall US population and for major ethnic groups, as is done with DNA analysis.

(c) large enough to provide appropriate estimates of the error rates.

(2) The empirical studies should be conducted so that neither the examiner nor those with whom the examiner interacts have any information about the correct answer.

(3) The study design and analysis framework should be specified in advance. In validation studies, it is inappropriate to modify the protocol afterwards based on the results.[118]

[118] The analogous situation in medicine is a clinical trial to test the safety and efficacy of a drug for a particular application. In the design of clinical trials, FDA requires that criteria for analysis must be pre-specified and notes that *post hoc* changes to the analysis compromise the validity of the study. See: FDA Guidance: "Adaptive Designs for Medical Device Clinical Studies" (2016) Available at: www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm446729.pdf; Alosh, M., Fritsch, K., Huque, M., Mahjoob, K., Pennello, G., Rothmann, M., Russek-Cohen, E., Smith, F., Wilson, S., and L. Yue. "Statistical considerations on subgroup analysis in clinical trials." *Statistics in Biopharmaceutical Research*, Vol. 7 (2015): 286-303; FDA Guidance: "Design Considerations for Pivotal Clinical Investigations for Medical Devices" (2013) (available at:

(4) The empirical studies should be conducted or overseen by individuals or organizations that have no stake in the outcome of the studies.[119]

(5) Data, software and results from validation studies should be available to allow other scientists to review the conclusions.

(6) To ensure that conclusions are reproducible and robust, there should be multiple studies by separate groups reaching similar conclusions.

An empirical measurement of error rates is not simply a desirable feature; it is *essential* for determining whether a method is foundationally valid. In science, a testing procedure—such as testing whether a person is pregnant or whether water is contaminated—is not considered valid until its reliability has been *empirically* measured. For example, we need to know how often the pregnancy test declares a pregnancy when there is none, and *vice versa*. The same scientific principles apply no less to forensic tests, which may contribute to a defendant losing his life or liberty.

Importantly, error rates cannot be inferred from casework, but rather must be determined based on samples where the correct answer is known. For example, the former head of the FBI's fingerprint unit testified that the FBI had "an error rate of one per every 11 million cases" based on the fact that the agency was known to have made only one mistake over the past 11 years, during which time it had made 11 million identifications.[120] The fallacy is obvious: the expert simply *assumed without evidence* that every error in casework had come to light.

Why is it essential to know a method's false positive rate and sensitivity? Because without appropriate empirical measurement of a method's accuracy, the fact that two samples in a particular case show similar features has *no probative value*—and, as noted above, it may have considerable prejudicial impact because juries will likely incorrectly attach meaning to the observation.[121]

www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm373750.htm); FDA Guidance for Industry: E9 Statistical Principles for Clinical Trials (September 1998) (available at: www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm073137.pdf); Pocock, S.J. Clinical trials: a practical approach. Wiley, Chichester (1983).

[119] In the setting of clinical trials, the sponsor of the trial (a pharmaceutical, device or biotech company or, in some cases, an academic institutions) funds and initiates the study, but the trial is conducted by individuals who are independent of the sponsor (often, academic physicians), in order to ensure the reliability of the data generated by the study and minimize the potential for bias. See, for example, 21 C.F.R. § 312.3 and 21 C.F.R. § 54.4(a).

[120] *U.S. v. Baines* 573 F.3d 979 (2009) at 984.

[121] Under Fed. R. Evid., Rule 403, evidence should be excluded "if its probative value is substantially outweighed by the danger of unfair prejudice."

The absolute need, from a scientific perspective, for empirical data is elegantly expressed in an analogy by U.S. District Judge John Potter in his opinion in *U.S. v. Yee (1991),* an early case on the use of DNA analysis:

> *Without the probability assessment, the jury does not know what to make of the fact that the patterns match: the jury does not know whether the patterns are as common as pictures with two eyes, or as unique as the Mona Lisa.*[122,123]

4.3 Foundational Validity: Requirement for Scientifically Valid Testimony

It should be obvious—but it bears emphasizing—that once a method has been established as foundationally valid based on appropriate empirical studies, claims about the method's accuracy and the probative value of proposed identifications, in order to be valid, must be based on such empirical studies. *Statements claiming or implying greater certainty than demonstrated by empirical evidence are scientifically invalid*. Forensic examiners should therefore report findings of a proposed identification with clarity and restraint, explaining in each case that the fact that two samples satisfy a method's criteria for a proposed match does not necessarily imply that the samples come from a common source. If the false positive rate of a method has been found to be 1 in 50, experts should not imply that the method is able to produce results at a higher accuracy.

Troublingly, expert witnesses sometimes go beyond the empirical evidence about the frequency of features—even to the extent of claiming or implying that a sample came from a specific source with near-certainty or even absolute certainty, despite having no scientific basis for such opinions.[124] From the standpoint of scientific validity, experts should never be permitted to state or imply in court that they can draw conclusions with certainty or near-certainty (such as "zero," "vanishingly small," "essentially zero," "negligible," "minimal," or "microscopic" error rates; "100 percent certainty" or "to a reasonable degree of scientific certainty;" or identification "to the exclusion of all other sources."[125]

The scientific inappropriateness of such testimony is aptly captured by an analogy by District of Columbia Court of Appeals Judge Catharine Easterly in her concurring opinion in *Williams v. United States*, a case in which an examiner testified that markings on certain bullets were unique to a gun recovered from a defendant's apartment:

[122] *U.S. v. Yee,* 134 F.R.D. 161 (N.D. Ohio 1991).

[123] Some courts have ruled that there is no harm in admitting feature-comparison evidence on the grounds that jurors can see the features with their own eyes and decide for themselves about whether features are shared. *U.S. v. Yee* shows why this reasoning is fallacious: jurors have no way to know how often two different samples would share features, and to what level of specificity.

[124] As noted above, the long history of exaggerated claims for the accuracy of forensic methods includes the DOJ's own prior statement that latent fingerprint analysis was "infallible," which the DOJ has judged to have been inappropriate. www.justice.gov/olp/file/861906/download.

[125] Cole, S.A. "Grandfathering evidence: Fingerprint admissibility rulings from Jennings to Llera Plaza and back again." *41 American Criminal Law Review, 1189* (2004). See also: National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (NRC Report, 2009): 87, 104, and 143.

> *As matters currently stand, a certainty statement regarding toolmark pattern matching has the same probative value as the vision of a psychic: it reflects nothing more than the individual's foundationless faith in what he believes to be true. This is not evidence on which we can in good conscience rely, particularly in criminal cases, where we demand proof—real proof—beyond a reasonable doubt, precisely because the stakes are so high.*[126]

In science, assertions that a metrological method is more accurate than has been empirically demonstrated are rightly regarded as mere speculation, not valid conclusions that merit credence.

4.4 Neither Experience nor Professional Practices Can Substitute for Foundational Validity

In some settings, an expert may be scientifically capable of rendering judgments based primarily on his or her "experience" and "judgment." Based on experience, a surgeon might be scientifically qualified to offer a judgment about whether another doctor acted appropriately in the operating theater or a psychiatrist might be scientifically qualified to offer a judgment about whether a defendant is mentally competent to assist in his or her defense.

By contrast, "experience" or "judgment" cannot be used to establish the scientific validity and reliability of a metrological method, such as a forensic feature-comparison method. The frequency with which a particular pattern or set of features will be observed in different samples, which is an essential element in drawing conclusions, is not a matter of "judgment." It is an empirical matter for which only empirical evidence is relevant. Moreover, a forensic examiner's "experience" from extensive casework is not informative—because the "right answers" are not typically known in casework and thus examiners cannot accurately know how often they erroneously declare matches and cannot readily hone their accuracy by learning from their mistakes in the course of casework.

Importantly, good professional practices—such as the existence of professional societies, certification programs, accreditation programs, peer-reviewed articles, standardized protocols, proficiency testing, and codes of ethics—cannot substitute for actual evidence of scientific validity and reliability.[127]

Similarly, an expert's expression of *confidence* based on personal professional experience or expressions of *consensus* among practitioners about the accuracy of their field is no substitute for error rates estimated from relevant studies. For a method to be *reliable*, empirical evidence of validity, as described above, is required.

Finally, the points above underscore that scientific validity of a method must be assessed within the framework of the broader scientific field of which it is a part (e.g., measurement science in the case of feature-comparison methods). The fact that bitemark examiners defend the validity of bitemark examination means little.

[126] *Williams v. United States,* DC Court of Appeals, decided January 21, 2016, (Easterly, concurring).

[127] For example, both scientific and pseudoscientific disciplines employ such practices.

4.5 Validity as Applied: Key Elements

Foundational validity means that a method can, *in principle,* be reliable. Validity as applied means that the method has been reliably applied *in practice.* It is the *scientific* concept we mean to correspond to the legal requirement, in Rule 702(d), that an expert "has reliably applied the principles and methods to the facts of the case."

From a scientific standpoint, certain criteria are essential to establish that a forensic practitioner has reliably applied a method to the facts of a case. These elements are described in Box 5.

BOX 5. Key criteria for validity as applied

(1) The forensic examiner must have been shown to be *capable* of reliably applying the method and must *actually* have done so. Demonstrating that an examiner is *capable* of reliably applying the method is crucial—especially for subjective methods, in which human judgment plays a central role. From a scientific standpoint, the ability to apply a method reliably can be demonstrated only through empirical testing that measures how often the expert reaches the correct answer. (Proficiency testing is discussed more extensively on p. 57-59.) Determining whether an examiner has *actually* reliably applied the method requires that the procedures actually used in the case, the results obtained, and the laboratory notes be made available for scientific review by others.

(2) Assertions about the probability of the observed features occurring by chance must be scientifically valid.

(a) The forensic examiner should report the overall false positive rate and sensitivity for the method established in the studies of foundational validity and should demonstrate that the samples used in the foundational studies are relevant to the facts of the case.[128]

(b) Where applicable, the examiner should report the random match probability based on the specific features observed in the case.

(c) An expert should not make claims or implications that go beyond the empirical evidence and the applications of valid statistical principles to that evidence.

[128] For example, for DNA analysis, the frequency of genetic variants is known to vary among ethnic groups; it is thus important that the sample collection reflect relevant ethnic groups to the case at hand. For latent fingerprints, the risk of falsely declaring an identification may be higher when latent fingerprints are of lower quality; so, to be relevant, the sample collections used to estimate accuracy should be based on latent fingerprints comparable in quality and completeness to the case at hand.

4.6 Validity as Applied: Proficiency Testing

Even when a method is foundationally valid, there are many reasons why examiners may not always get the right result.[129] As discussed above, the *only* way to establish scientifically that an examiner is capable of applying a foundationally valid method is through appropriate empirical testing to measure how often the examiner gets the correct answer.

Such empirical testing is often referred to as "proficiency testing." We note that term "proficiency testing" is sometimes used to refer to many different other types of testing—such as (1) tests to determine whether a practitioner reliably follows the steps laid out in a protocol, without assessing the *accuracy* of their conclusions, and (2) practice exercises that help practitioners improve their skills by highlighting their errors, without accurately reflect the circumstances of actual casework.

In this report, we use the term proficiency testing to mean ongoing empirical tests to "evaluate the capability and performance of analysts."[130, 131, 132]

Proficiency testing should be performed under conditions that are representative of casework and on samples, for which the true answer is known, that are representative of the full range of sample types and quality likely to be encountered in casework in the intended application. (For example, the fact that an examiner passes a proficiency test involving DNA analysis of simple, single-source samples does not demonstrate that they are capable of DNA analysis of complex mixtures of the sort encountered in casework; see p. 76-81.)

To ensure integrity, proficiency testing should be overseen by a disinterested third party that has no institutional or financial incentive to skew performance. We note that testing services have stated that forensic community prefers that tests not be too challenging.[133]

[129] J.J. Koehler has enumerated a number of possible problems that could, in principle, occur: features may be mismeasured; samples may be interchanged, mislabeled, miscoded, altered, or contaminated; equipment may be miscalibrated; technical glitches and failures may occur without warning and without being noticed; and results may be misread, misinterpreted, misrecorded, mislabeled, mixed up, misplaced, or discarded. Koehler, J.J. "Forensics or fauxrensics? Ascertaining accuracy in the forensic sciences." papers.ssrn.com/sol3/papers.cfm?abstract_id=2773255 (accessed June 28, 2016).

[130] ASCLD/LAB Supplemental Requirements for Accreditation of Forensic Testing Laboratories. des.wa.gov/SiteCollectionDocuments/About/1063/RFP/Add7_Item4ASCLD.pdf.

[131] We note that proficiency testing is not intended to estimate the inherent error rates of a method; these rates should be assessed from foundational validity studies.

[132] Proficiency testing should also be distinguished from "competency testing," which is "the evaluation of a person's knowledge and ability prior to performing independent work in forensic casework." des.wa.gov/SiteCollectionDocuments/About/1063/RFP/Add7_Item4ASCLD.pdf.

[133] Christopher Czyryca, the president of Collaborative Testing Services, Inc., the leading proficiency testing firm in the U.S., has publicly stated that "Easy tests are favored by the community." August 2015 meeting of the National Commission on Forensic Science, a presentation at the Accreditation and Proficiency Testing Subcommittee. www.justice.gov/ncfs/file/761061/download.

As noted previously, false positive rates consist of both coincidental match rates and technical/human failure rates. For some technologies (such as DNA analysis), the latter may be hundreds of times higher than the former.

Proficiency testing is especially critical for subjective methods: because the procedure is not based solely on objective criteria but relies on human judgment, it is inherently vulnerable to error and inter-examiner variability. Each examiner should be tested, because empirical studies have noted considerable differences in accuracy across examiners.[134,135]

The test problems used in proficiency tests should be publicly released after the test is completed, to enable scientists to assess the appropriateness and adequacy of the test for their intended purpose.

Finally, proficiency testing should *ideally* be conducted in a 'test-blind' manner—that is, with samples inserted into the flow of casework such that examiners do not know that they are being tested. (For example, the Transportation Security Administration conducts blind tests by sending weapons and explosives inside luggage through screening checkpoints to see how often TSA screeners detect them.) It has been established in many fields (including latent fingerprint analysis) that, when individuals are aware that they are being tested, they perform differently than they do in the course of their daily work (referred to as the "Hawthorne Effect").[136,137]

While test-blind proficiency testing is ideal, there is disagreement in the forensic community about its feasibility in all settings. On the one hand, laboratories vary considerably as to the type of cases they receive, how evidence is managed and processed, and what information is provided to an analyst about the evidence or the case in question. Accordingly, blinded, inter-laboratory proficiency tests may be difficult to design and

[134] For example, a 2011 study on latent fingerprint decisions observed that examiners frequently differed on whether fingerprints were suitable for reaching a conclusion. Ulery, B.T., Hicklin, R.A., Buscaglia, J., and M.A. Roberts. "Accuracy and reliability of forensic latent fingerprint decisions." *Proceedings of the National Academy of Sciences*, Vol. 108, No. 19 (2011): 7733-8.

[135] It is not sufficient to point to proficiency testing on volunteers in a laboratory, because better performing examiners are more likely to participate. Koehler, J.J. "Forensics or fauxrensics? Ascertaining accuracy in the forensic sciences." papers.ssrn.com/sol3/papers.cfm?abstract_id=2773255 (accessed June 28, 2016).

[136] Concerning the Hawthorne effect, see, for example: Bracht, G.H., and G.V. Glass. "The external validity of experiments." *American Educational Research Journal,* Vol. 5, No. 4 (1968): 437-74; Weech, T.L. and H. Goldhor. "Obtrusive versus unobtrusive evaluation of reference service in five Illinois public libraries: A pilot study." *Library Quarterly: Information, Community, Policy*, Vol. 52, No. 4 (1982): 305-24; Bouchet, C., Guillemin, F., and S. Braincon. "Nonspecific effects in longitudinal studies: impact on quality of life measures." *Journal of Clinical Epidemiology,* Vol. 49, No. 1 (1996): 15-20; Mangione-Smith, R., Elliott, M.N., McDonald, L., and E.A. McGlynn. "An observational study of antibiotic prescribing behavior and the Hawthorne Effect." *Health Services Research,* Vol. 37, No. 6 (2002): 1603-23; Mujis, D. "Measuring teacher effectiveness: Some methodological reflections." *Educational Research and Evaluation*, Vol. 12, No. 1 (2006): 53–74; and McCarney, R., Warner, J., Iliffe, S., van Haselen, R., Griffin, M., and P. Fisher. "The Hawthorne Effect: a randomized, controlled trial." *BMC Medical Research Methodology*, Vol. 7, No. 30 (2007).

[137] For demonstrations that forensic examiners change their behavior when they know their performance is being monitored in particular ways, see Langenburg, G. "A performance study of the ACE-V process: A pilot study to measure the accuracy, precision, reproducibility, repeatability, and biasability of conclusions resulting from the ACE-V process." *Journal of Forensic Identification*, Vol. 59, No. 2 (2009).

orchestrate on a large scale.[138] On the other hand, test-blind proficiency tests have been used for DNA analysis,[139] and select labs have begun to implement this type of testing, in-house, as part of their quality assurance programs.[140] We note that test-blind proficiency testing is much easier to adopt in laboratories that have adopted "context management procedures" to reduce contextual bias.[141]

PCAST believes that test-blind proficiency testing of forensic examiners should be vigorously pursued, with the expectation that it should be in wide use, at least in large laboratories, within the next five years. However, PCAST believes that it is not yet realistic to require test-blind proficiency testing because the procedures for test-blind proficiency tests have not yet been designed and evaluated.

While only non-test-blind proficiency tests are used to support validity as applied, it is scientifically important to report this limitation, including to juries—because, as noted above, non-blind proficiency tests are likely to overestimate the accuracy because the examiners knew they were being tested.

4.7 Non-Empirical Views in the Forensic Community

While the scientific validity of metrological methods requires empirical demonstration of accuracy, there have historically been efforts in the forensic community to justify non-empirical approaches. This is of particular concern because such views are sometimes mistakenly codified in policies or practices. These heterodox views typically involve four recurrent themes, which we review below.

"Theories" of Identification

A common argument is that forensic practices should be regarded as valid because they rest on scientific "theories" akin to the fundamental laws of physics, that should be accepted because they have been tested and not "falsified."[142]

An example is the "Theory of Identification as it Relates to Toolmarks," issued in 2011 by the Association of Firearm and Tool Mark Examiners.[143,144] It states in its entirety:

[138] Some of the challenges associated with designing blind inter-laboratory proficiency tests may be addressed if the forensic laboratories were to move toward a system where an examiner's knowledge of a case were limited to domain-relevant information.

[139] See: Peterson, J.L., Lin, G., Ho, M., Chen, Y., and R.E. Gaensslen. "The feasibility of external blind DNA proficiency testing. II. Experience with actual blind tests." *Journal of Forensic Science,* Vol. 48, No. 1 (2003): 32-40.

[140] For example, the Houston Forensic Science Center has implemented routine, blind proficiency testing for its firearms examiners and chemistry analysis unit, and is planning to carry out similar testing for its DNA and latent print examiners.

[141] For background, see www.justice.gov/ncfs/file/888586/download.

[142] See: www.swggun.org/index.php?option=com_content&view=article&id=66:the-foundations-of-firearm-and-toolmark-identification&catid=13:other&Itemid=43 and www.justice.gov/ncfs/file/888586/download.

[143] 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.

[144] Firearms analysis is considered in detail in Chapter 5.

> *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 compare to the corresponding features in the second set of surface contours. Agreement is significant when the agreement in individual characteristics exceeds the best agreement demonstrated between toolmarks known to have been produced by different tools and is consistent with agreement demonstrated by toolmarks known to have been produced by the same tool. The statement that "sufficient agreement" exists between two toolmarks means that the agreement of individual characteristics is of a quantity and quality that the likelihood another tool could have made the mark is so remote as to be considered a practical impossibility.*
>
> *3. Currently the interpretation of individualization/identification is subjective in nature, founded on scientific principles and based on the examiner's training and experience.*

The statement is clearly not a scientific theory, which the National Academy of Sciences has defined as "a comprehensive explanation of some aspect of nature that is supported by a vast body of evidence."[145] Rather, it is a claim that examiners applying a subjective approach can accurately individualize the origin of a toolmark. Moreover, a "theory" is not what is needed. What is needed are empirical tests to see how well the method performs.

More importantly, the stated method is circular. It declares that an examiner may state that two toolmarks have a "common origin" when their features are in "sufficient agreement." It then defines "sufficient agreement" as occurring when the examiner considers it a "practical impossibility" that the toolmarks have different origins. (In response to PCAST's concern about this circularity, the FBI Laboratory replied that: "'Practical impossibility' is the certitude that exists when there is sufficient agreement in the quality and quantity of individual characteristics."[146] This answer did not resolve the circularity.)

Focus on 'Training and Experience' Rather Than Empirical Demonstration of Accuracy

Many practitioners hold an honest belief that they are able to make accurate judgments about identification based on their training and experience. This notion is explicit in the AFTE's *Theory of Identification*, which notes that interpretation is subjective in nature, "based on an examiner's training and experience." Similarly, the leading textbook on footwear analysis states,

> *Positive identifications may be made with as few as one random identifying characteristic, but only if that characteristic is confirmable; has sufficient definition, clarity, and features; is in the same location and*

[145] See: www.nas.edu/evolution/TheoryOrFact.html.
[146] Communication from FBI Laboratory to PCAST (June 6, 2016).

> *orientation on the shoe outsole; and in the opinion of an experienced examiner, would not occur again on another shoe.*[147] *[emphasis added]*

In effect, it says, positive identification depends on the examiner being *positive* about the identification.

"Experience" is an inadequate foundation for drawing judgments about whether two sets of features could have been produced by (or found on) different sources. Even if examiners could recall in sufficient detail all the patterns or sets of features that they have seen, they would have no way of knowing accurately in which cases two patterns actually came from different sources, because the correct answers are rarely known in casework.

The fallacy of relying on "experience" was evident in testimony by a former head of the FBI's fingerprint unit (discussed above) that the FBI had "an error rate of one per every 11 million cases," based on the fact that the agency was only aware of one mistake.[148] By contrast, recent empirical studies by the FBI Laboratory (discussed in Chapter 5) indicate error rates of roughly one in several hundred.

"Training" is an even weaker foundation. The mere fact that an individual has been trained in a method does not mean that the method itself is scientifically valid nor that the individual is capable of producing reliable answers when applying the method.

Focus on 'Uniqueness' Rather Than Accuracy

Many forensic feature-comparison disciplines are based on the premise that various sets of features (for example, fingerprints, toolmarks on bullets, human dentition, and so on) are "unique."[149]

[147] Bodziak, W. J. *Footwear Impression Evidence: Detection, Recovery, and Examination*. 2nd ed. CRC Press-Taylor & Francis, Boca Raton, Florida (2000).

[148] *U.S. v. Baines* 573 F.3d 979 (2009) at 984.

[149] For fingerprints, see, for example: Wertheim, Kasey. "Letter re: ACE-V: Is it scientifically reliable and accurate?" *Journal of Forensic Identification*, Vol. 52 (2002): 669 ("The law of biological uniqueness states that exact replication of any given organism cannot occur (nature never repeats itself), and, therefore, no biological entity will ever be exactly the same as another") and Budowle, B., Buscaglia, J., and R.S. Perlman. "Review of the scientific basis for friction ridge comparisons as a means of identification: committee findings and recommendations." *Forensic Science Communications*, Vol. 8 (2006) ("The use of friction ridge skin comparisons as a means of identification is based on the assumptions that the pattern of friction ridge skin is both unique and permanent"). For firearms, see, for example, Riva, F., and C. Christope. "Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases." *Journal of Forensic Sciences*, Vol. 59, (2014): 637 ("The ability to identify a firearm as the source of a questioned cartridge case or bullet is based on two tenets constituting the scientific foundation of the discipline. The first assumes the uniqueness of impressions left by the firearms") and SWGGUN Admissibility Resource Kit (ARK): Foundational Overview of Firearm/Toolmark Identification. available at: afte.org/resources/swggun-ark ("The basis for identification in Toolmark Identification is founded on the principle of uniqueness . . . wherein, all objects are unique to themselves and thus can be differentiated from one another"). For bitemarks, see, for example, Kieser, J.A., Bernal, V., Neil Waddell, J., and S. Raju. "The uniqueness of the human anterior dentition: a geometric morphometric analysis." *Journal of Forensic Sciences,* Vol. 52 (2007): 671-7 ("There are two postulates that underlie all bitemark analyses: first, that the characteristics of the anterior teeth involved in the bite are unique, and secondly, that this uniqueness is accurately recorded in the material bitten.") and Pretty, I.A. "Resolving Issues in Bitemark Analysis" in *Bitemark Evidence: A Color Atlas* R.B.J Dorian, Ed. CRC Press. Chicago (2011) ("Bitemark

The forensics science literature contains many "uniqueness" studies that go to great lengths to try to establish the correctness of this premise.[150] For example, a 2012 paper studied 39 Adidas Supernova Classic running shoes (size 12) worn by a single runner over 8 years, during which time he kept a running journal and ran over the same types of surfaces. [151] After applying black shoe polish to the soles of the shoes, the author asked the runner to carefully produce tread marks on sheets of legal paper on a hardwood floor. The author showed that it was possible to identify small identifying differences between the tread marks produced by different pairs of shoes.

Yet, uniqueness studies miss the fundamental point. The issue is not whether *objects* or *features* differ; they surely do if one looks at a fine enough level. The issue is how well and under what circumstances *examiners* applying a given metrological method can reliably *detect* relevant differences in features to reliably identify whether they share a common source. Uniqueness studies, which focus on the properties of features themselves, can therefore never establish whether a particular *method* for measuring and comparing features is foundationally valid. Only empirical studies can do so.

Moreover, it is not *necessary* for features to be unique in order for them to be useful in narrowing down the source of a feature. Rather, it is essential that there be empirical evidence about how often a method incorrectly attributes the source of a feature.

Decoupling Conclusions about Identification from Estimates of Accuracy

Finally, some hold the view that, when the application of a scientific method leads to a conclusion of an association or proposed identification, it is *unnecessary* to report in court the reliability of the method.[152] As a rationale, it is sometimes argued that it is impossible to measure error rates perfectly or that it is impossible to know the error rate in the *specific* case at hand.

This notion is contrary to the fundamental principle of scientific validity in metrology—namely, that the claim that two objects have been compared and found to have the same property (length, weight, or fingerprint pattern) is meaningless without quantitative information about the reliability of the comparison process.

It is standard practice to study and report error rates in medicine—both to establish the reliability of a method in principle and to assess its implementation in practice. No one argues that measuring or reporting clinical error rates is inappropriate because they might not perfectly reflect the situation for a *specific* patient. If

analysis is based on two postulates: (a) the dental characteristics of anterior teeth involved in biting are unique among individuals, and (b) this asserted uniqueness is transferred and recorded in the injury.").

[150] Some authors have criticized attempts to affirm the uniqueness proposition based on observations, noting that they rest on pure inductive reasoning, a method for scientific investigation that "fell out of favour during the epoch of Sir Francis Bacon in the 16th century." Page, M., Taylor, J., and M. Blenkin. "Uniqueness in the forensic identification sciences—fact or fiction?" *Forensic Science International*, Vol. 206 (2011): 12-8.

[151] Wilson, H.D. "Comparison of the individual characteristics in the outsoles of thirty-nine pairs of Adidas Supernova Classic shoes." *Journal of Forensic Identification*, Vol. 62, No. 3 (2012): 194-204.

[152] See: www.justice.gov/olp/file/861936/download.

transparency about error rates is appropriate for matching blood types before a transfusion, it is appropriate for matching forensic samples—where errors may have similar life-threatening consequences.

We return to this topic in Chapter 8, where we observe that the DOJ's recent proposed guidelines on expert testimony are based, in part, on this scientifically inappropriate view.

4.8 Empirical Views in the Forensic Community

Although some in the forensic community continue to hold views such as those described in the previous section, a growing segment of the forensic science community has responded to the 2009 NRC report with an increased recognition of the need for empirical studies and with initial efforts to undertake them. Examples include published research studies by forensic scientists, assessments of research needs by Scientific Working Groups and OSAC committees, and statements from the NCFS.

Below we highlight several examples from recent papers by forensic scientists:

- *Researchers at the National Academy of Sciences and elsewhere (e.g., Saks & Koehler, 2005; Spinney, 2010) have argued that there is an urgent need to develop objective measures of accuracy in fingerprint identification. Here we present such data.*[153]

- *Tool mark impression evidence, for example, has been successfully used in courts for decades, but its examination has lacked scientific, statistical proof that would independently corroborate conclusions based on morphology characteristics (2–7). In our study, we will apply methods of statistical pattern recognition (i.e., machine learning) to the analysis of toolmark impressions.*[154]

- *The NAS report calls for further research in the area of bitemarks to demonstrate that there is a level of probative value and possibly restricting the use of analyses to the exclusion of individuals. This call to respond must be heard if bite-mark evidence is to be defensible as we move forward as a discipline.*[155]

- *The National Research Council of the National Academies and the legal and forensic sciences communities have called for research to measure the accuracy and reliability of latent print examiners' decisions, a challenging and complex problem in need of systematic analysis. Our research is focused on the development of empirical approaches to studying this problem.*[156]

[153] Tangen, J.M., Thompson, M.B., and D.J. McCarthy. "Identifying fingerprint expertise." *Psychological Science*, Vol. 22, No. 8 (2011): 995-7.

[154] Petraco, N.D., Shenkin, P., Speir, J., Diaczuk, P., Pizzola, P.A., Gambino, C., and N. Petraco. "Addressing the National Academy of Sciences' Challenge: A Method for Statistical Pattern Comparison of Striated Tool Marks." *Journal of Forensic Sciences*, Vol. 57 (2012): 900-11.

[155] Pretty, I.A., and D. Sweet. "A paradigm shift in the analysis of bitemarks." *Forensic Science International*, Vol. 201 (2010): 38-44.

[156] Ulery, B.T., Hicklin, R.A., Buscaglia, J., and M.A., Roberts. "Accuracy and reliability of forensic latent fingerprint decisions." *PNAS*, Vol. 108, No. 19 (2011): 7733-8.

- *We believe this report should encourage the legal community to require that the emerging field of forensic neuroimaging, including fMRI based lie detection, have a proper scientific foundation before being admitted in courts.*[157]

- *An empirical solution which treats the system [referring to voiceprints] as a black box and its output as point values is therefore preferred.*[158]

Similarly, the OSAC and other groups have acknowledged critical research gaps in the evidence supporting various forensic science disciplines and have begun to develop plans to close some of these gaps. We highlight several examples below:

- *While validation studies of firearms and toolmark analysis schemes have been conducted, most have been relatively small data sets. If a large study were well designed and has sufficient participation, it is our anticipation that similar lessons could be learned for the firearms and toolmark discipline.*[159]

- *We are unaware of any study that assesses the overall firearm and toolmark discipline's ability to correctly/consistently categorize evidence by class characteristics, identify subclass marks, and eliminate items using individual characteristics.*[160]

- *Currently there is not a reliable assessment of the discriminating strength of specific friction ridge feature types.*[161]

- *To date there is little scientific data that quantifies the overall risk of close non-matches in AFIS databases. It is difficult to create standards regarding sufficiency for examination or AFIS search searching without this type of research.*[162]

[157] Langleben, D.D., and J.C. Moriarty. "Using brain imaging for lie detection: Where science, law, and policy collide." *Psychology, Public Policy, and Law*, Vol. 19, No. 2 (2013): 222–34.

[158] Morrison, G.S., Zhang, C., and P. Rose. "An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system." *Forensic Science International*, Vol. 208, (2011): 59–65.

[159] OSAC Research Needs Assessment Form. "Study to Assess The Accuracy and Reliability of Firearm and Toolmark." Issued October 2015 (Approved January 2016). Available at: www.nist.gov/forensics/osac/upload/FATM-Research-Needs-Assessment_Blackbox.pdf.

[160] OSAC Research Needs Assessment Form. "Assessment of Examiners' Toolmark Categorization Accuracy." Issued October 2015 (Approved January 2016). Available at: www.nist.gov/forensics/osac/upload/FATM-Research-Needs-Assessment_Class-and-individual-marks.pdf.

[161] OSAC Research Needs Assessment Form. "Assessing the Sufficiency and Strength of Friction Ridge Features." Issued October 2015. Available at: www.nist.gov/forensics/osac/upload/FRS-Research-Need-Assessment-of-Features.pdf.

[162] OSAC Research Needs Assessment Form. "Close Non-Match Assessment." Issued October 2015. Available at: www.nist.gov/forensics/osac/upload/FRS-Research-Need-Close-Non-Match-Assessment.pdf.

- *Research is needed that studies whether sequential unmasking reduces the negative effects of bias during latent print examination.*[163]

- *The IAI has, for many years, sought support for research that would scientifically validate many of the comparative analyses conducted by its member practitioners. While there is a great deal of empirical evidence to support these exams, independent validation has been lacking.*[164]

The National Commission on Forensic Science has similarly recognized the need for rigorous empirical evaluation of forensic methods in a Views Document approved by the commission:

> *All forensic science methodologies should be evaluated by an independent scientific body to characterize their capabilities and limitations in order to accurately and reliably answer a specific and clearly defined forensic question.*[165]

PCAST applauds this growing focus on empirical evidence. We note that increased research funding will be needed to achieve these critical goals (see Chapter 6).

4.9 Summary of Scientific Findings

We summarize our scientific findings concerning the scientific criteria for foundational validity and validity as applied.

Finding 1: Scientific Criteria for Scientific Validity of a Forensic Feature-Comparison Method

(1) Foundational validity. To establish foundational validity for a forensic feature-comparison method, the following elements are required:

(a) a reproducible and consistent procedure for (i) identifying features in evidence samples; (ii) comparing the features in two samples; and (iii) determining, based on the similarity between the features in two sets of features, whether the samples should be declared to be likely to come from the same source ("matching rule"); and

(b) empirical estimates, from appropriately designed studies from multiple groups, that establish (i) the method's false positive rate—that is, the probability it declares a proposed identification between samples that actually come from different sources and (ii) the method's sensitivity—that is, the probability it declares a proposed identification between samples that actually come from the same source.

[163] OSAC Research Needs Assessment Form. "ACE-V Bias." Issued October 2015. Available at: www.nist.gov/forensics/osac/upload/FRS-Research-Need-ACE-V-Bias.pdf.
[164] International Association for Identification. Letter to Patrick J. Leahy, Chairman, Senate Committee on the Judiciary, March 18, 2009. Available at: www.theiai.org/current_affairs/nas_response_leahy_20090318.pdf.
[165] National Commission on Forensic Science: "Views of the Commission Technical Merit Evaluation of Forensic Science Methods and Practices." Available at: www.justice.gov/ncfs/file/881796/download.

As described in Box 4, scientific validation studies should satisfy a number of criteria: (a) they should be based on sufficiently large collections of known and representative samples from relevant populations; (b) they should be conducted so that the examinees have no information about the correct answer; (c) the study design and analysis plan should be specified in advance and not modified afterwards based on the results; (d) the study should be conducted or overseen by individuals or organizations with no stake in the outcome; (e) data, software and results should be available to allow other scientists to review the conclusions; and (f) to ensure that the results are robust and reproducible, there should be multiple independent studies by separate groups reaching similar conclusions.

Once a method has been established as foundationally valid based on adequate empirical studies, claims about the method's accuracy and the probative value of proposed identifications, in order to be valid, must be based on such empirical studies.

For objective methods, foundational validity can be established by demonstrating the reliability of each of the individual steps (feature identification, feature comparison, matching rule, false match probability, and sensitivity).

For subjective methods, foundational validity can be established *only* through black-box studies that measure how often many examiners reach accurate conclusions across many feature-comparison problems involving samples representative of the intended use. In the absence of such studies, a subjective feature-comparison method cannot be considered scientifically valid.

Foundational validity is a *sine qua non*, which can only be shown through empirical studies. Importantly, good professional practices—such as the existence of professional societies, certification programs, accreditation programs, peer-reviewed articles, standardized protocols, proficiency testing, and codes of ethics—cannot substitute for empirical evidence of scientific validity and reliability.

(2) Validity as applied. Once a forensic feature-comparison method has been established as foundationally valid, it is necessary to establish its validity as applied in a given case.

As described in Box 5, validity as applied requires that: (a) the forensic examiner must have been shown to be *capable* of reliably applying the method, as shown by appropriate proficiency testing (see Section 4.6), and must *actually* have done so, as demonstrated by the procedures actually used in the case, the results obtained, and the laboratory notes, which should be made available for scientific review by others; and (b) assertions about the probative value of proposed identifications must be scientifically valid—including that examiners should report the overall false positive rate and sensitivity for the method established in the studies of foundational validity; demonstrate that the samples used in the foundational studies are relevant to the facts of the case; where applicable, report probative value of the observed match based on the specific features observed in the case; and not make claims or implications that go beyond the empirical evidence.

5. Evaluation of Scientific Validity for Seven Feature-Comparison Methods

In the previous chapter, we described the scientific criteria that a forensic feature-comparison method must meet to be considered scientifically valid and reliable, and we underscored the need for empirical evidence of accuracy and reliability.

In this chapter, we illustrate the meaning of these criteria by applying them to six specific forensic feature-comparison methods: (1) DNA analysis of single-source and simple-mixture samples, (2) DNA analysis of complex-mixture samples, (3) bitemarks, (4) latent fingerprints, (5) firearms identification, and (6) footwear analysis.[166] For a seventh forensic feature- comparison method, hair analysis, we do not undertake a full evaluation, but review a recent evaluation by the DOJ.

We evaluate whether these methods have been established to be foundationally valid and reliable and, if so, what estimates of accuracy should accompany testimony concerning a proposed identification, based on current scientific studies. We also briefly discuss some issues related to validity as applied.

PCAST compiled a list of 2019 papers from various sources—including bibliographies prepared by the National Science and Technology Council's Subcommittee on Forensic Science, the relevant Scientific Working Groups (predecessors to the current OSAC),[167] and the relevant OSAC committees; submissions in response to PCAST's request for information from the forensic-science stakeholder community; and our own literature searches.[168] PCAST members and staff identified and reviewed those papers that were relevant to establishing scientific validity. After reaching a set of initial conclusions, input was obtained from the FBI Laboratory and individual scientists at NIST, as well as other experts—including asking them to identify additional papers supporting scientific validity that we might have missed.

For each of the methods, we provide a brief overview of the methodology, discuss background information and studies, and review evidence for scientific validity.

As discussed in Chapter 4, objective methods have well-defined procedures to (1) identify the features in samples, (2) measure the features, (3) determine whether the features in two samples match to within a stated measurement tolerance (matching rule), and (4) estimate the probability that samples from different sources would match (false match probability). It is possible to examine each of these separate steps for their validity

[166] The American Association for the Advancement of Science (AAAS) is conducting an analysis of the underlying scientific bases for the forensic tools and methods currently used in the criminal justice system. As of September 1, 2016 no reports have been issued. See: www.aaas.org/page/forensic-science-assessments-quality-and-gap-analysis.

[167] See: www.nist.gov/forensics/workgroups.cfm.

[168] See: www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_references.pdf.

and reliability. Of the six methods considered in this chapter, only the first two methods (involving DNA analysis) employ objective methods. The remaining four methods are subjective.

For subjective methods, the procedures are not precisely defined, but rather involve substantial expert human judgment. Examiners may focus on certain features while ignoring others, may compare them in different ways, and may have different standards for declaring proposed identification between samples. As described in Chapter 4, the sole way to establish foundational validity is through multiple independent "black-box" studies that measure how often examiners reach accurate conclusions across many feature-comparison problems involving samples representative of the intended use. In the absence of such studies, a feature-comparison method cannot be considered scientifically valid.

PCAST found few black-box studies appropriately designed to assess scientific validity of subjective methods. Two notable exceptions, discussed in this chapter, were a study on latent fingerprints conducted by the FBI Laboratory and a study on firearms identification sponsored by the Department of Defense and conducted by the Department of Energy's Ames Laboratory.

We considered whether proficiency testing, which is conducted by commercial organizations for some disciplines, could be used to establish foundational validity. We concluded that it could not, at present, for several reasons. First, proficiency tests are not intended to establish foundational validity. Second, the test problems or test sets used in commercial proficiency tests are not at present routinely made public—making it impossible to ascertain whether the tests appropriately assess the method across the range of applications for which it is used. The publication and critical review of methods and data is an essential component in establishing scientific validity. Third, the dominant company in the market, Collaborative Testing Services, Inc. (CTS), explicitly states that its proficiency tests are not appropriate for estimating error rates of a discipline, because (a) the test results, which are open to anyone, may not reflect the skills of forensic practitioners and (b) "the reported results do not reflect 'correct' or 'incorrect' answers, but rather responses that agree or disagree with the consensus conclusions of the participant population."[169] Fourth, the tests for forensic feature-comparison methods typically consist of only one or two problems each year. Fifth, "easy tests are favored by the community," with the result that tests that are too challenging could jeopardize repeat business for a commercial vendor.[170]

[169] See: www.ctsforensics.com/assets/news/CTSErrorRateStatement.pdf.

[170] PCAST thanks Collaborative Testing Services, Inc. (CTS) President Christopher Czyryca for helpful conversations concerning proficiency testing. Czyryca explained that that (1) CTS defines consensus as at least 80 percent agreement among respondents and (2) proficiency testing for latent fingerprints only occasionally involves a problem in which a questioned print matches *none* of the possible answers. Czyryca noted that the forensic community disfavors more challenging tests—and that testing companies are concerned that they could lose business if their tests are viewed as too challenging. An example of a "challenging" test is the very important scenario in which *none* of the questioned samples match any of the known samples: because examiners may expect they should find *some* matches, such scenarios provide an opportunity to assess how often examiners declare false-positive matches. (See also presentation to the National Commission on Forensic Science by CTS President Czyryca, noting that "Easy tests are favored by the community." www.justice.gov/ncfs/file/761061/download.)

PCAST's observations and findings below are largely consistent with the conclusions of earlier NRC reports.[171]

5.1 DNA Analysis of Single-source and Simple-mixture samples

DNA analysis of single-source and simple mixture samples includes excellent examples of objective methods whose foundational validity has been properly established.[172]

Methodology

DNA analysis involves comparing DNA profiles from different samples to see if a known sample may have been the source of an evidentiary sample.

To generate a DNA profile, DNA is first chemically *extracted* from a sample containing biological material, such as blood, semen, hair, or skin cells. Next, a predetermined set of DNA segments ("loci") containing small repeated sequences[173] are *amplified* using the Polymerase Chain Reaction (PCR), an enzymatic process that replicates a targeted DNA segment over and over to yield millions of copies. After amplification, the lengths of the resulting DNA fragments are *measured* using a technique called capillary electrophoresis, which is based on the fact that longer fragments move more slowly than shorter fragments through a polymer solution. The raw data collected from this process are analyzed by a software program to produce a graphical image (an electropherogram) and a list of numbers (the DNA profile) corresponding to the sizes of the each of fragments (by comparing them to known "molecular size standards").

As currently practiced, the method uses 13 specific loci and the amplification process is designed so that the DNA fragments corresponding to different loci occupy different size ranges—making it simple to recognize which fragments come from each locus.[174] At each locus, every human carries two variants (called "alleles")—one inherited from his or her mother, one from his or her father—that may be of different lengths or the same length.[175]

[171] National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (2009). National Research Council, *Ballistic Imaging*. The National Academies Press. Washington DC. (2008).

[172] Forensic DNA analysis belongs to two parent disciplines—metrology and human molecular genetics—and has benefited from the extensive application of DNA technology in biomedical research and medical application.

[173] The repeats, called short tandem repeats (STRs), consist of consecutive repeated copies of a segments of 2-6 base pairs.

[174] The current kit used by the FBI (Identifiler Plus) has 16 total loci: 15 STR loci and the amelogenin locus. A kit that will be implemented later this year has 24 loci.

[175] The FBI announced in 2015 that it plans to expand the core loci by adding seven additional loci commonly used in databases in other countries. (Population data have been published for the expanded set, including frequencies in 11 ethnic populations www.fbi.gov/about-us/lab/biometric-analysis/codis/expanded-fbi-str-2015-final-6-16-15.pdf.) Starting in 2017, these loci will be required for uploading and searching DNA profiles in the national system. The expanded data in each profile are expected to provide greater discrimination potential for identification, especially in matching samples with only partial DNA profiles, missing person inquiries, and international law enforcement and counterterrorism cases.

Analysis of single-source samples

DNA analysis of a sample from a single individual is an objective method. In addition to the laboratory protocols being precisely defined, the interpretation also involves little or no human judgment.

An examiner can assess if a sample came from a single source based on whether the DNA profile typically contains, for each locus, exactly one fragment from each chromosome containing the locus—which yields one or two distinct fragment lengths from each locus.[176] The DNA profile can then be compared with the DNA profile of a known suspect. It can also be entered into the FBI's National DNA Index System (NDIS) and searched against a database of DNA profiles from convicted offenders (and arrestees in more than half of the states) or unsolved crimes.

Two DNA profiles are declared to match if the lists of alleles are the same.[177] The probability that two DNA profiles from *different* sources would have the same DNA profile (the random match probability) is then calculated based on the empirically measured frequency of each allele and established principles of population genetics (see p. 53).[178]

Analysis of simple mixtures

Many sexual assault cases involve DNA mixtures of two individuals, where one individual (i.e., the victim) is known. DNA analysis of these simple mixtures is also relatively straightforward. Methods have been used for 30 years to differentially extract DNA from sperm cells vs. vaginal epithelial cells, making it possible to generate DNA profiles from the two sources. Where the two cell types are the same but one contributor is known, the alleles of the known individual can be subtracted from the set of alleles identified in the mixture.[179]

Once the known source is removed, the analysis of the unknown sample then proceeds as above for single-source samples. Like the analysis of single-source samples, the analysis of simple mixtures is a largely objective method.

[176] The examiner reviews the electropherogram to determine whether each of the peaks is a true allelic peak or an artifact (e.g., background noise in the form of stutter, spikes, and other phenomena) and to determine whether more than one individual could have contributed to the profile. In rare cases, an individual may have two fragments at a locus due to rare copy-number variation in the human genome.

[177] When only a partial profile could be generated from the evidence sample (for example, in cases with limited quantities of DNA, degradation of the sample, or the presence of PCR inhibitors), an examiner may also report an "inclusion" if the partial profile is *consistent* with the DNA profile obtained from a reference sample. An examiner may also report an inclusion when the DNA results from a reference sample are present in a mixture. These cases generally require significantly more human analysis and interpretation than single-source samples.

[178] Random match probabilities can also be expressed in terms of a likelihood ratio (LR), which is the ratio of (1) the probability of observing the DNA profile if the individual in question is the source of the DNA sample and (2) the probability of observing the DNA profile if the individual in question is *not* the source of the DNA sample. In the situation of a single-source sample, the LR should be simply the reciprocal of the random match probability (because the first probability in the LR is 1 and the second probability is the random match probability).

[179] In many cases, DNA will be present in the mixture in sufficiently different quantities so that the peak heights in the electropherogram from the two sources will be distinct, allowing the examiner to more readily separate out the sources.

Foundational Validity

To evaluate the foundational validity of an objective method (such as single-source and simple mixture analysis), one can examine the reliability of each of the individual steps rather than having to rely on black-box studies.

Single-source samples

Each step in the analysis is objective and involves little or no human judgment.

(1) *Feature identification*. In contrast to the other methods discussed in this report, the features used in DNA analysis (the fragments lengths of the loci) are defined *in advance*.

(2) *Feature measurement and comparison*. PCR amplification, invented in 1983, is widely used by tens of thousands of molecular biology laboratories, including for many medical applications in which it has been rigorously validated. Multiplex PCR kits designed by commercial vendors for use by forensic laboratories must be validated both externally (through developmental validation studies published in peer reviewed publication) and internally (by each lab that wishes to use the kit) before they may be used.[180] Fragment sizes are measured by an automated procedure whose variability is well characterized and small; the standard deviation is approximately 0.05 base pairs, which provides highly reliable measurements.[181,182] Developmental validation studies were performed—including by the FBI—to verify the accuracy, precision, and reproducibility of the procedure.[183,184]

[180] Laboratories that conduct forensic DNA analysis are required to follow FBI's Quality Assurance Standards for DNA Testing Laboratories as a condition of participating in the National DNA Index System (www.fbi.gov/about-us/lab/biometric-analysis/codis/qas-standards-for-forensic-dna-testing-laboratories-effective-9-1-2011). FBI's Scientific Working Group on DNA Analysis Methods (SWGDAM) has published guidelines for laboratories in validating procedures consistent the FBI's Quality Assurance Standards (QAS). SWGDAM Validation Guidelines for DNA Analysis Methods, December 2012. See: media.wix.com/ugd/4344b0_cbc27d16dcb64fd88cb36ab2a2a25e4c.pdf.

[181] Forensic laboratories typically use genetic analyzer systems developed by the Applied Biosystems group of Thermo-Fisher Scientific (ABI 310, 3130, or 3500).

[182] To incorrectly estimate a fragment length by 1 base pair (the minimum size difference) requires a measurement error of 0.5 base pair, which corresponds to 10 standard deviations. Moreover, alleles typically differ by at least 4 base pairs (although some STR loci have fairly common alleles that differ by 1 or 2 nucleotides).

[183] For examples of these studies see: Budowle, B., Moretti, T.R., Keys, K.M., Koons, B.W., and J.B. Smerick. "Validation studies of the CTT STR multiplex system." *Journal of Forensic Sciences,* Vol. 42, No. 4 (1997): 701-7; Kimpton, C.P., Oldroyd, N.J., Watson, S.K., Frazier, R.R., Johnson, P.E., Millican, E.S., Urguhart, A., Sparkes, B.L., and P. Gill. "Validation of highly discriminating multiplex short tandem repeat amplification systems for individual identification." *Electrophoresis*, Vol. 17, No. 8 (1996): 1283-93; Lygo, J.E., Johnson, P.E., Holdaway, D.J., Woodroffe, S., Whitaker, J.P., Clayton, T.M., Kimpton, C.P., and P. Gill. "The validation of short tandem repeat (STR) loci for use in forensic casework." *International Journal of Legal Medicine,* Vol. 107, No. 2 (1994): 77-89; and Fregeau, C.J., Bowen, K.L., and R.M. Fourney. "Validation of highly polymorphic fluorescent multiplex short tandem repeat systems using two generations of DNA sequencers." *Journal of Forensic Sciences*, Vol. 44, No. 1 (1999): 133-66.

[184] For example, a 2001 study that compared the performance characteristics of several commercially available STR testing kits tested the consistency and reproducibility of results using previously typed case samples, environmentally insulted samples, and body fluid samples deposited on various substrates. The study found that all of the kits could be used to amplify and type STR loci successfully and that the procedures used for each of the kits were robust and valid. No evidence

(3) *Feature comparison*. For single-source samples, there are clear and well-specified "matching rules" for declaring whether the DNA profiles match. When complete DNA profiles are searched against the NDIS at "high stringency," a "match" is returned only when each allele in the unknown profile is found to match an allele of the known profile, and *vice versa*. When partial DNA profiles obtained from a partially degraded or contaminated sample are searched at "moderate stringency," candidate profiles are returned if each of the alleles in the unknown profile is found to match an allele of the known profile.[185,186]

(4) *Estimation of random match probability*. The process for calculating the random match probability (that is, the probability of a match occurring by chance) is based on well-established principles of population genetics and statistics. The frequencies of the individual alleles were obtained by the FBI based on DNA profiles from approximately 200 unrelated individuals from each of six population groups and were evaluated prior to use.[187] The frequency of an overall pattern of alleles—that is, the random match probability—is typically estimated by multiplying the frequencies of the individual loci, under the assumption that the alleles are independent of one another.[188] The resulting probability is typically less than 1 in 10 billion, excluding the possibility of close relatives.[189] (Note: Multiplying the frequency of alleles can overstates the rarity of a pattern because the alleles are not completely independent, owing

of false positive or false negative results and no substantial evidence of preferential amplification within a locus were found for any of the testing kits. Moretti, T.R., Baumstark, A.L., Defenbaugh, D.A., Keys, K.M., Smerick, J.B., and B. Budowle. "Validation of Short Tandem Repeats (STRs) for forensic usage: performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples." *Journal of Forensic Sciences*, Vol. 46, No. 3 (2001): 647-60.

[185] See: FBI's Frequently Asked Questions (FAQs) on the CODIS Program and the National DNA Index System. www.fbi.gov/about-us/lab/biometric-analysis/codis/codis-and-ndis-fact-sheet.

[186] Contaminated samples are not retained in NDIS.

[187] The initial population data generated by FBI included data for 6 ethnic populations with database sizes of 200 individuals. See: Budowle, B., Moretti, T.R., Baumstark, A.L., Defenbaugh, D.A., and K.M. Keys. "Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians." *Journal of Forensic Sciences*, Vol. 44, No. 6 (1999): 1277-86 and Budowle, B., Shea, B., Niezgoda, S., and R. Chakraborty. "CODIS STR loci data from 41 sample populations." *Journal of Forensic Sciences,* Vol. 46, No. 3 (2001): 453-89. Errors in the original database were reported in July 2015 (Erratum, *Journal of Forensic Sciences*, Vol. 60, No. 4 (2015): 1114-6, the impact of these discrepancies on profile probability calculations were assessed (and found to be less than a factor of 2 in a full profile), and the allele frequency estimates were amended accordingly. At the same time as amending the original datasets, the FBI Laboratory also published expanded datasets in which the original samples were retyped for additional loci. In addition, the population samples that were originally studied at other laboratories were typed for additional loci, so the full dataset includes 9 populations. These "expanded" datasets are in use at the FBI Laboratory and can be found at www.fbi.gov/about-us/lab/biometric-analysis/codis/expanded-fbi-str-final-6-16-15.pdf.

[188] More precisely, the frequency at each locus is calculated first. If the locus has two copies of the same allele with frequency p, the frequency is calculated as p^2. If the locus has two different alleles with respective frequencies p and q, the frequency is calculated as 2pq. The frequency of the overall pattern is calculated by multiplying together the values for the individual loci.

[189] The random match probability will be higher for close relatives. For identical twins, the DNA profiles are expected to match perfectly. For first degree relatives, the random match probability may be on the order of 1 in 100,000 when examining the 13 CODIS core STR loci. See: Butler, J.M. "The future of forensic DNA analysis." *Philosophical Transactions of the Royal Society B,* 370: 20140252 (2015).

to population substructure. A 1996 NRC report concluded that the effect of population substructure on the calculated value was likely to be within a factor of 10 (for example, for a random match probability estimate of 1 in 10 million, the true probability is highly likely to be between 1 in 1 million and 1 in 100 million).[190] However, a recent study by NIST scientists suggests that the variation may be substantially greater than 10-fold.[191] The random match probability should be calculated using an appropriate statistical formula that takes account of population substructure.[192])

Simple mixtures

The steps for analyzing simple mixtures are the same as for analyzing single-source samples, up until the point of interpretation. DNA profiles that contain a mixture of two contributors, where one contributor is known, can be interpreted in much the same way as single-source samples. This occurs frequently in sexual assault cases, where a DNA profile contains a mixture of DNA from the victim and the perpetrator. Methods that are used to differentially extract DNA from sperm cells vs. vaginal epithelial cells in sexual assault cases are well-established.[193] Where the two cell types are the same, one DNA source may be dominant, resulting in a distinct contrast in peak heights between the two contributors; in these cases, the alleles from both the major contributor (corresponding to the larger allelic peaks) and the minor contributor can usually be reliably interpreted, provided the proportion of the minor contributor is not too low.[194]

Validity as Applied

While DNA analysis of single-source samples and simple mixtures is a foundationally valid and reliable method, it is not infallible in practice. Errors can and do occur in DNA testing. Although the probability that two samples from different sources have the same DNA profile is tiny, the chance of human error is much higher. Such errors may stem from sample mix-ups, contamination, incorrect interpretation, and errors in reporting.[195]

[190] National Research Council. *The Evaluation of Forensic DNA Evidence.* The National Academies Press. Washington DC. (1996). Goode, M. “Some observations on evidence of DNA frequency.” *Adelaide Law Review,* Vol. 23 (2002): 45-77.

[191] Gittelson, S. and J. Buckleton. “Is the factor of 10 still applicable today?” Presentation at the 68th Annual American Academy of Forensic Sciences Scientific Meeting, 2016. See: www.cstl.nist.gov/strbase/pub_pres/Gittelson-AAFS2016-Factor-of-10.pdf.

[192] Balding, D.J., and R.A. Nichols. “DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands.” *Forensic Science International*, Vol. 64 (1994): 125-140.

[193] Gill, P., Jeffreys, A.J., and D.J. Werrett. “Forensic application of DNA ‘fingerprints.’” *Nature*, Vol. 318, No. 6046 (1985): 577-9.

[194] Clayton, T.M., Whitaker, J.P., Sparkes, R., and P. Gill. “Analysis and interpretation of mixed forensic stains using DNA STR profiling.” *Forensic Science International*, Vol. 91, No. 1 (1998): 55-70.

[195] Krimsky, S., and T. Simoncelli. *Genetic Justice: DNA Data Banks, Criminal Investigations, and Civil Liberties.* Columbia University Press, (2011). Perhaps the most spectacular human error to date involved the German government’s investigation of the “Phantom of Heilbronn,” a woman whose DNA appeared at the scenes of more than 40 crimes in three countries, including 6 murders, several muggings and dozens of break-ins over the course of more than a decade. After an effort that included analyzing DNA samples from more than 3,000 women from four countries and that cost $18 million, authorities discovered that the woman of interest was a worker in the Austrian factory that fabricated the swabs used in DNA collection. The woman had inadvertently contaminated a large number of swabs with her own DNA, which was thus found in many DNA tests.

To minimize human error, the FBI requires, as a condition of participating in NDIS, that laboratories follow the FBI's Quality Assurance Standards (QAS).[196] Before the results of the DNA analysis can be compared, the examiner is required to run a series of controls to check for possible contamination and ensure that the PCR process ran properly. The QAS also requires semi-annual proficiency testing of all DNA analysts that perform DNA testing for criminal cases. The results of the tests do not have to be published, but the laboratory must retain the results of the tests, any discrepancies or errors made, and corrective actions taken.[197]

Forensic practitioners in the U.S. do not typically report quality issues that arise in forensic DNA analysis. By contrast, error rates in medical DNA testing are commonly measured and reported.[198] Refreshingly, a 2014 paper from the Netherlands Forensic Institute (NFI), a government agency, reported a comprehensive analysis of all "quality issue notifications" encountered in casework, categorized by type, source and impact.[199,200] The authors call for greater "transparency" and "culture change," writing that:

> *Forensic DNA casework is conducted worldwide in a large number of laboratories, both private companies and in institutes owned by the government. Quality procedures are in place in all laboratories, but the nature of the quality system varies a lot between the different labs. In particular, there are many forensic DNA laboratories that operate without a quality issue notification system like the one described in this paper. In our experience, such a system is extremely important for the detection and proper handling of errors. This is crucial in forensic casework that can have a major impact on people's lives. We therefore propose that the implementation of a quality issue notification system is necessary for any laboratory that is involved in forensic DNA casework.*
>
> *Such system can only work in an optimal way, however, when there is a blame-free culture in the laboratory that extends to the police and the legal justice system. People have a natural tendency to hide their mistakes, and it is essential to create an atmosphere where there are no adverse personal consequences when mistakes are reported. The management should take the lead in this culture change...*
>
> *As far as we know, the NFI is the first forensic DNA laboratory in the world to reveal such detailed data and reports. It shows that this is possible without any disasters or abuse happening, and there are no*

[196] FBI. "Quality assurance standards for forensic DNA testing laboratories." (2011). See: www.fbi.gov/about-us/lab/biometric-analysis/codis/qas-standards-for-forensic-dna-testing-laboratories-effective-9-1-2011.

[197] Ibid., Sections 12, 13, and 14.

[198] See, for example: Plebani, M., and P. Carroro. "Mistakes in a stat laboratory: types and frequency." *Clinical Chemistry,* Vol. 43 (1997): 1348-51; Stahl, M., Lund, E.D., and I. Brandslund. "Reasons for a laboratory's inability to report results for requested analytical tests." *Clinical Chemistry,* Vol. 44 (1998): 2195-7; Hofgartner, W.T., and J.F. Tait. "Frequency of problems during clinical molecular-genetic testing." *American Journal of Clinical Pathology,* Vol. 112 (1999): 14-21; and Carroro, P., and M. Plebani. "Errors in a stat laboratory: types and frequencies 10 years later." *Clinical Chemistry,* Vol. 53 (2007): 1338-42.

[199] Kloosterman, A., Sjerps, M., and A. Quak. "Error rates in forensic DNA analysis: Definition, numbers, impact and communication." *Forensic Science International: Genetics*, Vol. 12 (2014): 77-85 and J.M. Butler "DNA Error Rates" presentation at the International Forensics Symposium, Washington, D.C. (2015). www.cstl.nist.gov/strbase/pub_pres/Butler-ErrorManagement-DNA-Error.pdf.

[200] The Netherlands uses an "inquisitorial" approach to method of criminal justice rather than the adversarial system used in the U.S. Concerns about having to explain quality issues in court may explain in part why U.S. laboratories do not routinely report quality issues.

> *reasons for nondisclosure. As mentioned in the introduction, in laboratory medicine publication of data on error rates has become standard practice. Quality failure rates in this domain are comparable to ours.*

Finally, we note that there is a need to improve proficiency testing. There are currently no requirements concerning how challenging the proficiency tests should be. The tests should be representative of the full range of situations likely to be encountered in casework.

Finding 2: DNA Analysis

Foundational validity. PCAST finds that DNA analysis of single-source samples or simple mixtures of two individuals, such as from many rape kits, is an objective method that has been established to be foundationally valid.

Validity as applied. Because errors due to human failures will dominate the chance of coincidental matches, the scientific criteria for validity as applied require that an expert (1) should have undergone rigorous and relevant proficiency testing to demonstrate their ability to reliably apply the method, (2) should routinely disclose in reports and testimony whether, when performing the examination, he or she was aware of any facts of the case that might influence the conclusion, and (3) should disclose, upon request, all information about quality testing and quality issues in his or her laboratory.

5.2 DNA Analysis of Complex-mixture Samples

Some investigations involve DNA analysis of complex mixtures of biological samples from multiple unknown individuals in unknown proportions. Such samples might arise, for example, from mixed blood stains. As DNA testing kits have become more sensitive, there has been growing interest in "touch DNA"—for example, tiny quantities of DNA left by multiple individuals on a steering wheel of a car.

Methodology

The fundamental difference between DNA analysis of complex-mixture samples and DNA analysis of single-source and simple mixtures lies not in the laboratory processing, but in the interpretation of the resulting DNA profile.

DNA analysis of complex mixtures—defined as mixtures with more than two contributors—is inherently difficult and even more for small amounts of DNA.[201] Such samples result in a DNA profile that superimposes multiple individual DNA profiles. Interpreting a mixed profile is different for multiple reasons: each individual may contribute two, one or zero alleles at each locus; the alleles may overlap with one another; the peak heights may differ considerably, owing to differences in the amount and state of preservation of the DNA from each source; and the "stutter peaks" that surround alleles (common artifacts of the DNA amplification process) can

[201] See, for example, SWGDAM document on interpretation of DNA mixtures. www.swgdam.org/#!public-comments/c1t82.

obscure alleles that are present or suggest alleles that are not present.[202] It is often impossible to tell with certainty which alleles are present in the mixture or how many separate individuals contributed to the mixture, let alone accurately to infer the DNA profile of each individual.[203]

Instead, examiners must ask: "Could a suspect's DNA profile be present *within* the mixture profile? And, what is the probability that such an observation might occur by chance?" The questions are challenging for the reasons given above. Because many different DNA profiles may fit within some mixture profiles, the probability that a suspect "cannot be excluded" as a possible contributor to complex mixture may be *much higher* (in some cases, millions of times higher) than the probabilities encountered for matches to single-source DNA profiles. As a result, proper calculation of the statistical weight is critical for presenting accurate information in court.

Subjective Interpretation of Complex Mixtures

Initial approaches to the interpretation of complex mixtures relied on subjective judgment by examiners, together with the use of simplified statistical methods such as the "Combined Probability of Inclusion" (CPI). These approaches are problematic because subjective choices made by examiners, such as about which alleles to include in the calculation, can dramatically alter the result and lead to inaccurate answers.

The problem with subjective analysis of complex-mixture samples is illustrated by a 2003 double-homicide case, *Winston v. Commonwealth*.[204] A prosecution expert reported that the defendant could not be excluded as a possible contributor to DNA on a discarded glove that contained a mixed DNA profile of at least three contributors; the defendant was convicted and sentenced to death. The prosecutor told the jury that the chance the match occurred by chance was 1 in 1.1 billion. A 2009 paper, however, makes a reasonable scientific case that that the chance is closer to 1 in 2—that is, 50 percent of the relevant population could not be excluded.[205] Such a large discrepancy is unacceptable, especially in cases where a defendant was sentenced to death.

Two papers clearly demonstrate that these commonly used approaches for DNA analysis of complex mixtures can be problematic. In a 2011 study, Dror and Hampikian tested whether irrelevant contextual information biased their conclusions of examiners, using DNA evidence from an actual adjudicated criminal case (a gang rape case in Georgia).[206] In this case, one of the suspects implicated another in connection with a plea bargain. The two experts who examined evidence from the crime scene were aware of this testimony against the suspect and knew that the plea bargain testimony could be used in court only with corroborating DNA evidence. Due to the

[202] Challenges with "low-template" DNA are described in a recent paper, Butler, J.M. "The future of forensic DNA analysis." *Philosophical Transactions of the Royal Society B,* 370: 20140252 (2015).

[203] See: Buckleton, J.S., Curran, J.M., and P. Gill. "Towards understanding the effect of uncertainty in the number of contributors to DNA stains." *Forensic Science International Genetics*, Vol. 1, No. 1 (2007): 20-8 and Coble, M.D., Bright, J.A., Buckleton, J.S., and J.M. Curran. "Uncertainty in the number of contributors in the proposed new CODIS set." *Forensic Science International Genetics*, Vol. 19 (2015): 207-11.

[204] *Winston v. Commonwealth,* 604 S.E.2d 21 (Va. 2004).

[205] Thompson, W.C. "Painting the target around the matching profile: the Texas sharpshooter fallacy in forensic DNA interpretation." *Law, Probability and Risk*, Vol. 8, No. 3 (2009): 257-76.

[206] Dror, I.E., and G. Hampikian. "Subjectivity and bias in forensic DNA mixture interpretation." *Science & Justice*, Vol. 51, No. 4 (2011): 204-8.

complex nature of the DNA mixture collected from the crime scene, the analysis of this evidence required judgment and interpretation on the part of the examiners. The two experts both concluded that the suspect could not be excluded as a contributor.

Dror and Hampikian presented the original DNA evidence from this crime to 17 expert DNA examiners, but without any of the irrelevant contextual information. They found that only 1 out of the 17 experts agreed with the original experts who were exposed to the biasing information (in fact, 12 of the examiners *excluded* the suspect as a possible contributor).

In another paper, de Keijser and colleagues presented 19 DNA experts with a mock case involving an alleged violent robbery outside a bar:

> *There is a male suspect, who denies any wrongdoing. The items that were sampled for DNA analysis are the shirt of the (alleged) female victim (who claims to have been grabbed by her assailant), a cigarette butt that was picked up by the police and that was allegedly smoked by the victim and/or the suspect, and nail clippings from the victim, who claims to have scratched the perpetrator.* [207]

Although all the experts were provided the same DNA profiles (prepared from the three samples above and the two people), their conclusions varied wildly. One examiner excluded the suspect as a possible contributor, while another examiner declared a match between the suspect's profile and a few minor peaks in the mixed profile from the nails—reporting a random match probability of roughly 1 in 209 million. Still other examiners declared the evidence inconclusive.

In the summer of 2015, a remarkable chain of events in Texas revealed that the problems with subjective analysis of complex DNA mixtures were not limited to a few individual cases: they were systemic.[208] The Texas Department of Public Safety (TX-DPS) issued a public letter on June 30, 2015 to the Texas criminal justice community noting that (1) the FBI had recently reported that it had identified and corrected minor errors in its population databases used to calculate statistics in DNA cases, (2) the errors were not expected to have any significant effect on results, and (2) the TX-DPS Crime Laboratory System would, upon request, recalculate statistics previously reported in individual cases.

When several prosecutors submitted requests for recalculation to TX-DPS and other laboratories, they were stunned to find that the statistics had changed dramatically—e.g., *from 1 in 1.4 billion to 1 in 36 in one case, from 1 in 4000 to inconclusive in another*. These prosecutors sought the assistance of the Texas Forensic Science Commission (TFSC) in understanding the reason for the change and the scope of potentially affected cases.

[207] de Keijser, J.W., Malsch, M., Luining, E.T., Kranenbarg, M.W., and D.J.H.M. Lenssen. "Differential reporting of mixed DNA profiles and its impact on jurists' evaluation of evidence: An international analysis." *Forensic Science International: Genetics*, Vol. 23 (2016): 71-82.

[208] Relevant documents and further details can be found at www.fsc.texas.gov/texas-dna-mixture-interpretation-case-review. Lynn Garcia, General Counsel for the Texas Forensic Science Commission, also provided a helpful summary to PCAST.

In consultation with forensic DNA experts, the TFSC determined that the large shifts observed in some cases were unrelated to the minor corrections in the FBI's population database, but rather were due to the fact that forensic laboratories had changed the way in which they calculated the CPI statistic—especially how they dealt with phenomena such as "allelic dropout" at particular DNA loci.

The TFSC launched a statewide DNA Mixture Notification Subcommittee, which included representatives of conviction integrity units, district and county attorneys, defense attorneys, innocence projects, the state attorney general, and the Texas governor. By September 2015, the TX-DPS had generated a county-by-county list of more than 24,000 DNA mixture cases analyzed from 1999-2015. Because TX-DPS is responsible for roughly half of the casework in the state, the total number of Texas DNA cases requiring review may exceed 50,000. (Although comparable efforts have not been undertaken in other states, the problem is likely to be national in scope, rather than specific to forensic laboratories in Texas.)

The TFSC also convened an international panel of scientific experts—from the Harvard Medical School, the University of North Texas Health Science Center, New Zealand's forensic research unit, and NIST—to clarify the proper use of CPI. These scientists presented observations at a public meeting, where many attorneys learned for the first time the extent to which DNA-mixture analysis involved subjective interpretation. Many of the problems with the CPI statistic arose because existing guidelines did not clearly, adequately, or correctly specify the proper use or limitations of the approach.

In summary, the interpretation of complex DNA mixtures with the CPI statistic has been an inadequately specified—and thus inappropriately subjective—method. As such, the method is clearly not foundationally valid.

In an attempt to fill this gap, the experts convened by TFSC wrote a joint scientific paper, which was published online on August 31, 2016.[209] The paper underscores the "pressing need . . . for standardization of an approach, training and ongoing testing of DNA analysts." The authors propose a set of specific rules for the use of the CPI statistic.

The proposed rules are clearly *necessary* for a scientifically valid method for the application of CPI. Because the paper appeared just as this report was being finalized, PCAST has not had adequate time to assess whether the rules are also *sufficient* to define an objective and scientifically valid method for the application of CPI.

Current Efforts to Develop Objective Methods

Given these problems, several groups have launched efforts to develop "probabilistic genotyping" computer programs that apply various algorithms to interpret complex mixtures. As of March 2014, at least 8 probabilistic genotyping software programs had been developed (called LRmix, Lab Retriever, likeLTD, FST, Armed Xpert, TrueAllele, STRmix, and DNA View Mixture Solution), with some being open source software and some being

[209] 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*. bmcgenet.biomedcentral.com/articles/10.1186/s12863-016-0429-7.

commercial products.[210] The FBI Laboratory began using the STRmix program less than a year ago, in December 2015, and is still in the process of publishing its own internal developmental validation.

These probabilistic genotyping software programs clearly represent a major improvement over purely subjective interpretation. However, they still require careful scrutiny to determine (1) whether the methods are scientifically valid, including defining the limitations on their reliability (that is, the circumstances in which they may yield unreliable results) and (2) whether the software correctly implements the methods. This is particularly important because the programs employ different mathematical algorithms and can yield different results for the same mixture profile.[211]

Appropriate evaluation of the proposed methods should consist of studies by multiple groups, *not associated with the software developers*, that investigate the performance and define the limitations of programs by testing them on a wide range of mixtures with different properties. In particular, it is important to address the following issues:

(1) How well does the method perform as a function of the number of contributors to the mixture? How well does it perform when the number of contributors to the mixture is *unknown*?

(2) How does the method perform as a function of the number of alleles shared among individuals in the mixture? Relatedly, how does it perform when the mixtures include related individuals?

(3) How well does the method perform—and how does accuracy degrade—as a function of the absolute and relative amounts of DNA from the various contributors? For example, it can be difficult to determine whether a small peak in the mixture profile represents a true allele from a minor contributor or a stutter peak from a nearby allele from a different contributor. (Notably, this issue underlies a current case that has received considerable attention.[212])

[210] The topic is reviewed in Butler, J.M. "Chapter 13: Coping with Potential Missing Alleles." *Advanced Topics in Forensic DNA Typing: Interpretation*. Waltham, MA: Elsevier/Academic, (2015): 333-48.

[211] Some programs use discrete (semi-continuous) methods, which use only allele information in conjunction with probabilities of allelic dropout and dropin, while other programs use continuous methods, which also incorporate information about peak height and other information. Within these two classes, the programs differ with respect to how they use the information. Some of the methods involve making assumptions about the number of individuals contributing to the DNA profile, and use this information to clean up noise (such as "stutter" in DNA profiles).

[212] In this case, examiners used two different DNA software programs (STRMix and TrueAllele) and obtained different conclusions concerning whether DNA from the defendant could be said to be included within the low-level DNA mixture profile obtained from a sample collected from one of the victim's fingernails. The judge ruled that the DNA evidence implicating the defendant was inadmissible. McKinley, J. "Potsdam Boy's Murder Case May Hinge on Minuscule DNA Sample From Fingernail." *New York Times.* See: www.nytimes.com/2016/07/25/nyregion/potsdam-boys-murder-case-may-hinge-on-statistical-analysis.html (accessed August 22, 2016). Sommerstein, D. "DNA results will not be allowed in Hillary murder trail." North Country Public Radio (accessed September 1, 2016). The decision can be found here: www.northcountrypublicradio.org/assets/files/08-26-16DecisionandOrder-DNAAnalysisAdmissibility.pdf.

(4) Under what circumstances—and why—does the method produce results (random inclusion probabilities) that differ substantially from those produced by other methods?

A number of papers have been published that analyze known mixtures in order to address some of these issues.[213] Two points should be noted about these studies. First, most of the studies evaluating software packages have been undertaken by the software developers themselves. While it is completely appropriate for method developers to evaluate their own methods, establishing scientific validity also requires scientific evaluation by other scientific groups that did not develop the method. Second, there have been few comparative studies across the methods to evaluate the differences among them—and, to our knowledge, no comparative studies conducted by independent groups.[214]

Most importantly, current studies have adequately explored only a limited range of mixture types (with respect to number of contributors, ratio of minor contributors, and total amount of DNA). The two most widely used methods (STRMix and TrueAllele) appear to be reliable within a certain range, based on the available evidence and the inherent difficulty of the problem.[215] Specifically, these methods appear to be reliable for three-person mixtures in which the minor contributor constitutes at least 20 percent of the intact DNA in the mixture and in which the DNA amount exceeds the minimum level required for the method.[216]

[213] For example: Perlin, M.W., Hornyak, J.M., Sugimoto, G., and K.W.P. Miller. "TrueAllele genotype identification on DNA mixtures containing up to five unknown contributors." *Journal of Forensic Sciences*, Vol. 60, No. 4 (2015): 857-868; Greenspoon S.A., Schiermeier-Wood L., and B.C. Jenkins. "Establishing the limits of TrueAllele® Casework: A validation study." *Journal of Forensic Sciences*. Vol. 60, No. 5 (2015):1263–76; Bright, J.A., Taylor, D., McGovern, C., Cooper, S., Russell, L., Abarno, D., and J.S. Buckleton. "Developmental validation of STRmix^{TM}, expert software for the interpretation of forensic DNA profiles." *Forensic Science International: Genetics*. Vol. 23 (2016): 226-39; Bright, J-A., Taylor D., Curran, J.S., and J.S. Buckleton. "Searching mixed DNA profiles directly against profile databases." *Forensic Science International: Genetics*. Vol. 9 (2014):102-10; Taylor D., Buckleton J, and I. Evett. "Testing likelihood ratios produced from complex DNA profiles." *Forensic Science International: Genetics.* Vol. 16 (2015): 165-171; Taylor D. and J.S. Buckleton. "Do low template DNA profiles have useful quantitative data?" *Forensic Science International: Genetics,* Vol. 16 (2015): 13-16.

[214] Bille, T.W., Weitz, S.M., Coble, M.D., Buckleton, J., and J.A. Bright. "Comparison of the performance of different models for the interpretation of low level mixed DNA profiles." *Electrophoresis*. Vol. 35 (2014): 3125–33.

[215] The interpretation of DNA mixtures becomes increasingly challenging as the number of contributors increases. See, for example: Taylor D., Buckleton J, and I. Evett. "Testing likelihood ratios produced from complex DNA profiles." *Forensic Science International: Genetics.* Vol. 16 (2015): 165-171; Bright, J.A., Taylor, D., McGovern, C., Cooper, S., Russell, L., Abarno, D., and J.S. Buckleton. "Developmental validation of STRmix^{TM}, expert software for the interpretation of forensic DNA profiles." *Forensic Science International: Genetics*. Vol. 23 (2016): 226-39; Bright, J-A., Taylor D., Curran, J.S., and J.S. Buckleton. "Searching mixed DNA profiles directly against profile databases." *Forensic Science International: Genetics*. Vol. 9 (2014):102-10; 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*. bmcgenet.biomedcentral.com/articles/10.1186/s12863-016-0429-7.

[216] Such three-person samples involving similar proportions are more straightforward to interpret owing to the limited number of alleles and relatively similar peak height. The methods can also be reliably applied to single-source and simple-mixture samples, provided that, in cases where the two contributions cannot be separated by differential extraction, the proportion of the minor contributor is not too low (e.g., at least 10 percent).

For more complex mixtures (e.g. more contributors or lower proportions), there is relatively little published evidence.[217] In human molecular genetics, an experimental validation of an important diagnostic method would typically involve hundreds of distinct samples.[218] One forensic scientist told PCAST that many more distinct samples have, in fact, been analyzed, but that the data have not yet been collated and published.[219] Because empirical evidence is essential for establishing the foundational validity of a method, PCAST urges forensic scientists to submit and leading scientific journals to publish high-quality validation studies that properly establish the range of reliability of methods for the analysis of complex DNA mixtures.

When further studies are published, it will likely be possible to extend the range in which scientific validity has been established to include more challenging samples. As noted above, such studies should be performed by or should include independent research groups not connected with the developers of the methods and with no stake in the outcome.

Conclusion

Based on its evaluation of the published literature to date, PCAST reached several conclusions concerning the foundational validity of methods for the analysis of complex DNA mixtures. We note that foundational validity must be established with respect to a specified method applied to a specified range. In addition to forming its own judgment, PCAST also consulted with John Butler, Special Assistant to the Director for Forensic Science at NIST and Vice Chair of the NCFS.[220] Butler concurred with PCAST's finding.

[217] For four-person mixtures, for example, papers describing experimental validations with known mixtures using TrueAllele involve 7 and 17 distinct mixtures, respectively, with relatively large amounts of DNA (at least 200 pg), while those using STRMix involve 2 and 3 distinct mixtures, respectively, but use much lower amounts of DNA (in the range of 10 pg). Greenspoon S.A., Schiermeier-Wood L., and B.C. Jenkins. "Establishing the limits of TrueAllele® Casework: A validation study." *Journal of Forensic Sciences*. Vol. 60, No. 5 (2015):1263–76; Perlin, M.W., Hornyak, J.M., Sugimoto, G., and K.W.P. Miller. "TrueAllele genotype identification on DNA mixtures containing up to five unknown contributors." *Journal of Forensic Sciences*, Vol. 60, No. 4 (2015): 857-868; Taylor, D. "Using continuous DNA interpretation methods to revisit likelihood ratio behavior." *Forensic Science International: Genetics,* Vol. 11 (2014): 144-153; Taylor D., Buckleton J, and I. Evett. "Testing likelihood ratios produced from complex DNA profiles." *Forensic Science International: Genetics.* Vol. 16 (2015): 165-171; Taylor D. and J.S. Buckleton. "Do low template DNA profiles have useful quantitative data?" *Forensic Science International: Genetics,* Vol. 16 (2015): 13-16; Bright, J.A., Taylor, D., McGovern, C., Cooper, S., Russell, L., Abarno, D., J.S. Buckleton. "Developmental validation of STRmix^{TM}, expert software for the interpretation of forensic DNA profiles." *Forensic Science International: Genetics*. Vol. 23 (2016): 226-39.

[218] Preparing and performing PCR amplication on hundreds of DNA mixtures is straightforward; it can be accomplished within a few weeks or less.

[219] PCAST interview with John Buckleton, Principal Scientist at New Zealand's Institute of Environmental Science and Research and a co-developer of STRMix.

[220] Butler is a world authority on forensic DNA analysis, whose Ph.D. research, conducted at the FBI Laboratory, pioneered techniques of modern forensic DNA analysis and who has written five widely acclaimed textbooks on forensic DNA typing. See: Butler, J.M. *Forensic DNA Typing: Biology and Technology behind STR Markers*. Academic Press, London (2001); Butler, J.M. *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers (2nd Edition)*. Elsevier Academic Press, New York (2005); Butler, J.M. *Fundamentals of Forensic DNA Typing.* Elsevier Academic Press, San Diego (2010); Butler, J.M. *Advanced Topics in Forensic DNA Typing: Methodology.* Elsevier Academic Press, San Diego (2012); Butler, J.M. *Advanced Topics in Forensic DNA Typing: Interpretation*. Elsevier Academic Press, San Diego (2015).

Finding 3: DNA analysis of complex-mixture samples

Foundational validity. PCAST finds that:

(1) Combined-Probability-of-Inclusion (CPI)-based methods. DNA analysis of complex mixtures based on CPI-based approaches has been an inadequately specified, subjective method that has the potential to lead to erroneous results. As such, it is not foundationally valid.

A very recent paper has proposed specific rules that address a number of problems in the use of CPI. These rules are clearly *necessary*. However, PCAST has not adequate time to assess whether they are also *sufficient* to define an objective and scientifically valid method. If, for a limited time, courts choose to admit results based on the application of CPI, validity as applied would require that, at a minimum, they be consistent with the rules specified in the paper.

DNA analysis of complex mixtures should move rapidly to more appropriate methods based on probabilistic genotyping.

(2) Probabilistic genotyping. Objective analysis of complex DNA mixtures with probabilistic genotyping software is relatively new and promising approach. Empirical evidence is required to establish the foundational validity of each such method within specified ranges. At present, published evidence supports the foundational validity of analysis, with some programs, of DNA mixtures of 3 individuals in which the minor contributor constitutes at least 20 percent of the intact DNA in the mixture and in which the DNA amount exceeds the minimum required level for the method. The range in which foundational validity has been established is likely to grow as adequate evidence for more complex mixtures is obtained and published.

Validity as applied. For methods that are foundationally valid, validity as applied involves similar considerations as for DNA analysis of single-source and simple-mixtures samples, with a special emphasis on ensuring that the method was applied correctly and within its empirically established range.

The Path Forward

There is a clear path for extending the range over which objective methods have been established to be foundationally valid—specifically, through the publication of appropriate scientific studies.

Such efforts will be aided by the creation and dissemination (under appropriate data-use and data-privacy restrictions) of large collections of hundreds of DNA profiles created from known mixtures—representing widely varying complexity with respect to (1) the number of contributors, (2) the relationships among contributors, (3) the absolute and relative amounts of materials, and (4) the state of preservation of materials—that can be used by independent groups to evaluate and compare the methods. Notably, the PROVEDIt Initiative (Project Research Openness for Validation with Experimental Data) at Boston University has made available a resource of

25,000 profiles from DNA mixtures.[221,222] In addition to scientific studies on common sets of samples for the purpose of evaluating foundational validity, individual forensic laboratories will want to conduct their own internal developmental validation studies to assess the validity of the method in their own hands.[223]

NIST should play a leadership role in this process, by ensuring the creation and dissemination of materials and stimulating studies by independent groups through grants, contracts, and prizes; and by evaluating the results of these studies.

5.3 Bitemark Analysis

Methodology

Bitemark analysis is a subjective method. It typically involves examining marks left on a victim or an object at the crime scene, and comparing those marks with dental impressions taken from a suspect.[224] Bitemark comparison is based on the premises that (1) dental characteristics, particularly the arrangement of the front teeth, differ substantially among people and (2) skin (or some other marked surface at a crime scene) can reliably capture these distinctive features.

Bitemark analysis begins with an examiner deciding whether an injury is a mark caused by human teeth.[225] If so, the examiner creates photographs or impressions of the questioned bitemark and of the suspect's dentition; compares the bitemark and the dentition; and determines if the dentition (1) cannot be excluded as having made the bitemark, (2) can be excluded as having made the bitemark, or (3) is inconclusive. The bitemark standards do not provide well-defined standards concerning the degree of similarity that must be identified to support a reliable conclusion that the mark could have or could not have been created by the dentition in question. Conclusions about all these matters are left to the examiner's judgment.

Background Studies

Before turning to the question of foundational validity, we discuss some background studies (concerning such topics as uniqueness and consistency) that shed some light on the field. These studies cast serious doubt on the fundamental premises of the field.

[221] See: www.bu.edu/dnamixtures.
[222] The collection contains DNA samples with 1- to 5-person DNA mixtures, amplified with targets ranging from 1 to 0.007 ng. In the multi-person mixtures, the ratio of contributors range from 1:1 to 1:19. Additionally, the profiles were generated using a variety of laboratory conditions from samples containing pristine DNA; UV damaged DNA; enzymatically or sonically degraded DNA; and inhibited DNA.
[223] The FBI Laboratory has recently completed a developmental validation study and is preparing it for publication.
[224] Less frequently, marks are found on a suspected perpetrator that may have come from a victim.
[225] ABFO Bitemark Methodology Standards and Guidelines, abfo.org/wp-content/uploads/2016/03/ABFO-Bitemark-Standards-03162016.pdf (accessed July 2, 2016).

A widely cited 1984 paper claimed that “human dentition was unique beyond any reasonable doubt.”[226] The study examined 397 bitemarks carefully made in a wax wafer, measured 12 parameters from each, and—assuming, without any evidence, that the parameters were uncorrelated with each other—suggested that the chance of two bitemarks having the same parameters is less than one in six trillion. The paper was theoretical rather than empirical: it did not attempt to actually compare the bitemarks to one another.

A 2010 paper debunked these claims.[227] By empirically studying 344 human dental casts and measuring them by three-dimensional laser scanning, these authors showed that matches occurred vastly more often than expected under the theoretical model. For example, the theoretical model predicted that the probability of finding *even a single* five-tooth match among the collection of bitemarks is less than one in one million; yet, the empirical comparison revealed 32 such matches.

Notably, these studies examined human dentition patterns measured under idealized conditions. By contrast, skin has been shown to be an unreliable medium for recording the precise pattern of teeth. Studies that have involved inflicting bitemarks either on living pigs[228] (used as a model of human skin) or human cadavers[229] have demonstrated significant distortion in all directions. A 2010 study of experimentally created bitemarks produced by known biters concluded that skin deformation distorts bitemarks so substantially and so variably that current procedures for comparing bitemarks are unable to reliably exclude or include a suspect as a potential biter (“The data derived showed no correlation and was not reproducible, that is, the same dentition could not create a measurable impression that was consistent in all of the parameters in any of the test circumstances.”)[230] Such distortion is further complicated in the context of criminal cases, where biting often occurs during struggles, in which skin may be stretched and contorted at the time a bitemark is created.

Empirical research suggests that forensic odontologists do not consistently agree even on whether an injury is a human bitemark at all. A study by the American Board of Forensic Odontology (AFBO)[231] involved showing photos of 100 patterned injuries to ABFO board-certified bitemark analysts, and asking them to answer three basic questions concerning (1) whether there was sufficient evidence to render an opinion as to whether the patterned injury is a human bitemark; (2) whether the mark is a human bitemark, suggestive of a human

[226] Rawson, R.D., Ommen, R.K., Kinard, G., Johnson, J., and A. Yfantis. “Statistical evidence for the individuality of the human dentition.” *Journal of Forensic Sciences*, Vol. 29, No. 1 (1984): 245-53.

[227] Bush, M.A., Bush, P.J., and H.D. Sheets. “Statistical evidence for the similarity of the human dentition.” *Journal of Forensic Sciences*, Vol. 56, No. 1 (2011): 118-23.

[228] Dorion, R.B.J., ed. *Bitemark Evidence: A Color Atlas and Text*. 2nd ed. CRC Press-Taylor & Francis, Boca Raton, Florida (2011).

[229] Sheets, H.D., Bush, P.J., and M.A. Bush. “Bitemarks: distortion and covariation of the maxillary and mandibular dentition as impressed in human skin.” *Forensic Science International*, Vol. 223, No. 1-3 (2012): 202-7. Bush, M.A., Miller, R.G., Bush, P.J., and R.B. Dorion. “Biomechanical factors in human dermal bitemarks in a cadaver model.” *Journal of Forensic Sciences,* Vol. 54, No. 1 (2009): 167-76.

[230] Bush, M.A., Cooper, H.I., and R.B. Dorion. “Inquiry into the scientific basis for bitemark profiling and arbitrary distortion compensation.” *Journal of Forensic Sciences*, Vol. 55, No. 4 (2010): 976-83.

[231] Adam Freeman and Iain Pretty “Construct validity of bitemark assessments using the ABFO decision tree,” presentation at the 2016 Annual Meeting of the American Academy of Forensic Sciences. See: online.wsj.com/public/resources/documents/ConstructValidBMdecisiontreePRETTYFREEMAN.pdf.

bitemark, or not a human bitemark; and (3) whether distinct features (arches and toothmarks) were identifiable.[232] Among the 38 examiners who completed the study, it was reported that there was unanimous agreement on the first question in only 4 of the 100 cases and agreement of at least 90 percent in only 20 of the 100 cases. Across all three questions, there was agreement of at least 90 percent in only 8 of the 100 cases.

In a similar study in Australia, 15 odontologists were shown a series of six bitemarks from contemporary cases, five of which were marks confirmed by living victims to have been caused by teeth, and were asked to explain, in narrative form, whether the injuries were, in fact, bitemarks.[233] The study found wide variability among the practitioners in their conclusions about the origin, circumstance, and characteristics of the patterned injury for all six images. Surprisingly, those with the most experience (21 or more years) tended to have the widest range of opinions as to whether a mark was of human dental origin or not.[234] Examiners' opinions varied considerably as to whether they thought a given mark was suitable for analysis, and individual practitioners demonstrated little consistency in their approach in analyzing one bitemark to the next. The study concluded that this "inconsistency indicates a fundamental flaw in the methodology of bitemark analysis and should lead to concerns regarding the reliability of any conclusions reached about matching such a bitemark to a dentition."[235]

Studies of Scientific Validity and Reliability

As discussed above, the foundational validity of a subjective method can only be established through multiple independent black-box studies.

The 2009 NRC report found that the scientific validity of bitemark analysis had not been established.[236] In its own review of the literature PCAST found few empirical studies that attempted to study the validity and reliability of the methods to identify the source of a bitemark.

In a 1975 paper, two examiners were asked to match photographs of bitemarks made by 24 volunteers in skin from freshly slaughtered pigs with dental models from these same volunteers. [237] The photographs were taken at 0, 1, and 24 hours after the bitemark was produced. Examiners' performance was poor and deteriorated with

[232] The raw data are made available by the authors upon request. They were reviewed by Professor Karen Kafadar, a member of the panel of Senior Advisors for this study.

[233] Page, M., Taylor, J., and M. Blenkin. "Expert interpretation of bitemark injuries – a contemporary qualitative study." *Journal of Forensic Sciences*, Vol. 58, No. 3 (2013): 664-72.

[234] For example, one examiner expressed certainty that one of the images was a bitemark, stating, "I know from experience that that's teeth because I did a case at the beginning of the year, that when I first looked at the images I didn't think they were teeth, because the injuries were so severe. But when I saw the models, and scratched them down my arm, they looked just like that." Another expressed doubt that the same image was a bitemark, also based on his or her experience: "Honestly I don't think it's a bite mark… there could be any number of things that could have caused that. Whether this is individual tooth marks here I doubt. I've never seen anything like that." Ibid., 666.

[235] Ibid., 670.

[236] "There is continuing dispute over the value and scientific validity of comparing and identifying bite marks." National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (2009): 151.

[237] Whittaker, D.K. "Some laboratory studies on the accuracy of bitemark comparison." *International Dental Journal,* Vol. 25, No. 3 (1975): 166–71.

time following the bite. The proportion of photographs incorrectly attributed was 28 percent, 65 percent, and 84 percent at the 0, 1, and 24 hour time points.

In a 1999 paper, 29 forensic dental experts—as well as 80 others, including general dentists, dental students, and lay participants—were shown color prints of human bitemarks from 50 court cases and asked to decide whether each bitemark was made by an adult or a child.[238] The decisions were compared to the verdict from the cases. All groups performed poorly.[239]

In a 2001 paper, 32 AFBO-certified diplomates were asked to report their certainty that 4 specific bitemarks might have come from each of 7 dental models, consisting of the four correct sources and three unrelated samples.[240,241] Such a "closed-set" design (where the correct source is present for each questioned samples) is inappropriate for assessing reliability, because it will tend to underestimate the false positive rate.[242] Even with this closed-set design, 11 percent of comparisons to the incorrect source were declared to be "probable," "possible," or "reasonable medical certainty" matches.

In another 2001 paper, 10 AFBO-certified diplomates were given 10 independent tests, each consisting of bitemark evidence and two possible sources. The evidence was produced by clamping a dental model onto freshly slaughtered pigs, subjectively confirming that "sufficient detail was recorded," and photographing the bitemark. The correct source was present in all but two of the tests (mostly closed-set design). The mean false positive rate was 15.9 percent—that is, roughly 1 in 6.

In a 2010 paper, 29 examiners with various levels of training (including 9 AFBO-certified diplomates) were provided with photographs of 18 human bitemarks and dentition from three human individuals (A, B, C) and were asked to decide whether the bitemarks came from A, B, C, or none of the above. The bitemarks had been produced in live pigs, using a biting machine with dentition from individuals A, B, and D (for which the dentition was not provided to the examiners). For bitemarks produced by D, the diplomates erroneously declared a match to A, B, or C in 17 percent of cases—again, roughly 1 in 6.

[238] Whittaker, D.K., Brickley, M.R., and L. Evans. "A comparison of the ability of experts and non-experts to differentiate between adult and child human bite marks using receiver operating characteristic (ROC) analysis." *Forensic Science International*, Vol. 92, No. 1 (1998): 11-20.

[239] The authors asked observers to indicate how certain they were a bitemark was made by an adult, using a 6 point scale. Receiver-Operator Characteristic (ROC) curves were derived from the data. The Area under the Curve (AUC) was calculated for each group (where AUC = 1 represents perfect classification and AUC = 0.5 is equivalent to random decision-making). The Area under the Curve (AUC) was between 0.62-0.69, which is poor.

[240] Arheart, K.L., and I.A. Pretty. "Results of the 4th AFBO Bitemark Workshop-1999." *Forensic Science International*, Vol. 124, No. 2-3 (2001): 104-11.

[241] The four bitemarks consisted of three from criminal cases and one produced by an individual deliberately biting into a block of cheese. The seven dental models corresponded to the three defendants convicted in the criminal cases (presumed to be the biters), the individual who bit the cheese, and three unrelated individuals.

[242] In closed-set tests, examiners will perform well as long as they choose the closest matching dental model. In an open-set design in which none of models may be correct, the opportunity for false positives is higher. The open-set design resembles the application in casework. See the extensive discussion of closed-set designs in firearms analysis (Section 5.5).

Conclusion

Few empirical studies have been undertaken to study the ability of examiners to accurately identify the source of a bitemark. Among those studies that have been undertaken, the observed false positive rates were so high that the method is clearly scientifically unreliable at present. (Moreover, several of these studies employ inappropriate closed-set designs that are likely to *under*estimate the false-positive rate.)

Finding 4: Bitemark analysis

Foundational validity. PCAST finds that bitemark analysis does not meet the scientific standards for foundational validity, and is far from meeting such standards. To the contrary, available scientific evidence strongly suggests that examiners cannot consistently agree on whether an injury is a human bitemark and cannot identify the source of bitemark with reasonable accuracy.

The Path Forward

Some practitioners have expressed concern that the exclusion of bitemarks in court could hamper efforts to convict defendants in some cases.[243] If so, the correct solution, from a scientific perspective, would not be to admit expert testimony based on invalid and unreliable methods, but rather to attempt to develop scientifically valid methods.

However, PCAST considers the prospects of developing bitemark analysis into a scientifically valid method to be low. We advise against devoting significant resources to such efforts.

5.4 Latent Fingerprint Analysis

Latent fingerprint analysis was first proposed for use in criminal identification in the 1800s and has been used for more than a century. The method was long hailed as infallible, despite the lack of appropriate studies to assess its error rate. As discussed above, this dearth of empirical testing indicated a serious weakness in the scientific culture of forensic science—where validity was assumed rather than proven. Citing earlier guidelines now acknowledged to have been inappropriate,[244] the DOJ recently noted,

> *Historically, it was common practice for an examiner to testify that when the … methodology was correctly applied, it would always produce the correct conclusion. Thus any error that occurred would be human error and the resulting error rate of the methodology would be zero. This view was described by the Department of Justice in 1984 in the publication The Science of Fingerprints, where it states, "Of all the methods of identification, fingerprinting alone has proved to be both infallible and feasible."* [245]

In response to the 2009 NRC report, the latent print analysis field has made progress in recognizing the need to perform empirical studies to assess foundational validity and measure reliability. Much credit goes to the FBI

[243] The precise proportion of cases in which bitemarks play a key role is unclear, but is clearly small.
[244] Federal Bureau of Investigation. *The Science of Fingerprints*. U.S. Government Printing Office. (1984): iv.
[245] See: www.justice.gov/olp/file/861906/download.

Laboratory, which has led the way in performing both black-box studies, designed to measure reliability, and "white-box studies," designed to understand the factors that affect examiners' decisions.[246] PCAST applauds the FBI's efforts. There are also nascent efforts to begin to move the field from a purely subjective method toward an objective method—although there is still a considerable way to go to achieve this important goal.

Methodology

Latent fingerprint analysis typically involves comparing (1) a "latent print" (a complete or partial friction-ridge impression from an unknown subject) that has been developed or observed on an item) with (2) one or more "known prints" (fingerprints deliberately collected under a controlled setting from known subjects; also referred to as "ten prints"), to assess whether the two may have originated from the same source. (It may also involve comparing latent prints with one another.)

It is important to distinguish latent prints from known prints. A known print contains fingerprint images of up to ten fingers captured in a controlled setting, such as an arrest or a background check.[247] Because known prints tend to be of high quality, they can be searched automatically and reliably against large databases. By contrast, latent prints in criminal cases are often incomplete and of variable quality (smudged or otherwise distorted), with quality and clarity depending on such factors as the surface touched and the mechanics of touch.

An examiner might be called upon to (1) compare a latent print to the fingerprints of a known suspect that has been identified by other means ("identified suspect") or (2) search a large database of fingerprints to identify a suspect ("database search").

[246] See: Hicklin, R.A., Buscaglia, J., Roberts, M.A., Meagher, S.B., Fellner, W., Burge, M.J., Monaco, M., Vera, D., Pantzer, L.R., Yeung, C.C., and N. Unnikumaran. "Latent fingerprint quality: a survey of examiners." *Journal of Forensic Identification*. Vol. 61, No. 4 (2011): 385-419; Hicklin, R.A., Buscaglia, J., and M.A. Roberts. "Assessing the clarity of friction ridge impressions." *Forensic Science International*, Vol. 226, No. 1 (2013): 106-17; Ulery, B.T., Hicklin, R.A., Kiebuzinski, G.I., Roberts, M.A., and J. Buscaglia. "Understanding the sufficiency of information for latent fingerprint value determinations." *Forensic Science International*, Vol. 230, No. 1-3 (2013): 99-106; Ulery, B.T., Hicklin, R.A., and J. Buscaglia. "Repeatability and reproducibility of decisions by latent fingerprint examiners." *PLoS ONE*, (2012); and Ulery, B.T., Hicklin, R.A., Roberts, M.A., and J. Buscaglia. "Changes in latent fingerprint examiners' markup between analysis and comparison." *Forensic Science International*, Vol. 247 (2015): 54-61.

[247] See: Committee on Science, Subcommittee on Forensic Science of the National Science and Technology Council. "Achieving Interoperability for Latent Fingerprint Identification in the United States." (2014). www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/afis_10-20-2014_draftforcomment.pdf.

Examiners typically follow an approach called "ACE" or "ACE-V," for Analysis, Comparison, Evaluation, and Verification.[248,249] The approach calls on examiners to make a series of subjective assessments. An examiner uses subjective judgment to select particular regions of a latent print for analysis. If there are no identified persons of interest, the examiner will run the latent print against an Automated Fingerprint Identification System (AFIS),[250] containing large numbers of known prints, which uses non-public, proprietary image-recognition algorithms[251] to generate a list of potential candidates that share similar fingerprint features.[252] The examiner then manually compares the latent print to the fingerprints from the specific person of interest or from the closest candidate matches generated by the computer by studying selected features[253] and then comes to a subjective decision as to whether they are similar enough to declare a proposed identification.

ACE-V adds a verification step. For the verification step, implementation varies widely.[254] In many laboratories, only identifications are verified, because it is considered too burdensome, in terms of time and cost, to conduct

[248] "A latent print examination using the ACE-V process proceeds as follows: *Analysis* refers to an initial information-gathering phase in which the examiner studies the unknown print to assess the quality and quantity of discriminating detail present. The examiner considers information such as substrate, development method, various levels of ridge detail, and pressure distortions. A separate analysis then occurs with the exemplar print. *Comparison* is the side-by-side observation of the friction ridge detail in the two prints to determine the agreement or disagreement in the details. In the *Evaluation* phase, the examiner assesses the agreement or disagreement of the information observed during Analysis and Comparison and forms a conclusion. *Verification* in some agencies is a review of an examiner's conclusions with knowledge of those conclusions; in other agencies, it is an independent re-examination by a second examiner who does not know the outcome of the first examination." National Institute of Standards and Technology. "*Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach."* (2012), available at: www.nist.gov/oles/upload/latent.pdf.
[249] Reznicek, M., Ruth, R.M., and D.M. Schilens. "ACE-V and the scientific method." *Journal of Forensic Identification*, Vol. 60, No. 1 (2010): 87-103.
[250] State and local jurisdictions began purchasing AFIS systems in the 1970s and 1980s from private vendors, each with their own proprietary software and searching algorithms. In 1999, the FBI launched the Integrated Automated Fingerprint Identification System (IAFIS), a national fingerprint database that houses fingerprints and criminal histories on more than 70 million subjects submitted by state, local and federal law enforcement agencies (recently replaced by the Next Generation Identification (NGI) System). Some criminal justice agencies have the ability to search latent prints not only against their own fingerprint database but also against a hierarchy of local, state, and federal databases. System-wide interoperability, however, has yet to be achieved. See: Committee on Science, Subcommittee on Forensic Science of the National Science and Technology Council. "Achieving Interoperability for Latent Fingerprint Identification in the United States." (2014). www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/afis_10-20-2014_draftforcomment.pdf.
[251] The algorithms used in generating candidate matches are proprietary and have not been made publicly available.
[252] The FBI Laboratory requires examiners to complete and document their analysis of the latent fingerprint before reviewing any known fingerprints or moving to the comparison and evaluation phase, this this requirement is not shared by all labs.
[253] Fingerprint features are compared at three levels of detail—level 1 ("ridge flow"), level 2 ("ridge path"), and level 3 ("ridge features" or "shapes"). "Ridge flow" refers to classes of pattern types shared by many individuals, such as loop or whorl formations; this level is only sufficient for exclusions, not for declaring identifications. "Ridge path" refers to minutiae that can be used for declaring identifications, such as bifurcations or dots. "Ridge shapes" include the edges of ridges and location of pores. See: National Institute of Standards and Technology. "*Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach."* (2012), available at: www.nist.gov/oles/upload/latent.pdf.
[254] Black, J.P. "Is there a need for 100% verification (review) of latent print examination conclusions?" *Journal of Forensic Identification*, Vol. 62, No.1 (2012): 80-100.

independent examinations in all cases (for example, exclusions). This procedure is problematic because it is not blind: the second examiner knows the first examiner reached a conclusion of proposed identification, which creates the potential for confirmation bias. In the aftermath of the Madrid train bombing case misidentification (see below), the FBI Laboratory adopted requirements to conduct, in certain cases, "independent application of ACE to a friction ridge print by another qualified examiner, who does not know the conclusion of the primary examiner."[255] In particular, the FBI Laboratory uses blind verification in cases considered to present the greatest risk of error, such as where a single fingerprint is identified, excluded, or deemed inconclusive.[256]

As noted in Chapter 2, earlier concerns[257] about the reliability of latent fingerprint analysis increased substantially following a prominent misidentification of a latent fingerprint recovered from the 2004 bombing of the Madrid commuter train system. An FBI examiner concluded with "100 percent certainty" that the fingerprint matched Brandon Mayfield, an American in Portland, Oregon, even though Spanish authorities were unable to confirm the identification. Reviewers believe the misidentification resulted in part from "confirmation bias" and "reverse reasoning"—that is, going from the known print to the latent image in a way that led to overreliance on apparent similarities and inadequate attention to differences.[258] As described in a recent paper by scientists at the FBI Laboratory,

> *A notable example of the problem of bias from the exemplar resulting in circular reasoning occurred in the Madrid misidentification, in which the initial examiner reinterpreted five of the original seven analysis points to be more consistent with the (incorrect) exemplar: ''Having found as many as 10 points of unusual similarity, the FBI examiners began to 'find' additional features in LFP 17 [the latent print] that were not really there, but rather suggested to the examiners by features in the Mayfield prints.''*[259]

In contrast to DNA analysis, the rules for declaring an identification that were historically used in fingerprint analysis were not set in advance nor uniform among examiners. As described by a February 2012 report from an Expert Working Group commissioned by NIST and NIJ:

[255] U.S. Department of Justice, Office of the Inspector General. "A Review of the FBI's Progress in Responding to the Recommendations in the Office of the Inspector General Report on the Fingerprint Misidentification in the Brandon Mayfield Case." (2011). www.oig.justice.gov/special/s1105.pdf. See also: Federal Bureau of Investigation. Laboratory Division. *Latent Print Operations Manual: Standard Operating Procedures for Examining Friction Ridge Prints*. FBI Laboratory, Quantico, Virginia, 2007 (updated May 24, 2011).

[256] Federal Bureau of Investigation. Laboratory Division. *Latent Print Operations Manual: Standard Operating Procedures for Examining Friction Ridge Prints*. FBI Laboratory, Quantico, Virginia, 2007 (updated May 24, 2011).

[257] Faigman, D.L., Kaye, D.H., Saks, M.J., and J. Sanders (Eds). *Modern Scientific Evidence: The Law and Science of Expert Testimony, 2015-2016 ed.* Thomson/West Publishing (2016). Saks, M.J. "Implications of *Daubert* for forensic identification science." *Shepard's Expert and Science Evidence Quarterly* 427, (1994).

[258] A Review of the FBI's handling of the Brandon Mayfield Case. U.S. Department of Justice, Office of the Inspector General (2006). oig.justice.gov/special/s0601/final.pdf.

[259] Ulery, B.T., Hicklin, R.A., Roberts, M.A., and J. Buscaglia. "Changes in latent fingerprint examiners' markup between analysis and comparison." *Forensic Science International*, Vol. 247 (2015): 54-61. The internal quotation is from U.S. Department of Justice Office of the Inspector General: A review of the FBI's handling of the Brandon Mayfield case (March 2006), www.justice.gov/oig/special/s0601/PDF_list.htm. US Department of Justice Office of the Inspector General: A review of the FBI's handling of the Brandon Mayfield case (March 2006), www.justice.gov/oig/special/s0601/PDF_list.htm.

> *The thresholds for these decisions can vary among examiners and among forensic service providers. Some examiners state that they report identification if they find a particular number of relatively rare concurring features, for instance, eight or twelve. Others do not use any fixed numerical standard. Some examiners discount seemingly different details as long as there are enough similarities between the two prints. Other examiners practice the one-dissimilarity rule, excluding a print if a single dissimilarity not attributable to perceptible distortion exists. If the examiner decides that the degree of similarity falls short of satisfying the standard, the examiner can report an inconclusive outcome. If the conclusion is that the degree of similarity satisfies the standard, the examiner reports an identification.* [260]

In September 2011, the Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST) issued "Standards for Examining Friction Ridge Impressions and Resulting Conclusions (Latent/Tenprint)" that begins to move latent print analysis in the direction of an objective framework. In particular, it suggests criteria concerning what combination of image quality and feature quantity (for example, the number of "minutiae" shared between two fingerprints) would be sufficient to declare an identification. The criteria are not yet fully objective, but they are a step in the right direction. The Friction Ridge Subcommittee of the OSAC has recognized the need for objective criteria in its identification of "Research Needs."[261] We note that the black-box studies described below did not set out to test these specific criteria, and so they have not yet been scientifically validated.

Studies of Scientific Validity and Reliability

As discussed above, the foundational validity of a subjective method can *only* be established through multiple independent black-box studies appropriately designed to assess validity and reliability.

Below, we discuss various studies of latent fingerprint analysis. The first five studies were not intended as validation studies, although they provide some incidental information about performance. Remarkably, there have been only two black-box studies that were intentionally and appropriately designed to assess validity and reliability—the first published by the FBI Laboratory in 2011; the second completed in 2014 but not yet published. Conclusions about foundational validity thus must rest on these two recent studies.

In summarizing these studies, we apply the guidelines described earlier in this report (see Chapter 4 and Appendix A). First, while we note (1) both the estimated false positive rates and (2) the upper 95 percent confidence bound on the false positive rate, we focus on the latter as, from a scientific perspective, the appropriate rate to report to a jury—because the primary concern should be about underestimating the false positive rate and the true rate could reasonably be as high as this value.[262] Second, while we note both the false positive rate among *conclusive* examinations (identifications or exclusions) or among *all* examinations (including inconclusives) are relevant, we focus primarily on the former as being, from a scientific perspective, the

[260] See: NIST. "*Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach.*" (2012), available at: www.nist.gov/oles/upload/latent.pdf.

[261] See: workspace.forensicosac.org/kws/groups/fric_ridge/documents.

[262] By convention, the 95 percent confidence bound is most widely used in statistics as reflecting the range of plausible values (see Appendix A).

appropriate rate to report to a jury—because fingerprint evidence used against a defendant in court will typically be the result of a conclusive examination.

Evett and Williams (1996)

This paper is a discursive historical review essay that contains a brief description of a small "collaborative study" relevant to the accuracy of fingerprint analysis.[263] In this study, 130 highly experienced examiners in England and Wales, each with at least ten years of experience in forensic fingerprint analysis, were presented with ten latent print-known pairs. Nine of the pairs came from past casework at New Scotland Yard and were presumed to be 'mated pairs' (that is, from the same source). The tenth pair was a 'non-mated pair' (from different sources), involving a latent print deliberately produced on a "dimpled beer mug." For the single non-mated pair, the 130 experts made no false identifications. Because the paper does not distinguish between exclusions and inconclusive examinations (and the authors no longer have the data),[264] it is impossible to infer the upper 95 percent confidence bound.[265]

Langenburg (2009a)

In a small pilot study, the author examined the performance of six examiners on 60 tests each.[266] There were only 15 conclusive examinations involving non-mated pairs (see Table 1 of the paper). There was one false positive, which the author excluded because it appeared to be a clerical error and was not repeated on subsequent retest. Even if this error is excluded, the tiny sample size results in a huge confidence interval (upper 95 percent confidence bound of 19 percent), with this upper bound corresponding to 1 error in 5 cases.

Langenburg (2009b)

In this small pilot study for the following paper, the author tested examiners in a conference room at a convention of forensic identification specialists.[267] The examiners were divided into three groups: high-bias (n=16), low-bias (n=12), and control (n=15). Each group was presented with 6 latent-known pairs, consisting of 3 mated and 3 non-mated pairs. The first two groups received information designed to bias their judgment by heightening their attention, while the control group received a generic description. For the non-mated pairs, the control group had 1 false positive among 43 conclusive examinations. The false positive rate was 2.3

[263] Evett, I.W., and R.L. Williams. "Review of the 16 point fingerprint standard in England and Wales." *Forensic Science International*, Vol. 46, No. 1 (1996): 49-73.

[264] I.W. Evett, personal communication.

[265] For example, the upper 95 percent confidence bound would be 1 in 44 if all 130 examinations were conclusive and 1 in 22 if half of the examinations were conclusive.

[266] Langenburg, G. "A performance study of the ACE-V Process: A pilot study to measure the accuracy, precision, reproducibility, repeatability, and biasability of conclusions resulting from the ACE-V process." *Journal of Forensic Identification*, Vol. 59, No. 2 (2009): 219–57.

[267] Langenburg, G., Champod, C., and P. Wertheim. "Testing for potential contextual bias effects during the verification stage of the ACE-V methodology when conducting fingerprint comparisons." *Journal of Forensic Sciences,* Vol. 54, No. 3 (2009): 571-82.

percent (upper 95 percent confidence bound of 11 percent), with the upper bound corresponding to 1 error in 9 cases.[268,269]

Langenburg, Champod, and Genessay (2012)

This study was not designed to assess the accuracy of latent fingerprint analysis, but rather to explore how fingerprint analysts would incorporate information from newly developed tools (such as a quality tool to aid in the assessment of the clarity of the friction ridge details; a statistical tool to provide likelihood ratios representing the strength of the corresponding features between compared fingerprints; and consensus information from a group of trained fingerprint experts) into their decision making processes.[270] Nonetheless, the study provided some information on the accuracy of latent print analysis. Briefly, 158 experts (as well as some trainees) were asked to analyze 12 latent print-exemplar pairs, consisting of 7 mated and 5 non-mated pairs. For the non-mated pairs, there were 17 false positive matches among 711 conclusive examinations by the experts.[271] The false positive rate was 2.4 percent (upper 95 percent confidence bound of 3.5 percent). The estimated error rate corresponds to 1 error in 42 cases, with an upper bound corresponding to 1 error in 28 cases.[272]

Tangen et al. (2011)

This Australian study was designed to study the reliability of latent fingerprint analysis by fingerprint experts.[273] The authors asked 37 fingerprint experts, as well as 37 novices, to examine 36 latent print-known pairs—consisting of 12 mated pairs, 12 non-mated pairs chosen to be "similar" (the most highly ranked exemplar from a different source in the Australian National Automated Fingerprint Identification System), and 12 "non-similar" non-mated pairs (chosen at random from the other prints). Examiners were asked to rate the likelihood they came from the same source on a scale from 1 to 12. The authors chose to define scores of 1-6 as identifications and scores of 7-12 as exclusions.[274] This approach does not correspond to the procedures used in conventional fingerprint examination.

For the "similar" non-mated pairs, the experts made 3 errors among 444 comparisons; the false positive rate was 0.68 percent (upper 95 percent confidence bound of 1.7 percent), with the upper bound corresponding to 1 error in 58 cases. For the "non-similar" non-mated pairs, the examiners made no errors in 444 comparisons; the

[268] If the two inconclusive examinations are included, the values are only slightly different: 2.2 percent (upper 95 percent confidence bound of 10.1 percent), with the odds being 1 in 10.

[269] The biased groups made no errors among 69 conclusive examinations.

[270] Langenburg, G., Champod, C., and T. Genessay. "Informing the judgments of fingerprint analysts using quality metric and statistical assessment tools." *Forensic Science International*, Vol. 219, No. 1-3 (2012): 183-98.

[271] We thank G. Langenburg for providing the data for the experts alone.

[272] If the 79 inconclusive examinations are included, the false positive rate was 2.15 percent (upper 95 percent confidence bound of 3.2 percent). The estimated false positive rate corresponds to 1 error in 47 cases, with the upper bound corresponding to 1 in 31.

[273] Tangen, J.M., Thompson, M.B., and D.J. McCarthy. "Identifying fingerprint expertise." *Psychological Science*, Vol. 22, No. 8 (2011): 995-7.

[274] There were thus no inconclusive results in this study.

false positive rate was thus 0 percent (upper 95 percent confidence bound of 0.62 percent), with the upper bound corresponding to 1 error in 148 cases. The experts substantially outperformed the novices.

Although interesting, the study does not constitute a black-box validation study of latent fingerprint analysis because its design did not resemble the procedures used in forensic practice (in particular, the process of assigning rating on a 12-point scale that the authors subsequently converted into identifications and exclusions).

FBI studies

The first study designed to test foundational validity and measure reliability of latent fingerprint analysis was a major black-box study conducted by FBI scientists and collaborators. Undertaken in response to the 2009 NRC report, the study was published in 2011 in a leading international science journal, *Proceedings of the National Academy of Sciences*.[275] The authors assembled a collection of 744 latent-known pairs, consisting of 520 mated pairs and 224 non-mated pairs. To attempt to ensure that the non-mated pairs were representative of the type of matches that might arise when police identify a suspect by searching fingerprint databases, the known prints were selected by searching the latent prints against the 58 million fingerprints in the AFIS database and selecting one of the closest matching hits. Each of 169 fingerprint examiners was shown 100 pairs and asked to classify them as an identification, an exclusion, or inconclusive. The study reported 6 false positive identifications among 3628 nonmated pairs that examiners judged to have "value for identification." The false positive rate was thus 0.17 percent (upper 95 percent confidence bound of 0.33 percent). The estimated rate corresponds to 1 error in 604 cases, with the upper bound indicating that the rate could be as high as 1 error in 306 cases.[276,277]

In 2012, the same authors reported a follow-up study testing repeatability and reproducibility. After a period of about seven months, 75 of the examiners from the previous study re-examined a subset of the latent-known comparisons from the previous study. Among 476 nonmated pairs leading to conclusive examinations (including 4 of the pairs that led to false positives in the initial study and were reassigned to the examiner who had made the erroneous decision), there were no false positives. These results (upper 95 percent confidence bound of 0.63 percent, corresponding to 1 error in 160) are broadly consistent with the false positive rate measured in the previous study.[278]

Miami-Dade study (Pacheco et al. (2014))

The Miami-Dade Police Department Forensic Services Bureau, with funding from the NIJ, conducted a black-box study designed to assess foundational validity and measure reliability; the results were reported to the sponsor

[275] Ulery, B.T., Hicklin, R.A., Buscaglia, J., and M.A. Roberts. "Accuracy and reliability of forensic latent fingerprint decisions." *Proceedings of the National Academy of Sciences*, Vol. 108, No. 19 (2011): 7733-8.

[276] If one includes the 455 inconclusive results for latent prints judged to have "value for identification," the false positive rate is 0.15 percent (upper 95 percent confidence bound of 0 of 0.29 percent). The estimated false positive rate corresponds to 1 error in 681 cases, with the upper bound corresponding to 1 in 344.

[277] The sensitivity (proportion of mated samples that were correctly declared to match) was 92.5 percent.

[278] Overall, 85-90 percent of the conclusive results were unchanged, with roughly 30 percent of false exclusions being repeated.

and posted on the internet, but they have not yet published in a peer-reviewed scientific journal.[279] The study differed significantly from the 2011 FBI black-box study in important respects, including that the known prints were not selected by means of a large database search to be similar to the latent prints (which should, in principle, have made it easier to declare exclusions for the non-mated pairs). The study found 42 false positives among 995 conclusive examinations. The false positive rate was 4.2 percent (upper 95 percent confidence bound of 5.4 percent). The estimated rate corresponds to 1 error in 24 cases, with the upper bound indicating that the rate could be as high as 1 error in 18 cases.[280] (Note: The paper observes that "in 35 of the erroneous identifications the participants appeared to have made a clerical error, but the authors could not determine this with certainty." In validation studies, it is inappropriate to exclude errors in a *post hoc* manner (see Box 4). However, if these 35 errors were to be excluded, the false positive rate would be 0.7 percent (confidence interval 1.4 percent), with the upper bound corresponding to 1 error in 73 cases.)

Conclusions from the studies

While it is distressing that meaningful studies to assess foundational validity and reliability did not begin until recently, we are encouraged that serious efforts are now being made to try to put the field on a solid scientific foundation—including by measuring accuracy, defining quality of latent prints, studying the reason for errors, and so on. Much credit belongs to the FBI Laboratory, as well as to academic researchers who had been pressing the need for research. Importantly, the FBI Laboratory is responsible for the only black-box study to date that has been *published* in a peer-reviewed journal.

The studies above cannot be directly compared for many reasons—including differences in experimental design, selection and difficulty level of latent-known pairs, and degree to which they represent the circumstances, procedures and pressures found in casework. Nonetheless, certain conclusions can be drawn from the results of the studies (summarized in Table 1 below):

(1) The studies collectively demonstrate that many examiners can, under *some* circumstances, produce correct answers at *some* level of accuracy.

(2) The empirically estimated false positive rates are *much higher* than the general public (and, by extension, most jurors) would likely believe based on longstanding claims about the accuracy of fingerprint analysis.[281,282]

[279] Pacheco, I., Cerchiai, B., and S. Stoiloff. "Miami-Dade research study for the reliability of the ACE-V process: Accuracy & precision in latent fingerprint examinations." (2014). www.ncjrs.gov/pdffiles1/nij/grants/248534.pdf.

[280] If the 403 inconclusive examinations are included, the false positive rate was 3.0 percent (upper 95 percent confidence bound of 3.9 percent). The estimated false positive rate corresponds to 1 error in 33 cases, with the upper bound corresponding to 1 in 26.

[281] The conclusion holds regardless of whether the rates are based on the point estimates or the 95 percent confidence bound, and on conclusive examinations or all examinations.

[282] These claims include the DOJ's own longstanding previous assertion that fingerprint analysis is "infallible" (www.justice.gov/olp/file/861906/download); testimony by a former head of the FBI's fingerprint unit testified that the FBI had "an error rate of one per every 11 million cases" (see p. 53); and a study finding that mock jurors estimated that the false positive rate for latent fingerprint analysis is 1 in 5.5 million (see p. 45). Koehler, J.J. "Intuitive error rate estimates for the forensic sciences." (August 2, 2016). Available at: papers.ssrn.com/sol3/papers.cfm?abstract_id=2817443.

(3) Of the two appropriately designed black-box studies, the larger study (FBI 2011 study) yielded a false positive rate that is unlikely to exceed 1 in 306 conclusive examinations while the other (Miami-Dade 2014 study) yielded a considerably higher false positive rate of 1 in 18.[283] (The earlier studies, which were not designed as validation studies, also yielded high false positive rates.)

Overall, it would be appropriate to inform jurors that (1) only two properly designed studies of the accuracy of latent fingerprint analysis have been conducted and (2) these studies found false positive rates that could be as high as 1 in 306 in one study and 1 in 18 in the other study. This would appropriately inform jurors that errors occur at detectable frequencies, allowing them to weigh the probative value of the evidence.

It is likely that a properly designed program of systematic, blind verification would decrease the false-positive rate, because examiners in the studies tend to make *different* mistakes.[284] However, there has not been empirical testing to obtain a quantitative estimate of the false positive rate that might be achieved through such a program.[285] And, it would not be appropriate simply to *infer* the impact of independent verification based on the theoretical assumption that examiners' errors are uncorrelated.[286]

It is important to note that, for a verification program to be truly blind and thereby avoid cognitive bias, examiners cannot only verify individualizations. As the authors of the FBI black-box study propose, "this can be ensured by performing verifications on a mix of conclusion types, not merely individualizations"—that is, a mix that ensures that verifiers cannot make inferences about the conclusions being verified.[287] We are not aware of any blind verification programs that currently follow this practice.

At present, testimony asserting any specific level of increased accuracy (beyond that measured in the studies) due to blind independent verification would be scientifically inappropriate, as speculation unsupported by empirical evidence.

[283] As noted above, the rate is 1 in 73 if one ignores the presumed clerical errors—although such *post hoc* adjustment is not appropriate in validation studies.

[284] The authors of the FBI black-box study note that five of the false positive occurred on test problem where a large majority of examiners correctly declared an exclusion, while one occurred on a test problem where the majority of examiners made inconclusive decisions. They state that "this suggests that these erroneous individualizations would have been detected if blind verification were routinely performed." Ulery, B.T., Hicklin, R.A., Buscaglia, J., and M.A. Roberts. "Accuracy and reliability of forensic latent fingerprint decisions." *Proceedings of the National Academy of Sciences*, Vol. 108, No. 19 (2011): 7733-8.

[285] The Miami-Dade study involved a small test of verification step, involving verification of 15 of the 42 false positives. In these 15 cases, the second examiner declared 13 cases to be exclusions and 2 to be inconclusive. The sample size is too small to draw a meaningful conclusion. And, the paper does not report verification results for the other 27 false positives.

[286] The DOJ has proposed to PCAST that "basic probability states that given an error rate for one examiner, the likelihood of a second examiner making the exact same error (verification/blind verification), would dictate that the rates should be multiplied." However, such a theoretical model would assume that errors by different examiners will be uncorrelated; yet they may depend on the difficulty of the problem and thus be correlated. Empirical studies are necessary to estimate error rates under blind verification.

[287] Ulery, B.T., Hicklin, R.A., Buscaglia, J., and M.A. Roberts. "Accuracy and reliability of forensic latent fingerprint decisions." *Proceedings of the National Academy of Sciences*, Vol. 108, No. 19 (2011): 7733-8.

We note that the DOJ believes that the high false positive rate observed in the Miami-Dade study (1 in 24, with upper confidence limit of 1 in 18) is unlikely to apply to casework at the FBI Laboratory, because it believes such a high rate would have been detected by the Laboratory's verification procedures. An independent evaluation of the verification protocols could shed light on the extent to which such inferences could be drawn based on the current Laboratory's verification procedures.

We also note it is conceivable that the false-positive rate in real casework could be higher than that observed in the experimental studies, due to exposure to potentially biasing information in the course of casework. Introducing test samples blindly into the flow of casework could provide valuable insight about the actual error rates in casework.

In conclusion, the FBI Laboratory black-box study has significantly advanced the field. There is a need for ongoing studies of the reliability of latent print analysis, building on its study design. Studies should ideally estimate error rates for latent prints of varying "quality" levels, using well defined measures (ideally, objective measures implemented by automated software[288]). As noted above, studies should be designed and conducted in conjunction with third parties with no stake in the outcome. This important feature was not present in the FBI study.

[288] An example is the Latent Quality Assessment (LQAS), which is designed as a proof-of-concept tool to evaluate the clarity of prints. Studies have found that error rates are correlated to the quality of the print. The software provides a manual and automated definitions of clarity maps, functions to process clarity maps, and annotation of corresponding points providing a method for overlapping of impression areas. Hicklin, R.A., Buscaglia, J., and M.A. Roberts. "Assessing the clarity of friction ridge impressions." *Forensic Science International*, Vol. 226, No. 1 (2013): 106-17. Another example is the Picture Annotation System (PiAnoS), developed by the University of Lausanne, which is being tested as a quality metric and statistical assessment tool for analysts. This platform uses tools that (1) assess the clarity of the friction ridge details, (2) provide likelihood ratios representing the strength of corresponding features between fingerprints, and (3) gives consensus information from a group of trained fingerprint experts. PiAnoS is an open-source software package available at: ips-labs.unil.ch/pianos.

Table 1: Error Rates in Studies of Latent Print Analysis*

Study	False Positives			
	Raw Data	Freq. (Confidence bound)	Estimated Rate	Bound on Rate
Early studies				
Langenburg (2009a)	0/14	0% (19%)	1 in ∞	1 in 5
Langenburg (2009b)	1/43	2.3% (11%)	1 in 43	1 in 9
Langenburg et al. (2012)	17/711	2.4% (3.5%)	1 in 42	1 in 28
Tangen et al. (2011) ("similar pairs")	3/444	0.68% (1.7%)	1 in 148	1 in 58
Tangen et al. (2011) ("dissimilar pairs")	0/444	0% (0.67%)	1 in ∞	1 in 148
Black-box studies				
Ulery et al. 2011 (FBI)**	6/3628	0.17% (0.33%)	1 in 604	1 in 306
Pacheco et al. 2014 (Miami-Dade)	42/995	4.2% (5.4%)	1 in 24	1 in 18
Pacheco et al. 2014 (Miami-Dade) (excluding clerical errors)	7/960	0.7% (1.4%)	1 in 137	1 in 73

* "Raw Data": Number of false positives divided by number of conclusive examinations involving non-mated pairs. "Freq. (Confidence Bound)": Point estimate of false positive frequency, and upper 95 percent confidence bound. "Estimated Rate": The odds of a false positive occurring, based on the observed proportion of false positives. "Bound on Rate": The odds of a false positive occurring, based on the upper 95 percent confidence bound—that is, the rate could reasonably be as high as this value.
** If inconclusive examinations are included for the FBI study, the rates are 1 in 681 and 1 in 344, respectively.

Scientific Studies of How Latent-print Examiners Reach Conclusions

Complementing the black-box studies, various studies have shed important light on how latent fingerprint examiners reach conclusions and how these conclusions may be influenced by extraneous factors. These studies underscore the serious risks that may arise in subjective methods.

Cognitive-bias studies

Itiel Dror and colleagues have done pioneering work on the potential role of cognitive bias in latent fingerprint analysis.[289] In an exploratory study in 2006, they demonstrated that examiners' judgments can be influenced by knowledge about other forensic examiners' decisions (a form of "confirmation bias").[290] Five fingerprint examiners were given fingerprint pairs that they had studied five years earlier in real cases and had judged to "match." They were asked to re-examine the prints, but were led to believe that they were the pair of prints that had been erroneously matched by the FBI in a high-profile case. Although they were instructed to ignore this information, four out of five examiners no longer judged the prints to "match." Although these studies are

[289] Dror, I.E., Charlton, D., and A.E. Peron. "Contextual information renders experts vulnerable to making erroneous identifications." *Forensic Science International*, Vol. 156 (2006): 74-878. Dror, I.E., and D. Charlton. "Why experts make errors." *Journal of Forensic identification*, Vol. 56, No.4 (2006): 600-16.
[290] Dror, I.E., Charlton, D., and A.E. Peron. "Contextual information renders experts vulnerable to making erroneous identifications." *Forensic Science International*, Vol. 156 (2006): 74-878.

too small to provide precise estimates of the impact of cognitive bias, they have been instrumental in calling attention to the issue.

Several strategies have been proposed for mitigating cognitive bias in forensic laboratories, including managing the flow of information in a crime laboratory to minimize exposure of the forensic analyst to irrelevant contextual information (such as confessions or eyewitness identification) and ensuring that examiners work in a linear fashion, documenting their finding about evidence from crime science *before* performing comparisons with samples from a suspect.[291,292]

FBI white-box studies

In the past few years, FBI scientists and their collaborators have also undertaken a series of "white-box" studies to understand the factors underlying the process of latent fingerprint analysis. These studies include analyses of fingerprint quality,[293,294] examiners' processes to determine the value of a latent print for identification or exclusion,[295] the sufficiency of information for identifications,[296] and how examiners' assessments of a latent print change when they compare it with a possible match.[297]

Among work on subjective feature-comparison methods, this series of papers is unique in its breadth, rigor and willingness to explore challenging issues. We could find no similarly self-reflective analyses for other subjective disciplines.

The two most recent papers are particularly notable because they involve the serious issue of confirmation bias. In a 2014 paper, the FBI scientists wrote

> *ACE distinguishes between the Comparison phase (assessment of features) and Evaluation phase (determination), implying that determinations are based on the assessment of features. However, our results suggest that this is not a simple causal relation: examiners' markups are also influenced by their determinations. How this reverse influence occurs is not obvious. Examiners may subconsciously reach a*

[291] Kassin, S.M., Dror, I.E., and J. Kakucka. "The forensic confirmation bias: Problems, perspectives, and proposed solutions." *Journal of Applied Research in Memory and Cognition*, Vol. 2, No. 1 (2013): 42-52. See also: Krane, D.E., Ford, S., Gilder, J., Iman, K., Jamieson, A., Taylor, M.S., and W.C. Thompson. "Sequential unmasking: A means of minimizing observer effects in forensic DNA interpretation." *Journal of Forensic Sciences*, Vol. 53, No. 4 (July 2008): 1006-7.

[292] Irrelevant contextual information could, depending on its nature, bias an examiner toward an incorrect identification or an incorrect exclusion. Either outcome is undesirable.

[293] Hicklin, R.A., Buscaglia, J., Roberts, M.A., Meagher, S.B., Fellner, W., Burge, M.J., Monaco, M., Vera, D., Pantzer, L.R., Yeung, C.C., and N. Unnikumaran. "Latent fingerprint quality: a survey of examiners." *Journal of Forensic Identification*. Vol. 61, No. 4 (2011): 385-419.

[294] Hicklin, R.A., Buscaglia, J., and M.A. Roberts. "Assessing the clarity of friction ridge impressions." *Forensic Science International*, Vol. 226, No. 1 (2013): 106-17.

[295] Ulery, B.T., Hicklin, R.A., Kiebuzinski, G.I., Roberts, M.A., and J. Buscaglia. "Understanding the sufficiency of information for latent fingerprint value determinations." *Forensic Science International*, Vol. 230, No. 1-3 (2013): 99-106.

[296] Ulery, B.T., Hicklin, R.A., and J. Buscaglia. "Repeatability and reproducibility of decisions by latent fingerprint examiners." *PLoS ONE*, (2012).

[297] Ulery, B.T., Hicklin, R.A., Roberts, M.A., and J. Buscaglia. "Changes in latent fingerprint examiners' markup between analysis and comparison." *Forensic Science International*, Vol. 247 (2015): 54-61.

> *preliminary determination quickly and this influences their behavior during Comparison (e.g., level of effort expended, how to treat ambiguous features). After making a decision, examiners may then revise their annotations to help document that decision, and examiners may be more motivated to provide thorough and careful markup in support of individualizations than other determinations. As evidence in support of our conjecture, we note in particular the distributions of minutia counts, which show a step increase associated with decision thresholds: this step occurred at about seven minutiae for most examiners, but at 12 for those examiners following a 12-point standard.*[298]

Similar observations had been made by Dror et al., who noted that the number of minutiae marked in a latent print was greater when a matching exemplar was present. [299] In addition, Evett and Williams described how British examiners, who used a 16-point standard for declaring identifications, used an exemplar to ‘‘tease the points out’’ of the latent print after they had reached an ‘‘inner conviction’’ that the prints matched.[300]

In a follow-up paper in 2015, the FBI scientists carefully studied how examiners analyzed prints and confirmed that, in the vast majority (>90 percent) of identification decisions, examiners modified the features marked in the latent fingerprint in response to an apparently matching known fingerprint (more often adding than subtracting features).[301] (The sole false positive in their study was an extreme case in which the conclusion was based almost entirely on subsequent marking of minutiae that had not been initially found and deletion of features that had been initially marked.)

The authors concluded that "there is a need for examiners to have some means of unambiguously documenting what they see during analysis and comparison (in the ACE-V process)" and that "rigorously defined and consistently applied methods of performing and documenting ACE-V would improve the transparency of the latent print examination process."

PCAST compliments the FBI scientists for calling attention to the risk of confirmation bias arising from circular reasoning. As a matter of scientific validity, examiners must be required to "complete and document their analysis of a latent fingerprint before looking at any known fingerprint" and "must separately document any data relied upon during comparison or evaluation that differs from the information relied upon during analysis."[302] The FBI adopted these rules following the Madrid train bombing case misidentification; they need to be universally adopted by all laboratories.

[298] Ulery, B.T., Hicklin, R.A., Roberts, M.A., and J. Buscaglia. "Measuring what latent fingerprint examiners consider sufficient information for individualization determinations." *PLoS ONE*, (2014).

[299] Dror, I.E., Champod, C., Langenburg, G., Charlton, D., Hunt, H., and R. Rosenthal. "Cognitive issues in fingerprint analysis: Inter- and intra-expert consistency and the effect of a 'target' comparison." *Forensic Science International*, Vol. 208, No. 1-3 (2011): 10-7.

[300] Evett, I.W., and R.L. Williams. "Review of the 16 point fingerprint standard in England and Wales." *Forensic Science International*, Vol. 46, No. 1 (1996): 49–73.

[301] Ulery, B.T., Hicklin, R.A., Roberts, M.A., and J. Buscaglia. "Changes in latent fingerprint examiners' markup between analysis and comparison." *Forensic Science International*, Vol. 247 (2015): 54-61.

[302] U.S. Department of Justice, Office of the Inspector General. "A Review of the FBI's Progress in Responding to the Recommendations in the Office of the Inspector General Report on the Fingerprint Misidentification in the Brandon Mayfield Case." (2011): 5, 27. www.oig.justice.gov/special/s1105.pdf.

Validity as Applied

Foundational validity means that a large group of examiners analyzing a specific type of sample can, under test conditions, produce correct answers at a known and useful frequency. It does not mean that a particular examiner has the ability to reliably apply the method; that the samples in the foundational studies are representative of the actual evidence of the case; or that the circumstances of the foundational study represent a reasonable approximation of the circumstances of casework.

To address these matters, courts should take into account several key considerations.

(1) Because latent print analysis, as currently practiced, depends on subjective judgment, it is scientifically unjustified to conclude that a particular examiner is capable of reliably applying the method unless the examiner has undergone regular and rigorous proficiency testing. Unfortunately, it is not possible to assess the appropriateness of current proficiency testing because the test problems are not publically released. (As emphasized previously, training and experience are no substitute, because neither provides any assurance that the examiner can apply the method reliably.)

(2) In any given case, it must be established that the latent print(s) are of the quality and completeness represented in the foundational validity studies.

(3) Because contextual bias may have an impact on experts' decisions, courts should assess the measures taken to mitigate bias during casework—for example, ensuring that examiners are not exposed to potentially biasing information and ensuring that analysts document ridge features of an unknown print before referring to the known print (a procedure known as "linear ACE-V"[303]).

Finding 5: Latent fingerprint analysis

Foundational validity. Based largely on two recent appropriately designed black-box studies, PCAST finds that latent fingerprint analysis is a foundationally valid subjective methodology—albeit with a false positive rate that is substantial and is likely to be higher than expected by many jurors based on longstanding claims about the infallibility of fingerprint analysis.

Conclusions of a proposed identification may be scientifically valid, provided that they are accompanied by accurate information about limitations on the reliability of the conclusion—specifically, that (1) only two properly designed studies of the foundational validity and accuracy of latent fingerprint analysis have been conducted, (2) these studies found false positive rates that could be as high as 1 error in 306 cases in one study and 1 error in 18 cases in the other, and (3) because the examiners were aware they were being tested, the actual false positive rate in casework may be higher. At present, claims of higher accuracy are

[303] U.S. Department of Justice, Office of the Inspector General. "A Review of the FBI's Progress in Responding to the Recommendations in the Office of the Inspector General Report on the Fingerprint Misidentification in the Brandon Mayfield Case." (2011): 27. www.oig.justice.gov/special/s1105.pdf.

not warranted or scientifically justified. Additional black-box studies are needed to clarify the reliability of the method.

Validity as applied. Although we conclude that the method is foundationally valid, there are a number of important issues related to its validity as applied.

> **(1) Confirmation bias.** Work by FBI scientists has shown that examiners typically alter the features that they initially mark in a latent print based on comparison with an apparently matching exemplar. Such circular reasoning introduces a serious risk of confirmation bias. Examiners should be required to complete and document their analysis of a latent fingerprint *before* looking at any known fingerprint and should separately document any additional data used during their comparison and evaluation.
>
> **(2) Contextual bias**. Work by academic scholars has shown that examiners' judgments can be influenced by irrelevant information about the facts of a case. Efforts should be made to ensure that examiners are not exposed to potentially biasing information.
>
> **(3) Proficiency testing**. Proficiency testing is essential for assessing an examiner's capability and performance in making accurate judgments. As discussed elsewhere in this report, proficiency testing needs to be improved by making it more rigorous, by incorporating it within the flow of casework, and by disclosing tests for evaluation by the scientific community.

From a scientific standpoint, validity as applied requires that an expert: (1) has undergone appropriate proficiency testing to ensure that he or she is capable of analyzing the full range of latent fingerprints encountered in casework and reports the results of the proficiency testing; (2) discloses whether he or she documented the features in the latent print in writing before comparing it to the known print; (3) provides a written analysis explaining the selection and comparison of the features; (4) discloses whether, when performing the examination, he or she was aware of any other facts of the case that might influence the conclusion; and (5) verifies that the latent print in the case at hand is similar in quality to the range of latent prints considered in the foundational studies.

The Path Forward

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

One direction is to continue to improve latent print analysis as a subjective method. With only two black-box studies so far (with very different error rates), there is a need for additional black-box studies building on the study design of the FBI black-box study. Studies should estimate error rates for latent prints of varying quality and completeness, using well-defined measures. As noted above, the studies should be designed and conducted in conjunction with third parties with no stake in the outcome.

A second—and more important—direction is to convert latent print analysis from a subjective method to an objective method. The past decade has seen extraordinary advances in automated image analysis based on machine learning and other approaches—leading to dramatic improvements in such tasks as face recognition.[304,305] In medicine, for example, it is expected that automated image analysis will become the gold standard for many applications involving interpretation of X-rays, MRIs, fundoscopy, and dermatological images.[306]

Objective methods based on automated image analysis could yield major benefits—including greater efficiency and lower error rates; it could also enable estimation of error rates from millions of pairwise comparisons. Initial efforts to develop automated systems could not outperform humans.[307] However, given the pace of progress in image analysis and machine learning, we believe that fully automated latent print analysis is likely to be possible in the near future. There have already been initial steps in this direction, both in academia and industry.[308]

The most important resource to propel the development of objective methods would be the creation of huge databases containing known prints, each with many corresponding ”simulated” latent prints of varying qualities and completeness, which would be made available to scientifically-trained researchers in academia and industry. The simulated latent prints could be created by “morphing” the known prints, based on transformations derived from collections of actual latent print-record print pairs.[309]

[304] See: cs.stanford.edu/people/karpathy/cvpr2015.pdf.

[305] Lu, C., and X. Tang. “Surpassing human-level face verification performance on LFW with GaussianFace.” arxiv.org/abs/1404.3840 (accessed July 2, 2016). Taigman, Y., Yang, M., Ranzato, M., and L. Wolf. “Deepface: Closing the gap to human-level performance in face verification.” www.cs.toronto.edu/~ranzato/publications/taigman_cvpr14.pdf (accessed July 2, 2016) and Schroff, F., Kalenichenko, D., and J. Philbin. “FaceNet: A unified embedding for face recognition and clustering.” arxiv.org/abs/1503.03832 (accessed July 2, 2016).

[306] Doi, K. “Computer-aided diagnosis in medical imaging: historical review, current status and future potential.” *Computerized Medical Imaging and Graphics*, Vol. 31, No. 4-5 (2007): 198-211 and Shiraishi, J., Li, Q., Appelbaum, D., and K. Doi. “Computer-aided diagnosis and artificial intelligence in clinical imaging.” *Seminars in Nuclear Medicine*, Vol. 41, No. 6 (2011): 449-62.

[307] For example, a study in 2010 reported that that humans outperformed an automated program for toolmark comparisons. See: Chumbley, L.S., Morris, M.D., Kreiser, M.J., Fisher, C., Craft J., Genalo, L.J., Davis, S., Faden, D., and J. Kidd. “Validation of Tool Mark Comparisons Obtained Using a Quantitative, Comparative, Statistical Algorithm." *Journal of Forensic Sciences*, Vol. 55, No. 4 (2010): 953-961.

[308] Arunalatha, J.A., Tejaswi, V., Shaila, K., Anvekar, D., Venugopal, K.R., Iyengar, S.S., and L.M. Patnaik. “FIVDL: Fingerprint Image Verification using Dictionary Learning.” *Procedia Computer Science*, Vol. 54 (2015): 482-490 and Srihari, S.N. “Quantitative Measures in Support of Latent Print Comparison: Final Technical Report.” NIJ Award Number: 2009-DN-BX-K208, University at Buffalo, SUNY, 2013. www.crime-scene-investigator.net/QuantitativeMeasuresinSupportofLatentPrint.pdf. In addition, Christophe Champod’s group at Université de Lausanne has an active program in this area.

[309] For privacy, fingerprints from deceased individuals could be used.

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* x *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) that 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/partly 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 examiners[332] 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-postive-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 images[339]—could attract significant interest from academic and industry.

5.6 Footwear Analysis: Identifying Characteristics

Methodology

Footwear analysis is a process that typically involves comparing a known object, such as a shoe, to a complete or partial impression found at a crime scene, to assess whether the object is likely to be the source of the impression. The process proceeds in a stepwise manner, beginning with a comparison of "class characteristics" (such as design, physical size, and general wear) and then moving to "identifying characteristics" or "randomly acquired characteristics (RACs)" (such as marks on a shoe caused by cuts, nicks, and gouges in the course of use).[340]

In this report, we do not address the question of whether examiners can reliably determine class characteristics—for example, whether a particular shoeprint was made by a size 12 shoe of a particular make. While it is important that that studies be undertaken to estimate the reliability of footwear analysis aimed at

[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/NRBTD). The database contains reflectance microscopy images and three-dimensional surface topography data acquired by NIST or submitted by users.

[340] See: SWGTREAD Range of Conclusions Standards for Footwear and Tire Impression Examinations (2013). SWGTREAD Guide for the Examination of Footwear and Tire Impression Evidence (2006) and Bodziak W. J. *Footwear Impression Evidence: Detection, Recovery, and Examination*. 2nd ed. CRC Press-Taylor & Francis, Boca Raton, Florida (2000): p 347.

determining class characteristics, PCAST chose not to focus on this aspect of footwear examination because it is not *inherently* a challenging measurement problem to determine class characteristics, to estimate the frequency of shoes having a particular class characteristic, or (for jurors) to understand the nature of the features in question.

Instead, PCAST focused on the reliability of conclusions, based on RACs, that an impression was likely to have come from a specific piece of footwear. This is a much harder problem, because it requires knowing how accurately examiners identify specific features shared between a shoe and an impression, how often they fail to identify features that would distinguish them, and what probative value should be ascribed to a particular RAC.

Despite the absence of empirical studies that measure examiners' accuracy, authorities in the footwear field express confidence that they can identify the source of an impression based on a single RAC.

As described in a 2009 article by an FBI forensic examiner published in the FBI's Forensic Science Communications:

> *An examiner first determines whether a correspondence of class characteristics exists between the questioned footwear impression and the known shoe. If the examiner deems that there are no inconsistencies in class characteristics, then the examination progresses to any identifying characteristics in the questioned impression. The examiner compares these characteristics with any identifying characteristics observed on the known shoe. Although unpredictable in their occurrence, the size, shape, and position of these characteristics have a low probability of recurrence in the same manner on a different shoe. Thus, combined with class characteristics, even one identifying characteristic is extremely powerful evidence to support a conclusion of identification.* [341]

In support, the article cites a leading textbook on footwear identification:

> *According to William J. Bodziak (2000), "Positive identifications may be made with as few as one random identifying characteristic, but only if that characteristic is confirmable; has sufficient definition, clarity, and features; is in the same location and orientation on the shoe outsole; and in the opinion of an experienced examiner, would not occur again on another shoe."* [342]

The article points to a mathematical model by Stone that claims that the chance is 1 in 16,000 that two shoes would share one identifying characteristics and 1 in 683 billion that they would share three characteristics.[343]

Such claims for "identification" based on footwear analysis are breathtaking—but lack scientific foundation.

The statement by Bodziak has two components: (1) that the examiner consistently observes a demonstrable RAC in a set of impressions and (2) that the examiner is positive that the RAC would not occur on another shoe. The

[341] Smith, M.B. *The Forensic Analysis of Footwear Impression Evidence*. www.fbi.gov/about-us/lab/forensic-science-communications/fsc/july2009/review/2009_07_review02.htm

[342] Bodziak W.J. *Footwear Impression Evidence: Detection, Recovery, and Examination*. 2nd ed. CRC Press-Taylor & Francis, Boca Raton, Florida (2000).

[343] Stone, R.S. "Footwear examinations: Mathematical probabilities of theoretical individual characteristics." *Journal of Forensic Identification*, Vol. 56, No. 4 (2006): 577-99.

first part is not unreasonable, but the second part is deeply problematic: It requires the examiner to rely on recollections and guesses about the frequency of features.

The model by Stone is entirely theoretical: it makes many unsupported assumptions (about the frequency and statistical independence of marks) that it does not test in any way.

The entire process—from choice of features to include (and ignore) and the determination of rarity—relies entirely on an examiner's subjective judgment. Under such circumstances, it is essential that the scientific validity of the method and estimates of its reliability be established by multiple, appropriate black-box studies.[344]

Background

The 2009 NRC report cited some papers that cast doubt on whether footwear examiners reach consistent conclusions when presented with the same evidence. For example, the report contained a detailed discussion of a 1996 European paper that presented examiners with six mock cases—two involving worn shoes from crime scenes, four with new shoes in which specific identifying characteristics had been deliberately added; the paper reported considerable variation in their answers.[345] PCAST also notes a 1999 Israeli study involving two cases from crime scenes that reached similar conclusions.[346]

In response to the 2009 NRC report, a 2013 paper claimed to demonstrate that American and Canadian footwear analysts exhibit greater consistency than seen in the 1996 European study.[347] However, this study differed substantially because the examiners in this study did not conduct their own examinations. For example, the photographs were pre-annotated to call out all relevant features for comparison—that is, the examiners were not asked to identify the features.[348] Thus, the study, by virtue of its design, cannot address the consistency of the examination process.

Moreover, the fundamental issue is not one of *consistency* (whether examiners give the *same* answer) but rather of *accuracy* (whether they give the *right* answer). Accuracy can be evaluated only from large, appropriately designed black-box studies.

[344] In addition to black-box studies, white-box studies are also valuable to identify the sources of errors.
[345] Majamma, H., and A. Ytti. "Survey of the conclusions drawn of similar footwear cases in various crime laboratories." *Forensic Science International*. Vol. 82, No. 1 (1996): 109-20.
[346] Shor, Y., and S. Weisner. "Survey on the conclusions drawn on the same footwear marks obtained in actual cases by several experts throughout the world." *Journal of Forensic Science,* Vol. 44, No. 2 (1999): 380-4384.
[347] Hammer, L., Duffy, K., Fraser, J., and N.N. Daeid. "A study of the variability in footwear impression comparison conclusions." *Journal of Forensic Identification*, Vol. 63, No. 2 (2013): 205-18.
[348] The paper states that "All characteristics and observations that were to be considered by the examiners during the comparisons were clearly identified and labeled for each impression."

Studies of Scientific Validity and Reliability

PCAST could find no black-box studies appropriately designed to establish the foundational validity of identifications based on footwear analysis.

Consistent with our conclusion, the OSAC Footwear and Tire subcommittee recently identified the need for both black-box and white-box examiner reliability studies—citing it as a "major gap in current knowledge" in which there is "no or limited current research being conducted."[349]

Finding 7: Footwear analysis

Foundational validity. PCAST finds there are no appropriate empirical studies to support the foundational validity of footwear analysis to associate shoeprints with particular shoes based on specific identifying marks (sometimes called "randomly acquired characteristics). Such conclusions are unsupported by any meaningful evidence or estimates of their accuracy and thus are not scientifically valid.

PCAST has not evaluated the foundational validity of footwear analysis to identify class characteristics (for example, shoe size or make).

The Path Forward

In contrast to latent fingerprint analysis and firearms analysis, there is little research on which to build with respect to conclusions that seek to associate a shoeprint with a particular shoe (identification conclusions).

New approaches will be needed to develop paradigms. As an initial step, the FBI Laboratory is engaging in a study examining a set of 700 similar boots that were worn by FBI Special Agent cadets during their 16-week training program. The study aims to assess whether RACs are observed on footwear from different individuals. While such "uniqueness" studies (i.e., demonstrations that many objects have distinct features) cannot establish foundational validity (see p. 42), the impressions generated from the footwear could provide an initial dataset for (1) a pilot black-box study and (2) a pilot database of feature frequencies. Importantly, NIST is beginning a study to see if it is possible to quantify the footwear examination process, or at minimum aspects of the process, in an effort to increase the objectivity of footwear analysis.

Separately, evaluations should be undertaken concerning the accuracy and reliability of determinations about class characteristics, a topic that is not addressed in this report.

[349] See: www.nist.gov/forensics/osac/upload/SAC-Phy-Footwear-Tire-Sub-R-D-001-Examiner-Reliability-Study_Revision_Feb_2016.pdf (accessed on May, 12, 2016).

5.7 Hair Analysis

Forensic hair examination is a process by which examiners compare microscopic features of hair to determine whether a particular person may be the source of a questioned hair. As PCAST was completing this report, the DOJ released for comment guidelines concerning testimony on hair examination that included supporting documents addressing the validity and reliability of the discipline.[350] While PCAST has not undertaken a comprehensive review of the discipline, we undertook a review of the supporting document in order to shed further light on the standards for conducting a scientific evaluation of a forensic feature-comparison discipline.

The supporting document states that "microscopic hair comparison has been demonstrated to be a valid and reliable scientific methodology," while noting that "microscopic hair comparisons alone cannot lead to personal identification and it is crucial that this limitation be conveyed both in the written report and in testimony."

Foundational Studies of Microscopic Hair Examination

In support of its conclusion that hair examination is valid and reliable, the DOJ supporting document discusses five studies of human hair comparison. The primary support is a series of three studies by Gaudette in 1974, 1976 and 1978.[351] The 1974 and 1976 studies focus, respectively, on head hair and pubic hair. Because the designs and results are similar, we focus on the head hair study.

The DOJ supporting document states that "In the head hair studies, a total of 370,230 intercomparisons were conducted, with only nine pairs of hairs that could not be distinguished"—corresponding to a false positive rate of less than 1 in 40,000. More specifically, the design of this 1974 study was as follows: a single examiner (1) scored between 6 and 11 head hairs from each of 100 individuals (a total of 861 hairs) with respect to 23 distinct categories (with a total of 96 possible values); (2) compared the hairs from *different* individuals, to identify those pairs of hairs with fewer than four differences; and (3) compared these pairs of hairs microscopically to see if they could be distinguished.

The DOJ supporting document fails to note that these studies were strongly criticized by other scientists for flawed methodology.[352] The most serious criticism was that Gaudette compared only hairs from *different* individuals, but did not look at hairs from the *same* individual. As pointed out by a 1990 paper by two authors at the Hair and Fibre Unit of the Royal Canadian Mounted Police Forensic Laboratory (as well as in other papers),

[350] See: Department of Justice Proposed Uniform Language for Testimony and Reports for the Forensic Hair Examination Discipline, available at: www.justice.gov/dag/file/877736/download and Supporting Documentation for Department of Justice Proposed Uniform Language for Testimony and Reports for the Forensic Hair Examination Discipline, available at: www.justice.gov/dag/file/877741/download.

[351] Gaudette, B.D., and E.S. Keeping. "An attempt at determining probabilities in human scalp hair comparisons." *Journal of Forensic Sciences*, Vol. 19 (1974): 599-606; Gaudette, B.D. "Probabilities and Human Pubic Hair Comparisons." *Journal of Forensic Science*, Vol. 21 (1976): 514-517; Gaudette, B.D. "Some further thoughts on probabilities and human hair comparisons." *Journal of Forensic Sciences,* Vol. 23 (1978): 758–763.

[352] Wickenheiser, R. A. and D.G. Hepworth, D.G. "Further evaluation of probabilities in human scalp hair comparisons." *Journal of Forensic Sciences*, Vol. 35 (1990): 1323-29. See also Barnett, P.D. and R.R. Ogle. "Probabilities and human hair comparison." *Journal of Forensic Sciences*, Vol. 27 (1982): 272–278 and Gaudette, B.D. "A Supplementary Discussion of Probabilities and Human Hair Comparisons." *Journal of Forensic Sciences*, Vol. 27, No. 2, (1982): 279-89.

the apparently low false positive rate could have resulted from examiner bias—that is, that the examiner explicitly knew that all hairs being examined came from *different* individuals and thus could be inclined, consciously or unconsciously, to search for differences.[353] In short, one cannot appropriately assess a method's false-positive rate without simultaneously assessing its *true*-positive rate (sensitivity). In the 1990 paper, the authors used a similar study design, but employed *two* examiners who examined *all* pairs of hairs. They found non-repeatability for the individual examiners ("each examiner had considerable day-to-day variation in hair feature classification") and non-reproducibility between the examiners ("in many cases, the examiners classified the same hairs differently"). Most notably, they found that, while the examiners found no matches between hairs from *different* individuals, they also found almost no consistent matches among hairs from the *same* person. Of 15 pairs of same-source hairs that the authors determined *should* have been declared to match, *only two* were correctly called by both examiners.

In Gaudette's 1978 study, the author gave a different hair to each of three examiner trainees, who had completed one year of training, and asked them to identify any matching samples among a reference set of 100 hairs (which, unbeknownst to the examiners, came from 100 different people, including the sources of the hairs). The three examiners reported 1, 1 and 4 matches, consisting of 3 correct and 3 incorrect answers. Of the declared matches, 50 percent were thus false positive associations. Among the 300 total comparisons, the overall false positive rate was 1 percent, which notably is 400-fold higher than the rate estimated in the 1974 study.

Interestingly, we noted that the DOJ supporting document wrongly reports the results of the study—*claiming that the third examiner trainee made only 1 error, rather than 3 errors*. The explanation for this discrepancy is found in a remarkably frank passage of the text, which illustrates the need for employing rigorous protocols in evaluating the results of experiments:

> *"Two trainees correctly identified one hair and only one hair as being similar to the standard. The third trainee first concluded that there were four hairs similar to the standard. Upon closer examination and consultation with the other examiners, he was easily able to identify one of his choices as being incorrect. However, he was still convinced that there were three hairs similar to the standard, the correct one and two others. Examination by the author brought the opinion that one of these two others could be eliminated but that the remaining one was indistinguishable from hairs in the standard. Another experienced examiner then studied the hairs and also concluded that one of the two others could be eliminated. This time, however, it was the opposite to the one picked by the author!"*[354]

Ex post facto reclassification of errors is generally not advisable in studies pertaining to validity and reliability.

[353] In addition, inconsistency in scoring features would add random noise to any structure in the data (e.g., feature correlations) and thereby decrease the frequency of matches occurring by chance.

[354] Gaudette, B.D. "Some further thoughts on probabilities and human hair comparisons." *Journal of Forensic Sciences* Vol. 23, (1978): 758–763.

The two other human-hair studies discussed in the DOJ supporting document are also problematic. A 1983 paper involved hair samples from 100 individuals, classified into three racial groups.[355] After the author had extensively studied the hairs, she asked a neutral party to set up seven "blind" challenge problems for her—by selecting 10 questioned hairs and 10 known hairs (across groups in three cases, within a group in four cases).[356] The results consist of a single sentence in which the author simply states that she performed with "100 percent accuracy." Self-reported performance on a test is not generally regarded as appropriate scientific methodology.

A 1984 paper studied hairs from 17 pairs of twins (9 fraternal, 6 identical and 2 unknown zygosity) and one set of identical triplets.[357] Interestingly, the hairs from identical twins showed no greater similarity than the hairs from fraternal twins. In the sole test designed to simulate forensic casework, two examiners were given seven challenge problems, each consisting of comparing a questioned hair to between 5 and 10 known hairs. The false positive rate was 1 in 12, which is roughly 3300-fold higher than in Gaudette's 1974 study of hair from unrelated individuals.[358]

PCAST finds that, based on their methodology and results, the papers described in the DOJ supporting document do not provide a scientific basis for concluding that microscopic hair examination is a valid and reliable process.

After describing the scientific papers, the DOJ document goes on to discuss the conclusions that can be drawn from hair comparison:

> *These studies have also shown that microscopic hair comparison alone cannot lead to personal identification and it is crucial that this limitation be conveyed both in the written report and in testimony.*
>
> *The science of microscopic hair comparison acknowledges that the microscopic characteristics exhibited by a questioned hair may be encompassed by the range of characteristics exhibited by known hair samples of more than one person. If a questioned hair is associated with a known hair sample that is truly not the source, it does not mean that the microscopic hair association is in error. Rather, it highlights the limitation of the science in that there is an unknown pool of people who could have contributed the questioned hair. However, studies have not determined the number of individuals who share hairs with the same or similar characteristics.*

The passage violates fundamental scientific principles in two important ways. The first problem is that it uses the fact that the method's accuracy is not *perfect* to dismiss the need to know the method's accuracy *at all*. According to the supporting document, it is not an "error" but simply a "limitation of the science" when an examiner associates a hair with an individual who was not actually the source of the hair. This is disingenuous. When an expert witness tells a jury that a hair found at the scene of a crime is microscopically indistinguishable

[355] Strauss, M.T. "Forensic characterization of human hair." *The Microscope*, Vol. 31, (1983): 15-29.

[356] The DOJ supporting document mistakenly reports that the comparison-microscopy test involved comparing 100 questioned hairs with 100 known hairs.

[357] Bisbing, R.E. and M.F. Wolner. "Microscopical Discrimination of Twins' Head Hair." *Journal of Forensic Sciences*, Vol. 29, (1984): 780-786.

[358] The DOJ supporting document describes the results in positive terms: "In the seven tests, one examiners correctly excluded 47 of 52 samples, and a second examiner correctly excluded 49 of 52 samples." It does not specify whether the remaining results are inconclusive results or false positives.

from a defendant's hair, the expert and the prosecution intend the statement to carry weight. Yet, the document goes on to say that no information is available about the proportion of individuals with similar characteristics. As Chapter 4 makes clear, this is scientifically unacceptable. 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. In short, if scientific hair analysis is to *mean* something, there must be actual *empirical evidence* about its meaning.

The second problem with the passage is its implication that there is no relevant empirical evidence about the accuracy of hair analysis. In fact, such evidence was generated by the FBI Laboratory. We turn to this point next.

FBI Study Comparing Microscopic Hair Examination and DNA Analysis

A particularly concerning aspect of the DOJ supporting document is its treatment of the FBI study on hair examination discussed in Chapter 2. In that 2002 study, FBI personnel used mitochondrial DNA analysis to re-examine 170 samples from previous cases in which the FBI Laboratory had performed microscopic hair examination. The authors found that, in 9 of 80 cases (11 percent) in which the FBI Laboratory had found the hairs to be microscopically indistinguishable, the DNA analysis showed that the hairs actually came from *different* individuals.

The 2002 FBI study is a landmark in forensic science because it was the first study to systematically and comprehensively analyze a large collection of previous casework to measure the frequency of false-positive associations. Its conclusion is of enormous importance to forensic science, to police, to courts and to juries: *When hair examiners conclude in casework that two hair samples are microscopically indistinguishable, the hairs often (1 in 9 times) come from different sources.*

Surprisingly, the DOJ document completely ignores this key finding. Instead, it references the FBI study only to support the proposition that DNA analysis "can be used in conjunction with microscopic hair comparison," citing "a 2002 study, which indicated that out of 80 microscopic associations, approximately 88 percent were also included by additional mtDNA testing." The document fails to acknowledge that the remaining cases were found to be false associations—that is, results that, if presented as evidence against a defendant, would mislead a jury about the origins of the hairs.[359]

Conclusion

Our brief review is intended simply to illustrate potential pitfalls in evaluations of the foundational validity and reliability of a method. PCAST is mindful of the constraints that DOJ faces in undertaking scientific evaluations of

[359] In a footnote, the document also takes pains to note that paper cannot be taken to provide an estimate of the *false-positive rate* for microscopic hair comparison, because it contains no data about the number of different-sources comparison that examiners correctly excluded. While this statement is correct, it is misleading—because the paper provides an estimate of a far more important quantity—namely, the frequency of false associations that occurred in actual casework.

the validity and reliability of forensic methods, because critical evaluations by DOJ might be taken as admissions that could be used to challenge past convictions or current prosecutions.

These issues highlight why it is important for evaluations of scientific validity and reliability to be carried out by a science-based agency that is not itself involved in the application of forensic science within the legal system (see Section 6.1).

They also underscore why it is important that *quantitative* information about the reliability of methods (e.g., the frequency of false associations in hair analysis) be stated clearly in expert testimony. We return to this point in Chapter 8, where we consider the DOJ's proposed guidelines, which would bar examiners from providing information about the statistical weight or probability of a conclusion that a questioned hair comes from a particular source.

5.8 Application to Additional Methods

Although we have undertaken detailed evaluations of only six specific methods and included a discussion of a seventh method, the basic analysis can be applied to assess the foundational validity of any forensic feature-comparison method—including traditional forensic disciplines (such as document examination) as well as methods yet to be developed (such as microbiome analysis or internet-browsing patterns).

We note that the evaluation of scientific validity is based on the available scientific evidence at a point in time. Some methods that have not been shown to be foundationally valid may ultimately be found to be reliable—although significant modifications to the methods may be required to achieve this goal. Other methods may not be salvageable—as was the case with compositional bullet lead analysis and is likely the case with bitemarks. Still others may be subsumed by different but more reliable methods, much as DNA analysis has replaced other methods in many instances.

5.9 Conclusion

As the chapter above makes clear, many forensic feature-comparison methods have historically been *assumed* rather than *established* to be foundationally valid based on appropriate empirical evidence. Only within the past decade has the forensic science community begun to recognize the need to empirically *test* whether specific methods meet the scientific criteria for scientific validity. Only in the past five years, for example, have there been appropriate studies that establish the foundational validity and measure the reliability of latent fingerprint analysis. For most subjective methods, there are no appropriate black-box studies with the result that there is no appropriate evidence of foundational validity or estimates of reliability.

The scientific analysis and findings in Chapters 4 and 5 are intended to help focus the relevant actors on *how* to ensure scientific validity, both for existing technologies and for technologies still to be developed.

PCAST expects that some forensic feature-comparison methods may be rejected by courts as inadmissible because they lack adequate evidence of scientific validity. We note that decisions to exclude unreliable methods have historically helped propel major improvements in forensic science—as happened in the early days of DNA evidence—with the result that some methods become established (possibly in revised form) as scientifically valid, while others are discarded.

In the remaining chapters, we offer recommendations on specific actions that could be taken by the Federal Government—including science-based agencies (NIST and OSTP), the FBI Laboratory, the Attorney General, and the Federal judiciary—to ensure the scientific validity and reliability of forensic feature-comparison methods and promote their more rigorous use in the courtroom.

6. Actions to Ensure Scientific Validity in Forensic Science: Recommendations to NIST and OSTP

Based on the scientific findings in Chapters 4 and 5, PCAST has identified actions that we believe should be taken by science-based Federal agencies—specifically, NIST and OSTP—to ensure the scientific validity of forensic feature-comparison methods.

6.1 Role for NIST in Ongoing Evaluation of Foundational Validity

There is an urgent need for ongoing evaluation of the foundational validity of important methods, to provide guidance to the courts, the DOJ, and the forensic science community. Evaluations should be undertaken of both existing methodologies that have not yet met the scientific standards for foundational validity and new methodologies that are being and will be developed in the years ahead. To ensure that the scientific judgments are unbiased and independent, such evaluations must clearly be conducted by a science agency with no stake in the outcome.[360]

This responsibility should be lodged with NIST. NIST is the world's leading metrological laboratory, with a long and distinguished history in the science and technology of measurement. It has tremendous experience in designing and carrying out validation studies, as well as assessing the foundational validity and reliability of laboratory techniques and practices. NIST's mission of advancing measurement science, technology, and standards has expanded from traditional physical measurement standards to respond to many other important societal needs, including those of forensic science, in which NIST has vigorous programs.[361] As described above, NIST has begun to lead a number of important efforts to strengthen the forensic sciences, including its roles with respect to NCFS and OSAC.

PCAST recommends that NIST be tasked with responsibility for preparing an annual report evaluating the foundational validity of key forensic feature-comparison methods, based on available, published empirical studies. These evaluations should be conducted under the auspices of NIST, with input from additional expertise as deemed necessary from experts outside forensic science, and overseen by an appropriate review panel. The reports should, as a minimum, produce assessments along the lines of those in this report, updated as appropriate. Our intention is not that NIST have a formal regulatory role with respect to forensic science, but rather that NIST's evaluations help inform courts, the DOJ, and the forensic science community.

[360] For example, agencies that apply forensic feature-comparison methods within the legal system have a clear stake in the outcome of such evaluations.

[361] See: www.nist.gov/forensics.

We do not expect NIST to take responsibility for *conducting* the necessary validation studies. However, NIST should advise on the design and execution of such studies. NIST could carry out some studies through its own intramural research program and through CSAFE. However, the majority of studies will likely be conducted by other groups—such as NSF's planned Industry/University Cooperative Research Centers; the FBI Laboratory; the U.S. national laboratories; other Federal agencies; state laboratories; and academic researchers.

We note that the NCFS has recently endorsed the need for independent scientific review of forensic science methods. A Views Document overwhelmingly approved by the commission in June 2016 stated that, "All forensic science methodologies should be evaluated by an independent scientific body to characterize their capabilities and limitations in order to accurately and reliably answer a specific and clearly defined forensic question" and that "The National Institute of Standards and Technology (NIST) should assume the role of independent scientific evaluator within the justice system for this purpose."[362]

Finally, we believe that the state of forensic science would be improved if papers on the foundational validity of forensic feature-comparison methods were published in leading scientific journals rather than in forensic-science journals, where, owing to weaknesses in the research culture of the forensic science community discussed in this report, the standards for peer review are less rigorous. Commendably, FBI scientists published its black-box study of latent fingerprints in the *Proceedings of the National Academy of Sciences*. We suggest that NIST explore with one or more leading scientific journals the possibility of creating a process for rigorous review and online publication of important studies of foundational validity in forensic science. Appropriate journals could include *Metrologia*, a leading international journal in pure and applied metrology, and the *Proceedings of the National Academy of Sciences*.

6.2 Accelerating the Development of Objective Methods

As described throughout the report, objective methods are generally preferable to subjective methods. The reasons include greater accuracy, greater efficiency, lower risk of human error, lower risk of cognitive bias, and greater ease of establishing foundational validity and estimating reliability. Where possible, vigorous efforts should be undertaken to transform subjective methods into objective methods.

Two forensic feature-comparison methods—latent fingerprint analysis and firearms analysis—are ripe for such transformation. As discussed in the previous chapter, there are strong reasons to believe that both methods can be made objective through automated image analysis. In addition, DNA analysis of complex mixtures has recently been converted into a foundationally valid objective method for a limited range of mixtures, but additional work will be needed to expand the limits of the range.

NIST, in conjunction with the FBI Laboratory, should play a leadership role in propelling this transformation by (1) the creation and dissemination of large datasets to support the development and testing of methods by both

[362] Views of the Commission: Technical Merit Evaluation of Forensic Science Methods and Practices. www.justice.gov/ncfs/file/881796/download.

companies and academic researchers, (2) grant and contract support, and (3) sponsoring processes, such as prize competitions, to evaluate methods.

6.3 Improving the Organization for Scientific Area Committees

The creation by NIST of OSAC was an important step in strengthening forensic science practice. The organizational design—which houses all of the subject area communities under one structure and encourages cross-disciplinary communication and coordination—is a significant improvement over the previous Scientific Working Groups (SWGs), which functioned less formally as stand-alone committees.

However, initial lessons from its first years of operation have revealed some important shortcomings. OSAC's membership includes relatively few independent scientists: it is dominated by forensic professionals, who make up more than two-thirds of its members. Similarly, it has few independent statisticians: while virtually all of the standards and guidelines evaluated by this body need consideration of statistical principles, OSAC's 600 members include only 14 statisticians spread across all four Science Area Committees and 23 subcommittees.

Restructuring

PCAST concludes that OSAC lacks sufficient independent scientific expertise and oversight to overcome the serious flaws in forensic science. Some restructuring is necessary to ensure that independent scientists and statisticians have a greater voice in the standards development process, a requirement for meaningful scientific validity. Most importantly, OSAC should have a formal committee—a Metrology Resource Committee—at the level of the other three Resource Committees (the Legal Resource Committee, the Human Factors Committee, and the Quality Infrastructure Committee). This Committee should be composed of laboratory scientists and statisticians from outside the forensic science community and charged with reviewing each standard and guideline that is recommended for registry approval by the Science Area Committees before it is sent for final review the Forensic Science Standards Board (FSSB).

Availability of OSAC Standards

OSAC is not a formal standard-setting body. It reviews and evaluates standards relevant to forensic science developed by standards developing organizations such as ASTM International, the National Fire Protection Association (NFPA) and the International Organization for Standardization (ISO) for inclusion on the OSAC Registries of Standards and Guidelines. The OSAC evaluation process includes a public comment period. OSAC, working with the standards developers, has arranged for the content of standards under consideration to be accessible to the public during the public comment period. Once approved by OSAC, a standard is listed, by title, on a public registry maintained by NIST. It is customary for some standards developing organization, including ASTM International, to charge a fee for a licensed copy of each copyrighted standard and to restrict users from distributing these standards.[363,364]

[363] For a list of ASTM's forensic science standards, see: www.astm.org/DIGITAL_LIBRARY/COMMIT/PAGES/E30.htm.

[364] The American Academy of Forensic Sciences (AAFS) will also become an accredited Standards Developing Organization (SDO) and could, in the future, develop standards for review and listing by OSAC.

NIST recently negotiated a licensing agreement with ASTM International that, for a fee, allows federal, state and local government employees online access to ASTM Committee E30 standards.[365] However, this list does not include indigent defendants, private defense attorneys, or large swaths of the academic research community. At present, contracts have been negotiated with the other SDOs that have standards currently under review by the OSAC. PCAST believes it is important that standards intended for use in the criminal justice system are widely available to all who may need access. It is important that the standards be readily available to defendants and to external observers, who have an important role to play in ensuring quality in criminal justice.[366]

NIST should ensure that the content of OSAC-registered standards and guidelines are freely available to any party that may desire them in connection with a legal case or for evaluation and research, including by aligning with the policies related to reasonable availability of standards in the Office of Management and Budget Circular A-119, Federal Participation in the Development and Use of Voluntary Consensus Standards and Conformity Assessment Activities and the Office of the Federal Register, IBR (incorporation by reference) Handbook.

6.4 Need for an R&D Strategy for Forensic Science

The 2009 NRC report found that there is an urgent need to strengthen forensic science, noting that, "Forensic science research is not well supported, and there is no unified strategy for developing a forensic science research plan across federal agencies."[367]

It is especially important to create and support a vibrant academic research community rooted in the scientific culture of universities. This will require significant funding to support academic research groups, but will pay big dividends in driving quality and innovation in both existing and entirely new methods.

Both NIST and NSF have recently taken initial steps to help bridge the significant gaps between the forensic practitioner and academic research communities through multi-disciplinary research centers. These centers promise to engage the broader research community in advancing forensic science and create needed links between the forensic science community and a broad base of research universities and could help drive forward critical foundational research.

Nonetheless, as noted in Chapter 2, the total level of Federal funding by NIJ, NIST, and NSF to the academic community for fundamental research in forensic science is extremely small. Substantially larger funding will be needed to develop a robust research community and to support the development and evaluation of promising new technologies.

[365] According to the revised contract, ASTM will provide unlimited web-based access for all ASTM committee E30 Forensic Science Standards to: OSAC members and affiliates; NIST and Federal/State/Local Crime Laboratories; Public Defenders Offices; Law Enforcement Agencies; Prosecutor Offices; and Medical Examiner/and Coroners Offices.

[366] PCAST expresses no opinion about the appropriateness of paywalls for standards in areas other than criminal justice.

[367] National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (2009): 78.

Federal R&D efforts in forensic science, both intramural and extramural, need to be better coordinated. No one agency has lead responsibility for ensuring that the forensic sciences are adequately supported. Greater coordination is needed across the relevant Federal agencies and laboratories to ensure that funding is directed to the highest priorities and that work is of high quality.

OSTP should convene relevant Federal agencies, laboratories, and stakeholders to develop a national research strategy and 5-year plan to ensure that foundational research in support of the forensic sciences is well-coordinated, solidify Federal agency commitments made to date, and galvanize further action and funding that could be taken to encourage additional foundational research, improve current forensic methods, support the creation of new research databases, and oversee the regular review and prioritization of research.

6.5 Recommendations

Based on its scientific findings, PCAST makes the following recommendations.

Recommendation 1. Assessment of foundational validity

It is important that scientific evaluations of the foundational validity be conducted, on an ongoing basis, to assess the foundational validity of current and newly developed forensic feature-comparison technologies. To ensure the scientific judgments are unbiased and independent, such evaluations must be conducted by a science agency which has no stake in the outcome.

(A) The National Institute of Standards and Technology (NIST) should perform such evaluations and should issue an annual public report evaluating the foundational validity of key forensic feature-comparison methods.

(i) The evaluations should (a) assess whether each method reviewed has been adequately defined and whether its foundational validity has been adequately established and its level of accuracy estimated based on empirical evidence; (b) be based on studies published in the scientific literature by the laboratories and agencies in the U.S. and in other countries, as well as any work conducted by NIST's own staff and grantees; (c) as a minimum, produce assessments along the lines of those in this report, updated as appropriate; and (d) be conducted under the auspices of NIST, with additional expertise as deemed necessary from experts outside forensic science.

(ii) NIST should establish an advisory committee of experimental and statistical scientists from outside the forensic science community to provide advice concerning the evaluations and to ensure that they are rigorous and independent. The members of the advisory committee should be selected jointly by NIST and the Office of Science and Technology Policy.

(iii) NIST should prioritize forensic feature-comparison methods that are most in need of evaluation, including those currently in use and in late-stage development, based on input from the Department of Justice and the scientific community.

(iv) Where NIST assesses that a method has been established as foundationally valid, it should (a) indicate appropriate estimates of error rates based on foundational studies and (b) identify any issues relevant to validity as applied.

(v) Where NIST assesses that a method has not been established as foundationally valid, it should suggest what steps, if any, could be taken to establish the method's validity.

(vi) NIST should not have regulatory responsibilities with respect to forensic science.

(vii) NIST should encourage one or more leading scientific journals outside the forensic community to develop mechanisms to promote the rigorous peer review and publication of papers addressing the foundational validity of forensic feature-comparison methods.

(B) The President should request and Congress should provide increased appropriations to NIST of (a) $4 million to support the evaluation activities described above and (b) $10 million to support increased research activities in forensic science, including on complex DNA mixtures, latent fingerprints, voice/speaker recognition, and face/iris biometrics.

Recommendation 2. Development of objective methods for DNA analysis of complex mixture samples, latent fingerprint analysis, and firearms analysis

The National Institute of Standards and Technology (NIST) should take a leadership role in transforming three important feature-comparison methods that are currently subjective—latent fingerprint analysis, firearms analysis, and, under some circumstances, DNA analysis of complex mixtures—into objective methods.

(A) NIST should coordinate these efforts with the Federal Bureau of Investigation Laboratory, the Defense Forensic Science Center, the National Institute of Justice, and other relevant agencies.

(B) These efforts should include (i) the creation and dissemination of large datasets and test materials (such as complex DNA mixtures) to support the development and testing of methods by both companies and academic researchers, (ii) grant and contract support, and (iii) sponsoring processes, such as prize competitions, to evaluate methods.

Recommendation 3. Improving the Organization for Scientific Area Committees process

(A) The National Institute of Standards and Technology (NIST) should improve the Organization for Scientific Area Committees (OSAC), which was established to develop and promulgate standards and guidelines to improve best practices in the forensic science community.

(i) NIST should establish a Metrology Resource Committee, composed of metrologists, statisticians, and other scientists from outside the forensic science community. A representative of the Metrology Resource Committee should serve on each of the Scientific Area Committees (SACs) to provide direct guidance on the application of measurement and statistical principles to the developing documentary standards.

(ii) The Metrology Resource Committee, as a whole, should review and publically approve or disapprove all standards proposed by the Scientific Area Committees before they are transmitted to the Forensic Science Standards Board.

(B) NIST should ensure that the content of OSAC-registered standards and guidelines are freely available to any party that may desire them in connection with a legal case or for evaluation and research, including by aligning with the policies related to reasonable availability of standards in the Office of Management and Budget Circular A-119, Federal Participation in the Development and Use of Voluntary Consensus Standards and Conformity Assessment Activities and the Office of the Federal Register, IBR (incorporation by reference) Handbook.

Recommendation 4. R&D strategy for forensic science

(A) The Office of Science and Technology Policy (OSTP) should coordinate the creation of a national forensic science research and development strategy. The strategy should address plans and funding needs for:

(i) major expansion and strengthening of the academic research community working on forensic sciences, including substantially increased funding for both research and training;

(ii) studies of foundational validity of forensic feature-comparison methods;

(iii) improvement of current forensic methods, including converting subjective methods into objective methods, and development of new forensic methods;

(iv) development of forensic feature databases, with adequate privacy protections, that can be used in research;

(v) bridging the gap between research scientists and forensic practitioners; and

(vi) oversight and regular review of forensic science research.

(B) In preparing the strategy, OSTP should seek input from appropriate Federal agencies, including especially the Department of Justice, Department of Defense, National Science Foundation, and National Institute of Standards and Technology; Federal and State forensic science practitioners; forensic science and non-forensic science researchers; and other stakeholders.

7. Actions to Ensure Scientific Validity in Forensic Science: Recommendation to the FBI Laboratory

Based on the scientific findings in Chapters 4 and 5, PCAST has identified actions that we believe should be taken by the FBI Laboratory to ensure the scientific validity of forensic feature-comparison methods.

We note that the FBI Laboratory has played an important role in recent years in undertaking high-quality scientific studies of latent fingerprint analysis. PCAST applauds these efforts and urges the FBI Laboratory to expand them.

7.1 Role for FBI Laboratory

The FBI Laboratory is a full-service, state-of-the-art facility that works to apply cutting-edge science to solve cases and prevent crime. Its mission is to apply scientific capabilities and technical services to the collection, processing, and exploitation of evidence for the Laboratory and other duly constituted law enforcement and intelligence agencies in support of investigative and intelligence priorities. Currently, the Laboratory employs approximately 750 employees and over 300 contractors to meet the broad scope of this mission.

Laboratory Capabilities and Services

The FBI has specialized capabilities and personnel to respond to incidents, collect evidence in their field, carry out forensic analyses, and provide expert witness testimony. The FBI Laboratory supports Evidence Response Teams in all 56 FBI field offices and has personnel who specialize in hazardous evidence and crime scene documentation and data collection. The Laboratory is responsible for training and supplying these response activities for FBI personnel across the U.S.[368] The Laboratory also manages the Terrorist Explosive Device Analytical Center (TEDAC), which received nearly 1,000 evidence submissions in FY 2015 and disseminated over 2,000 intelligence products.

The FBI Laboratory employs forensic examiners to carry out analyses in a range of disciplines, including chemistry, cryptanalysis, DNA, firearms and toolmarks, latent prints, questioned documents, and trace evidence. The FBI Laboratory received over 3875 evidence submissions and authored over 4850 laboratory reports in FY 2015. In addition to carrying out casework for federal cases, the Laboratory provides support to state and local laboratories and carries out testing in state and local cases for some disciplines.

[368] The FBI Laboratory supported 162 deployments and 168 response exercises, as well as delivering 239 training courses in FY 2015.

Research and Development Activities

In addition to its services, the FBI Laboratory carries out important research and development activities. The activities are critical for providing the Laboratory with the most advanced tools for advancing its mission. A strong research program and culture is also important to the Laboratory's ability to maintain excellence and to attract and retain highly qualified personnel.

Due to the expansive scope and many requirements on its operations, only about five percent of the FBI Laboratory's annual $100 million budget is available for research and development activities.[369] The R&D budget is stretched across a number of applied research activities, including validation studies (for new methods or commercial products, such as new DNA analyzers). For its internal research activities, the Laboratory relies heavily on its Visiting Scientist Program, which brings approximately 25 post docs, master's students, and bachelor's degree students into the laboratory each year. The Laboratory has worked to partner with other government agencies to provide more resources to its research priorities as a composite initiative, and has also been able to stretch available budgets by performing critical research studies incrementally over several years.

The FBI Laboratory's series of studies in latent print examination is an example of important foundational research that it was able to carry out incrementally over a five-year period. The work includes "black box" studies that evaluate the accuracy and reliability of latent print examiners' conclusions, as well as "white box" studies to evaluate how the quality and quantity of features relate to latent print examiners' decisions. These studies have resulted in a series of important publications that have helped to quantify error rates for the community of practice and assess the repeatability and reproducibility of latent fingerprint examiners' decisions. Indeed, PCAST's judgment that latent fingerprint analysis is foundationally valid rests heavily on the FBI black-box study. Similar lines of research are being pursued in some other disciplines, including firearms examination and questioned documents.

Unfortunately, the limited funding available for these studies—and for the intramural research program more generally—has hampered progress in testing the foundational validity of forensic science methods and in strengthening the forensic sciences. PCAST believes that the budget for the FBI Laboratory should be significantly increased, and targeted so as allow the R&D budget to be increased to a total of $20 million.

Access to databases

The FBI also has an important role to play in encouraging research by external scientists, by facilitating access, under appropriate conditions, to large forensic databases. Most of the databases routinely used in forensic analysis are not accessible for use by researchers, and the lack of access hampers progress in improving forensic science. For example, ballistic database systems such as the Bureau of Alcohol, Tobacco, Firearms and Explosives' National Integrated Ballistic Information System (NIBIN), which is searched by firearms examiners seeking to identify a firearm or cartridge case, cannot be assessed to study its completeness, relevance or

[369] In 2014, the FBI Laboratory spent $10.9 million on forensic science research and development, with roughly half from its own budget and half from grants from NIST and the Department of Homeland Security. See: National Academies of Sciences, Engineering, and Medicine. *Support for Forensic Science Research: Improving the Scientific Role of the National Institute of Justice***.** The National Academies Press. Washington DC. (2015): p. 31.

quality, and the search algorithm that is used to identify potential matches cannot be evaluated. The NGI (formerly IAFIS)[370] system that currently houses more than 70 million fingerprint entries would dramatically expand the data available for study; currently, there exists only one publicly available fingerprint database, consisting of 258 latent print-10 print pairs.[371] And, the FBI's NDIS system, which currently houses more than 14 million offender and arrestee DNA profiles. NIST has developed an inventory of all of the forensic databases that are heavily used by law enforcement and forensic scientists, with information as to their accessibility.

Substantial efforts are needed to make existing forensic databases more accessible to the research community, subject to appropriate protection of privacy, such as removal of personally identifiable information and data-use restrictions.

For some disciplines, such as firearms analysis and treadmarks, there are no significant privacy concerns.

For latent prints, privacy concerns might be ameliorated in variety of ways. For example, one might avoid the issue by (1) generating large collections of known-latent print pairs with varying quality and quantity of information through the touching and handling of natural items in a wide variety of circumstances (surfaces, pressure, distortion, etc.), (2) using software to automatically generate the "morphing transformations" from the known prints and the latent prints, and (3) applying these transformations to prints from deceased individuals to create millions of latent-known print pairs.[372]

For DNA, protocols have been developed in human genomic research, which poses similar or greater privacy concerns, to allow access to bona fide researchers.[373] Such policies should be feasible for forensic DNA databases as well. We note that the law that authorizes the FBI to maintain a national forensic DNA database explicitly contemplates allowing access to DNA samples and DNA analyses "if personally identifiable information is removed . . . for identification research and protocol development purposes."[374] Although the law does not contain an explicit statement on this point, DOJ interprets the law as allowing use for this purpose only by criminal justice agencies. It is reluctant, in the absence of statutory clarification, to provide even controlled access to other researchers. This topic deserves attention.

PCAST believes that the availability of data will speed the development of methods, tools, and software that will improve forensic science. For databases under its control, the FBI Laboratory should develop programs to make forensic databases (or subsets of those databases) accessible to researchers under conditions that protect

[370] NGI standards for "Next Generation Identification" and combines multiple biometric information systems, including IAFIS, iris and face recognition systems, and others.

[371] NIST Special Database 27A, available at: www.nist.gov/itl/iad/image-group/nist-special-database-27a-sd-27a.

[372] Medical examiners offices routinely collect fingerprints from deceased individuals as part of the autopsy process; these fingerprints could be collected and used to create a large database for research purposes.

[373] A number of models that have been developed in the biomedical research context that allow for tiered access to sensitive data while providing adequate privacy protection could be employed here. Researchers could be required to sign Non-Disclosure Agreements (NDAs) or enter into limited use agreements. Researchers could be required to access the data on site, so that data cannot be downloaded or shared, or could be permitted to download only aggregated or summary data.

[374] Federal DNA Identification Act, 42 U.S.C. §14132(b)(3)(D)).

privacy. For databases owned by others, the FBI Laboratory and NIST should each work with other agencies and companies that control the databases to develop programs providing appropriate access.

7.2 Recommendation

Based on its scientific findings, PCAST makes the following recommendation.

Recommendation 5. Expanded forensic-science agenda at the Federal Bureau of Investigation Laboratory

(A) ***Research programs.*** **The Federal Bureau of Investigation (FBI) Laboratory should undertake a vigorous research program to improve forensic science, building on its recent important work on latent fingerprint analysis.** The program should include:

> (i) conducting studies on the reliability of feature-comparison methods, in conjunction with independent third parties without a stake in the outcome;
>
> (ii) developing new approaches to improve reliability of feature-comparison methods;
>
> (iii) expanding collaborative programs with external scientists; and
>
> (iv) ensuring that external scientists have appropriate access to datasets and sample collections, so that they can carry out independent studies.

(B) ***Black-box studies***. **Drawing on its expertise in forensic science research, the FBI Laboratory should assist in the design and execution of additional black-box studies for subjective methods, including for latent fingerprint analysis and firearms analysis.** These studies should be conducted by or in conjunction with independent third parties with no stake in the outcome.

(C) ***Development of objective methods.*** **The FBI Laboratory should work with the National Institute of Standards and Technology to transform three important feature-comparison methods that are currently subjective—latent fingerprint analysis, firearm analysis, and, under some circumstances, DNA analysis of complex mixtures—into objective methods.** These efforts should include (i) the creation and dissemination of large datasets to support the development and testing of methods by both companies and academic researchers, (ii) grant and contract support, and (iii) sponsoring prize competitions to evaluate methods.

(D) ***Proficiency testing.*** **The FBI Laboratory, should promote increased rigor in proficiency testing by (i) within the next four years, instituting routine blind proficiency testing within the flow of casework in its own laboratory, (ii) assisting other Federal, State, and local laboratories in doing so as well, and (iii) encouraging routine access to and evaluation of the tests used in commercial proficiency testing.**

(E) *Latent fingerprint analysis*. The FBI Laboratory should vigorously promote the adoption, by all laboratories that perform latent fingerprint analysis, of rules requiring a "linear Analysis, Comparison, Evaluation" process—whereby examiners must complete and document their analysis of a latent fingerprint *before* looking at any known fingerprint and should separately document any additional data used during comparison and evaluation.

(F) *Transparency concerning quality issues in casework*. The FBI Laboratory, as well as other Federal forensic laboratories, should regularly and publicly report quality issues in casework (in a manner similar to the practices employed by the Netherlands Forensic Institute, described in Chapter 5), as a means to improve quality and promote transparency.

(G) *Budget*. The President should request and Congress should provide increased appropriations to the FBI to restore the FBI Laboratory's budget for forensic science research activities from its current level to $30 million and should evaluate the need for increased funding for other forensic-science research activities in the Department of Justice.

8. Actions to Ensure Scientific Validity in Forensic Science: Recommendations to the Attorney General

Based on the scientific findings in Chapters 4 and 5, PCAST has identified actions that we believe should be taken by the Attorney General to ensure the scientific validity of forensic feature-comparison methods and promote their more rigorous use in the courtroom.

8.1 Ensuring the Use of Scientifically Valid Methods in Prosecutions

The Federal Government has a deep commitment to ensuring that criminal prosecutions are not only fair in their process, but correct in their outcome—that is, that guilty individuals are convicted, while innocent individuals are not.

Toward this end, the DOJ should ensure that testimony about forensic evidence presented in court is scientifically valid. This report provides guidance to DOJ concerning the scientific criteria for both foundational validity and validity as applied, as well as evaluations of six specific forensic methods and a discussion of a seventh. Over the long term, DOJ should look to ongoing evaluations of forensic methods that should be performed by NIST (as described in Chapter 6).

In the interim, DOJ should undertake a review of forensic feature-comparison methods (beyond those reviewed in this report) to identify which methods used by DOJ lack appropriate black-box studies necessary to assess foundational validity. Because such subjective methods are presumptively not established to be foundationally valid, DOJ should evaluate (1) whether DOJ should present in court conclusions based on such methods and (2) whether black-box studies should be launched to evaluate those methods.

8.2 Revision of DOJ Recently Proposed Guidelines on Expert Testimony

On June 3, 2016, the DOJ released for comment a first set of proposed guidelines, together with supporting documents, on “Proposed Uniform Language for Testimony and Reports” on several forensic sciences, including latent fingerprint analysis and forensic footwear and tire impression analysis.[375] On July 21, 2016, the DOJ released for comment a second set of proposed guidelines and supporting documents for several additional forensic sciences, including microscopic hair analysis, certain types of DNA analysis, and other fields.

[375] See: www.justice.gov/dag/proposed-language-regarding-expert-testimony-and-lab-reports-forensic-science. A second set of proposed guidelines was released on July 21, 2016 including hair analysis and mitochondrial DNA and Y chromosome typing (www.justice.gov/dag/proposed-uniform-language-documents-anthropology-explosive-chemistry-explosive-devices-geology).

The guidelines represent an important step forward, because they instruct DOJ examiners not to make sweeping claims that they can identify the source of a fingerprint or footprint to the exclusion of all other possible sources. PCAST applauds DOJ's intention and efforts to bring uniformity and to prevent inaccurate testimony concerning feature comparisons.

Some aspects of the guidelines, however, are not scientifically appropriate and embody heterodox views of the kind discussed in Section 4.7. As an illustration, we focus on the guidelines for footwear and tire impression analysis and the guidelines for hair analysis.

Footwear and Tire Impression Analysis

Relevant portions of the guidelines for testimony and reports about forensic footwear and tire impression are shown in Box 6.

BOX 6. Excerpt from DOJ Proposed uniform language for testimony and reports for the forensic footwear and tire impression discipline[376]

Statements Approved for Use in Laboratory Reports and Expert Witness Testimony Regarding Forensic Examination of Footwear and Tire Impression Evidence

Identification

1. The examiner may state that it is his/her opinion that the shoe/tire is the source of the impression because there is sufficient quality and quantity of corresponding features such that the examiner would not expect to find that same combination of features repeated in another source. This is the highest degree of association between a questioned impression and a known source. This opinion requires that the questioned impression and the known source correspond in class characteristics and also share one or more randomly acquired characteristics. This opinion acknowledges that an identification to the exclusion of all others can never be empirically proven.

Statements Not Approved for Use in Laboratory Reports and Expert Witness Testimony Regarding Forensic Examination of Footwear and Tire Impression Evidence

Exclusion of All of Others

1. The examiner may not state that a shoe/tire is the source of a questioned impression to the exclusion of all other shoes/tires because all other shoes/tires have not been examined. Examining all of the shoes/tires in the world is a practical impossibility.

[376] See: www.justice.gov/olp/file/861936/download.

Error Rate

2. The examiner may not state a numerical value or percentage regarding the error rate associated with either the methodology used to conduct the examinations or the examiner who conducted the analyses.

Statistical Weight

3. The examiner may not state a numerical value or probability associated with his/her opinion. Accurate and reliable data and/or statistical models do not currently exist for making quantitative determinations regarding the forensic examination of footwear/tire impression evidence.

These proposed guidelines have serious problems.

An examiner may opine that a shoe is the source of an impression, but not that the shoe is the source of impression *to the exclusion of all other possible shoes*. But, as a matter of logic, there is no difference between these two statements. If an examiner believes that X is the source of Y, then he or she necessarily believes that *nothing else* is the source of Y. Any sensible juror should understand this equivalence.

What then is the goal of the guidelines? It appears to be to acknowledge the possibility of error. In effect, examiners should say, “I believe X is the source of Y, although I could be wrong about that.”

This is appropriate. But, the critical question is then: How likely is it that the examiner is wrong?

There’s the rub: the guidelines bar the examiner from discussing the likelihood of error, because there is no accurate or reliable information about accuracy. In effect, examiners are instructed to say, “I believe X is the source of Y, although I could be wrong about that. But, I have no idea how often I’m wrong because we have no reliable information about that.”

Such a statement does not meet any plausible test of scientific validity. As Judge Easterly wrote in *Williams v. United States*, a claim of identification under such circumstances:

> *has the same probative value as the vision of a psychic: it reflects nothing more than the individual’s foundationless faith in what he believes to be true. This is not evidence on which we can in good conscience rely, particularly in criminal cases, where we demand proof—real proof—beyond a reasonable doubt, precisely because the stakes are so high.* [377]

[377] *Williams v. United States,* DC Court of Appeals, Decided January 21, 2016, (Easterly, concurring). We cite the analogy for its expositional value concerning the scientific point; we express no position on the role of the case as legal authority.

Hair Analysis

Relevant portions of the guidelines for testimony and reports on forensic hair examination are shown in Box 7.

BOX 7. Excerpt from DOJ Proposed uniform language for testimony and reports for the forensic hair examination discipline[378]

Statements Not Approved for Use in Forensic Hair Examination Testimony and/or Laboratory Reports

Human Hair Comparisons

1. The examiner may state or imply that the questioned human hair is microscopically consistent with the known hair sample and accordingly, the source of the known hair sample can be included as a possible source of the questioned hair.

Statements Not Approved for Use in Forensic Hair Examination Testimony and/or Laboratory Reports

Individualization

1. The examiner may not state or imply that a hair came from a particular source to the exclusion of all others.

Statistical Weight

2. The examiner may not state or imply a statistical weight or probability to a conclusion or provide a likelihood that the questioned hair originated from a particular source.

Zero Error Rate

3. The examiner may not state or imply that the method used in performing microscopic hair examinations has a zero error rate or is infallible.

The guidelines appropriately state that examiners may not claim that they can individualize the source of a hair nor that they have a zero error rate. However, while examiners may "state or imply that the questioned human hair is microscopically consistent with the known hair sample and accordingly, the source of the known hair sample can be included as a possible source of the questioned hair," they are barred from providing accurate information about the reliability of such conclusions. This is contrary to the scientific requirement that forensic feature-comparison methods must be supported by and accompanied by appropriate empirical estimates of reliability.

In particular, as discussed in Section 5.7, a landmark study in 2002 by scientists at the FBI Laboratory showed that, among 80 instances in actual casework where examiners concluded that a questioned hair was microscopically consistent with the known hair sample, the hair were found by DNA analysis to have come from

[378] Department of Justice Proposed Uniform Language for Testimony and Reports for the Forensic Hair Examination Discipline, available at: www.justice.gov/dag/file/877736/download.

a different source in 11 percent of cases. The fact that such a significant proportion of conclusions were false associations is of tremendous importance in interpreting conclusions of hair examiners.

In cases of hair examination unaccompanied by DNA analysis, examiners should be required to disclose the high frequency of false associations seen in the FBI study so that juries can appropriately weigh conclusions.

Conclusion

The DOJ should revise the proposed guidelines, to bring them into alignment with scientific standards for scientific validity. The supporting documentation should also be revised, as discussed in Section 5.7.

8.3 Recommendations

Based on its scientific findings, PCAST makes the following recommendations.

Recommendation 6. Use of feature-comparison methods in Federal prosecutions

(A) The Attorney General should direct attorneys appearing on behalf of the Department of Justice (DOJ) to ensure expert testimony in court about forensic feature-comparison methods meets the scientific standards for scientific validity.

While pretrial investigations may draw on a wider range of methods, expert testimony in court about forensic feature-comparison methods in criminal cases—which can be highly influential and has led to many wrongful convictions—must meet a higher standard. In particular, attorneys appearing on behalf of the DOJ should ensure that:

> (i) the forensic feature-comparison methods upon which testimony is based have been established to be foundationally valid, as shown by appropriate empirical studies and consistency with evaluations by the National Institute of Standards and Technology (NIST), where available; and
>
> (ii) the testimony is scientifically valid, with the expert's statements concerning the accuracy of methods and the probative value of proposed identifications being constrained by the empirically supported evidence and not implying a higher degree of certainty.

(B) DOJ should undertake an initial review, with assistance from NIST, of subjective feature-comparison methods used by DOJ to identify which methods (beyond those reviewed in this report) lack appropriate black-box studies necessary to assess foundational validity. Because such subjective methods are presumptively not established to be foundationally valid, DOJ should evaluate whether it is appropriate to present in court conclusions based on such methods.

(C) Where relevant methods have not yet been established to be foundationally valid, DOJ should encourage and provide support for appropriate black-box studies to assess foundational validity and measure reliability. The design and execution of these studies should be conducted by or in conjunction with independent third parties with no stake in the outcome.

Recommendation 7. Department of Justice guidelines on expert testimony

(A) The Attorney General should revise and reissue for public comment the Department of Justice's (DOJ) proposed "Uniform Language for Testimony and Reports" and supporting documents to bring them into alignment with scientific standards for scientific validity.

(B) The Attorney General should issue instructions directing that:

(i) Where empirical studies and/or statistical models exist to shed light on the accuracy of a forensic feature-comparison method, an examiner should provide quantitative information about error rates, in accordance with guidelines to be established by DOJ and the National Institute of Standards and Technology, based on advice from the scientific community.

(ii) Where there are not adequate empirical studies and/or statistical models to provide meaningful information about the accuracy of a forensic feature-comparison method, DOJ attorneys and examiners should not offer testimony based on the method. If it is necessary to provide testimony concerning the method, they should clearly acknowledge to courts the lack of such evidence.

(iii) In testimony, examiners should always state clearly that errors can and do occur, due both to similarities between features and to human mistakes in the laboratory.

9. Actions to Ensure Scientific Validity in Forensic Science: Recommendations to the Judiciary

Based on the scientific findings in Chapters 4 and 5, PCAST has identified actions that we believe should be taken by the judiciary to ensure the scientific validity of evidence based on forensic feature-comparison methods and promote their more rigorous use in the courtroom.

9.1 Scientific Validity as a Foundation for Expert Testimony

In Federal courts, judges are assigned the critical role of "gatekeepers" charged with ensuring that expert testimony "rests on a reliable foundation."[379] Specifically, Rule 702 (c,d) of the Federal Rules of Evidence requires that (1) expert testimony must be the product of "reliable principles and methods" and (2) experts must have "reliably applied" the methods to the facts of the case.[380] The Supreme Court has stated that judges must determine "whether the reasoning or methodology underlying the testimony is scientifically valid."[381]

As discussed in Chapter 3, this framework establishes an important conversation between the judiciary and the scientific community. The admissibility of expert testimony depends on a threshold test of whether it meets certain *legal* standards for evidentiary reliability, which are exclusively the province of the judiciary. Yet, in cases involving scientific evidence, these legal standards are to be "based upon scientific validity."[382]

PCAST does not opine on the legal standards, but aims in this report to clarify the *scientific* standards that underlie them. To ensure that the distinction between scientific and legal concepts is clear, we have adopted specific terms to refer to *scientific* concepts (*foundational validity* and *validity as applied*) intended to parallel *legal* concepts expressed in Rule 702 (c,d).

As the Supreme Court has noted, the judge's inquiry under Rule 702 is a flexible one: there is no simple one-size-fits-all test that can be applied uniformly to all scientific disciplines.[383] Rather, the evaluation of scientific validity should be based on the appropriate scientific criteria for the scientific field. Moreover, the appropriate scientific field should be the larger scientific discipline to which it belongs.[384]

[379] *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993) at 597.
[380] See: www.uscourts.gov/file/rules-evidence.
[381] *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993) at 592.
[382] *Daubert*, at FN9 ("in a case involving scientific evidence, *evidentiary reliability* will be based on *scientific validity*." [emphasis in original]).
[383] *Daubert*, at 594.
[384] For example, in *Frye*, the court evaluated whether a proffered lie detector had gained "standing and scientific recognition among physiological and psychological authorities," rather than among lie detector experts. *Frye v. United*

In this report, PCAST has focused on forensic feature-comparison methods—which belong to the field of metrology, the science of measurement and its application.[385] We have sought—in a form usable by courts, as well as by scientists and others who seek to improve forensic science—to lay out the scientific criteria for foundational validity and validity as applied (Chapter 4) and to illustrate their application to specific forensic feature-comparison methods (Chapter 5).

The scientific criteria are described in Finding 1. PCAST's conclusions can be summarized as follows:

> *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 personal experience nor professional practices—can substitute for adequate empirical demonstration of accuracy.*

The applications to specific feature-comparison methods are described in Findings 2-7. The full set of scientific findings is collected in Chapter 10.

Finally, we note that the Supreme Court in *Daubert* suggested that judges should be mindful of Rule 706, which allows a court at its discretion to procure the assistance of an expert of its own choosing.[386] Such experts can provide independent assessments concerning, among other things, the validity of scientific methods and their applications.

9.2 Role of Past Precedent

One important issue that arose throughout our deliberations was the role of past precedents.

As discussed in Chapter 5, our scientific review found that most forensic feature-comparison methods (with the notable exception of DNA analysis of single-source and simple-mixture samples) have historically been *assumed* rather than *established* to be foundationally valid. Only after it became clear in recent years (based on DNA and other analysis) that there are fundamental problems with the reliability of some of these methods has the forensic science community begun to recognize the need to *empirically test* whether specific methods meet the scientific criteria for scientific validity.

This creates an obvious tension, because many courts admit forensic feature-comparison methods based on longstanding precedents that were set before these fundamental problems were discovered.

States, 293 F. 1013 (D.C. Cir. 1923). Similarly, the fact that bitemark examiners believe that bitemark examination is valid carries little weight.

[385] See footnote 93 on p.44.

[386] *Daubert*, at 595.

From a purely *scientific* standpoint, the resolution is clear. When new facts falsify old assumptions, courts should not be obliged to defer to past precedents: they should look afresh at the scientific issues. How are such tensions resolved from a legal standpoint? The Supreme Court has made clear that a court may overrule precedent if it finds that an earlier case was "erroneously decided and that subsequent events have undermined its continuing validity."[387]

PCAST expresses no view on the legal question of whether any past cases were "erroneously decided." However, PCAST notes that, from a *scientific* standpoint, subsequent events have indeed undermined the continuing validity of conclusions that were not based on appropriate empirical evidence. These events include (1) the recognition of systemic problems with some forensic feature-comparison methods, including through study of the causes of hundreds of wrongful convictions revealed through DNA and other analysis; (2) the 2009 NRC report from the National Academy of Sciences, the leading scientific advisory body established by the Legislative Branch, [388] that found that some forensic feature-comparison methods lack a scientific foundation; and (3) the scientific review in this report by PCAST, the leading scientific advisory body established by the Executive Branch,[389] finding that some forensic feature-comparison methods lack foundational validity.

9.3 Resources for Judges

Another important issue that arose frequently in our conversations with experts was the need for better resources for judges related to evaluation of forensic feature-comparison methods for use in the courts.

The most appropriate bodies to provide such resources are the Judicial Conference of the United States and the Federal Judicial Center.

The Judicial Conference of the United States is the national policy-making body for the federal courts.[390] Its statutory responsibility includes studying the operation and effect of the general rules of practice and procedure in the federal courts. The Judicial Conference develops best practices manuals and issues Advisory Committee notes to assist judges with respect to specific topics, including through its Standing Advisory Committee on the Federal Rules of Evidence.

The Federal Judicial Center is the research and education agency of the federal judicial system.[391] Its statutory duties include (1) conducting and promoting research on federal judicial procedures and court operations and

[387] *Boys Markets, Inc. v. Retails Clerks Union*, 398 U.S. 235, 238 (1970). See also: *Patterson v. McLean Credit Union*, 485 U.S. 617, 618 (1988) (noting that the Court has "overruled statutory precedents in a host of cases"). PCAST sought advice on this matter from its panel of Senior Advisors.

[388] The National Academy of Sciences was chartered by Congress in 1863 to advise the Federal government on matters of science (U.S. Code, Section 36, Title 1503).

[389] The President formally established a standing scientific advisory council soon after the launch of Sputnik in 1957. It is currently titled the President's Council of Advisors of Science and Technology (operating under Executive Order 13539, as amended by Executive Order 13596).

[390] Created in 1922 under the name the Conference of Senior Circuit Judges, the Judicial Conference of the United States is currently established under 28 U.S.C. § 331.

[391] The Federal Judicial Center was established by Congress in 1967 (28 U.S.C. §§ 620-629), on the recommendation of the Judicial Conference of the United States.

(2) conducting and promoting orientation and continuing education and training for federal judges, court employees, and others.

PCAST recommends that the Judicial Conference of the United States, through its Subcommittee on the Federal Rules of Evidence, develop best practices manuals and an Advisory Committee note and the Federal Judicial Center develop educational programs related to procedures for evaluating the scientific validity of forensic feature-comparison methods.

9.4 Recommendation

Based on its scientific findings, PCAST makes the following recommendation.

Recommendation 8. Scientific validity as a foundation for expert testimony

(A) When deciding the admissibility of expert testimony, Federal judges should take into account the appropriate scientific criteria for assessing scientific validity including:

> *(i) foundational validity,* with respect to the requirement under Rule 702(c) that testimony is the product of reliable principles and methods; and
>
> *(ii) validity as applied,* with respect to requirement under Rule 702(d) that an expert has reliably applied the principles and methods to the facts of the case.

These scientific criteria are described in Finding 1.

(B) Federal judges, when permitting an expert to testify about a foundationally valid feature-comparison method, should ensure that testimony about the accuracy of the method and the probative value of proposed identifications is scientifically valid in that it is limited to what the empirical evidence supports. Statements suggesting or implying greater certainty are not scientifically valid and should not be permitted. In particular, courts should never permit scientifically indefensible claims such as: "zero," "vanishingly small," "essentially zero," "negligible," "minimal," or "microscopic" error rates; "100 percent certainty" or proof "to a reasonable degree of scientific certainty;" identification "to the exclusion of all other sources;" or a chance of error so remote as to be a "practical impossibility."

(C) To assist judges, the Judicial Conference of the United States, through its Standing Advisory Committee on the Federal Rules of Evidence, should prepare, with advice from the scientific community, a best practices manual and an Advisory Committee note, providing guidance to Federal judges concerning the admissibility under Rule 702 of expert testimony based on forensic feature-comparison methods.

(D) To assist judges, the Federal Judicial Center should develop programs concerning the scientific criteria for scientific validity of forensic feature-comparison methods.

10. Scientific Findings

PCAST's scientific findings in this report are collected below. Finding 1, concerning the scientific criteria for scientific validity, is based on the discussion in Chapter 4. Findings 2–6, concerning foundational validity of six forensic feature-comparison methods, is based on the evaluations in Chapter 5.

Finding 1: Scientific Criteria for Scientific Validity of a Forensic Feature-Comparison Method

(1) Foundational validity. To establish foundational validity for a forensic feature-comparison method, the following elements are required:

(a) a reproducible and consistent procedure for (i) identifying features within evidence samples, (ii) comparing the features in two samples, and (iii) determining, based on the similarity between the features in two samples, whether the samples should be declared to be likely to come from the same source ("matching rule"); and

(b) empirical estimates, from appropriately designed studies from multiple groups, that establish (i) the method's false positive rate—that is, the probability it declares a proposed identification between samples that actually come from different sources, and (ii) the method's sensitivity—that is, the probability it declares a proposed identification between samples that actually come from the same source.

As described in Box 4, scientific validation studies should satisfy a number of criteria: (a) they should be based on sufficiently large collections of known and representative samples from relevant populations; (b) they should be conducted so that have no information about the correct answer; (c) the study design and analysis plan are specified in advance and not modified afterwards based on the results; (d) the study is conducted or overseen by individuals or organizations with no stake in the outcome; (e) data, software and results should be available to allow other scientists to review the conclusions; and (f) to ensure that the results are robust and reproducible, there should be multiple independent studies by separate groups reaching similar conclusions.

Once a method has been established as foundationally valid based on adequate empirical studies, claims about the method's accuracy and the probative value of proposed identifications, in order to be valid, must be based on such empirical studies.

For objective methods, foundational validity can be established by demonstrating the reliability of each of the individual steps (feature identification, feature comparison, matching rule, false match probability, and sensitivity).

For subjective methods, foundational validity can be established *only* through black-box studies that measure how often many examiners reach accurate conclusions across many feature-comparison problems involving samples representative of the intended use. In the absence of such studies, a subjective feature-comparison method cannot be considered scientifically valid.

Foundational validity is a *sine qua non*, which can only be shown through empirical studies. Importantly, good professional practices—such as the existence of professional societies, certification programs, accreditation programs, peer-reviewed articles, standardized protocols, proficiency testing, and codes of ethics—cannot substitute for empirical evidence of scientific validity and reliability.

(2) Validity as applied. Once a forensic feature-comparison method has been established as foundationally valid, it is necessary to establish its validity as applied in a given case.

> As described in Box 5, validity as applied requires that: (a) the forensic examiner must have been shown to be *capable* of reliably applying the method, as shown by appropriate proficiency testing (see Section 4.6), and must *actually* have done so, as demonstrated by the procedures actually used in the case, the results obtained, and the laboratory notes, which should be made available for scientific review by others; and (b) the forensic examiner's assertions about the probative value of proposed identifications must be scientifically valid—including that the expert should report the overall false positive rate and sensitivity for the method established in the studies of foundational validity; demonstrate that the samples used in the foundational studies are relevant to the facts of the case; where applicable, report probative value of the observed match based on the specific features observed in the case; and not make claims or implications that go beyond the empirical evidence.

Finding 2: DNA Analysis

Foundational validity. PCAST finds that DNA analysis of single-source samples or simple mixtures of two individuals, such as from many rape kits, is an objective method that has been established to be foundationally valid.

Validity as applied. Because errors due to human failures will dominate the chance of coincidental matches, the scientific criteria for validity as applied require that an expert (1) should have undergone rigorous and relevant proficiency testing to demonstrate their ability to reliably apply the method, (2) should routinely disclose in reports and testimony whether, when performing the examination, he or she was aware of any facts of the case that might influence the conclusion, and (3) should disclose, upon request, all information about quality testing and quality issues in his or her laboratory.

Finding 3: DNA analysis of complex-mixture samples

Foundational validity. PCAST finds that:

(1) Combined Probability of Inclusion-based methods. DNA analysis of complex mixtures based on CPI-based approaches has been an inadequately specified, subjective method that has the potential to lead to erroneous results. As such, it is not foundationally valid.

A very recent paper has proposed specific rules that address a number of problems in the use of CPI. These rules are clearly *necessary*. However, PCAST has not adequate time to assess whether they are also *sufficient* to define an objective and scientifically valid method. If, for a limited time, courts choose to admit results based on the application of CPI, validity as applied would require that, at a minimum, they be consistent with the rules specified in the paper.

DNA analysis of complex mixtures should move rapidly to more appropriate methods based on probabilistic genotyping.

(2) Probabilistic genotyping. Objective analysis of complex DNA mixtures with probabilistic genotyping software is relatively new and promising approach. Empirical evidence is required to establish the foundational validity of each such method within specified ranges. At present, published evidence supports the foundational validity of analysis, with some programs, of DNA mixtures of 3 individuals in which the minor contributor constitutes at least 20 percent of the intact DNA in the mixture and in which the DNA amount exceeds the minimum required level for the method. The range in which foundational validity has been established is likely to grow as adequate evidence for more complex mixtures is obtained and published.

Validity as applied. For methods that are foundationally valid, validity as applied involves similar considerations as for DNA analysis of single-source and simple-mixtures samples, with a special emphasis on ensuring that the method was applied correctly and within its empirically established range.

Finding 4: Bitemark analysis

Foundational validity. PCAST finds that bitemark analysis does not meet the scientific standards for foundational validity, and is far from meeting such standards. To the contrary, available scientific evidence strongly suggests that examiners cannot consistently agree on whether an injury is a human bitemark and cannot identify the source of bitemark with reasonable accuracy.

Finding 5: Latent fingerprint analysis

Foundational validity. Based largely on two recent appropriately designed black-box studies, PCAST finds that latent fingerprint analysis is a foundationally valid subjective methodology—albeit with a false positive rate that is substantial and is likely to be higher than expected by many jurors based on longstanding claims about the infallibility of fingerprint analysis.

Conclusions of a proposed identification may be scientifically valid, provided that they are accompanied by accurate information about limitations on the reliability of the conclusion—specifically, that (1) only two properly designed studies of the foundational validity and accuracy of latent fingerprint analysis have been conducted, (2) these studies found false positive rates that could be as high as 1 error in 306 cases in one study and 1 error in 18 cases in the other, and (3) because the examiners were aware they were being tested, the actual false positive rate in casework may be higher. At present, claims of higher accuracy are not warranted or scientifically justified. Additional black-box studies are needed to clarify the reliability of the method.

Validity as applied. Although we conclude that the method is foundationally valid, there are a number of important issues related to its validity as applied.

> **(1) Confirmation bias.** Work by FBI scientists has shown that examiners typically alter the features that they initially mark in a latent print based on comparison with an apparently matching exemplar. Such circular reasoning introduces a serious risk of confirmation bias. Examiners should be required to complete and document their analysis of a latent fingerprint *before* looking at any known fingerprint and should separately document any additional data used during their comparison and evaluation.
>
> **(2) Contextual bias**. Work by academic scholars has shown that examiners' judgments can be influenced by irrelevant information about the facts of a case. Efforts should be made to ensure that examiners are not exposed to potentially biasing information.
>
> **(3) Proficiency testing**. Proficiency testing is essential for assessing an examiner's capability and performance in making accurate judgments. As discussed elsewhere in this report, there is a need to improve proficiency testing, including making it more rigorous, incorporating it within the flow of casework, and disclosing test problems following a test so that they can evaluated for appropriateness by the scientific community.

From a scientific standpoint, validity as applied requires that an expert: (1) has undergone appropriate proficiency testing to ensure that he or she is capable of analyzing the full range of latent fingerprints encountered in casework and reports the results of the proficiency testing; (2) discloses whether he or she documented the features in the latent print in writing before comparing it to the known print; (3) provides a written analysis explaining the selection and comparison of the features; (4) discloses whether, when performing the examination, he or she was aware of any other facts of the case that might influence the conclusion; and (5) verifies that the latent print in the case at hand is similar in quality to the range of latent prints considered in the foundational studies.

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).

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 measure his or her accuracy 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.

Finding 7: Footwear analysis

Foundational validity. PCAST finds there are no appropriate empirical studies to support the foundational validity of footwear analysis to associate shoeprints with particular shoes based on specific identifying marks (sometimes called "randomly acquired characteristics). Such conclusions are unsupported by any meaningful evidence or estimates of their accuracy and thus are not scientifically valid.

PCAST has not evaluated the foundational validity of footwear analysis to identify class characteristics (for example, shoe size or make).

Appendix A: Statistical Issues

To enhance its accessibility to a broad audience, the main text of this report avoids, where possible, the use of mathematical and statistical terminology. However, for the actual implementation of some of the principles stated in the report, somewhat more precise descriptions are necessary. This Appendix summarizes the relevant concepts from elementary statistics.[392]

Sensitivity and False Positive Rate

Forensic feature-comparison methods typically aim to determine how likely it is that two samples came from the same source, given the result of a forensic test on the samples. Two possibilities are considered: the null hypothesis (H0) that they are from different sources (H0) and the alternative hypothesis (H1) that two samples are from the same source. The forensic test result may be summarized as match declared (M) or no match declared (O).

There are two necessary characterizations of a method's accuracy: Sensitivity (abbreviated SEN) and False Positive Rate (FPR).

Sensitivity is defined as the probability that the method declares a match between two samples when they are known to be from the same source (drawn from an appropriate population), that is, SEN = P(M|H1). For example, a value SEN = 0.95 would indicate that two samples from the same source will be declared as a match 95 percent of the time. In the statistics literature, SEN is sometimes also called the "true positive rate," "TPR," or "recall rate."[393]

False positive rate (abbreviated FPR) is defined as the probability that the method declares a match between two samples that are from different sources (again in an appropriate population), that is, FPR = P(M|H0). For example, a value FPR = 0.01 would indicate that two samples from different sources will be (mistakenly) called as a match 1 percent of the time.[394] Methods with a high FPR are scientifically unreliable for making important

[392] See, e.g.: Peter Amitage, G. Berry, JNS Matthews: Statistical Methods in Medical Research, 4th ed., Blackwell Science, 2002; George Snedecor, William G Cochran: Statistical Methods, 8th ed., Iowa State University Press, 1989; Gerald van Belle, Lloyd D Fisher, Patrick Heagerty, Thomas Lumley, Biostatistics: A Methodology for the Health Sciences, Wiley, 2004; Alan Agresti; Brent A. Coull: Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. The American Statistician 52(2), 119-126, 1998; Robert V Hogg, Elliot Tanis, Dale Zimmerman: Probability and Statistical Inference, 9th ed., Pearson, 2015; David Freedman, Roger Pisani, Roger Purves: Statistics. Norton, 2007; Lincoln E Moses: Think and Explain with Statistics, Addison-Wesley, 1986; David S Moore, George P McCabe, Bruce A Craig: Introduction to the Practice of Statistics. W.H. Freeman, 2009.

[393] The term false negative rate is sometimes used for the complement of SEN, that is, FNR = 1 – SEN.

[394] Statisticians may refer to a method's specificity (SPC) instead of its false positive rate (FPR). The two are related by the formula FPR = 1 – SPC. In the example given, FPR = 0.01 (1 percent) and SPC = 0.99 (99 percent).

judgments in court about the source of a sample. To be considered reliable, the FPR should certainly be less than 5 percent and it may be appropriate that it be considerably lower, depending on the intended application.

The results of a given empirical study can be summarized by four values: the number of occurrences in the study of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). (The matrix of these values is, perhaps oddly, referred to as the "confusion matrix.")

	Test Result	
	Match	No Match
H1: Truly from same source	TP	FN
H0: Truly from different sources	FP	TN

In this standard-but-confusing terminology, "true" and "false" refer to agreement or disagreement with the ground truth (either H0 or H1), while "positive" and "negative" refer to the test results (that is, results M and O, respectively).

A widely-used estimate, called the *maximum likelihood* estimate, of SEN is given by TP/(TP+FN), the fraction of events with ground truth H1 (same source) that are correctly declared as M (match). The maximum likelihood estimate of FPR is correspondingly FP/(FP+TN), the fraction events with ground truth H0 (different source) that are mistakenly declared as M (match).

Since the false positive rate will often be the mathematically determining factor in the method's probative value in a particular case (discussion below), it is particularly important that FPR be well measured empirically.

In addition, tests with very low sensitivity should be viewed with suspicion because rare positive test results may be matched or outweighed by the occurrence of false positive results.[395]

Confidence Intervals

As discussed in the main text, to be valid, empirical measurements of SEN and FPR must be based on large collections of known and representative samples from each relevant population, so as to reflect how often a given feature or combination of features occurs. (Other requirements for validity are also discussed in the main text.)

Since empirical measurements are based on a limited number of samples, SEN and FPR cannot be measured exactly, but only estimated. Because of the finite sample sizes, the maximum likelihood estimates thus do not tell the whole story. Rather, it is necessary and appropriate to quote confidence bounds within which SEN, and FPR, are highly likely to lie.

[395] The argument in favor of a test that "this test succeeds only occasionally, but in this case it did succeed" is thus a fallacious one

Because one should be primarily concerned about overestimating SEN or underestimating FPR, it is appropriate to use a *one-sided* confidence bound. By convention, a confidence level of 95 percent is most widely used—meaning that there is a 5 percent chance the true value exceeds the bound. Upper 95 percent one-sided confidence bounds should thus be used for assessing the error rates and the associated quantities that characterize forensic feature matching methods. (The use of lower values may rightly be viewed with suspicion as an attempt at obfuscation.)

The confidence bound for proportions depends on the sample size in the empirical study. When the sample size is small, the estimates may be far from the true value. For example, if an empirical study found no false positives in 25 individual tests, there is still a reasonable chance (at least 5 percent) that the true error rate might be as high as roughly 1 in 9.

For technical reasons, there is no single, universally agreed method for calculating these confidence intervals (a problem known as the "binomial proportion confidence interval"). However, the several widely used methods give very similar results, and should all be considered acceptable: the Clopper-Pearson/Exact Binomial method, the Wilson Score interval, the Agresti-Coull (adjusted Wald) interval, and the Jeffreys interval.[396] Web-based calculators are available for all of these methods.[397] For example, if a study finds zero false positives in 100 tries, the four methods mentioned give, respectively, the values 0.030, 0.026, 0.032, and 0.019 for the upper 95 percent confidence bound. From a scientific standpoint, any of these might appropriately be reported to a jury in the context "the false positive rate might be as high as." (In this report, we used the Clopper-Pearson/Exact Binomial method.)

Calculating Results for Conclusive Tests

For many forensic tests, examiners may reach a conclusion (e.g., match or no match) or declare that the test is inconclusive. SEN and FPR can thus be calculated based on the *conclusive* examinations or on *all* examinations. While both rates are of interest, from a scientific standpoint, the former rate should be used for reporting FPR to a jury. This is appropriate because evidence used against a defendant will typically be based on *conclusive*, rather than inconclusive, examinations. To illustrate the point, consider an extreme case in which a method had been tested 1000 times and found to yield 990 inconclusive results, 10 false positives, and no correct results. It would be misleading to report that the false positive rate was 1 percent (10/1000 examinations). Rather, one should report that 100 percent of the conclusive results were false positives (10/10 examinations).

Bayesian Analysis

In this appendix, we have focused on the Sensitivity and False Positives rates (SEN = P(M|H1) and FPR = P(M|H0)). The quantity of most interest in a criminal trial is P(H1|M), that is, "the probability that the samples are from the same source *given* that a match has been declared." This quantity is often termed the *positive predictive value* (PPV) of the test.

[396] Brown, L.D., Cai, T.T., and A. DasGupta. "Interval estimation for a binomial proportion." *Statistical Science*, Vol. 16, No. 2 (2001): 101-33.

[397] For example, see: epitools.ausvet.com.au/content.php?page=CIProportion.

The calculation of PPV depends on two quantities: the "Bayes factor" BF = SEN/FPR and a second quantity called the "prior odds ratio" (POR). This latter quantity is defined mathematically as POR = P(H0)/P(H1), where P(H0) and P(H1) are the prior (i.e., before doing the test) probabilities of the hypotheses H0 and H1.[398] The formula for PPV in terms of BF and POR is: PPV = BF / (BF + POR), a formula that follows from the statistical principle known as Bayes Theorem.[399]

Bayes Theorem offers a mathematical way to combine the test result with independent information—such as (1) one's prior probability that two samples came from the same source and (2) the number of samples searched. Some Bayesian statisticians would choose POR = 1 in the case of a match to single sample (implying that it is equally likely *a priori* that the samples came from the same source as from different sources) and POR = 100,000 for a match identified by comparing a sample to a database containing 100,000 samples. Others would set POR = (1-p)/p, where p is the *a priori* probability of same-source identity in the relevant population, given the other facts of the case.

The Bayesian approach is mathematically elegant. However, it poses challenges for use in courts: (1) different people may hold very different beliefs about POR and (2) many jurors may not understand how beliefs about POR affect the mathematical calculation of PPV. (Moreover, as noted previously, the empirical estimates of SEN and FPR have uncertainty, so the estimated BF = SEN/FPR also has uncertainty.)

Some commentators therefore favor simply reporting the empirically measured quantities (the sensitivity, the false positive rate of the test, and the probability of a false positive match given the number of samples searched against)and allowing a jury to incorporate them into their own intuitive Bayesian judgments. (For example, "*Yes, the test has a false positive rate of only 1 in 100, but two witnesses place the defendant 1000 miles from the crime scene, so the test result was probably one of those 1 in 100 false positives*.")

[398] That is, if p is the *a priori* probability of same-source identity in the population under examination then POR = (1-p)/p.

[399] In the main text, the phrase "appropriately correct for the size of the pool that was searched in identifying a suspect" refers to the use of this formula with an appropriate value for POR.

Appendix B. Additional Experts Providing Input

PCAST sought input from a diverse group of additional experts and stakeholders. PCAST expresses its gratitude to those listed here who shared their expertise. They did not have the opportunity to review drafts of the report, and their willingness to engage with PCAST on specific points does not imply endorsement of the views expressed therein. Responsibility for the opinions, findings, and recommendations in this report and for any errors of fact or interpretation rests solely with PCAST.

Richard Alpert
Assistant Criminal District Attorney Tarrant
County Criminal District Attorney's Office

Kareem Belt
Forensic Policy Analyst
Innocence Project

William Bodziak
Consultant
Bodziak Forensics

John Buckleton
Principal Scientist
Institute of Environment and Scientific Research
New Zealand

Bruce Budowle
Professor, Executive Director of Institute of Applied Genetics
University of North Texas Health Science Center

Mary A. Bush
Associate Professor
Department of Restorative Dentistry
University at Buffalo School of Dental Medicine

Peter Bush
Research Instructor
Director of the South Campus Instrument Center
University at Buffalo School of Dental Medicine

John Butler
Special Assistant to the Director for Forensic Science
Special Programs Office
National Institute of Standards and Technology

Arturo Casadevall
Professor
Department of Microbiology & Immunology and Department of Medicine
Albert Einstein College of Medicine

Alicia Carriquiry
Distinguished Professor at Iowa State and Director, Center for Statistics and Applications in Forensic Evidence
Iowa State University

Richard Cavanagh
Director
Special Programs Office
National Institute of Standards and Technology

Eleanor Celeste
Policy Analyst
Medical and Forensic Sciences
Office of Science and Technology Policy

Christophe Champod
Professor of Law, Criminal Science and Public Administration
University of Lausanne

Sarah Chu
Senior Forensic Policy Advocate
Innocence Project

Simon A. Cole
Professor of Criminology, Law and Society
School of Social Ecology
University of California Irvine

Kelsey Cook
Program Director
Chemical Measurement and Imaging
National Science Foundation

Patricia Cummings
Special Fields Bureau Chief
Dallas County District Attorney's Office

Christopher Czyryca
President
Collaborative Testing Services

Dana Delger
Staff Attorney
Innocence Project

Shari Diamond
Howard J. Trienens Professor of Law
Professor of Psychology
Pritzker School of Law
Northwestern University

Itiel Dror
Senior Cognitive Neuroscience Researcher
University College London

Meredith Drosback
Assistant Director
Education and Physical Sciences
Office Of Science and Technology Policy

Kimberly Edwards
Physical Scientist
Forensic Examiner
Federal Bureau of Investigation Laboratory

Ian Evett
Forensic Statistician
Principal Forensic Services

Chris Fabricant
Director, Strategic Litigation
Innocence Project

Kenneth Feinberg
Steven and Maureen Klinsky Visiting Professor of Practice for Leadership and Progress
Harvard Law School

Rebecca Ferrell
Program Director
Biological Anthropology
National Science Foundation

Jennifer Friedman
Forensic Science Coordinator
Los Angeles County Public Defender

Lynn Garcia
General Counsel
Texas Forensic Science Commission

Daniel Garner
Chief Executive Officer and President
Houston Forensic Science Center

Constantine A. Gatsonis
Henry Ledyard Goddard University Professor of Biostatistics
Chair of Biostatistics
Director of Center for Statistical Sciences
Brown University

Eric Gilkerson
Forensic Examiner
Federal Bureau of Investigation Laboratory

Brandon Giroux
President
Giroux Forensics, L.L.C.
President
Forensic Assurance

Catherine Grgicak
Assistant Professor
Anatomy and Neurobiology
Boston University School of Medicine

Austin Hicklin
Fellow
Noblis

Cindy Homer
Forensic Scientist
Maine State Police Crime Lab

Alice Isenberg
Deputy Assistant Director
Federal Bureau of Investigation Laboratory

Matt Johnson
Senior Forensic Specialist
Orange County Sheriff's Department

Jonathan Koehler
Beatrice Kuhn Professor of Law
Pritzker School of Law
Northwestern University

Glenn Langenburg
Forensic Science Supervisor
Minnesota Bureau of Criminal Apprehension

Gerald LaPorte
Director
Office of Investigative and Forensic Sciences
National Institute of Justice

Julia Leighton
General Counsel
Public Defender Service
District of Columbia

Alan I. Leshner
Chief Executive Officer, Emeritus
American Association for the Advancement of Science and Executive Publisher of the journal *Science*

Ryan Lilien
Chief Science Officer
Cadre Research Labs

Elizabeth Mansfield
Deputy Office Director
Personalized Medicine
Food and Drug Administration

Anne-Marie Mazza
Director
Committee on Science, Technology, and Law
The National Academies of Science, Engineering and Medicine

Willie E. May
Director
National Institute of Standards and Technology

Daniel MacArthur
Assistant Professor
Harvard Medical School
Co-Director of Medical and Population Genetics
Broad Institute of Harvard and MIT

Brian McVicker
Forensic Examiner
Federal Bureau of Investigation Laboratory

Stephen Mercer
Director
Litigation Support Group
Office of the Public Defender
State of Maryland

Melissa Mourges
Chief
Forensic Sciences/Cold Case Unit
New York County District Attorney's Office

Peter Neufeld
Co-Director and Co-Founder
Innocence Project

Steven O'Dell
Director
Forensic Services Division
Baltimore Police Department

Lynn Overmann
Senior Policy Advisor
Office of Science and Technology Policy

Skip Palenik
Founder
Microtrace

Matthew Redle
County and Prosecuting Attorney
Sheridan County Prosecutor's Office

Maria Antonia Roberts
Research Program Manager
Latent Print Support Unit
Federal Bureau of Investigation Laboratory

Walter F. Rowe
Professor of Forensic Sciences
George Washington University

Norah Rudin
President and CEO
Scientific Collaboration, Innovation & Education Group

Jeff Salyards
Director
Defense Forensic Science Center
The Defense Forensics and Biometrics Agency

Rodney Schenck
Defense Forensic Science Center
The Defense Forensics and Biometric Agency

David Senn
Director
Center for Education and Research in Forensics and the Southwest Symposium on Forensic Dentistry
University of Texas Health Science Center at San Antonio

Stephen Shaw
Trace Examiner
Federal Bureau of Investigation Laboratory

Andrew Smith
Supervisor Firearm/ Toolmark Unit
San Francisco Police Department

Erich Smith
Physical Scientist
Firearms-Toolmarks Unit
Federal Bureau of Investigation Laboratory

Tasha Smith
Firearm and Tool Mark Unit
Criminalistics Laboratory
San Francisco Police Department

Jeffrey Snipes
Associate Professor
Criminal Justice Studies
San Francisco State University

Jill Spriggs
Laboratory Director
Sacramento County District Attorney's Office

Harry Swofford
Chief, Latent Print Branch
Defense Forensics Science Center
The Defense Forensics and Biometric Agency

Robert Thompson
Program Manager Forensic Data Systems
Law Enforcement Standards Office
National Institute of Standards and Technology

William Thompson
Professor of Criminology, Law, and Society and Psychology & Social Behavior
Law School of Social Ecology
University of California, Irvine

Rick Tontarski
Chief Scientist
Defense Forensic Science Center

Jeremy Triplett
Laboratory Supervisor
Kentucky State Police Central Forensic Laboratory

Richard Vorder Bruegge
Senior Photographic Technologist
Federal Bureau of Investigation

Victor Weedn
Chair of Forensic Sciences
Department of Forensic Sciences
George Washington University

Robert Wood
Associate Professor and Head
Department of Dental Oncology
Dentistry, Ocular and Maxillofacial Prosthetics
Princess Margaret Cancer Centre
University of Toronto

Xiaoyu Alan Zheng
Mechanical Engineer
National Institute of Standards and Technology

President’s Council of Advisors on Science and Technology (PCAST)

www.whitehouse.gov/ostp/pcast

Report: Validation Study of the Accuracy, Repeatability, and Reproducibility of Firearm Comparisons

October 7, 2020

Stanley J. Bajic, L. Scott Chumbley, Max Morris, and Daniel Zamzow
Ames Laboratory-USDOE

Technical Report # ISTR-5220

Prepared for the Federal Bureau of Investigation

This work was supported by the Federal Bureau of Investigation (FBI) through contract with the Ames Laboratory, U.S. Department of Energy, under Interagency Agreement DJF-15-1200-0000667. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the FBI or the U.S. Government.

Exhibit D

Abstract: This report details the results of a black-box validation study designed to determine the accuracy, repeatability, and reproducibility of analyses performed by firearm examiners for cartridge case and bullet sample sets. The study was conducted between 2016 and 2020, in collaboration between the Federal Bureau of Investigation (FBI) and Ames Laboratory-USDOE. Fired cartridge case and bullet samples were collected by the FBI and provided to Ames Lab to distribute to participating firearm examiners, for analysis. The study was implemented in six mailings, with the intent that each examiner would receive two test packets for each of the three rounds of the study to assess accuracy, repeatability, and reproducibility.

Volunteer active examiners were provided with test packets that contained 15 comparison sets of 2 known + 1 unknown cartridge cases fired from a collection of Beretta and Jimenez firearms and 15 comparison sets of 2 known + 1 unknown bullets fired from Beretta and Ruger firearms. The ammunition was all Wolf Polyformance 9 mm Luger (9x19mm). Examiners were provided with a brief background survey, answer sheets with a rubric allowing for the AFTE Range of Conclusions, and return shipping materials. The participating examiners were expected to conduct their examinations as they would with real case specimens but were asked to follow the provided instructions for reporting their conclusions rather than adhere to their laboratory policies. In particular they were instructed not to discuss their results with anyone. Results are presented for cartridge case and bullet analyses performed by 173 examiners. The total number of comparisons reported is 20,130, for the first (8640), second (5700), and third (5790) rounds of the project.

The overall rate of false positives is 0.656 % and 0.933% for bullets and cartridge cases, respectively, and the rate of false-negative errors is 2.87% and 1.87% for bullets and cartridge cases, respectively. These estimates are based on the beta-binomial probability model and do not depend on an assumption of equal examiner-specific error rates. These are the error rates for all analyses conducted, including comparisons from barrels produced sequentially in time and separated in the manufacturing process, cartridges fired early in the life of a barrel and after many rounds had been fired, and rounds fired from both high and low cost-point firearms. Individual error rates within these categories are also calculated and are presented in this report. These numbers are generally consistent with the results of a prior study conducted by Baldwin et al. As in the earlier study, the majority of errors were produced by a relatively small number of examiners.

It should be pointed out that the firearms and ammunition selected for this study were chosen for their difficulty to test the boundaries of an examiners' pattern recognition ability. Laboratory error rates may be lower than these individual rates, provided quality assurance procedures are applied that can effectively manage to reduce or eliminate the propagation of false positives reported by individuals.

Table of Contents

List of Figures

List of Tables

Definitions

Breech Face: That part of the firearm which is against the head of the cartridge case or shotshell during firing.

Breech-face Marks: Marks characteristic of the breech that are impressed onto a cartridge case or shotshell during the firing process.

Case: Term used in table headings to denote cartridge cases.

Consecutively Matching Striae: A procedure supplementary to traditional pattern recognition that involves comparing the number and spacing of given series of impressed striae, common between known and questioned samples, to a threshold criteria.

Communication Group: Those researchers at Ames Laboratory responsible for mailing and receiving the test packets distributed to participants. Members of this group were unaware of what each packet contained in terms of known matches and nonmatches.

Comparison Set: A collection of 3 bullets or 3 cartridge cases containing two known (K) and one questioned (Q) sample. Also referred to more briefly as a **Set**.

Firing Pin Impression: The indentation left on the cartridge case when it is struck by the firing pin.

Experimental/Analysis Group: Those researchers at Ames Laboratory responsible for assembling the test packets distributed to participants and scoring the returned results. Members of this group were unaware of to whom and where each packet was sent.

Group Number: Unique primary identifier assigned to each test packet.

Hard Error: Declaring an Identification when it is in fact an Elimination or declaring an Elimination when in fact it is an Identification.

Mailing: Term used to describe the physical act (and contents) of boxes shipped to the examiners containing the sample packets. Examiners who completed the study received a total of 6 separate mailings. New mailings were not shipped until the previous mailed box had been returned.

Pooling: The combining of two or more AFTE categories for the purposes of statistical analysis

Round: One of three phases of the study determined by the study design. The analyses described here are each based on the data collected in specific subsets of rounds, as described in detail in the report.

> **Accuracy**: The ability of an examiner to correctly identify a known match or eliminate a known nonmatch.
>
> **Repeatability:** The ability of an examiner, when confronted with the exact same comparison once again, to reach the same determination as when first examined.

Reproducibility: The ability of a second examiner to evaluate a set previously viewed by a different examiner and reach the same conclusion.

Sample Packet: The collection of items mailed to each examiner. Consists of one test packet, answer sheets, instruction sheet, return envelope, and return mailing box. The first mailing also included the participant survey.

Set: Same as a Comparison Set.

Striations (striae): Microscopic markings, usually appearing as parallel lines, that result on cartridge cases or bullets as a result of the firing process. They differ from impression marks in that they are caused by a shear force applied parallel to the marked surface rather than a compressive force perpendicular to it.

Test Packet: Consisting of 15 bullet sets and 15 cartridge sets, a test packet represents the total number of sets provided to each examiner per mailing.

Unsuitable: AFTE definition stating that a comparison can not be made due to quality of the provided samples.

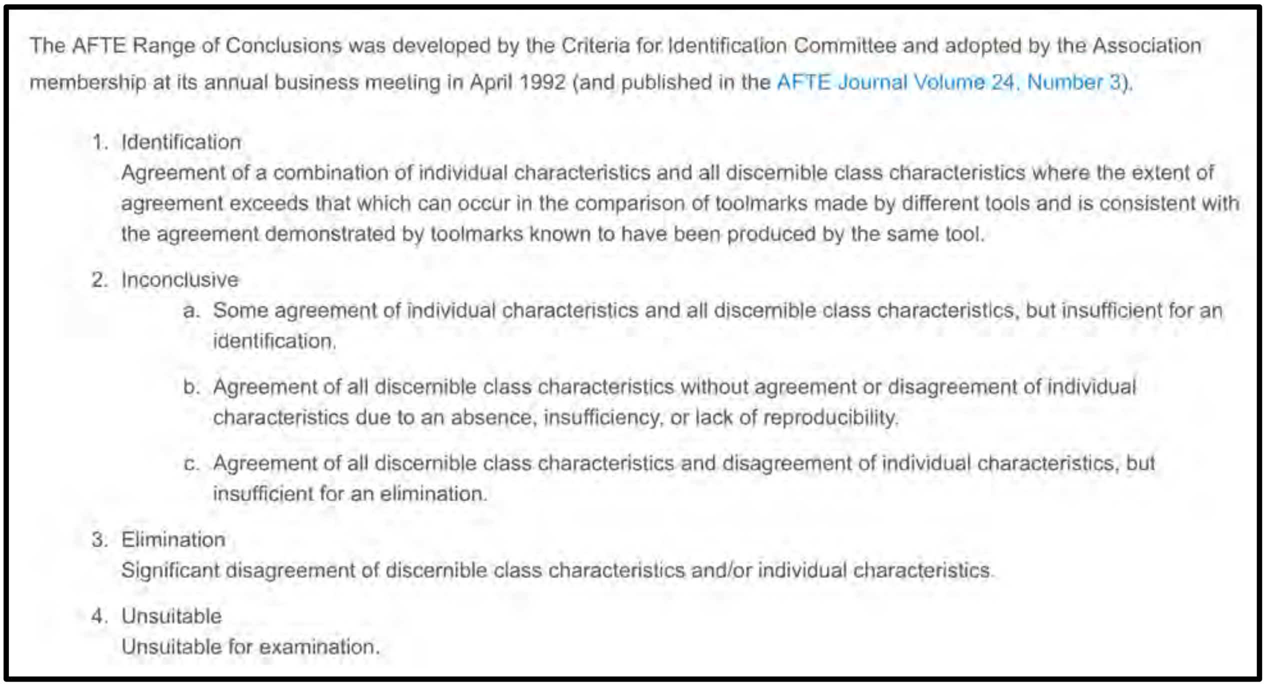
Unusable: In this study this describes a bullet or cartridge case sample, designated as a Known in a set, that does not have sufficient reproducible detail for comparison as evaluated by an examiner.

Introduction

In September 2016 the President's Council of Advisors on Science and Technology (PCAST) published a report titled "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods" [1]. This report examined the scientific validity of a number of feature-comparison analysis methods including DNA, bite-marks, latent fingerprints, firearms, footwear, and hair. Their conclusion, as stated in Chapter 5 of the report, is that "many forensic feature-comparison methods have historically been assumed rather than established" and that "For most subjective methods, there are no appropriate black-box studies" that provide "appropriate evidence of foundational validity or estimates of reliability" (*N.B. A "black-box" study as defined by the Organization of Scientific Area Committees for Forensic Science is one that assesses the accuracy of examiners' conclusions without considering how the conclusions were reached. The examiner is treated as a "black-box" and the researcher measures how the output of the "black-box", i.e. the examiner's conclusion, varies depending on the input, which in this case is the test specimens presented for analysis*). In the area of firearms, numerous studies have been conducted to test the validity of firearm cartridge case and bullet comparisons by firearm examiners and independent researchers, often with a high degree of scientific rigor [2-6]. However, given the guidelines applied by PCAST only a single study was found to have been conducted that adequately addressed their concerns, namely the 2014 report by Baldwin et al. conducted by Ames Laboratory-USDOE [7]. The PCAST report further stated that in order to establish foundational validity the principle of reproducibility needed to be satisfied by an additional study. The investigative work planned and discussed below was designed to provide that necessary information.

In this study contact between the participating examiners and the experimental team was restricted at all times to both preserve anonymity of the participants and prevent any interactions that might result in bias between participants and investigators. To maintain the double blind nature of the study, the FBI Laboratory produced the specimens for the study but awarded a contract to Ames Laboratory to perform the studies it designed, where duties were separated into two groups that performed different tasks. The first group (communication) had contact with the participating firearm examiners; maintained a list of names and addresses; collected consent forms; and arranged for shipping and receiving sample packets to and from the examiners. The second group (experimental/analysis) was responsible for constructing the cartridge case and bullet sample sets to be analyzed; scoring, database entry, and verification of the results; repackaging analyzed packets for subsequent mailings of the study; and performing statistical analysis of the reported results. Communication between the two groups was restricted to the exchange of packets, each possessing a unique three-digit code, that either needed to be sent to examiners or had been returned by examiners. None of the sample-packet specific information was shared between groups. Similarly, contact information was never shared between the groups. This arrangement ensured that only the Ames Lab experimental/analysis group knew the ground-truth information regarding the cartridge case and bullet sets analyzed; the identities of the examiners were known only by the Ames Lab communication group.

A pilot study was initially conducted to provide guidance as to the firearms and ammunition to use. Based on these results bullet and cartridge case samples were obtained using three different brands of firearms and a single standard ammunition. The FBI acquired new Beretta M9A3 and Ruger SR9c firearms for production of bullet specimens, and the same Berettas plus some Jimenez J.A. Nine firearms for production of cartridge case samples. Packets sent to the participating firearm examiners consisted of 15 comparison sets for both bullets and cartridge cases. The prior Ames study was limited to the analysis of cartridge case samples, in one mailing to examiners [7]. The current study was designed to measure accuracy, including false-positive and false-negative error rate determinations, as well as repeatability and reproducibility of analyses, for both bullets and cartridge cases over six mailings. The design of the current study resulted in a large increase in the total number of comparisons each examiner was asked to provide compared to the prior Ames study. Participants were specifically asked not to use their laboratory or agency peer-review process and to not discuss their conclusions with others. The reporting sheet provided for each bullet and cartridge case sample set analyzed listed four possible findings for the examination: Identification, Inconclusive, Elimination, and Unsuitable, with three possibilities for qualifying an Inconclusive decision. These choices are in accordance with those published by the Association of Firearm and Tool Mark Examiners (AFTE), shown in Figure 1 [8].



The AFTE Range of Conclusions was developed by the Criteria for Identification Committee and adopted by the Association membership at its annual business meeting in April 1992 (and published in the AFTE Journal Volume 24, Number 3).

1. Identification
   Agreement of a combination of individual characteristics and all discernible class characteristics where the extent of agreement exceeds that which can occur in the comparison of toolmarks made by different tools and is consistent with the agreement demonstrated by toolmarks known to have been produced by the same tool.
2. Inconclusive
   a. Some agreement of individual characteristics and all discernible class characteristics, but insufficient for an identification.
   b. Agreement of all discernible class characteristics without agreement or disagreement of individual characteristics due to an absence, insufficiency, or lack of reproducibility.
   c. Agreement of all discernible class characteristics and disagreement of individual characteristics, but insufficient for an elimination.
3. Elimination
   Significant disagreement of discernible class characteristics and/or individual characteristics.
4. Unsuitable
   Unsuitable for examination.

Figure 1: AFTE range of conclusions.

The ammunition chosen for the study consisted of copper-plated, steel jacketed bullets and steel cases with brass primers, which is generally believed by examiners to produce a more challenging test. Similarly, the firearms selected for this study were purposely chosen due to their propensity to produce challenging and ambiguous test samples, creating difficult comparisons for examiners. Thus, the study was designed to be a rigorous trial of examiner ability.

Relevant Literature

Two international compilations review recent (2013-2019) research studies and general discussions as applied to firearms and toolmark examinations [9,10]. The Association of Firearm and Tool Mark

Examiners (AFTE) also maintains lists of published work related to firearm examinations that can be found on the AFTE website [11, 12].

Since publication of the PCAST report in 2016 a number of studies have been published dealing with various aspects of comparative examinations. These include efforts to produce either automated or computer based objective determinations [13-20]; statistical evaluation methods in the identification of toolmarks [21-23]; and examination of examiner error rates [24, 25]. In these latter studies very low error rates have been reported, well under 1%.

Several additional studies are also on-going at this time aimed at investigating the validity of conclusions and reliability of firearm examinations including studies funded by the National Science Foundation (NSF) and the National Institute of Standards and Technology (NIST). The reader would be advised to consult the relevant web sites (https://www.nsf.gov and https://www.nist.gov) for the latest information about these efforts.

Experimental Procedure

Study Participants

Pilot Study

A pilot test was conducted prior to soliciting participants and assembling test materials for the full study. Volunteers were recruited through the FBI Laboratory. Test samples consisting of 100 fired cartridge cases and 100 fired bullets were collected by the FBI Laboratory using Beretta Model 92, Hi-Point Model C, and Ruger SR9c firearms. The caliber was 9mm Luger and consisted of brass and steel cartridge cases and copper- and steel-jacketed bullets. Pilot test packets were assembled containing both bullet and cartridge case comparisons and sent to four volunteers. Feedback was requested on study design, survey forms, consent forms, and any other information the participants wanted to provide. Improvements were made to the study materials based on the volunteers' comments and the basic outline for the full study was refined and decided upon by FBI personnel, who selected the firearms and ammunition to use, developed a firing plan for the collection of samples, and did initial testing on methods to mount and hold bullet samples.

Solicitation

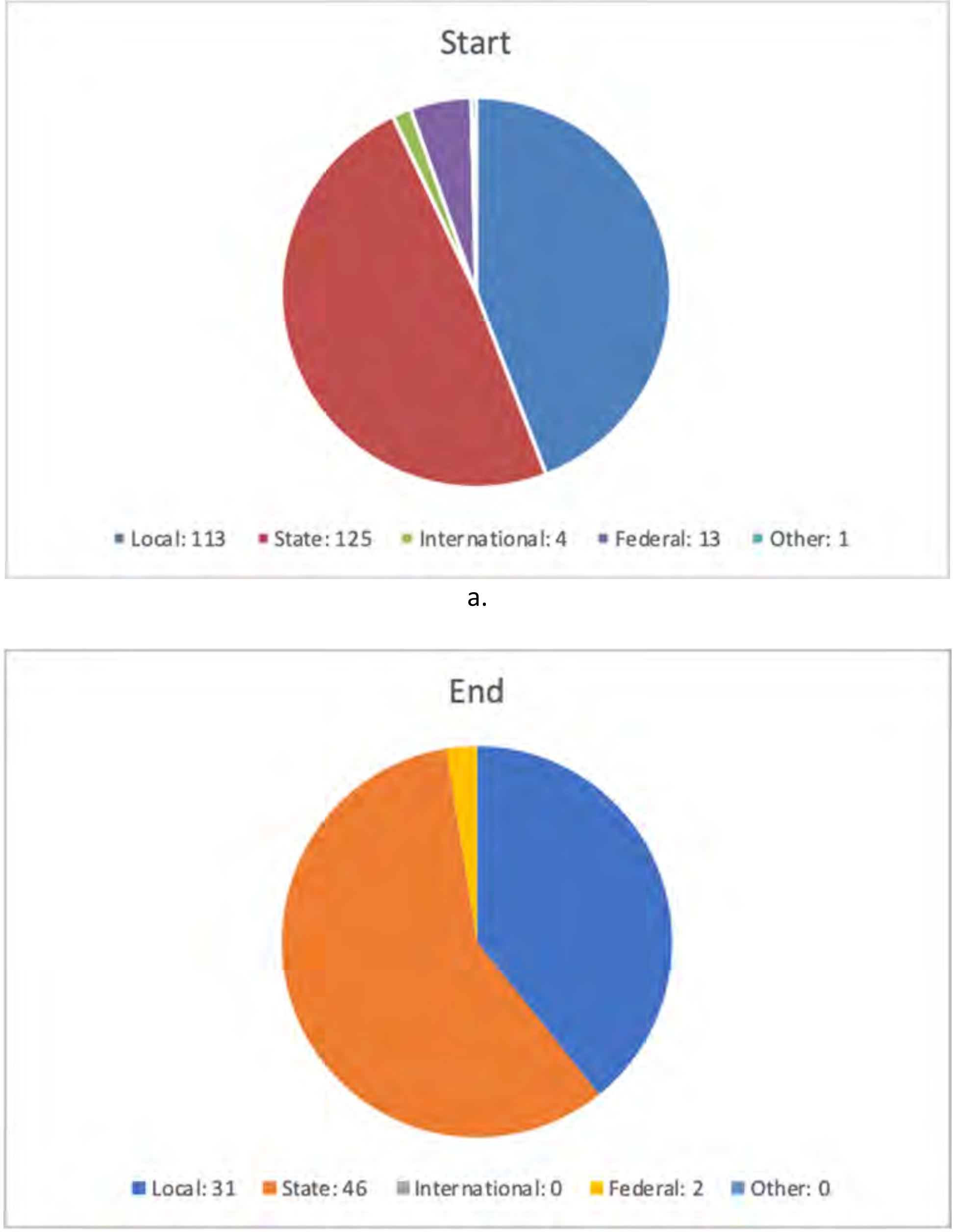
Due to the participation of human subjects in this study, the experimental program was subject to review and approval by the Institutional Review Boards of Ames Lab's contracting agency, Iowa State University, and the Federal Bureau of Investigation. The study was designed to protect the participants from risk to their professional standing and reputation by making all results anonymous. Broad calls for volunteers were made through the AFTE website, via announcements by FBI personnel at national forensic meetings, through e-mail lists maintained by AFTE, and through national / international listservs. The letter of invitation and the informed consent form are shown in Appendix A. These methods resulted in 270 respondents who contacted the Ames Lab Communication Group. A list of volunteers' names, mailing addresses, and other contact information was created; the list was neither an alphabetical listing nor organized according to laboratory or location. Only respondents who returned a signed consent form and were currently conducting firearm examinations and were members of AFTE, or else were employed in the firearms section of an accredited crime laboratory within the U.S. or a U.S. territory were accepted into the study. A decision was made to exclude any examiners currently employed by the FBI, to avoid any possibility of bias. This reduced the initial starting number of respondents to 256 participants. Once initial sample packets were distributed and volunteers became fully aware of the amount of work required, many examiners decided to drop out of the study without analyzing the first test packet. As a result, a total of 173 examiners returned evaluations and were active in the study.

Demographics

A breakdown of the initial 256 participants according to employment is shown in Figure 2, with the actual numbers shown in the legends. The majority of examiners are employed by state and local crime labs, with a smaller number at federal and “international” (U.S. territories) labs. As noted in the previous section, the number of examiners involved fell to 173, with additional examiners withdrawing from the study over the course of the data-collection period. Thus, at the conclusion of the study only 80 participants had finished all six mailings of test-packet analyses. The demographics of those starting, and those fully completing the study are shown in Figures 2a and 2b, respectively. Despite the high dropout rate, the average number of packets completed by a participant was between three and four; in other words, the average examiner contributed over 100 sample-set comparisons to the study. Figure 3 summarizes the number of examiners active after each mailing of the study.

While a high dropout rate has been considered a warning flag for possible test bias [26], comments from withdrawing examiners overwhelmingly stated that they simply did not have enough time to complete both their assigned duties and the additional work asked of them by the study. Universally the volunteers regretted that they were not able to continue with the study.

The distribution of years of experience for the participating examiners is shown in Figure 4. While participants were present from both ends of the career spectrum (less than 1 to 50 years), the average years of experience of the examiners was 10.7 years, whilst the median was 9.



a.

b.

Figure 2: Demographics of the participants (a) at the start of the study and (b) at the conclusion of the study. Note that "international" refers to participants in U.S. territories.

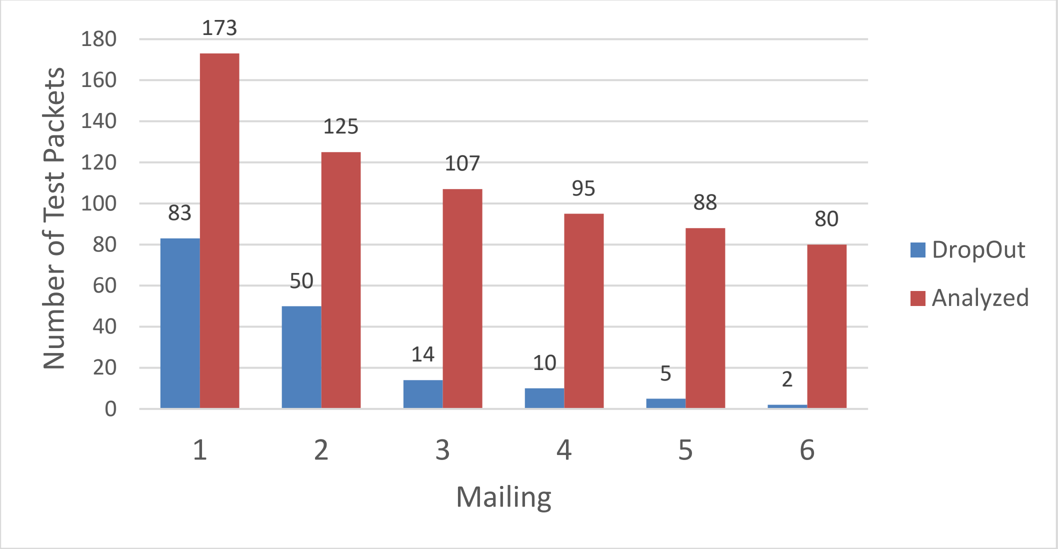
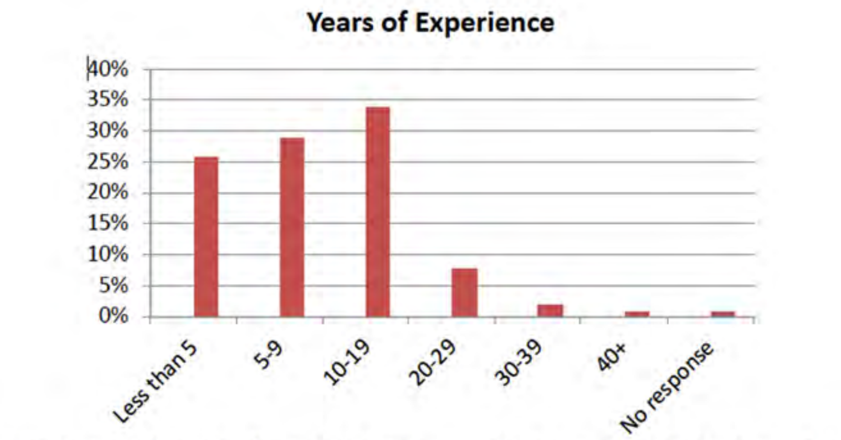


Figure 3: Numbers of test packets analyzed by examiners and numbers of examiners that withdrew from the study, in the six mailings.



Minimum: 0.1; Median: 9; Maximum: 50; Mean: 10.7; St. Dev.: 8.1

Figure 4: Distribution of Examiners' Years of Experience.

Appendix B contains the participant survey which was used to obtain information concerning the examiners themselves and the manner in which they typically carry out examinations at their place of business and a summary of their responses. This form was included in the initial mailing to all participants, along with an instruction sheet for the examination and an answer sheet. The instruction sheet is included as Appendix C.

Study Samples

Collection

Cartridge case and bullet samples were collected by the Federal Bureau of Investigation at the FBI Laboratory in Quantico, VA. For cartridge cases, 10 new Jimenez plus 1 Bryco replacement for a failed Jimenez and 27 Beretta firearms were used to collect 850 and 700 cartridge cases, respectively, resulting in the collection of 9350 Jimenez and 18,900 Beretta samples. (*N.B. The Bryco firearm is identical to the Jimenez in all aspects. The manufacturing process for this particular firearm used the exact same machinery but the name passed from Jennings to Bryco to Jimenez due to a series of name changes and company buy-outs, See [27]. The failed part was an internal part necessary for operation but it is not instrumental in producing identification marks*) While four Beretta firearms came from the FBI collection, 23 of the 27 Beretta barrels were newly manufactured, selected in groups of 4 or 5 that were consecutively produced using the same broach at different periods in the life of the broach. Prior to the collection of cartridge cases to be used in the study, 30 rounds were test-fired using each of the Jimenez guns, to “break in” the firearms. In a similar manner, 50 test-fires for each of the Beretta guns were done prior to the collection of samples; 10 test rounds had also been fired at the factory. All the ammunition used in this study was Wolf Polyformance 9mm Luger (9x19mm). The cartridges were polymer coated steel cartridge case 115 grain full metal jacket (FMJ) rounds, that came in boxes of 50 cartridges each. The Wolf bullet consists of a lead core with a copper plated steel jacket. The ammunition was fired and collected sequentially, in groups of 50 - (31-80, 81-130, … 831-880) for each Jimenez gun and (51-100, 101-150, … 701-750) for each Beretta gun. Firearms were cleaned after each 50 fired rounds during the collection process. Since steel cartridge cases obturate to a lesser extent than brass, carbon tends to deposit in the breech and bore. Cartridge cases that had double-strikes were not used in this study. A total of 17 different lot numbers for the Wolf ammunition were used, with 5 of the 17 used for both the Jimenez and Beretta cartridge case samples. The collected cartridge cases were returned to their original 50-holed plastic containers to prevent individual cartridges from contacting each other and potentially acquiring additional marks. The containers were labeled with the gun serial number and the sequential firing order (within a range of 50 fired samples) prior to shipment to Ames Lab.

Bullet samples were collected in a similar fashion, using the 11 Ruger and the 27 Beretta guns, to provide a total of 9350 Ruger and 18,900 Beretta bullets. For the Ruger guns, 60 rounds were test-fired prior to the collection of samples, so the sequential groups of 50 samples used in the study were the (61-110, 111-160, … 861-910) range of fired bullets. Beretta bullets were collected concurrently with the Beretta cartridge cases for the ranges listed above. Bullets that became deformed during collection were not used in this study. The collected bullets were placed in small manila envelopes that were labeled with the gun serial number and the sequential firing order (within a range of 50 fired samples) prior to shipment to Ames Lab.

Separate cardboard boxes, one for each of the Jimenez, Ruger, and Beretta guns, were used to contain all the samples for a given serial number firearm, so 11 boxes of Jimenez cartridge cases, 11 boxes of Ruger bullets, and 27 boxes of Beretta cartridge cases and bullets were shipped by the FBI to Ames Lab. Each Jimenez cardboard box contained 17 labeled boxes of 50 cartridge cases (850 rounds fired per gun), so the collection of Jimenez boxes totaled 187 boxes of 50 fired cartridge cases (9350 samples). Each Beretta box contained 14 labeled boxes of 50 cartridge cases (700 rounds fired per gun), so the collection of Beretta boxes totaled 378 boxes of 50 fired cartridge cases (18,900 samples). The same number of bullets were collected for use in the study, so there were 565 manila envelopes (187 Ruger and 378 Beretta) each containing 50 bullets (28,250 total samples), received at Ames Lab.

The firearms and ammunition discussed above were specifically chosen to present examiners with a difficult task. Due to its hardness the steel cartridge case and bullet jackets used may not significantly reproduce individual characteristics as opposed to softer materials, such as brass. This can reduce the production of individual characteristics and introduce confusion. Similarly, the firearms chosen have been shown to have the propensity for subclass characteristics [28-30]. Subclass characteristics are markings that are incidental to manufacture and change over time due to tool wear in the manufacturing process [31, 32]. Identifications can not be made on the basis of subclass markings as a number of barrels and / or slides may generate similar markings. The collection of barrels and slides during the manufacturing process close and far apart in the manufacturing order of the equipment allowed for comparisons that would test the ability of examiners to identify and exclude such subclass markings from their examinations.

Labeling and Tracking

Sample labels were generated by the FBI and provided to Ames Lab. Labels were printed that contained encoded Data Matrix two-dimensional bar codes having random and unique alphanumeric identifiers, on easy-peel label adhesive paper. The use of two-dimensional barcodes made it more difficult for participants to identify samples in the third and subsequent mailings of the study if provided again for repeatability testing. Each printed sheet contained 100 labels - 66 of these included a “K” on the label and were used for known samples, while the remaining 34 (with no additional letter) were used for questioned samples. (*N.B. While questioned samples were not further designated with a Q on the label, they are so identified in subsequent tables that describe how test packets were assembled.*) This was done so that if an examiner were to mix up the K’s and Q within a comparison set during analysis, the samples could still be distinguished. Nine sheets of labels were printed for each firearm and for each cartridge case or bullet examined.

Within a collection of 50 cartridge cases in an ammunition box, 33 of the cartridge cases were labeled as K’s with the remaining 17 as Q’s, by manually affixing labels to each individual cartridge case. Care was taken to place the label on an area on the cartridge case that had a minimum amount of marks present and in the same orientation on each cartridge case. Figure 5 shows an example of some labeled cartridge cases. As cartridge cases were labeled they were placed back into their original 50-holed plastic container to maintain the group firing-order information.

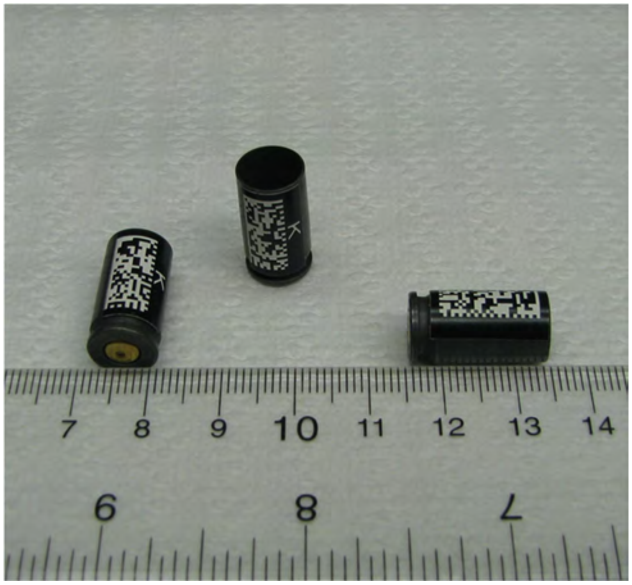


Figure 5: Labeled cartridge cases; two known (K) cartridge cases and one questioned cartridge case are shown.

Once all the ammunition boxes of cartridge cases were labeled for a particular firearm they were returned to their original cardboard shipping boxes. Throughout the labeling process only samples from one gun and from one box were handled at a time to ensure samples would not be erroneously mixed between boxes.

After the cartridge cases were labeled, they were inventoried by barcode-reading the labels using a Cognex DataMan 260 attached to a solid support, Figure 6, which allowed for fast and consistent reads. The information linked to each sample, in addition to the captured 2-D alphanumeric identifier, included the sample type (cartridge case or bullet), serial number of the firearm, firing-order range within 50 shots fired, and specimen designation (K or Q). All the linked information was saved to text files and converted into Excel files, so that for each ammunition box of labeled cartridge cases a file listing 33 K's and 17 Q's and their identifying alphanumeric barcodes was created. After all the boxes for a particular firearm were barcode-read, the individual Excel files were merged so that all the barcode labels for a particular firearm were combined into one file. Additional information for each sample added to each combined inventory file included early (E), middle (M), or late (L) firing-order designations and lot-number information, as discussed below. A cartridge case master list was created by combining all the Jimenez and Beretta firearm inventory files into one file. Additional details regarding the barcode reader, the interface, and the tracking of samples are provided in Appendix D.

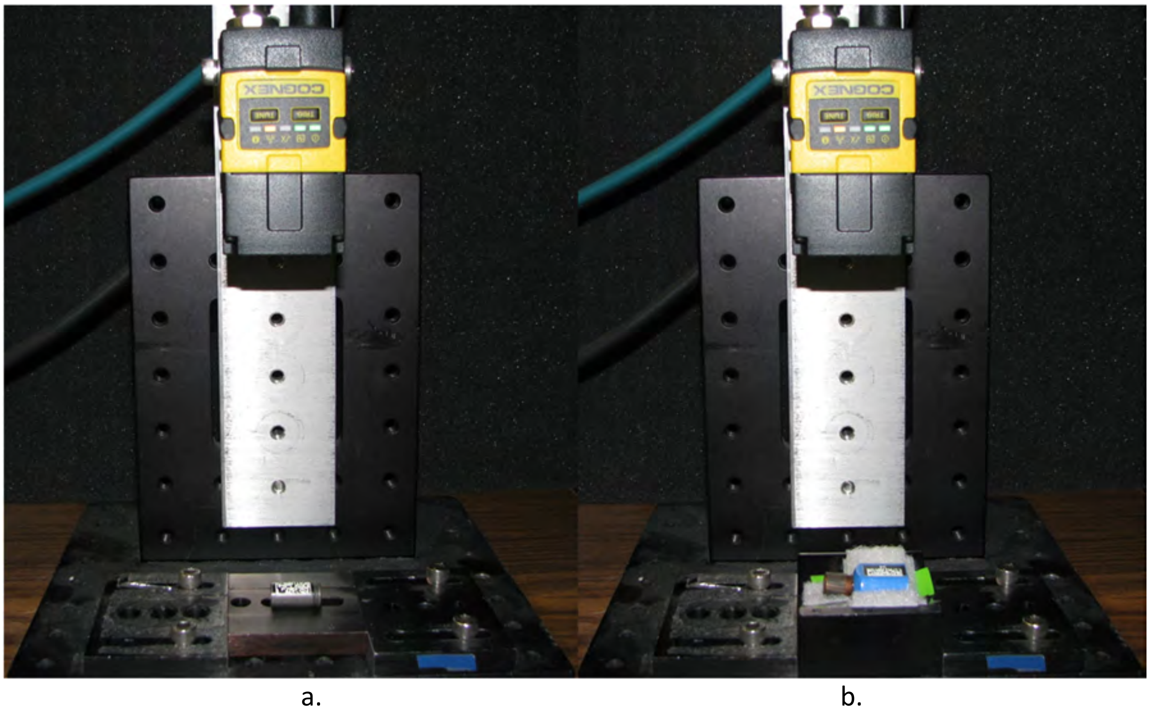


Figure 6: Cognex DataMan 260 Barcode reader used for inventorying samples. The photos show the cradles that were used for a) cartridge cases and b) mounted bullets, respectively.

The 2-D labels were too large to be placed directly on the bullets. Thus, to permit labeling, bullets were first epoxied to plastic sample mounts. Two colors, blue for Known samples and white for Q samples, were used. A specially-made plexiglass jig provided by the FBI was used to facilitate epoxying bullets to the plastic mounts, Figure 7. The base of the jig consisted of a plexiglass plate that had 50 recessed 2-cm holes that allowed as many as 50 plastic mounts to be positioned in the jig. The top of the jig was a plexiglass cover plate that had 50 holes with diameters slightly larger than 9-mm, that allowed bullets to fit through the cover to the mounts and supported the bullets in an upright position. The holes in the base and cover plates were center-aligned. Eight plexiglass jigs were provided by the FBI to use for bullet-mount epoxying.

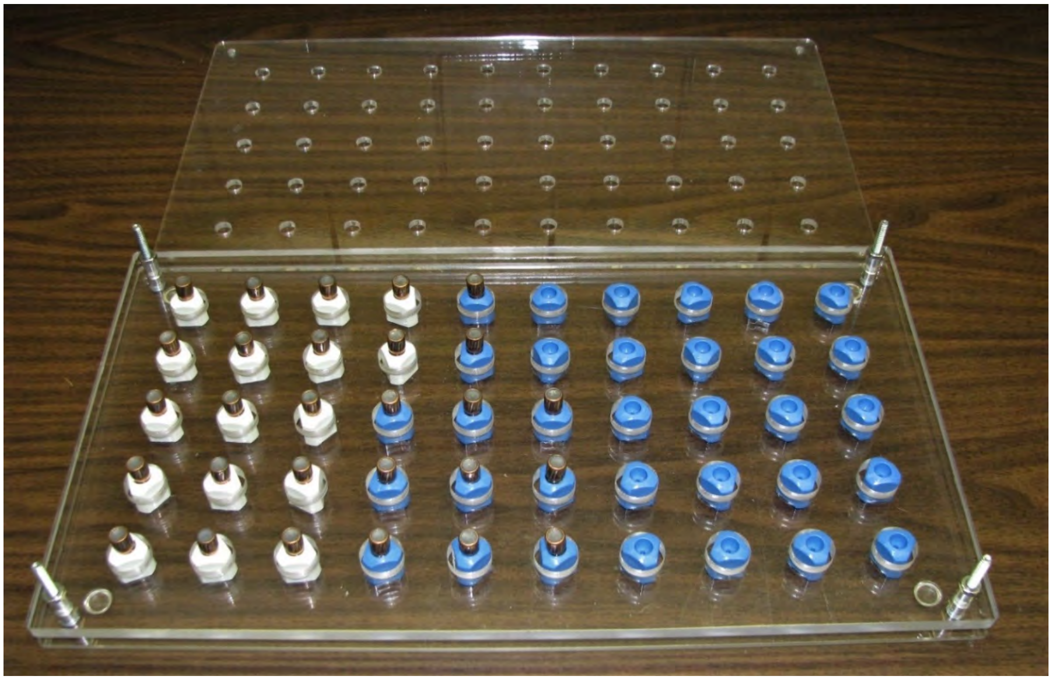


Figure 7: Plexiglass jig used for epoxying bullets to plastic mounts. The cover plate (not attached) is shown above the base that holds the plastic mounts. The left half of the jig has bullets already epoxied to the mounts, for illustrative purposes.

Bullets for one firing-order range of 50 samples, from one labeled manila envelope for a given firearm, were handled one at a time in order to prevent possible errors mixing firing ranges and samples. Mounts were positioned in the base of the jig and a small amount of a fast-curing two-part epoxy, DevCon 5-Minute Epoxy, was applied in the center depression in each mount. The appropriate amount of epoxy was applied to sufficiently secure the bullets to the mounts without the epoxy spreading and covering any land or groove markings on the bullets. The cover plate was then positioned over the mounts and 49 bullets were quickly placed, nose (pointed end) first, through the cover into the mounts. The epoxy was allowed to cure for about 2 to 3 hours before the mounts were removed from the jig, and the mounted bullets were subsequently stored in a small cardboard box that was labeled with the gun serial number and firing-order sequence. During this process all the bullets for a single firearm were epoxied to mounts in successive batches before bullets from another firearm were used; this was done to reduce the possibility of error in the mounting procedure.

For each firing-order range of 50 bullet samples only 49 were actually mounted, due to a shortage of white plastic mounts. Thus, each mounted firing-order range consisted of 33 known and 16 questioned samples. The mounted bullets were labeled with 2-D labels, placing K-labels on the blue known mounts and non-K-labels on the white questioned mounts, Figure 8. Care was taken to place the label in the same orientation on each mount. The bullets were subsequently inventoried by barcode-reading the labels, as described above for cartridge cases. The alpha-numeric identifier and additional linked information was saved to an electronic inventory file, as before. The separate inventory files were combined to create a master list for all the labeled-bullet samples.

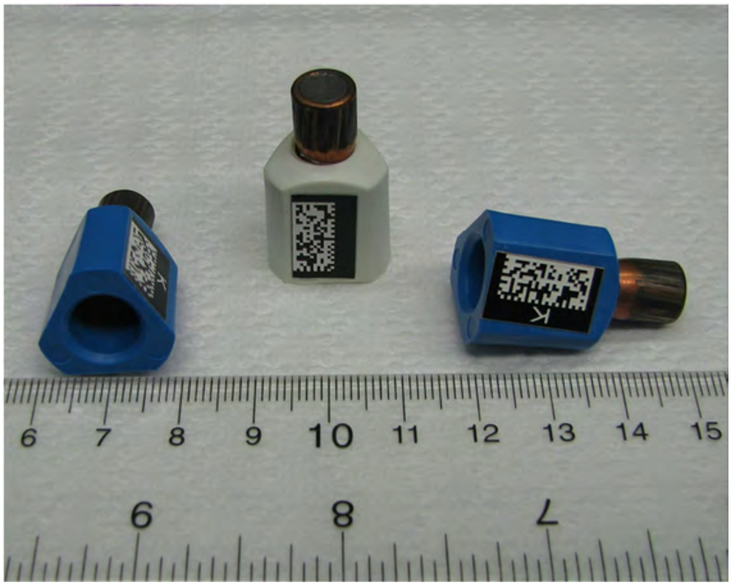


Figure 8: Mounted and labeled bullets, showing known samples epoxied to blue mounts and a questioned sample to a white mount.

The cartridge case and bullet master lists were checked to determine if there were any duplicate alphanumeric identifiers on any of the samples. The samples were then ready to be assembled into comparison sets and test packets.

Comparison Sets

Assembly

Each test packet that a firearm examiner received consisted of 30 comparison sample sets made up of 15 comparisons of 2 knowns to 1 questioned cartridge case and 15 comparisons of 2 knowns to 1 questioned bullet. The cartridge case comparisons consisted of 5 sets of Jimenez and 10 sets of Beretta cartridge cases. Bullet comparison sets were comprised of 5 sets of Ruger and 10 sets of Beretta bullets. The overall ratio of known same-source firearms to known different-source firearms was approximately 1 to 2 for both cartridge case and bullet comparison sets, but varied among the test packets. Participants were instructed not to share or discuss the contents of their packets or their reported results to minimize the risk of revealing details of the experimental design. Details of the design were not shared with anyone outside the group of researchers assembling the comparison sets and FBI project managers. Only those researchers in the experimental/analysis group at Ames Lab knew the ground-truth for the assembled test packets and the matrix devised was designed to make it difficult, if not impossible, for an examiner to deduce the correct test results.

The arrangement of known match and known nonmatch comparison sets in the first 25 test packets for cartridge cases and bullets is shown in Table I. The arrangement described below in detail was used to reduce the possibility that participants might deduce a pattern in the total number of known-match sets in a packet, confer, and therefore affect subsequent analyses. This 25-packet sequence was repeated in assembling the test packets of cartridge cases and bullets for analysis as needed to produce the required number of packets examined in this study.

Each packet shown in Table I has between 3 and 7 known-match cartridge case and between 3 and 7 known-match bullet sets, so each also has between 8 and 12 known nonmatch cartridge case and bullet sets. Within each group of 5 packets, the numbers of known-match sets vary from 0-4 Jimenez and from 1-5 Beretta cartridge case sets (top half of Table I), and from 0-4 Ruger and from 1-5 Beretta bullet sets (bottom half of Table I), such that each packet has between 3 and 7 known-match cartridge case and between 3 and 7 known-match bullet sets. The numbers of known-match cartridge case and known-match bullet sets are offset with respect to each other, in order to vary the total number of known-match sets in the packets. Each group of 5 cartridge case packets has 7, 6, 5, 4, and 3 known matches, respectively, in that same order, for all five groups of 5 packets. In contrast, the ordering of the numbers of known-match bullet sets for the five groups of 5 packets changes - 6, 5, 4, 3, and 7 known matches for the first group of 5 packets - 5, 4, 3, 7, and 6 known matches for the second group of 5 packets - and so on, as shown in Table I. Thus, the total number of known-match sets ranges from 6 to 14 but an overall ratio of 1:2 known match to known nonmatch sets for the group of 25 packets is maintained. In addition, all 25 possible permutations of the combination of 3-7 known-match cartridge case with 3-7 known-match bullet comparison sets is achieved, for these packets.

Table I: Numbers of Known Same-Source and Known Different-Source Comparison Sets in Test Packets.

	Test Packet																								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
Cartridge Cases																									
Known Match	7	6	5	4	3	7	6	5	4	3	7	6	5	4	3	7	6	5	4	3	7	6	5	4	3
Known Non-match	8	9	10	11	12	8	9	10	11	12	8	9	10	11	12	8	9	10	11	12	8	9	10	11	12
Jimenez Match	4	1	3	0	2	4	1	3	0	2	4	1	3	0	2	4	1	3	0	2	4	1	3	0	2
Jimenez Non-match	1	4	2	5	3	1	4	2	5	3	1	4	2	5	3	1	4	2	5	3	1	4	2	5	3
Beretta Match	3	5	2	4	1	3	5	2	4	1	3	5	2	4	1	3	5	2	4	1	3	5	2	4	1
Beretta Non-match	7	5	8	6	9	7	5	8	6	9	7	5	8	6	9	7	5	8	6	9	7	5	8	6	9
Bullets																									
Known Match	6	5	4	3	7	5	4	3	7	6	4	3	7	6	5	3	7	6	5	4	7	6	5	4	3
Known Non-match	9	10	11	12	8	10	11	12	8	9	11	12	8	9	10	12	8	9	10	11	8	9	10	11	12
Ruger Match	1	3	0	2	4	3	0	2	4	1	0	2	4	1	3	2	4	1	3	0	4	1	3	0	2
Ruger Non-match	4	2	5	3	1	2	5	3	1	4	5	3	1	4	2	3	1	4	2	5	1	4	2	5	3
Beretta Match	5	2	4	1	3	2	4	1	3	5	4	1	3	5	2	1	3	5	2	4	3	5	2	4	1
Beretta Non-match	5	8	6	9	7	8	6	9	7	5	6	9	7	5	8	9	7	5	8	6	7	5	8	6	9

Test packets and comparison sets were assembled using the following parameters:

1) A comparison set consisted of a single Questioned sample to be compared to two Knowns, the latter fired from the same firearm.

2)Only cartridge cases and bullets fired from the same make and model firearm were compared. This was not revealed to the examiners. There were no comparisons of samples from different manufacturers in any set of two known and one questioned item.

3) Each set represents an independent comparison unrelated to any other set in the test packet.

4) An open-set design was utilized, i.e., there was not necessarily a match for every Questioned sample.

5) The overall proportion of known matches in a test packet was approximately 33% but varied across test packets, as described earlier and shown in Table I.

6) The ratio of non-Beretta to Beretta specimens (for cartridge cases and bullets) in a test packet was 1:2.

Matrix templates were created and populated reflecting the parameters outlined above to aid in the generation of test packets and comparison sets, for both cartridge cases and bullets. The dimensions of the complete matrices to generate all the test packets were 15 rows, representing 15 comparison sets, by 600 columns, to accommodate a total of 600 possible test packets. This number was based on the initial assumption that as many as 300 examiners would participate in the study, with each examiner receiving two packets of new or initial samples, i.e., samples that had not been previously seen or analyzed by any examiner, in the first round of the study.

Small portions of the matrices are shown in Table II, for the first 10 test packets displayed in Table I, for both cartridge cases and bullets. The orange and blue color-coding of the cells matches the numbers of known match and known nonmatch comparison sets for the Jimenez and Beretta cartridge cases (top) and for the Ruger and Beretta bullets (bottom). Test packet 1 in Table II consists of 4 known-match Jimenez and 3 known-match Beretta cartridge case sets (orange cells, top) and 1 known-match Ruger and 5 known-match Beretta bullet sets (orange cells, bottom). The remaining cells for sample packet 1 in Table II are blue, indicating known nonmatch comparison sets for cartridge cases and bullets. The color-coding for the remaining packets in Table II shows the variation in numbers of known match and known nonmatch comparison sets, for cartridge cases and bullets, for test packets 2-10. The offset in the numbers of known-match cartridge case and known-match bullet sets for the two groups of 5 test packets is also evident in Table II, from the color-coding of the cells. For cartridge cases, the pattern is the same for packets 1-5 and 6-10; for bullets, the pattern is shifted so that packets 2-6, 3-7, 4-8, 5-9, and 1-10 have the same color-coding.

Cartridge cases fired from the 11 Jimenez and 27 Beretta slides were assigned either a number or letter, being 1 through 11 for Jimenez and A through AA for Beretta. Similarly, bullets from the 11 Ruger barrels were assigned numbers 1-11 and those from the Beretta barrels letters A-AA. Sample-set pairings for cartridge case and bullet specimens were established by populating the matrices in Table II in a serial fashion using these number and letter designations. Each test packet in Table II is split into K- and Q-columns, to show which specimens are the 2 Knowns and 1 Questioned samples in a comparison set.

To illustrate how the matrices in Table II were populated with assigned gun numbers and letters, consider packet 1. Cartridge cases from Jimenez guns 1-4 fill the four orange-shaded known match K- and Q-cells, and cartridge cases from guns 5 (K) and 6 (Q) fill the blue-shaded known nonmatch cells. Cartridge cases from Beretta slides A-C fill the three orange-shaded known match K- and Q-cells, with cartridge cases from guns D-J (K) and K-Q (Q), respectively, serially filling the remaining blue-shaded nonmatch cells. For bullets, Ruger gun 1 bullets fill the one known match K- and Q-cells, with bullets from guns 2-5 (K) and 6-9 (Q) serially filling the known nonmatch cells. Bullets from Beretta guns A-E fill the five known match K- and Q-cells, with bullets from guns F-J (K) and K-O (Q) serially filling the remaining known nonmatch cells.

This sequence of assigning gun numbers and letters for selection of cartridge cases and bullets continued until the matrices were filled, starting each successive test packet with the next number or letter in the sequence. For cartridge cases, since packet 1 ended with Jimenez gun 6 and Beretta gun Q, packet 2 started with Jimenez gun 7 and Berretta gun R. For bullets, packet 1 ended with Ruger gun 9 and Beretta gun O, so packet 2 started with Ruger gun 10 and Beretta gun P. This sequencing continued, so that the starting "K-guns" for Set 1 (Jimenez cartridge case), Set 6 (Beretta cartridge case), Set 16 (Ruger bullet), and Set 21 (Beretta bullet) vary, as shown in Table II. No firearm was included in more than one Q or K-pair in a set.

The serial numbers and assigned identifiers used for each handgun in this study are listed in Appendix E. Note that the assigned letter and number identifiers for the guns do not correlate to cartridge case and bullet samples generated from the same firearm; an identifier used for a slide on a particular firearm will be different from the identifier used for the barrel of that same firearm. This is evident by examination of the tables provided in Appendix E. The total possible K-Q pairings generated using these matrix combinations are also shown in Appendix E. The collection of possible pairings is the same for cartridge cases and bullets. For each individual firearm, sample cartridge case and bullet specimens were compared to samples from 5 different nonmatching firearms.

Table II: Sample Distribution Lists for Cartridge Cases and Bullets. The orange and blue color-coding of the cells indicate the numbers of known match and known nonmatch comparison sets (respectively) for the Jimenez and Beretta cartridge cases (top) and for the Ruger and Beretta bullets (bottom).

		Test Packet																			
		1		2		3		4		5		6		7		8		9		10	
	Set #	**K**	**Q**	**K**	**Q**	**K**	**Q**	**K**	**Q**	**K**	**Q**	**K**	**Q**	**K**	**Q**	**K**	**Q**	**K**	**Q**	**K**	**Q**
Jimenez Cartridge Cases	1	1	1	7	7	5	5	1	6	11	11	8	8	3	3	1	1	8	2	7	7
	2	2	2	8	1	6	6	2	7	1	1	9	9	4	8	2	2	9	3	8	8
	3	3	3	9	2	7	7	3	8	2	5	10	10	5	9	3	3	10	4	9	1
	4	4	4	10	3	8	10	4	9	3	6	11	11	6	10	4	6	11	5	10	2
	5	5	6	11	4	9	11	5	10	4	7	1	2	7	11	5	7	1	6	11	3
Beretta Cartridge Cases	6	A	A	R	R	F	F	X	X	M	M	E	E	V	V	J	J	A	A	Q	Q
	7	B	B	S	S	G	G	Y	Y	N	W	F	F	W	W	K	K	B	B	R	AA
	8	C	C	T	T	H	P	Z	Z	O	X	G	G	X	X	L	T	C	C	S	A
	9	D	K	U	U	I	Q	AA	AA	P	Y	H	O	Y	Y	M	U	D	D	T	B
	10	E	L	V	V	J	R	A	G	Q	Z	I	P	Z	Z	N	V	E	K	U	C
	11	F	M	W	A	K	S	B	H	R	AA	J	Q	AA	E	O	W	F	L	V	D
	12	G	N	X	B	L	T	C	I	S	A	K	R	A	F	P	X	G	M	W	E
	13	H	O	Y	C	M	U	D	J	T	B	L	S	B	G	Q	Y	H	N	X	F
	14	I	P	Z	D	N	V	E	K	U	C	M	T	C	H	R	Z	I	O	Y	G
	15	J	Q	AA	E	O	W	F	L	V	D	N	U	D	I	S	AA	J	P	Z	H

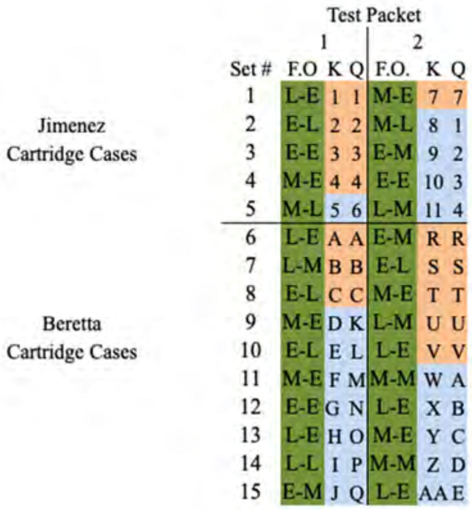
		Test Packet																			
		1		2		3		4		5		6		7		8		9		10	
	Set #	**K**	**Q**	**K**	**Q**	**K**	**Q**	**K**	**Q**	**K**	**Q**	**K**	**Q**	**K**	**Q**	**K**	**Q**	**K**	**Q**	**K**	**Q**
Ruger Bullets	16	1	1	10	10	6	11	5	5	2	2	8	8	4	9	3	3	11	11	6	6
	17	2	6	11	11	7	1	6	6	3	3	9	9	5	10	4	4	1	1	7	11
	18	3	7	1	1	8	2	7	10	4	4	10	10	6	11	5	8	2	2	8	1
	19	4	8	2	4	9	3	8	11	5	5	11	2	7	1	6	9	3	3	9	2
	20	5	9	3	5	10	4	9	1	6	7	1	3	8	2	7	10	4	5	10	3
Beretta Bullets	21	A	A	P	P	G	G	W	W	O	O	E	E	W	W	L	L	D	D	U	U
	22	B	B	Q	Q	H	H	X	F	P	P	F	F	X	X	M	V	E	E	V	V
	23	C	C	R	Z	I	I	Y	G	Q	Q	G	O	Y	Y	N	W	F	F	W	W
	24	D	D	S	AA	J	J	Z	H	R	Y	H	P	Z	Z	O	X	G	N	X	X
	25	E	E	T	A	K	Q	AA	I	S	Z	I	Q	AA	F	P	Y	H	O	Y	Y
	26	F	K	U	B	L	R	A	J	T	AA	J	R	A	G	Q	Z	I	P	Z	D
	27	G	L	V	C	M	S	B	K	U	A	K	S	B	H	R	AA	J	Q	AA	E
	28	H	M	W	D	N	T	C	L	V	B	L	T	C	I	S	A	K	R	A	F
	29	I	N	X	E	O	U	D	M	W	C	M	U	D	J	T	B	L	S	B	G
	30	J	O	Y	F	P	V	E	N	X	D	N	V	E	K	U	C	M	T	C	H

Tracking of Firing Order

Cartridge cases and bullets for each firearm were also divided into three groups, designating the firing order of the collected samples as early (E), middle (M), or late (L). For the Jimenez cartridge cases and the Ruger bullets, 17 boxes of 50 samples were collected; the E-M-L designations assigned were E for the first 6 boxes, M for the next 5 boxes, and L for the last 6 boxes. For the Beretta cartridge cases and bullets, 14 boxes of 50 samples were collected; for these the E-M-L designations were assigned to the first 5, next 4, and last 5 boxes, respectively. The EML firing order was tracked to enable possible conclusions to be drawn as to the effect of firearm wear on examiners' analysis results. A listing of firing order ranges for the samples used is shown in Appendix Table E1.

To determine whether the EML firing order of K- and Q-specimens affected examiners' analyses, comparison sets for cartridge cases and bullets in the test packets were randomly assigned one of the nine possible pairings (E-E, E-M, E-L; M-E, M-M, M-L; L-E, L-M, L-L), where the first letter for these combinations indicates the firing order for two K's, and the second letter the firing order for the Q sample in the set. Table III shows the firing-order EML assignments in the green-highlighted columns for just the cartridge case comparison sets of the first two packets shown in Table II. When test packet 1 was assembled the known-match cartridge cases included for comparison set 1 were two Late K's and 1 Early Q, from the collection of Jimenez gun 1 cartridge cases. For comparison set 2, two Early K's and 1 Late Q, from Jimenez gun 2 were selected and included; comparison set 3 had two Early K's and 1 Early Q from Jimenez gun 3; and so on, for the remaining cartridge case comparison sets in packets 1 and 2. In a similar fashion, randomized EML firing-order assignments for the group of 15 Ruger and Beretta bullet comparison sets in each test packet were used when assembling the sample sets to be analyzed.

Table III: Random Firing-Order Pairings for Cartridge Case Test Packets 1 and 2.



		Test Packet					
		1			2		
	Set #	F.O	K	Q	F.O.	K	Q
Jimenez Cartridge Cases	1	L-E	1	1	M-E	7	7
	2	E-L	2	2	M-L	8	1
	3	E-E	3	3	E-M	9	2
	4	M-E	4	4	E-E	10	3
	5	M-L	5	6	L-M	11	4
Beretta Cartridge Cases	6	L-E	A	A	E-M	R	R
	7	L-M	B	B	E-L	S	S
	8	E-L	C	C	M-E	T	T
	9	M-E	D	K	L-M	U	U
	10	E-L	E	L	L-E	V	V
	11	M-E	F	M	M-M	W	A
	12	E-E	G	N	L-E	X	B
	13	L-E	H	O	M-E	Y	C
	14	L-L	I	P	M-M	Z	D
	15	E-M	J	Q	L-E	AA	E

Tracking of Samples from Sequentially-Manufactured Components

It is known that bullets fired from sequentially-manufactured barrels produced using the same tool can result in the introduction of sub-class characteristics common to both barrels that can complicate source identification [33]. In the manufacture of a firearm pistol barrel a broaching process is the method often used and was used for the firearms employed in this study. A typical machining tool used in this process may last for the manufacture of hundreds of barrels, although shorter runs are more common. While a series of different broaches are used to produce the final barrel diameter, only patterns from the final broach tool that imparts the tooling marks transferred to the bullet are of interest. In addition to the influence of sub-class over numerous test fires, there may be considerable differences in sub-class characteristics between barrels made when this tool is first used versus barrels made near the end of the broach life. In order to obtain data related to this effect, sequentially-manufactured Beretta barrels were collected from the beginning, middle, and end of a single broach's lifetime. During this manufacturing run 66 barrels were made before the broach was exchanged. The Beretta barrels used in this study were partitioned into 5 groups for collections of barrels sequentially produced at different intervals in the lifetime of the broach. The manufacturing sequence of the barrel groups is shown in Appendix Table E2, where barrels produced within the same manufacturing interval are color-coded in the table. This allowed closest known different-source comparison sets of bullets (i.e., bullets from two different barrels made in close temporal sequence) to be included in the assembled test packets, to study the possible effect that sub-class characteristics that originate from the manufacturing process have on examiners' reported results. Samples fired using these barrels account for 22% of the total nonmatched Beretta bullet comparison sets analyzed.

The Ruger barrels used in this study were also produced by a broaching process and were consecutively manufactured. Thus, they too have the potential of subclass characteristics being introduced. The Ruger manufacturing sequence is indicated by the barrel serial number and is also shown in Appendix Table E2.

Similarly, the Beretta slides used in this study were consecutively hand finished. The Beretta slides were also partitioned into 5 groups for collections of slides sequentially produced at different intervals in the production process. The manufacturing sequence of the Beretta slide groups is shown in Appendix Table E3 and is similar to the barrel sequence, where slides produced within the same manufacturing interval are color-coded in the table. Samples fired using these slides correspondingly account for 22% of the total nonmatched Beretta cartridge case comparison sets analyzed. The Jimenez slides manufacturing sequence is indicated by the slide serial number and is shown in Appendix Table E3.

Preparation of Sample Packets

Given the information and experimental design described above, distribution tables were created for cartridge cases and for bullets that defined the sample-set pairings to be used for assembling the test packets. These distribution tables listed the guns designated for match and nonmatch sets, as well as the EML firing sequence ranges of the fired cartridge cases and bullets. The original experimental design called for 600 test packets to be assembled initially, but this was reduced to 480 to accommodate the number of examiners that agreed to participate. For the first mailing 256 packets were sent to the examiners who initially volunteered. The remaining new packets were mailed to examiners over the course of mailings 2-4. When packets were returned, the reported results were scored and recorded, and then sample sets were repackaged for use in subsequent mailings to test repeatability and reproducibility. Reproducibility test packets were mailed to examiners in mailings 2-6, and repeatability packets were mailed in mailings 3-6. At least one mailing separated the accuracy and repeatability analyses done by an examiner for a given packet.

The comparison set pairings within a given test packet were maintained throughout the study, however, new randomly-selected set numbers (see, http://www.random.org/lists/) were assigned before test packets were sent out for the repeatability and reproducibility analyses. New set numbers for cartridge cases and bullets were used so that the order of same-source-known-match and different-source-known-nonmatch sets was random for each test packet analyzed. As a result, no identifiable trend was discernable by an examiner analyzing a particular packet. Based on the study design, a given test packet could be analyzed as many as three times, by two different examiners - one that did the accuracy and repeatability analyses and one who performed the reproducibility analysis.

Throughout the course of the study, due to drop-out of participating examiners, 397 of the 480 original packets assembled were mailed. Examiners returned packets with analysis results for 288 of these. For the second and third rounds of the study, where repeatability and reproducibility were assessed, 189 and 191 packets, respectively, were returned. Of the 288 test packets analyzed, 138 were analyzed three times, resulting in data used for all three aspects of the study (accuracy, repeatability, and reproducibility); 108 were analyzed twice, contributing to accuracy and either repeatability or reproducibility analysis; and 42 were analyzed once and were used only in the accuracy analysis.

Mailing, Receiving, and Scoring of Sample Packets

Assembly

Using the distribution tables described in the preceding sections, individual cartridge cases and bullets were gathered into the proper sets and test packets. Once a packet was assembled, the cartridge case and bullet samples in the comparison sets were barcode-read again and verified against their respective distribution lists. Each assembled test packet was assigned a unique group number that was used as a primary sample-identifier code for tracking purposes by the Ames Lab experimental/analysis group, that was included on the individually-labeled bullet and cartridge case set bags, the survey form, and the answer sheets sent to examiners. The primary identifiers were not made known to the Ames Lab communication group.

Sample packets assembled consisted of the 30 bullet and cartridge case comparison sets to be analyzed (15 each), an instruction sheet, and answer sheets; the 3-page survey form was included in the first mailing, Figure 9. These items were placed into a Tyvek envelope, sealed, and labeled with a secondary examiner-identifier code, that was linked to the unique primary group number. The secondary code was a 2-letter examiner identifier - one for each of the participants - with a third letter or number added to indicate the mailing. A return Tyvek envelope, also labeled with the secondary-identifier code, was included in the assembled packet and used by examiners to return the samples and their analysis-results sheets. The examiner-identifier codes were used by the Ames Lab communication group to track shipping and receiving of sample packets; these codes linked packets to specific participants.

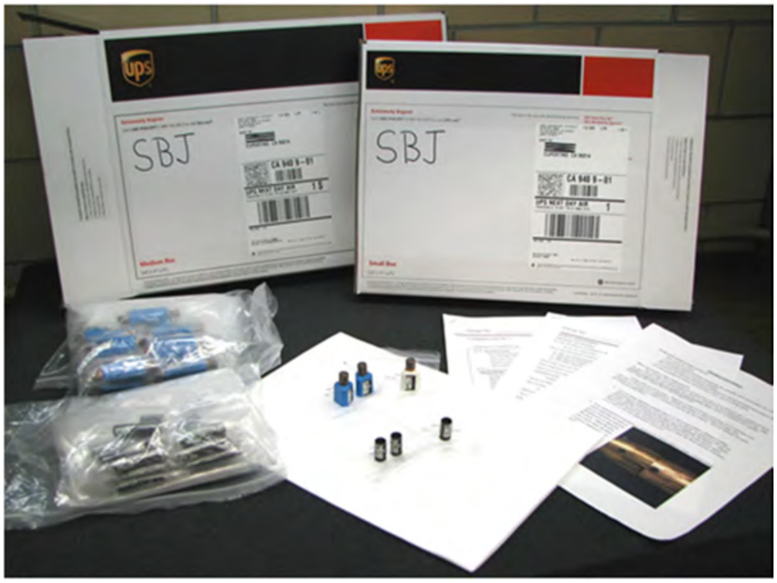


Figure 9: Sample packet components, including the bullet and cartridge case test sets, forms, and shipping and return boxes.

Handling and Shipping

The Tyvek-sealed envelopes were transferred by the experimental/analysis group to the communication group, who were responsible for shipping and receiving. The communication group never opened the sealed Tyvek envelopes nor inspected any of the contents. The communication group labeled the shipping and return boxes with the secondary-identifier codes (as well as the shipping and return addresses to be used), prior to mailing packets to examiners. Packets returned to Ames Lab were inspected upon arrival by the communication group for any examiner-specific identifying information, prior to transferring the sealed Tyvek bag containing the analysis results to the experimental/analysis group for scoring, database entry, and verification of the results.

Use of the secondary examiner-identifier codes provided the double blind separation between the Ames Lab experimental/analysis group (who knew the primary sample-identifier codes, ground-truth information for the assembled test packets, and the examiners' results) and the communication group (who knew examiner contact information). The only commonality between the two groups was the three-letter secondary codes, which were used for the different purposes delineated above, by each group. Thus, there was no link between examiners' identities and their analysis results.

When a participating examiner withdrew from the study, they typically made this decision known to the communication group, who then shared this information with the experimental/analysis group using the three-letter secondary identifier. In some cases, withdrawal by an examiner was signaled by the return of an unanalyzed test packet; in this case the secondary identifier was shared by the experimental / analysis group with the communication group, who then verified the withdrawal of the examiner either by e-mail or a phone call. If verified, no additional test packets were assembled for that examiner. At the conclusion of data collection, the communication group destroyed the examiner-identifier codes linked to examiners' identities. Through the course of the study, attrition continually reduced the number of participating examiners and the number of sample packets sent out in subsequent mailings.

A summary of the overall numbers is shown in Figure 10. Note that the number of scored packets plus the number of dropouts per mailing equals the number returned.

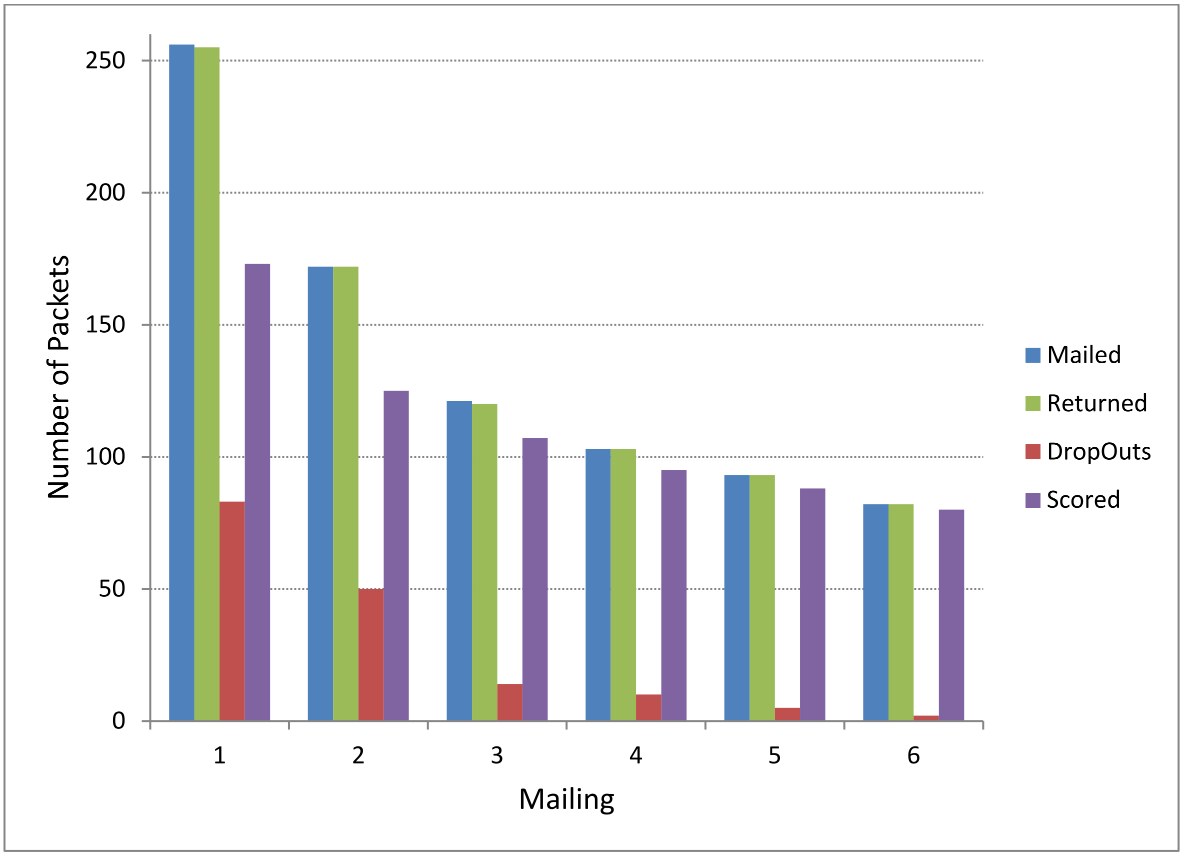


Figure 10: Numbers of test packets shipped, received, and scored, and the examiners that dropped-out, in the six mailings of the study.

Scoring

Answer sheets were scored, comparing examiners' reported classifications (Identification, Elimination, Inconclusive, or Unsuitable) for the analyzed sets to the ground-truth information, for each test packet returned. One member of the Ames Lab experimental/analysis group entered these results along with the additional information included on each answer sheet into an Access database file; a second person verified the accuracy of the entries. If a classification error was made, for example, a false-negative or false-positive error, the erroneous set was barcode-read and the information compared to that in the corresponding bullet or cartridge case distribution table to verify the error. For bullet false-positive errors, the lands or grooves on the K's and Q samples that the examiner used for identification, indicated by Sharpie marks (as directed in the instruction sheet), were recorded, photographed, and imaged using a Leica UFM4 comparison microscope for later analysis by a qualified examiner if desired by the funding agency.

Analyzed test packets that were needed in the second or third rounds of the study were repackaged. Each specimen in each comparison set was visually examined and gently cleaned of any debris or marks.

The K's and Q for a particular set remained together, but new random set numbers and group numbers were used. The repackaged test packet was barcode-read once again, to verify the ground-truth of the reassembled packet, prior to use in subsequent mailings.

Experimental Results

Overview of Data Sets and Reported Results

For purposes of material handling, test packets were constructed, each composed of 15 bullet comparison sets and 15 cartridge case comparison sets. A single mailing of material to a single examiner consisted of one test packet. The definition of test packets persisted throughout the study, i.e. comparison sets were not recombined in different combinations when returned test packets were distributed again to the same or different examiners. The overall design of the study was organized in three rounds, referenced by number (1, 2, and 3) in this report.

The 173 participating firearm examiners provided analysis results for a total of 668 test-packets, for cartridge cases and bullets, in the first (288), second (189) and third (191) rounds of this study. The total number of sample-set comparisons reported is 20,130. Slightly more cartridge case (10,110) than bullet (10,020) sets are reported because a small number of examiners returned partially-completed test packets that had results for all the cartridge case sets but not all the bullet sets. Specifically, in all three rounds of the study, 13 examiners returned bullet sets with incomplete evaluations, eight examiners returned cartridge case sets with incomplete evaluations, with one of these examiners returning incomplete evaluations for both kinds of sets. Examiners were asked to render a decision for each individual comparison set analyzed as either Identification (ID), Elimination, Inconclusive, or Unsuitable. Inconclusive analyses were categorized using the three AFTE choices [Figure 1], and three options for classifying Unsuitable comparison sets, namely, whether the K's or Q samples were unsuitable or if either of these were missing or damaged, were included. In addition, examiners were asked to report on a number of other factors associated with their examination. These included:

- the number of knowns from the two provided with sufficient reproduced detail for comparison [0, 1, or 2]
- the relative difficulty of the comparison [easy, average, or hard]
- the level of individual characteristics that were available [extensive, some, or limited]
- whether consecutive matching striae (CMS) were used in the analysis [yes / no]
- the amount of time required to conduct the comparison [estimated]

Appendix F contains various data acquired by the study including some of the bulleted items listed above. For Identifications, examiners were asked to indicate the areas used in making their decisions, with five areas for cartridge cases being possible (breech face marks, firing pin impression, chamber marks, extractor marks, and ejector marks) and two for bullets (land and groove impressions).

The numbers of bullet and cartridge case comparison sets analyzed in the six mailings are shown in Table IV. For the purpose of study objectives, the results from each comparison were included in one of the three rounds of the study, which were used in combinations to assess Accuracy, Repeatability, and Reproducibility. It is important to note the difference between "mailings" and "rounds" of the study as referred to in Table IV and the subsequent statistical analysis of the data. The six mailings pertain to the physical management, distribution, and collection of samples and results from examiners. Rounds are more directly related to study objectives. Examinations in Round 1 were the initial examinations of each comparison set. Comparison sets examined in Round 2 had been examined by the same examiner in Round 1. Comparison sets examined in Round 3 had been examined by a different examiner in Round 1. As described below, the data generated in different combinations of rounds were used to address the specific goals in this study. Note that data from partially-completed packets included as Drop-Outs in Figure 10 above are included in the results presented below.

Table IV: Numbers of Bullet and Cartridge case Sets Analyzed by Mailing and Round.

Comparison Sets Analyzed by Mailing and by Round							
	Mailing 1	Mailing 2	Mailing 3	Mailing 4	Mailing 5	Mailing 6	Total
Bullet							
Round 1	2,595	960	345	420	-	-	4,320
Round 2	-	-	750	765	645	675	2,835
Round 3	-	915	510	240	675	525	2,865
Total	2,595	1,875	1,605	1,425	1,320	1,200	10,020
Case							
Round 1	2,595	960	345	420	-	-	4320
Round 2	-	-	750	780	645	690	2865
Round 3	-	945	540	240	675	525	2925
Total	2,595	1,905	1,635	1,440	1,320	1,215	10,110

Analysis of Examiner Performance

The analyses focused on accuracy, repeatability, and reproducibility that follow are based on the subsets of the data defined as follows:

1) Accuracy is defined as the ability of an examiner to correctly identify a known match or eliminate a known nonmatch. The data used for this analysis includes only those evaluations made in the first round of the study.

2) Repeatability is defined as the ability of an examiner, when confronted with the exact same comparison once again, to reach the same conclusion as when first examined. The data used for this analysis includes only evaluations made in the first and second rounds of the study.

3) Reproducibility is defined as the ability of a second examiner to evaluate a comparison set previously viewed by a different examiner and reach the same conclusion. The data used for this analysis includes only evaluations made in the first and third rounds of the study.

It is important to understand that, while this study is designed to assess all three of these aspects of examiner performance, most analyses should not include data from all three rounds. For example, if data from Rounds 1 and either 2 or 3 were included in the assessment of accuracy, this would include multiple evaluations of some of the same comparison sets by the same or different examiner. The resulting double counting of some material would lead to statistical non-independence among data values being combined and potential bias in the estimates of interest. To avoid such problems, most of the analyses in this report other than those assessing repeatability and reproducibility are limited to the data collected in Round 1.

Accuracy

In the first round of the study each of 173 examiners evaluated **sets** of bullets and cartridge cases, each consisting of 2 known items and 1 questioned item. Individual examiners evaluated 15, 30, or (in one instance) 45 sets in the first round. A total of 4320 bullet set examinations and 4320 cartridge case set examinations were performed. A summary of the resulting evaluations, by ground truth status of each set, is given in Table V for bullets and cartridge cases, respectively.

Table V: First-Round Bullet and Cartridge case Summary Counts.

Bullet Evaluations by Set Type						
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Other
Matching	1076	127	125	36	**41**	24
Nonmatching	**20**	268	848	745	961	49
Cartridge case Evaluations by Set Type						
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Other
Matching	1056	177	140	22	**25**	25
Nonmatching	**26**	177	637	620	1375	40

Counts of “hard errors” are highlighted in bold. (*N.B. Throughout this report, a hard error is defined as an instance in which elimination was declared for a matching set, or Identification was declared for a nonmatching set.*) The final column labeled “other” in Table V includes records for which an evaluation was not coded or was recorded as Inconclusive without a level designation (A, B, or C), where multiple levels were recorded, or for which the examiner indicated that the material was Unsuitable for evaluation. (*N.B. Counts reflected in the “other” category are not included in this discussion of accuracy; unsuitable scores are included in some cases where indicated later in the report.*) Summary conclusion percentages are computed by dividing each of the entries in Table V by its corresponding row sum, again, excluding sets classified as “other”, and are presented in Table VI. Hence, for example, the proportion of incorrect identifications among nonmatching bullet sets (or false positives, F-Pos) is:

F-Pos = 100% x ID / (Identification + Inconclusive-A + Inconclusive-B + Inconclusive-C + Eliminations)
= 100% x 20 / (20 + 268 + 848 + 745 + 961) = 0.704%

and the proportion of eliminations among matching bullet sets (or false negatives, F-Neg) is

F-Neg = 100% x Eliminations / (Identification + Inconclusive-A + Inconclusive-B + Inconclusive-C + Eliminations)
= 100% x 41 / (1076 + 127 + 125 + 36 + 41) = 2.92%

after removal of the comparisons represented in the “other” column of Table V.

Table VI: First-Round Bullet and Cartridge case Summary Percentages.

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

Confidence Intervals for Hard Error Probabilities

As has been observed in other similar studies [7], the false positive and false negative errors were made by a relatively small subset of the examiners participating in the study, as documented in Table VII.

Table VII: Numbers of Examiners Making Hard Errors.

Bullet Evaluations in the First Round of the Study				
	No False Negatives	One False Negative	Two or More False Negatives	Total Examiners
No False Positives	139	17	7	163
One False Positive	3	1	1	5
Two or More False Positives	4	0	1	5
Total Examiners	146	18	9	173
Cartridge case Evaluations in the First Round of the Study				
	No False Negatives	One False Negative	Two or More False Negatives	Total Examiners
No False Positives	137	14	4	155
One False Positive	9	3	0	12
Two or More False Positives	6	0	0	6
Total Examiners	152	17	4	173

Hence, for example, of the 173 examiners, 139 made no hard errors of either kind when examining bullets, and three made both kinds of errors, i.e. a false positive and a false negative.

A natural concern that arises from the above observation is the possibility that error probabilities are actually different for different examiners. If this is true, regarding the entire collection of examinations of 1405 matching bullet sets (Table VI) as each having the same probability of being mistakenly labeled Elimination, is not an appropriate assumption. To examine this possibility chi-square tests for independence were performed on tables of counts with 173 rows (one for each examiner), and with columns for examination results. For matching sets the proportions of Identification evaluations versus pooled Elimination and Inconclusive evaluations were compared; for nonmatching sets the proportions of Elimination evaluations versus pooled Identification and Inconclusive evaluations were compared. (*N.B. Pooling of counts was used for these tests because hard errors are relatively rare and, if maintained as a separate category, would result in many zero counts, which are problematic in chi-square tests, e.g.* [34].) For each test, and for both bullets and cartridge cases, the hypothesis of independence was rejected (p value < 0.0001), strongly suggesting that the probabilities associated with each conclusion are not the same for each examiner. As a consequence, the most common methods of computing confidence intervals for proportions based on an assumption of equal probabilities for each evaluation category, e.g. the Clopper-Pearson intervals [35], are not appropriate.

A more appropriate procedure in this case is based on an assumption that each examiner has an error probability, that these probabilities are adequately represented by a beta distribution, which is a flexible two-parameter probability distribution on the unit interval, across the population of examiners, and that the number of errors made by each examiner follows a binomial distribution characterized by that examiner's individual probability. Usual confidence intervals, in contrast, are based on an assumption that there is only one relevant binomial distribution, and that all examiners operate with the same error probability – an assumption strongly contradicted by the chi-square tests noted above. Based on the beta-binomial model, maximum likelihood estimates and 95% confidence intervals for false positive and false negative error probabilities, integrated over all examiners, were calculated using the R statistics package, including the VGAM package [36, 37], and are summarized in Table VIII. A more extensive description of how these confidence intervals are constructed is presented in Appendix G.

Table VIII: Maximum Likelihood Estimates and 95% Confidence Intervals for Error Probabilities.

Bullet Comparisons			
	Point estimate	Lower 95% confidence limit	Upper 95% confidence limit
False Positive Probability	0.656%	0.305%	1.423%
False Negative Probability	2.87%	1.89%	4.26%
Cartridge case Comparisons			
	Point estimate	Lower 95% confidence limit	Upper 95% confidence limit
False Positive Probability	0.933%	0.548%	1.574%
False Negative Probability	1.87%	1.16%	2.99%

It should be noted that even with the allowance for inconsistent probabilities across examiners, this result should still be considered as approximate since, as will be demonstrated in following sections, the model of handgun and the positioning of known and questioned rounds in the firing sequence for a firearm also appear to affect error probabilities, and these considerations are not taken into account in this calculation. Still, as strongly suggested by the analysis above, differences among examiners are associated with differences in error probabilities, and the assumptions underlying the method used here are more appropriate than those upon which simpler methods are based.

Repeatability

In addition to the first round of testing summarized above, two additional rounds of examiner evaluations were undertaken which, in combination with the examinations from Round 1, allow us to evaluate the repeatability and reproducibility of examiner evaluations. As defined in [38] for the setting of industrial measurement:

> *Variation in measurement typical of that seen in the … measurements for a particular operator on a particular part is called the* ***repeatability*** *variation of the gauge. Variation which can be attributed to differences between the … operators is called* ***reproducibility*** *variation of the measurement system.*

In the present context, ``measurement'' refers to evaluation of a test set and "operator" refers to examiner. In the second round of evaluation, examiners were asked to re-examine test sets they had examined in the first round, to provide data relevant for assessment of repeatability, i.e. the extent to which an examiner's repeated evaluations of the same material are consistent. In the third round, test sets that had been examined in the first round were re-examined by different examiners, to provide data relevant for assessment of reproducibility, i.e. the extent to which the evaluations by two examiners of the same material are consistent.

A summary across examiners of the raw counts of bullet set classifications as Identification, Inconclusive-A, Inconclusive-B, Inconclusive-C, Elimination, and Unsuitable by the same examiner in Rounds 1 and 2 of the study, is presented in Table IX. Analogous counts for cartridge case comparisons are presented in Table X. These counts pertain to the idea of repeatability, i.e. the agreement or disagreement, by the same examiner, in two evaluations of the same test set. If the examinations were perfectly repeatable, all off-diagonal cells of Tables IX and X would contain zeroes.

Table IX: Paired Classifications by the Same Examiner (Repeatability) for Bullets.

Matching Sets

Classification on First Evaluation	Classification on Second Evaluation					
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Unsuitable
ID	665	27	26	14	**8**	2
Inconclusive-A	31	28	12	6	**2**	0
Inconclusive-B	13	14	45	5	**2**	2
Inconclusive-C	2	3	3	5	**3**	0
Elimination	**8**	**7**	**3**	**2**	**13**	**0**
Unsuitable	1	3	3	0	**0**	2

Nonmatching Sets

Classification on First Evaluation	Classification on Second Evaluation					
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Unsuitable
ID	**2**	**3**	**6**	**2**	**6**	**0**
Inconclusive-A	**0**	52	37	42	27	0
Inconclusive-B	**5**	31	341	98	45	7
Inconclusive-C	**1**	32	109	284	53	1
Elimination	**1**	20	35	66	514	4
Unsuitable	**0**	0	13	6	4	8

Table X: Paired Classifications by the Same Examiner (Repeatability) for Cartridge cases.

Matching Sets

Classification on First Evaluation	Classification on Second Evaluation					
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Unsuitable
ID	620	47	30	4	**5**	1
Inconclusive-A	43	33	14	4	**3**	1
Inconclusive-B	19	20	40	4	**0**	1
Inconclusive-C	3	6	2	3	**0**	1
Elimination	**1**	**3**	**2**	**2**	**5**	**1**
Unsuitable	6	2	2	1	**0**	5

Nonmatching Sets

Classification on First Evaluation	Classification on Second Evaluation					
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Unsuitable
ID	**2**	**4**	**4**	**1**	**2**	**0**
Inconclusive-A	**5**	21	57	19	10	1
Inconclusive-B	**5**	37	242	98	52	4
Inconclusive-C	**2**	19	91	209	99	5
Elimination	**2**	25	72	90	718	6
Unsuitable	**0**	1	5	7	3	4

The shaded cells along the diagonals of Tables IX-X are counts of instances in which an examiner reported the same score twice for a set; their sum as a proportion of the total number of comparisons is a rough measure of reliability agreement, as displayed in Table XI for bullets and cartridge cases, respectively.

Table XI: Proportion of Paired Classifications in Agreement/Disagreement by the Same Examiner (Repeatability).

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

Tables IX – XI contain raw counts and proportions of scores in the first round versus scores in the second round, accumulating over examiners. Because examiners apparently have different classification probabilities, it is clear that the entries in these tables "average over" the results of those differences, and so can't be interpreted as reflecting the repeatability of evaluation by any one examiner, or even the average of examiner-specific repeatability indicators. To separate the examiner-specific effects of repeatability, separate 6x6 tables (as in Tables IX and X) were constructed for each of the 105 examiners who performed repeated examinations of the same sets of bullets and cartridge cases in Rounds 1 and 2. Two statistics were computed from each of these individualized tables:

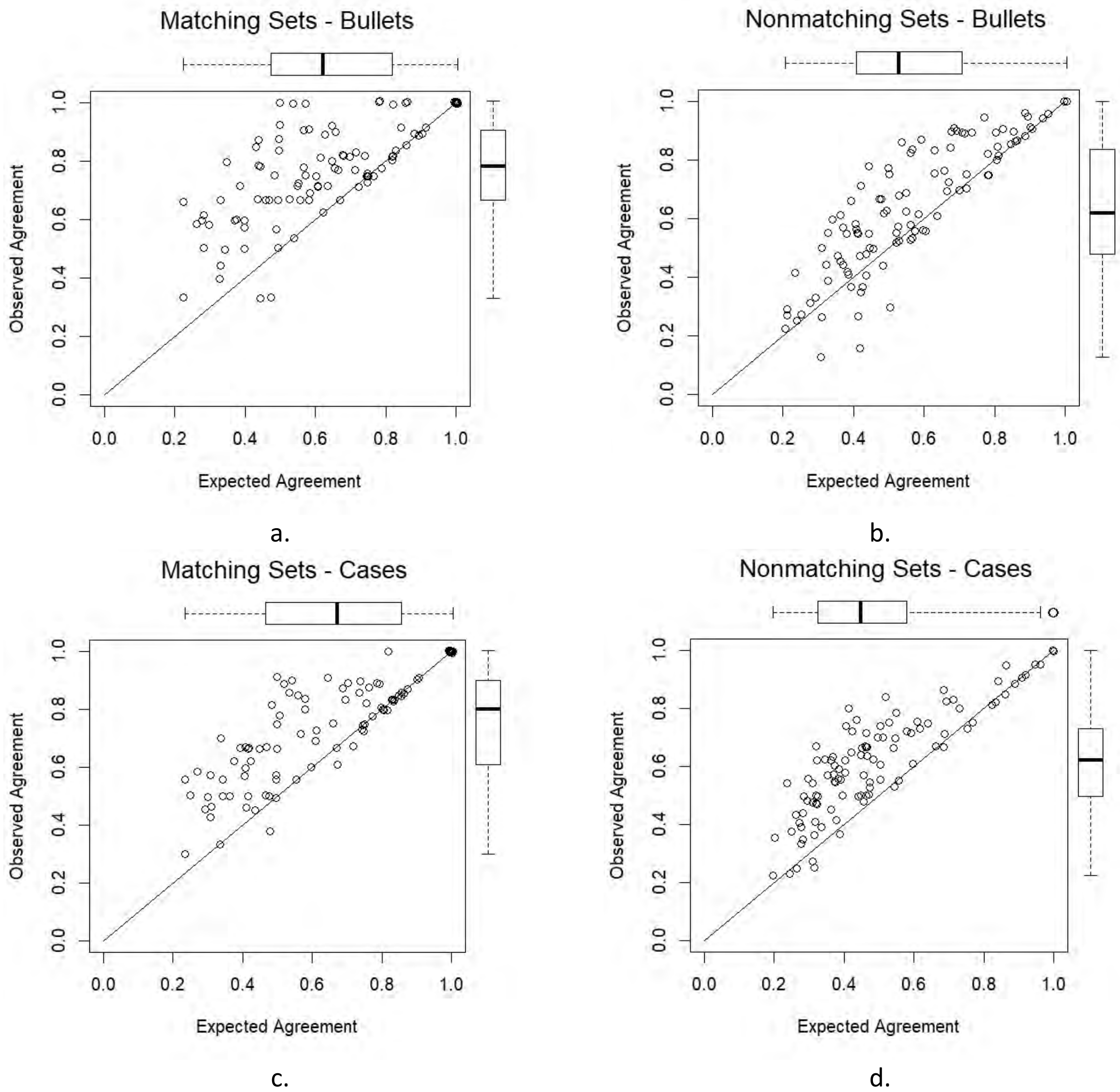
1. The proportion of *Observed Agreement*, i.e. the proportion of counts in the table that fall in the shaded cells along the top-left-to-bottom-right diagonal, when both examinations resulted in the same conclusion. These are the count totals re-expressed as the proportion of paired agreements as in Table XI, but for individual examiners.

2. The proportion of *Expected Agreement*, computed as the sum of corresponding marginal proportions, i.e. the proportion of Identification determinations in the first round times the proportion of Identification determinations in the second round, plus the proportion of Inconclusive-A determinations in the first round times the proportion of Inconclusive-A determinations in the second stage, etc.

The second of these computed statistics requires a bit of explanation. If each examination were entirely *independent* of the others – i.e. if the probability of concluding Identification in the second examination were exactly the same regardless of how the set was evaluated in the first examination – the proportion of Expected Agreement would be an estimate of the number of sets on which the examiner should be expected to agree with him/herself. Hence, if the proportion of Observed Agreement regularly exceeds the proportion of Expected Agreement, this is an indication of examiner repeatability. On the other hand, if these proportions tend to be about the same, this is an indication that an examiner "takes the same chance" of being right or wrong each time he or she examines the same test set, interpreted as a lack of repeatability. Figure 11 displays Observed Agreement versus Expected Agreement proportions for the 105 examiners for matching and nonmatching test sets, for bullets and cartridge cases, respectively.

In all panels of Figures 11 the general trend is toward larger Observed Agreement proportions than would be expected if classifications were made independently on each examination. Even for examiners for which Expected Agreement proportions are less than 1, i.e., for whom first- and second-round determinations were more variable, agreement on any particular set tends to be larger than would be attributed to chance. In Figure 11 and similar figures that follow, the vertical boxplot on the right side displays the distribution of observed agreement proportions, and the horizontal boxplot at the top displays the distribution of expected agreement proportions. These show, for example, that the median observed agreement proportion for matching sets of bullets, denoted by the heavy line in the middle of the box, is approximately 0.8, compared to the median expected proportion of about 0.6 in Figure 11.a. A detailed example of how expected and observed agreement proportions are computed for two-way contingency tables is presented in Appendix H.

Because there is some ambiguity involved in how the three Inconclusive categories are used, it is reasonable to ask how these results would change if the three were pooled into a single category. There is also interest in knowing how repeatability might be affected if Identification and Inconclusive-A categories were artificially pooled and Elimination and Inconclusive-C categories were pooled, leaving only Inconclusive-B as indeterminate. Table XII contains the simple proportions of agreement and disagreement under these two alternative scoring schemes, analogous to Table XI. Figures 12 and 13 display Expected versus Observed Agreement values for these pooled-score systems, analogous to Figure 11.



a.

b.

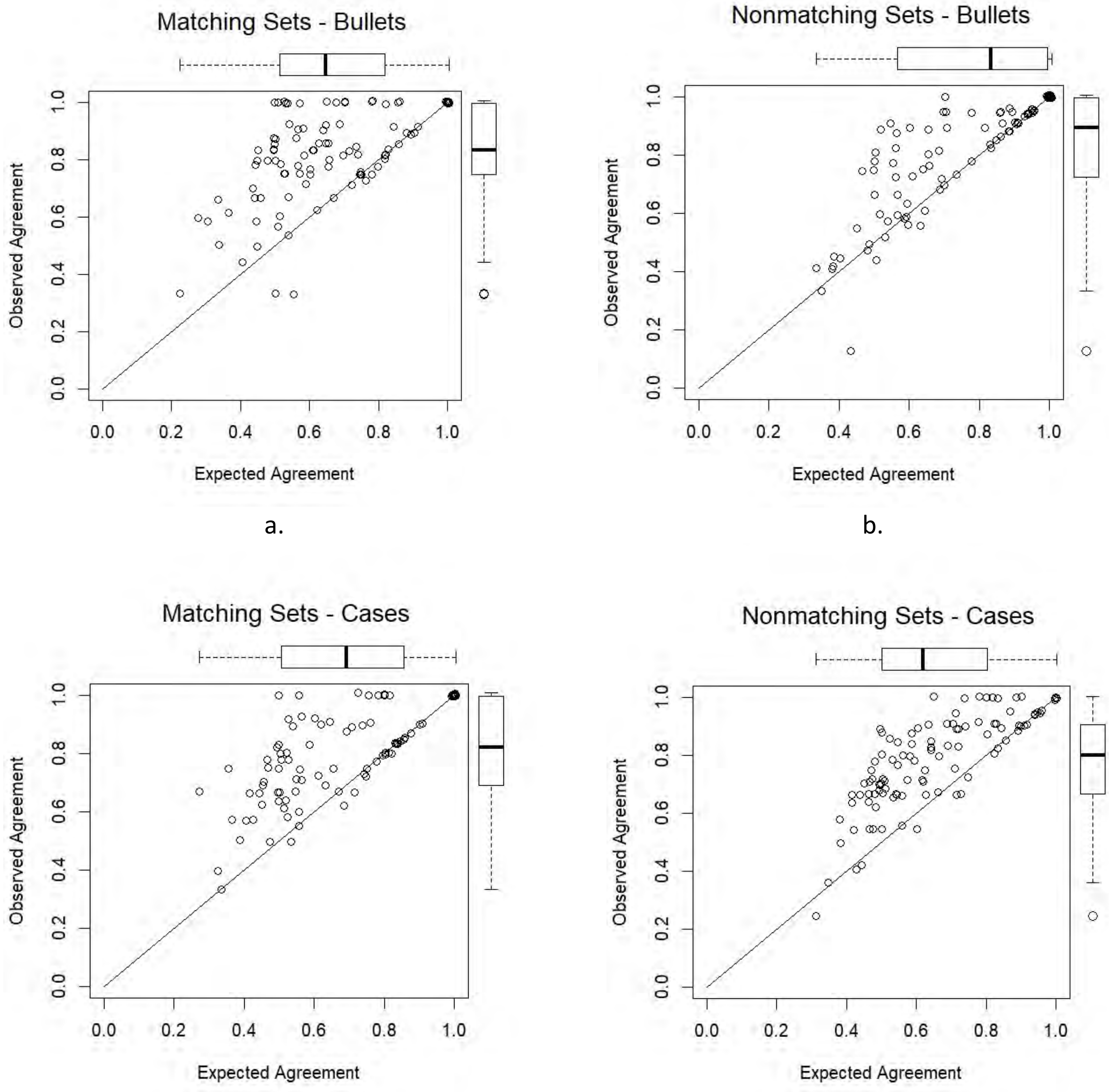
c.

d.

Figure 11: Observed versus Expected agreement for repeated examinations by the same examiner (repeatability) for a) matching bullet; b) nonmatching bullet; c) matching cartridge cases; and d) nonmatching cartridge case sets (repeatability) with no pooling of the five categories.

Table XII: Proportion of Paired Classifications in Agreement and Disagreement by the Same Examiner (Repeatability) when Inconclusive Categories are Pooled / ID and Inconclusive-A are Pooled and Elimination and Inconclusive-C are Pooled.

Paired Bullet Classifications		
	Proportion of paired agreements	Proportion of paired disagreements
Matching Sets	83.4% / 85.5%	16.6% / 14.5%
Nonmatching Sets	83.6% / 71.3%	16.4% / 28.7%
Paired Cartridge case Classifications		
	Proportion of paired agreements	Proportion of paired disagreements
Matching Sets	80.9% / 85.4%	19.1% / 14.6%
Nonmatching Sets	78.9% / 72.5%	21.1% / 27.5%



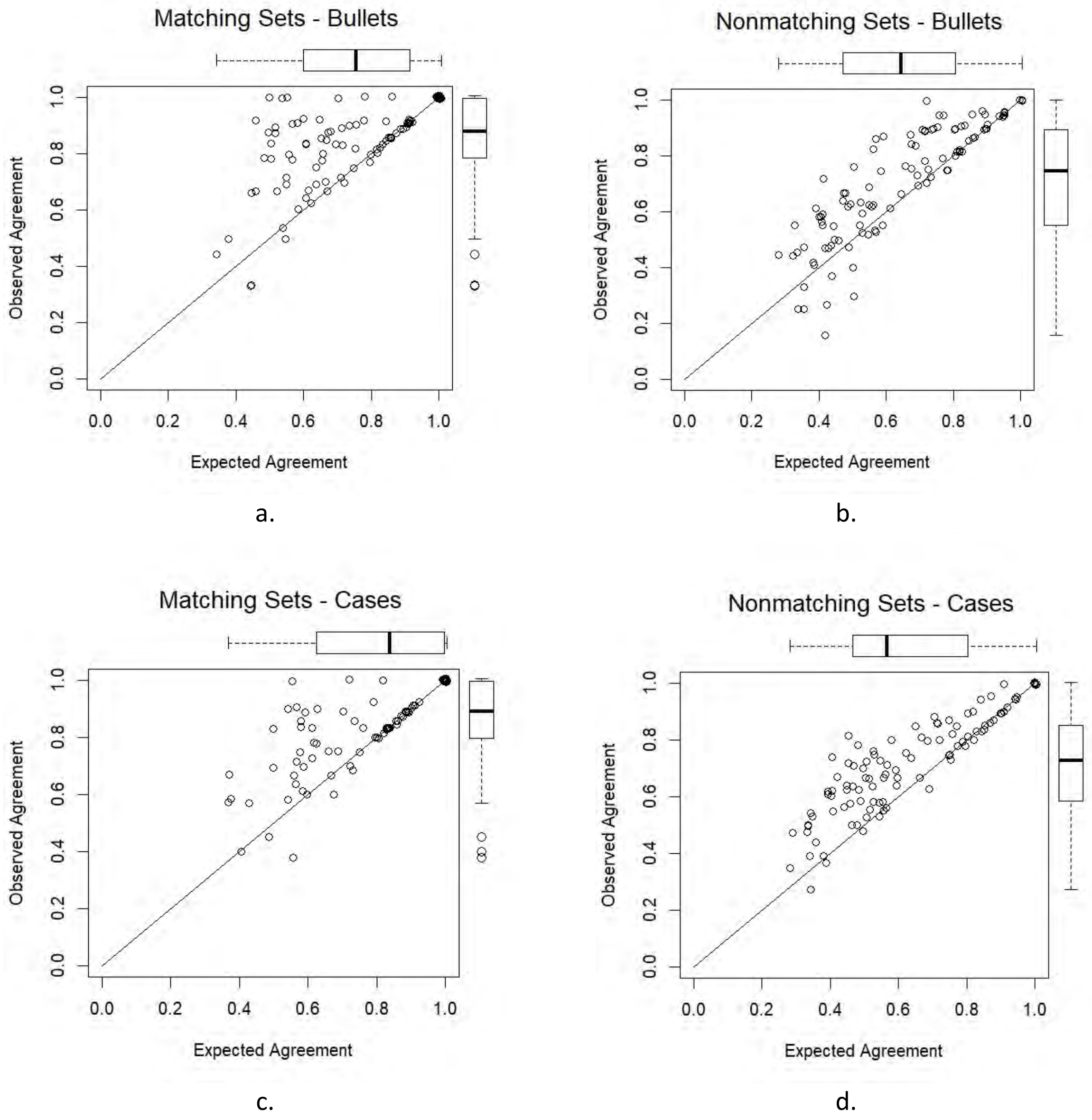
a.

b.

c.

d.

Figure 12: Observed versus Expected agreement for repeated examinations by the same examiner (repeatability) for a) matching bullet; b) nonmatching bullet; c) matching cartridge case; d) nonmatching cartridge case sets. The three Inconclusive categories are pooled and accounted as a single category.



a.

b.

c.

d.

Figure 13: Observed versus Expected agreement for repeated examinations by the same examiner (repeatability) for a) matching bullet; b) nonmatching bullet; c) matching cartridge case and d) nonmatching cartridge case sets. ID and Inconclusive-A results are pooled, and Elimination and Inconclusive-C are pooled.

Table XIII summarizes the average Expected and Observed agreement proportions for both bullets and cartridge cases, for the 5-category scoring system and the two alternative 3-category systems. A nonparametric sign test of the null hypothesis that observed frequency minus expected frequency has a median of zero, versus the alternative that observed frequency exceeds expected frequency more than half the time, is rejected with a p-value of 0.0001 or less for all data sets represented in Figures 11-13 (i.e. bullets and cartridge cases, matching and nonmatching sets, and all three scoring schemes), indicating “better than chance” repeatability. In most cases, the observed proportion of matches is from 10% to 15% greater than the corresponding expected proportion. Both proportions increase as less detailed (pooled) classification scales are used, and for both bullets and cartridge cases, agreement is somewhat better for matching than for nonmatching sets.

Table XIII: Average (Over Examiners) Observed and Expected Proportions of Agreement for Two Examinations of the Same Comparison Set by the Same Examiner (Repeatability), for 5-Category and Pooled Scoring Systems.

Proportions of Agreement: Bullet Sets				
	Matches		**Nonmatches**	
Scoring	Observed Agreement	Expected Agreement	Observed Agreement	Expected Agreement
ID, Inc-A, Inc-B, Inc-C, Elim	77.7%	63.7%	63.4%	55.8%
ID, (Inc-A & Inc-B & Inc-C), Elim	82.0%	66.4%	82.5%	77.0%
(ID & Inc-A), Inc-B, (Inc-C & Elim)	85.0%	75.7%	70.4%	63.9%
Proportions of Agreement: Cartridge case Sets				
	Matches		**Nonmatches**	
Scoring	Observed Agreement	Expected Agreement	Observed Agreement	Expected Agreement
ID, Inc-A, Inc-B, Inc-C, Elim	76.1%	66.3%	61.3%	48.7%
ID, (Inc-A & Inc-B & Inc-C), Elim	81.2%	69.4%	78.3%	65.0%
(ID & Inc-A), Inc-B, (Inc-C & Elim)	86.3%	80.7%	72.0%	63.2%

Reproducibility

A summary of the raw counts of set classifications as Identification, Inconclusive-A, Inconclusive-B, Inconclusive-C, Elimination, and Unsuitable *by different examiners* in Rounds 1 and 3 of the study is presented for bullet comparisons in Table XIV and for cartridge case comparisons in Table XV. These counts pertain to the idea of reproducibility, i.e. the agreement or disagreement, by different examiners, in two evaluations of the same set.

Table XIV: Paired Classifications by Different Examiners (Reproducibility) for Bullets.

Matching Sets of Bullets

Classification by First Round Examiner	Classification by Third Round Examiner					
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Unsuitable
ID	601	38	39	14	**12**	5
Inconclusive-A	42	18	7	6	**6**	0
Inconclusive-B	34	15	22	4	**6**	0
Inconclusive-C	9	7	5	2	**6**	0
Elimination	**13**	**5**	**14**	**6**	**3**	**0**
Unsuitable	3	2	8	1	**1**	2

Nonmatching Sets of Bullets

Classification by First Round Examiner	Classification by Third Round Examiner					
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Unsuitable
ID	**0**	**5**	**8**	**5**	**1**	**0**
Inconclusive-A	**1**	15	58	33	60	0
Inconclusive-B	**5**	61	180	125	159	10
Inconclusive-C	**2**	35	134	114	142	4
Elimination	**1**	71	162	193	274	0
Unsuitable	**0**	1	13	5	9	0

Table XV: Paired Classifications by Different Examiners (Reproducibility) for Cartridge cases.

Matching Sets of Cartridge cases

Classification by First Round Examiner	Classification by Third Round Examiner					
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Unsuitable
ID	582	56	56	11	**5**	7
Inconclusive-A	66	12	24	2	**5**	2
Inconclusive-B	30	25	14	4	**2**	1
Inconclusive-C	6	3	6	1	**0**	0
Elimination	**15**	**3**	**4**	**1**	**2**	**0**
Unsuitable	5	4	7	0	**0**	1

Nonmatching Sets of Cartridge cases

Classification by First Round Examiner	Classification by Third Round Examiner					
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Unsuitable
ID	**0**	**2**	**12**	**7**	**5**	**0**
Inconclusive-A	**1**	5	34	31	45	1
Inconclusive-B	**2**	29	115	88	159	9
Inconclusive-C	**1**	24	75	118	146	4
Elimination	**1**	51	159	220	530	12
Unsuitable	**0**	2	9	5	9	2

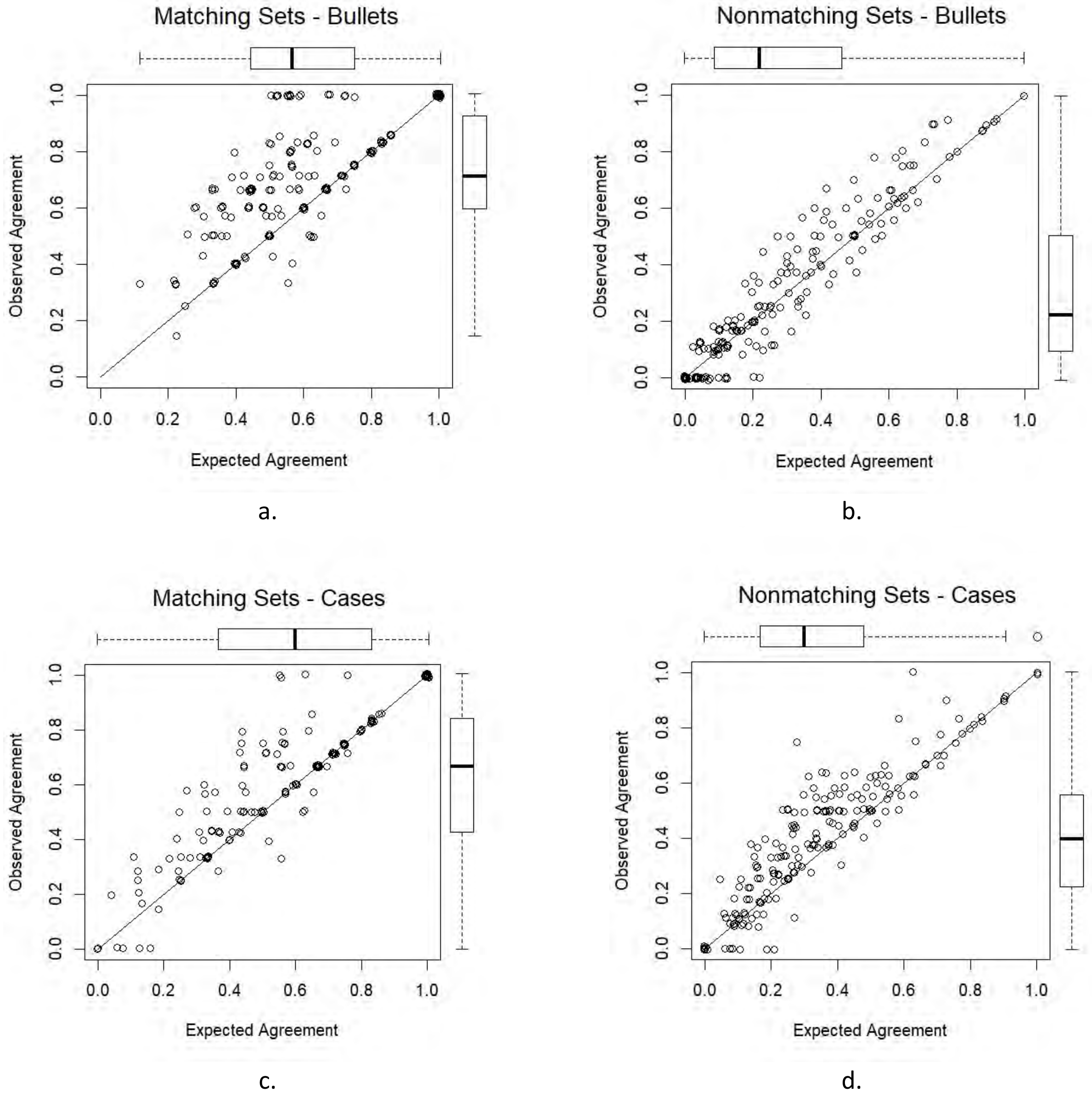
The shaded cells along the diagonals (upper-left to lower-right) are counts of instances in which two examiners reported the same score twice for a set, i.e., they have perfect agreement; their sum (as a proportion of the total number of comparisons) is a rough measure of reproducibility agreement, as displayed in Table XVI.

Table XVI: Proportion of Paired Classifications in Agreement/Disagreement by Different Examiners (Reproducibility).

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

Relative to the summaries reported for Repeatability (Table XII), the striking feature of these results is the larger proportions of examiner disagreements for nonmatching sets. Of course, as can be seen in the more detailed Tables XIV and XV, many of these disagreements reflect different scores in the Inconclusive range, i.e. do not involve evaluation pairs where one or both of the evaluations are scored as ID or Elimination.

As with the examination of repeatability, these counts are interesting overall summaries of agreement or disagreement in how sets were classified in Rounds 1 and 3 of the study, but they do not offer a reliable indication of how the scores of any two examiners might coincide. Proportions of Observed Agreement and Expected Agreement (under a hypothesis of independent evaluations) were computed for the 191/193 bullet/cartridge case test packets evaluated by two different examiners, with results displayed in Figure 14 for bullet and cartridge case sets, respectively.



a.

b.

c.

d.

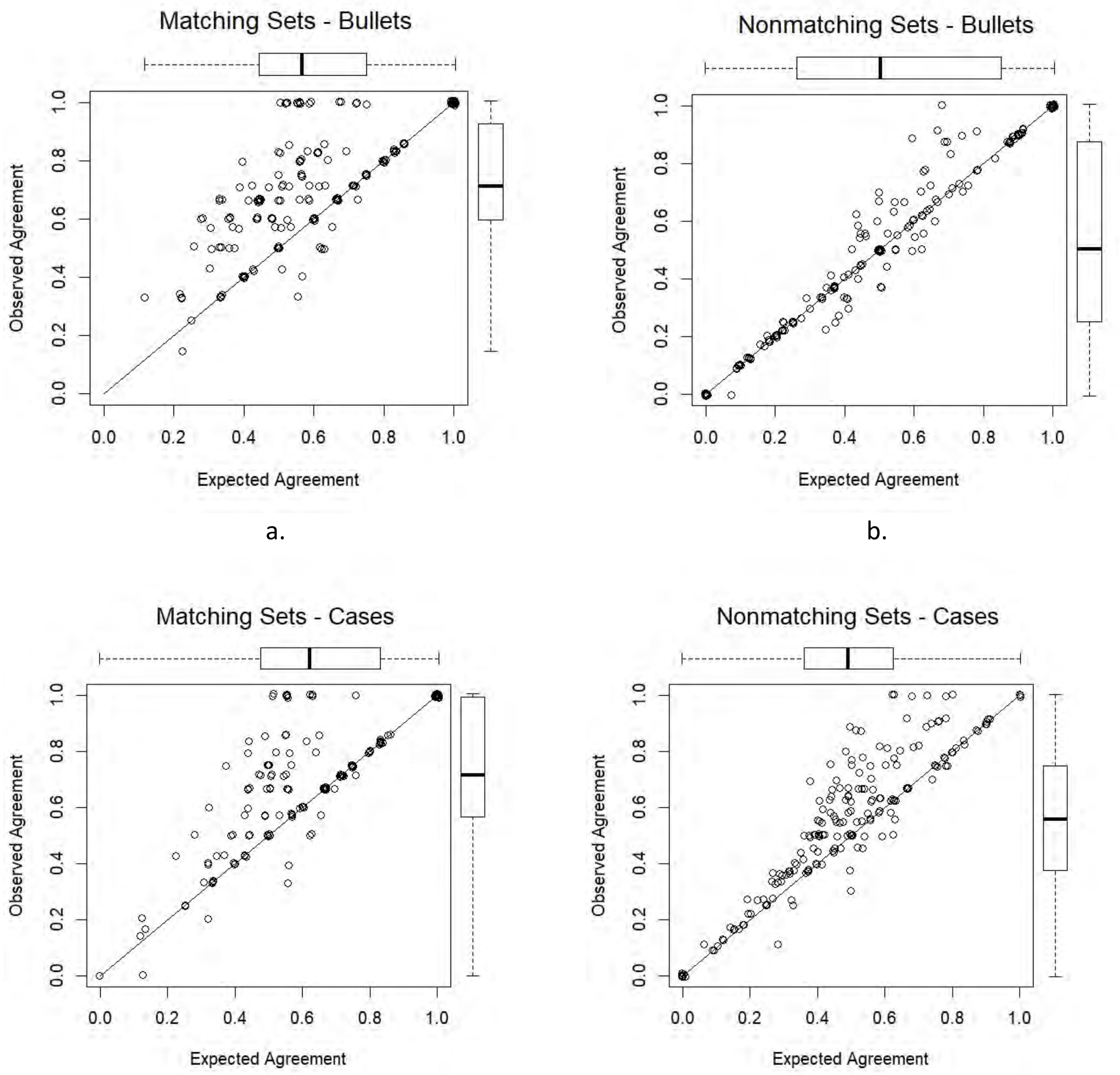
Figure 14: Observed versus Expected agreement for examinations of a) matching bullet; b) nonmatching bullet; c) matching cartridge case and d) nonmatching cartridge case sets by the different examiners (reproducibility) with no pooling of the five categories.

Consistent with the overall proportions of agreement given in Table XVI, these graphs suggest more limited reproducibility (as compared to repeatability), especially for nonmatching sets.

Table XVII displays overall percentages of inter-examiner agreement when the three Inconclusive categories are pooled, and when ID and Inconclusive-A are pooled and when Elimination and Inconclusive-C are pooled. Figures 15 and 16 display Expected versus Observed Agreement values for pairs of examiners who evaluated the same material, under these two pooled-scoring schemes.

Table XVII: Proportion of Paired Classifications in Agreement and Disagreement by Different Examiners (Reproducibility) when Inconclusive Categories are Pooled / ID and Inconclusive-A are Pooled and Elimination and Inconclusive-C are Pooled.

<table>
<tr><th colspan="3">Paired Bullet Classifications</th></tr>
<tr><td></td><td>Proportion of paired agreements</td><td>Proportion of paired disagreements</td></tr>
<tr><td><b>Matching Sets</b></td><td>72.4% / 77.4%</td><td>27.6% / 22.6%</td></tr>
<tr><td><b>Nonmatching Sets</b></td><td>54.6% / 49.0%</td><td>45.4% / 51.0%</td></tr>
<tr><th colspan="3">Paired Cartridge case Classifications</th></tr>
<tr><td></td><td>Proportion of paired agreements</td><td>Proportion of paired disagreements</td></tr>
<tr><td><b>Matching Sets</b></td><td>70.3% / 76.4%</td><td>29.7% / 23.6%</td></tr>
<tr><td><b>Nonmatching Sets</b></td><td>54.9% / 59.5%</td><td>45.1% / 40.5%</td></tr>
</table>



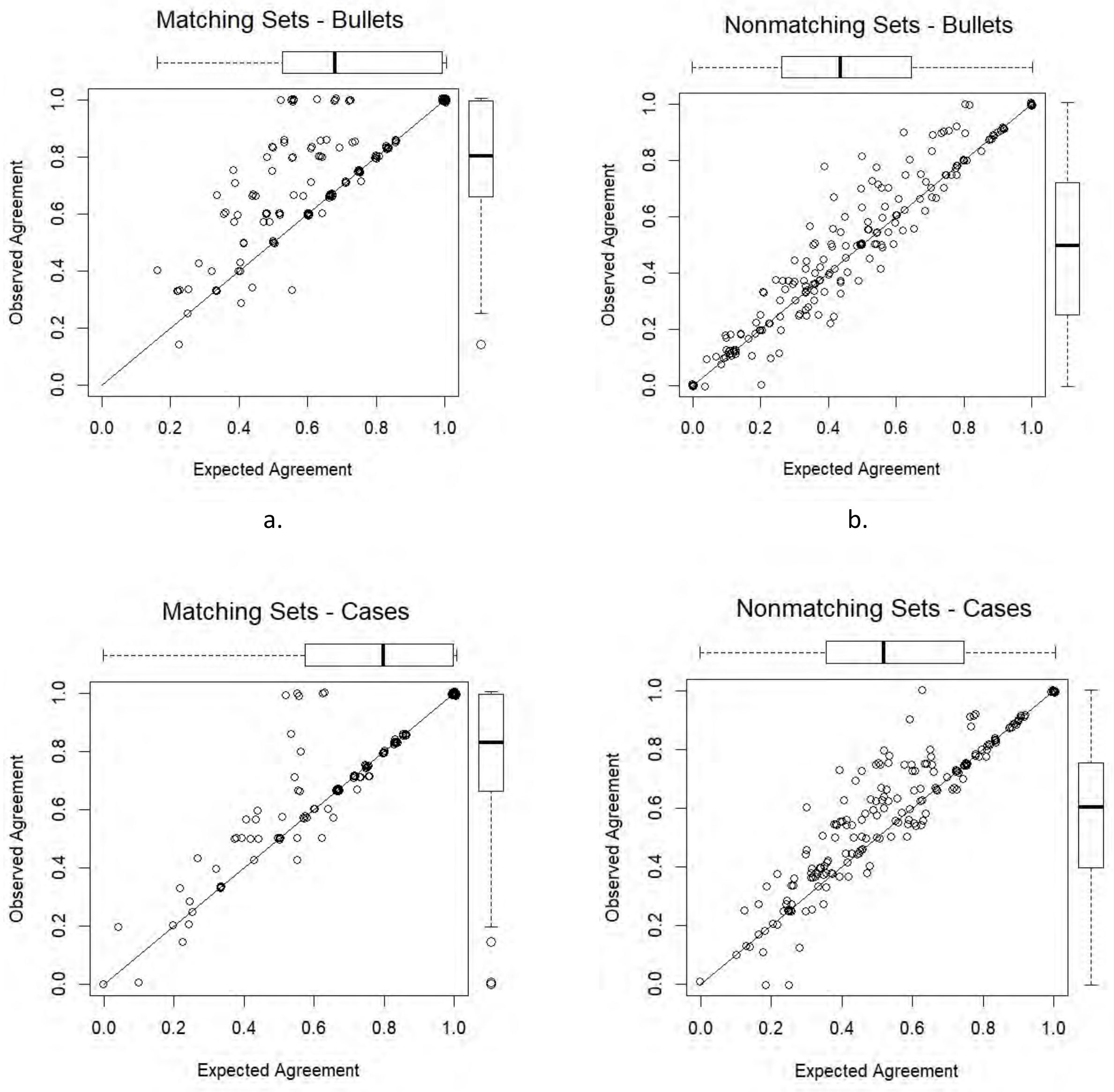
a.

b.

c.

d.

Figure 15: Observed versus Expected agreement for examinations by different examiners (reproducibility) with the three Inconclusive categories pooled into a single category.



a. b. c. d.

Figure 16: Observed versus Expected agreement for examinations by different examiners (reproducibility) with ID and Inconclusive-A pooled, and Elimination and Inconclusive-C pooled for a) matching bullet; b) nonmatching bullet; c) matching cartridge case and d) nonmatching cartridge case sets.

Table XVIII summarizes the average Expected and Observed Agreement proportions, by different examiners, for both bullets and cartridge cases, for the 5-category scoring system and the two alternative 3-category systems. Trends in this table are similar to those for Repeatability (Table XIII); both Observed and Expected Agreement proportions tend to be larger for matching sets than for nonmatching sets, and either of the two pooling schemes increases both Observed and Expected Agreement relative to the standard 5-category scale. Perhaps not surprisingly, the values for both

Observed and Expected Agreement are lower than the corresponding figures for Repeatability, but Observed Agreement is at least somewhat greater than Expected Agreement in each case. A nonparametric sign test of the null hypothesis that observed frequency minus expected frequency has a median of zero, versus the alternative that observed frequency exceeds expected frequency more than half the time is rejected with a p-value of 0.01 except for nonmatching bullets when the 5-category classification scheme is used, and when all Inconclusive scores are pooled; and for cases when Inconclusive-A is pooled with ID and Inconclusive-C is pooled with Exclusion. Note that the size of p-values should not be regarded as an indicator of the size (in percentage points) of a difference between observed and expected agreement.

Table XVIII: Average (Over Examiners) Observed and Expected Proportions of Agreement for Two Examinations of the Same Comparison Set by Different Examiners (Reproducibility), for 5-Category and Pooled Scoring Systems.

Bullet Set				
	Matches		**Nonmatches**	
Scoring	Observed Agreement	Expected Agreement	Observed Agreement	Expected Agreement
ID, Inc-A, Inc-B, Inc-C, Elim	67.6%	58.3%	30.5%	28.8%
ID, (Inc-A & Inc-B & Inc-C), Elim	72.1%	61.3%	54.4%	53.2%
(ID & Inc-A), Inc-B, (Inc-C & Elim)	77.2%	70.4%	48.7%	46.2%
Cartridge case Set				
	Matches		**Nonmatches**	
Scoring	Observed Agreement	Expected Agreement	Observed Agreement	Expected Agreement
ID, Inc-A, Inc-B, Inc-C, Elim	64.3%	60.1%	40.1%	34.1%
ID, (Inc-A & Inc-B & Inc-C), Elim	70.7%	64.1%	55.2%	48.3%
(ID & Inc-A), Inc-B, (Inc-C & Elim)	77.5%	75.5%	59.2%	54.8%

Effects Related to Firearm Type and Wear

Firearm Make

Comparison bullet sets were produced using Beretta and Ruger handguns, and comparison cartridge case sets were produced using Beretta and Jimenez handguns. The two known and one questioned bullets in nonmatching sets were produced using two different handguns of the same make. The summary accuracy proportions of Table VI are recalculated for each handgun make separately, and are displayed in Table XIX.

Table XIX: First Round Summary Percentages of Bullet and Cartridge case Set Evaluations by Firearm Type and Test Set Type.

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

Error proportions are relatively smaller, and correct conclusion proportions relatively larger, for comparisons of bullets fired using the Beretta handguns. In addition, the proportions of Inconclusive findings are smaller for Beretta handguns than for Ruger handguns. Trends are similar for cartridge cases, here with the number of correct conclusions being relatively larger, and inconclusive proportions relatively smaller, for cartridge cases produced with Beretta handguns relative to Jimenez handguns.

Firing-Sequence Separation

Each handgun employed in the study was used to fire between 700 and 850 rounds. The ordered sequence of rounds fired using each handgun was recorded, and these were divided into "Early", "Middle" and "Late" positions in the sequence. In each set, the two known bullets were taken from the same sequence group, but the questioned bullet could come from any group. A set might be classified nine different ways: Early, Middle, or Late classification for the known rounds, and Early, Middle, or Late classification for the questioned round. In particular, if the known and questioned rounds are fired at opposite ends of the test sequence, there might be an effect associated with the mechanical wear experienced by the handgun(s) over the course of use. Table XX displays summary proportions of bullet and cartridge case evaluation results conditioned on sets for which known and questioned bullets were produced in the same sequence groups ("EE-E", "MM-M", and "LL-L"), and at opposite end of the sequence ("EE-L" or "LL-E").

Perhaps not surprisingly, evaluation proportions for the nonmatching sets are not substantially different for either bullets or cartridge cases. For matching sets, however, the proportion of correct Identifications is substantially larger, and the proportion of each category of Inconclusive and (false) Eliminations is smaller, when both known and questioned bullets or cartridge cases are fired in the same third of the sequence. This would apparently support the hypothesis of an effect of tool wear on classification probabilities. Chi-square tests comparing EE-L and LL-E evaluation frequencies were not significant (p > 0.1) for both bullets and cartridge cases, and both matching and nonmatching sets; there is little evidence to suggest meaningful differences between these.

Table XX: First Round Summary Percentages of Evaluations by Set Type and Firing Sequence Separation.

			Bullet Evaluations			
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Total Sets
Matching						
EE-E, MM-M, LL-L	91.5%	3.34%	4.23%	0.445%	**0.45%**	449
EE-L	60.2%	12.4%	14.9%	4.35%	**8.07%**	161
LL-E	52.8%	11.8%	16.1%	8.07%	**11.2%**	161
Nonmatching						
EE-E, MM-M, LL-L	**1.20%**	9.63%	33.6%	25.3%	30.3%	914
EE-L	**0.97%**	9.68%	25.8%	29.7%	33.9%	310
LL-E	**0%**	6.69%	23.1%	32.2%	38.0%	329
			Cartridge case Evaluations			
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Total Sets
Matching						
EE-E, MM-M, LL-L	84.8%	7.41%	6.21%	0.80%	**0.80%**	499
EE-L	61.5%	19.0%	14.4%	2.30%	**2.87%**	174
LL-E	62.4%	20.4%	13.4%	1.27%	**2.55%**	157
Nonmatching						
EE-E, MM-M, LL-L	**1.18%**	7.82%	20.3%	23.6%	47.1%	934
EE-L	**0.59%**	4.42%	22.4%	20.6%	51.9%	339
LL-E	**0.64%**	5.41%	23.9%	17.2%	52.9%	314

There is also interest in knowing whether effects associated with the firing-sequence are related to the make of the handgun, i.e. whether use-related wear has more influence on examiner evaluations for some makes of handguns than for others. With this in mind, firing-sequence conclusion percentages were recomputed, further divided by Beretta or Ruger/Jimenez (bullet/cartridge case) sets. Because there was no apparent difference between EE-L and LL-E comparisons above, these were pooled here to prevent estimated proportions based on very small sample sizes. Results are displayed in Table XXI for bullets and cartridge cases, respectively.

Table XXI: First Round Summary Percentages by Set Type, Firing Sequence Separation, and Weapon Manufacturer.

Bullet Evaluations							
	Make	ID	Inc-A	Inc-B	Inc-C	Elim	Total Sets
Matching							
EE-E, MM-M, LL-L	Beretta	98.5%	1.15%	0.39%	0%	**0%**	260
EE-E, MM-M, LL-L	Ruger	82.0%	6.35%	9.52%	1.06%	**1.06%**	189
EE-L, LL-E	Beretta	71.7%	10.1%	7.07%	3.54%	**7.58%**	198
EE-L, LL-E	Ruger	32.3%	15.3%	29.0%	10.5%	**12.9%**	124
Nonmatching							
EE-E, MM-M, LL-L	Beretta	**0.79%**	9.42%	24.5%	27.6%	37.7%	637
EE-E, MM-M, LL-L	Ruger	**2.17%**	10.1%	54.5%	19.9%	13.4%	227
EE-L, LL-E	Beretta	**0.44%**	8.75%	18.2%	33.7%	38.9%	457
EE-L, LL-E	Ruger	**0.55%**	6.59%	40.1%	24.2%	28.6%	182
Cartridge case Evaluations							
	Make	ID	Inc-A	Inc-B	Inc-C	Elim	Total Sets
Matching							
EE-E, MM-M, LL-L	Beretta	85.3%	6.69%	6.69%	0.33%	**1.00%**	299
EE-E, MM-M, LL-L	Jimenez	84.0%	8.50%	5.50%	1.50%	**0.50%**	200
EE-L, LL-E	Beretta	70.8%	16.3%	11.4%	0.50%	**0.99%**	202
EE-L, LL-E	Jimenez	48.1%	24.8%	17.8%	3.88%	**5.43%**	129
Nonmatching							
EE-E, MM-M, LL-L	Beretta	**0.78%**	7.80%	21.7%	24.2%	45.6%	641
EE-E, MM-M, LL-L	Jimenez	**2.05%**	7.85%	17.4%	22.2%	50.5%	293
EE-L, LL-E	Beretta	**0.87%**	6.10%	25.7%	18.7%	48.6%	459
EE-L, LL-E	Jimenez	**0%**	2.06%	17.0%	19.6%	61.3%	194

The largest model-related differences apparent in these tables are for bullet examinations for matching sets, where correct decisions are made more frequently for Beretta sets, and Inconclusive-B determinations are made more frequently for Ruger sets. Based on the fairly limited sample sizes produced in this breakdown, apparent differences among hard-error proportions should not be interpreted as definitive.

Proportions of Unsuitable Evaluations

Most of the comparisons in this report do not include counts of sets classified as Unsuitable; proportions of examiner determinations, and especially "hard-error" percentages, are computed using only sets classified as ID, Inconclusive (A, B, or C), or Elimination. However, there is also interest in understanding whether the weapon model is associated with the proportion of sets labeled as unsuitable by the examiners. Table XXII summarizes the proportions of such determinations for matching and nonmatching sets by manufacturer, for bullets and cartridge cases, respectively. Despite these determinations being relatively rare, the apparent trend is that fewer bullet sets produced with Beretta weapons are deemed unsuitable, and more cartridge case sets produced with Beretta weapons are deemed unsuitable, than with the alternative models used. (*N.B. A reminder: evaluations from Rounds 2* *and 3 are not included here to avoid the bias associated with double counting of some, but not all, material.*)

Table XXII: First Round Summary Percentages of Evaluations Rated Unsuitable, by Set Type and Manufacturer.

Bullet Sets			Cartridge case Sets		
	Unsuitable	Total Sets		Unsuitable	Total Sets
Matching			**Matching**		
Beretta	0% (0 sets)	848	Beretta	1.95% (17 sets)	874
Ruger	3.80% (22 sets)	579	Jimenez	1.05% (6 sets)	569
Nonmatching			**Nonmatching**		
Beretta	0.05% (1 set)	2023	Beretta	1.45% (29 sets)	2000
Ruger	4.09% (35 sets)	855	Jimenez	0.35% (3 sets)	867

Effects Associated with Manufacturing

Of the 27 Beretta handguns used in the study, 23 were from a single recent manufacturing run, and four were guns produced in separate earlier manufacturing runs. That is, these handguns represented five different manufacturing runs, for which the barrels were produced with different broaches, etc. If there are differences between the conclusions reached by examiners in evaluating nonmatching bullets fired from Beretta barrels from the same manufacturing run, compared to the conclusions they reach in evaluating nonmatching bullets fired from Beretta barrels from different manufacturing runs, this might be related to different subclass characteristics associated with the individual runs. Similarly, if there are differences between the conclusions reached in evaluating nonmatching cartridge cases fired with slides from different manufacturing runs, versus from those produced in the same manufacturing run, this might also be related to subclass characteristics associated with different runs. Table XXIII displays the proportions of examiner determinations for nonmatching sets produced by firearms within a common run, and firearms from different runs. For bullets, a chi-square test comparing same-group and different-group comparisons does not support a hypothesis of difference ($p > 0.5$). However, a similar chi-square test for cartridge case comparisons is highly significant ($p < 0.0001$) indicating strong support a difference between conclusions for same-group and different-group examinations. This difference is most easily seen in Eliminations – far more Elimination determinations are made when comparing cartridge cases fired from guns from different production runs than from guns produced in the same run, while percentages of all other determinations are higher for same-run comparisons.

Table XXIII: Summary of First-Round Evaluations of Nonmatch Bullet/Cartridge case Sets for Beretta Barrels/Slides from Same and Different Manufacturing Runs.

Bullets/Barrels						
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Total Sets
Same Run	**0.60%**	9.52%	23.4%	28.4%	38.1%	1502
Different Runs	**0.38%**	9.81%	21.5%	27.9%	40.4%	520
Cartridge cases/Slides						
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Total Sets
Same Run	**1.04%**	8.03%	26.7%	24.4%	39.8%	1444
Different Runs	**0.38%**	2.85%	16.9%	18.0%	61.9%	527

Most of the Beretta handguns used (23 of 27) were assembled from barrels produced in a single manufacturing run, employing a single broach. These barrels used can be separated into five groups that were sequentially manufactured, with each group containing either four or five barrels. At least 10 barrels were produced (and not used) between any two of these groups. For example, the first group (in sequential order of manufacture) contains five consecutively manufactured barrels, the next 10 barrels manufactured in the run were not used in this study, the second group contains the next four barrels manufactured, etc. If there are differences between the conclusions reached by examiners in evaluating nonmatching bullets fired from Beretta barrels in the same group, compared to the conclusions they reach in evaluating nonmatching bullets fired from Beretta barrels in different groups, this might be related to changing subclass characteristics related to tool (broach) wear within a single run of the manufacturing process. Similarly, the slides of these 23 Beretta handguns can be grouped into five sequentially manufactured sets from one continuous run. If there are differences between the conclusions reached in evaluating nonmatching cartridge cases fired with slides from the same group, relative to those fired from slides in different groups, this might also be related to wear-related changes in subclass characteristics associated with the manufacturing process. Table XXIV displays the proportions of determinations for nonmatching sets produced by firearms within a common group, and firearms from different groups. (*N.B. Evaluations of nonmatching sets produced using the four additional guns from different manufacturing runs, noted above, are not used in this analysis.*) For each of bullets and cartridge cases, a chi-square test does not support the hypothesis of a difference between same-group and different-group comparisons (p > 0.2 in each case); that is, there is little evidence of tool wear effect on the examination results over the length of the manufacturing run reported here.

Table XXIV: Summary of First-Round Evaluations of Nonmatch Bullet/Cartridge case Sets for Beretta Barrels/Slides from Same and Different Sequential Groups of a Single Manufacturing Run.

Bullets/Barrels						
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Total Sets
Same Group	**0.87%**	10.2%	23.9%	30.8%	34.3%	461
Different Groups	**0.48%**	9.22%	23.2%	27.3%	39.9%	1041
Cartridge cases/Slides						
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Total Sets
Same Group	**1.14%**	8.18%	28.9%	25.0%	36.8%	440
Different Groups	**1.00%**	7.97%	25.7%	24.2%	41.1%	1004

Effects Related to Examination Procedure

Examiner's Evaluation of Difficulty

Examination sets included two known and one questioned bullets or cartridge cases. For each set, examiners were asked to indicate how many of the known items had sufficient reproduced detail to support comparison. Table XXV contains counts of the number of sets for which 0, 1, or 2 knowns were sufficiently marked, subdivided by make of handgun.

Table XXV: First Round Summary of Bullet and Cartridge case Sets with 0, 1, and 2 Knowns Judged to Have Sufficient Reproducible Detail for Comparison.

Bullet Sets				
	0 Knowns	1 Known	2 Knowns	Total Sets
Beretta	6	37	2824	2867
Ruger	155	125	1148	1428
Overall	161	162	3972	4295
Cartridge case Sets				
	0 Knowns	1 Known	2 Knowns	Total Sets
Beretta	179	178	2499	2856
Jimenez	75	61	1292	1428
Overall	254	239	3791	4284

As part of each set evaluation, examiners were asked to rate the "Degree of Difficulty" of the evaluation as "Easy", "Average", or "Hard". Table XXVI summarizes the percentage of bullet and cartridge case comparison sets classified for each of these categories, for both matching and nonmatching sets, both overall and subdivided by the make of handgun. Most evaluations were rated as being of Average difficulty. For bullet comparisons, more Ruger sets were judged to be Hard and more Beretta sets Easy. Trends were more complex for cartridge case examinations, where Beretta/Jimenez differences depended on whether sets are matching or nonmatching.

Table XXVI: First Round Summary Percentages of Degree of Difficulty Evaluations by Type of Set and Manufacturer.

Bullets				
	Easy	Average	Hard	Total Sets
Matching				
Beretta	43.8%	50.3%	5.92%	845
Ruger	11.8%	53.6%	34.6%	549
Overall	31.2%	51.6%	17.2%	1394
Nonmatching				
Beretta	9.12%	77.7%	13.2%	2007
Ruger	4.05%	61.5%	34.4%	814
Overall	7.66%	73.0%	19.3%	2821
Cartridge cases				
	Easy	Average	Hard	Total Sets
Matching				
Beretta	26.7%	53.6%	19.7%	849
Jimenez	17.5%	53.0%	29.5%	559
Overall	23.1%	53.3%	23.6%	1408
Nonmatching				
Beretta	15.1%	61.4%	23.5%	1955
Jimenez	24.3%	57.7%	18.0%	855
Overall	17.9%	60.3%	21.8%	2810

The percentages of Identification, Inconclusive-A, Inconclusive-B, Inconclusive-C, and Elimination determinations, separated by the true status of the set (matching or nonmatching) and Degree of Difficulty categorization is summarized in Table XXVII for bullets and cartridge cases, respectively. For both matching and Nonmatching sets, the percentage of correct evaluations (Identification for matching, Elimination for Nonmatching) decreases substantially with increasing difficulty, while the percentage of each category of Inconclusive evaluation generally increases with increasing difficulty. The frequency of hard errors (declaring Elimination for matching, Identification for nonmatching) also increases with difficulty for bullets, to 1.65% for Nonmatching sets (i.e. false positive) and 4.17% for matching sets (false negative), but is more consistent across difficulty levels for cartridge cases. Differences in false negative and false positive percentages over degree of difficulty are significant for bullets (p=.004 and .009, respectively), but are not for cartridge cases (p>.2).

Table XXVII: First Round Summary Percentages of Evaluations by Type of Set and Degree of Difficulty.

			Bullets			
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Total Sets
Matching						
Easy	96.8%	0.46%	1.61%	0.46%	**0.69%**	435
Average	73.6%	8.48%	11.0%	3.06%	**3.89%**	719
Hard	49.6%	26.7%	15.0%	4.58%	**4.17%**	240
Nonmatching						
Easy	**0.00%**	3.70%	7.41%	16.7%	72.2%	216
Average	**0.53%**	8.20%	29.1%	27.5%	34.7%	2060
Hard	**1.65%**	16.5%	40.0%	25.7%	16.1%	545
			Cartridge cases			
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Total Sets
Matching						
Easy	97.5%	0.31%	0.31%	0%	**1.85%**	325
Average	78.2%	10.1%	8.66%	1.33%	**1.73%**	751
Hard	43.4%	29.2%	22.0%	3.61%	**1.81%**	332
Nonmatching						
Easy	**0.40%**	0.40%	1.19%	4.37%	93.6%	503
Average	**1.18%**	5.61%	22.4%	25.3%	45.5%	1694
Hard	**0.65%**	12.4%	39.6%	26.3%	21.0%	613

Examiners were also asked to rate “Individual Characteristics” as “Extensive”, “Some”, or “Limited” in each examination. Table XXVIII displays the percentage of bullet and cartridge case comparison sets classified for each of these categories, for both matching and nonmatching sets, both overall and divided by the make of handgun. For bullet comparisons, more Ruger sets were judged to have Limited characteristics, and more Beretta sets Extensive. As with Degree of Difficulty evaluations, patterns for cartridge case comparisons are more complicated.

Table XXVIII: First Round Summary Percentages of Individual Characteristics Evaluations by Type of Set and Manufacturer.

Bullets				
	Extensive	Some	Limited	Total Sets
Matching				
Beretta	65.8%	33.2%	1.07%	844
Ruger	13.4%	49.5%	37.0%	551
Overall	45.1%	39.6%	15.3%	1395
Nonmatching				
Beretta	53.4%	44.1%	2.59%	2011
Ruger	12.4%	43.7%	43.9%	814
Overall	41.6%	44.0%	14.5%	2825
Cartridge cases				
	Extensive	Some	Limited	Total Sets
Matching				
Beretta	35.3%	47.3%	17.4%	850
Jimenez	28.7%	55.2%	16.1%	558
Overall	32.7%	50.4%	16.9%	1408
Nonmatching				
Beretta	21.1%	57.3%	21.6%	1954
Jimenez	32.8%	52.5%	14.7%	857
Overall	24.7%	55.8%	19.5%	2811

The relationship of individual categorization to the distribution of evaluation scores is summarized in Table XXIX for bullets and cartridge cases, for matching and nonmatching sets. Overall patterns in this table are similar to those seen for Degree of Difficulty (Table XXVII), with proportions of correct evaluations decreasing, and Inconclusive determinations generally increasing, as the degree of Individual Characteristics is more limited.

Table XXIX: First Round Summary Percentages of Evaluations by Type of Set and Individual Characteristics.

Bullets						
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Total Sets
Matching						
Extensive	91.6%	2.86%	1.27%	0.95%	**3.34%**	629
Some	74.9%	9.76%	9.04%	3.44%	**2.89%**	553
Limited	37.1%	25.8%	30.0%	5.16%	**1.88%**	213
Nonmatching						
Extensive	**0.34%**	8.43%	12.8%	27.9%	50.5%	1174
Some	**1.13%**	10.7%	34.3%	27.5%	26.4%	1242
Limited	**0.49%**	8.56%	63.6%	18.3%	9.05%	409
Cartridge cases						
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Total Sets
Matching						
Extensive	93.5%	2.17%	2.17%	0.65%	**1.52%**	460
Some	78.5%	13.1%	4.65%	1.69%	**2.11%**	710
Limited	25.2%	30.3%	40.3%	2.94%	**1.26%**	238
Nonmatching						
Extensive	**0.29%**	3.17%	7.36%	15.0%	74.2%	693
Some	**1.27%**	7.07%	20.0%	24.1%	47.5%	1569
Limited	**0.55%**	6.92%	48.1%	24.0%	20.4%	549

Frequency of Indicators Used in Evaluation

In their evaluation of bullet sets, examiners were asked to indicate whether land impressions and/or groove impressions were informative when they classified a comparison as an Identification (either correctly or incorrectly). There is interest in knowing how often these features play a role in evaluations, in particularly in whether these frequencies are different for different weapon models. Table XXX summarizes the frequencies, among the evaluations made in the first round of the study, of ID determinations in which the examiner’s classification relied on land impressions, groove impressions, or both land and groove impressions. Note that percentage values in this row do not total 100% because the categories are not exclusive, e.g. some examinations cite both land and groove impressions, and because examiners may not have indicated either in some cases. The number of nonmatching sets classified as Identification is very small, so the percentages in this section of the table are based on a small number of sample sets, which is shown in parentheses in the table.

Table XXX: First Round Summary Percentages of Bullet Set "ID" Evaluations Referencing Land Impressions and Groove Impressions, by Set Type and Manufacturer.

	Land Impressions	Groove Impressions	Land and Groove Impressions	Sets Classified "ID"
Matching				
Beretta	98.3%	7.10%	5.78%	761
Ruger	98.1%	23.8%	21.9%	315
Nonmatching				
Beretta	100%	9.09% (1 set)	9.09% (1 set)	11
Ruger	100%	22.2% (2 sets)	22.2% (2 sets)	9

For cartridge case sets, examiners were asked to indicate whether Breech Face Marks, Firing Pin Impressions, Chamber Marks, Extractor Marks, and/or Ejector Marks were definitive when they classified a comparison as an Identification. (*N.B. Again, percentages in a row do not total 100% since multiple types of marks/impressions could be cited.*) There is interest in how often these features play a role in evaluations, in particularly in whether these frequencies are different for different weapon models. Table XXXI summarizes frequencies for each of these characteristics individually, and for both breech face impression and firing pin marks jointly, for the evaluations made in the first round of the study. The number of nonmatching sets classified as Identification is very small, with the actual number of samples sets upon which these percentages is based being shown in parentheses.

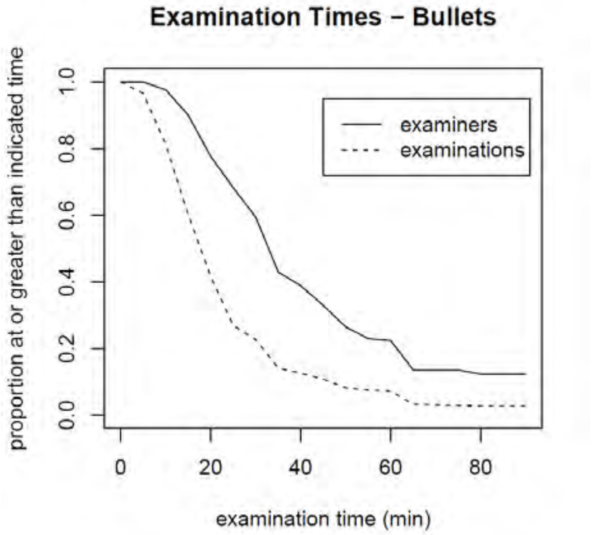
Table XXXI: First Round Summary Percentages of Cartridge case Set Identification Evaluations Referencing Five Specific Observations, by Set Type and Manufacturer.

	Breech Face Marks	Firing Pin Impressions	Breech Face and Firing Pin Impressions	Chamber Marks	Extractor Marks	Ejector Marks	Sets Classified "ID"
Matching							
Beretta	97.3%	49.4%	47.1%	0.43%	2.60%	6.36%	692
Jimenez	85.2%	61.5%	47.3%	0%	1.37%	2.47%	364
Nonmatching							
Beretta	88.2% (15 sets)	47.1% (8 sets)	41.2% (7 sets)	0%	23.5% (4 sets)	17.6% (3 sets)	17
Jimenez	77.8% (7 sets)	88.9% (8 sets)	66.7% (6 sets)	0%	0%	11.1% (1 set)	9

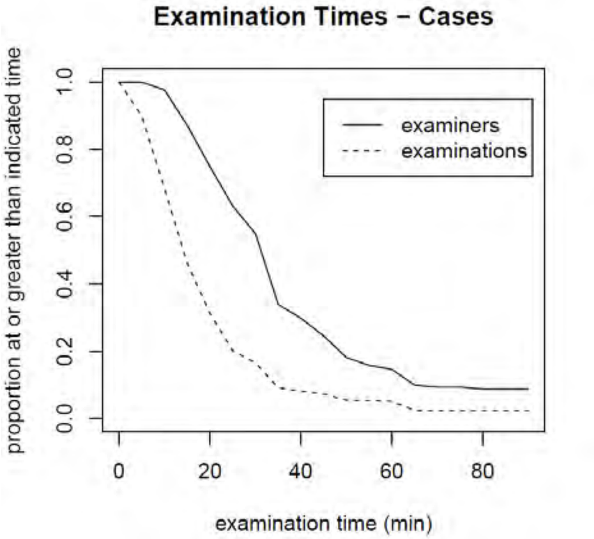
Time of Evaluation

Examiners were asked to record the amount of time in minutes they spent on each evaluation. Figure 17 displays the proportion of bullet and cartridge case comparisons, respectively, from Round 1 for which the reported time of examination was as indicated or longer, along with the proportion of examiners making these evaluations. For example, approximately 23% of bullet comparisons required 30 minutes or longer, with these examinations being reported by approximately 59% of examiners who reported examination times. Approximately 12% and 9% of examiners were responsible for reported examination times in excess of 90 minutes for bullets and cartridge cases, respectively.

Summaries of the distribution of reported times for examinations in the first round of the study that resulted in Identification, Inconclusive (pooled into a single category), and Elimination determinations are displayed as box plots in Figure 18 and given numerically in Table XXXII for bullets and cartridge cases.

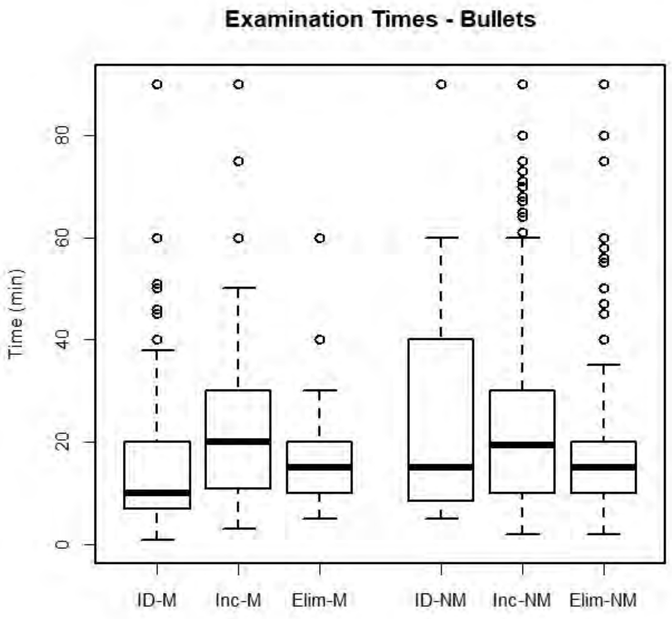


a.

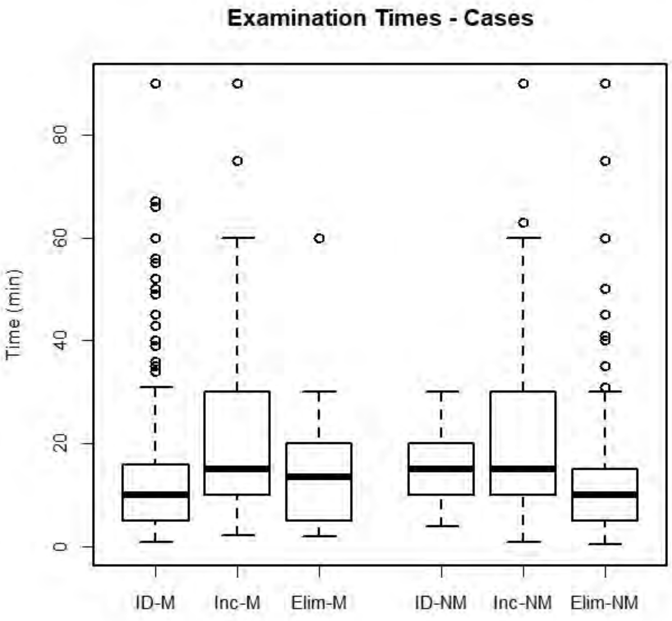


b.

Figure 17: The proportion examinations of a) bullet and b) cartridge case sets from Round 1 for which the reported examination time was as indicated or longer (dashed line), and the proportion of examiners responsible for these examinations (solid line).



a.



b.

Figure 18: Reported examination times (minutes) for a) bullet and b) cartridge case sets. Times greater than 90 minutes are not included (See Table XXXII).

Table XXXII: First Round Summary of Evaluation Times (in Minutes) for Comparisons by Set Type and Examiner's Evaluation.

Bullets						
	Matching Sets			Nonmatching Sets		
	ID	Inconclusive	Elimination	ID	Inconclusive	Elimination
Maximum	300	420	60	90	1260	540
75th Percentile	20	30	20	42.5	30	20
Median	10	20	15	15	19.5	15
25th Percentile	7	11	10	8.25	10	10
Minimum	1	3	5	5	2	2
Number of Sets	1062	280	40	20	1824	945
Cartridge cases						
	Matching Sets			Nonmatching Sets		
	ID	Inconclusive	Elimination	ID	Inconclusive	Elimination
Maximum	450	240	240	30	1140	270
75th Percentile	16	30	20	20	30	15
Median	10	15	13.5	15	15	10
25th Percentile	5	10	5	10	10	5
Minimum	1	2.25	2	4	1	1
Number of Sets	1038	334	24	26	1410	1357

Median examination times are approximately the same for (definitive) correct and incorrect determinations, ranging from 10 to 15 minutes in each case. Examinations that resulted in Inconclusive determinations took slightly more time for both matching and nonmatching bullet sets. The most extreme times recorded were for nonmatching sets that resulted in Inconclusive determinations. Of all examination times reported, 7% of those for bullets and 5% of those for cartridge cases were one hour or more. The most extreme single examination times reported were 1260 minutes (21 hours!).

Use of Consecutively Matching Striae

Examiners reported using consecutive matching striae (CMS) infrequently and inconsistently in their analysis of bullets and cartridge cases. In Round 1 of the analysis, 10 examiners (of 173) reported using CMS in examining bullets for between 1 and 29 examinations, and eight reported using CMS in examining cartridge cases for between 1 and 3 examinations. Five examiners reported some use of CMS for both bullets and cartridge cases.

Setting aside the comparison of examiners, it is of interest to compare, for examinations in which CMS was and was not used, the number of Round 1 examinations that resulted in hard errors. Table XXXIII displays the number of evaluations in which CMS was or was not used, by type of evaluation set, matching or nonmatching, and whether a hard error was made. Except in the case of matching cartridge case sets (where no false negative errors were associated with the use of CMS), the proportion of false negative determinations was greater in examinations where CMS was used, than in examinations where it was not. These differences are statistically significant for comparisons of matching bullets and nonmatching cartridge cases (Fisher's exact test, p= 0.017 and 0.045, respectively), but not for nonmatching bullets and matching cartridge cases. Even with technically significant differences, the relatively rare use of CMS in these examinations (particularly for cartridge cases) suggests that these observations should be taken only as suggestive.

Table XXXIII: Numbers of Set Evaluations in Round 1 in which CMS was or was not used, by Set Type, Matching or Nonmatching, and Hard Error or Correct/Inconclusive.

Bullet Evaluations in the First Round of the Study					
Nonmatching Sets			**Matching Sets**		
Evaluation	CMS Used	CMS Not Used	Evaluation	CMS Used	CMS Not Used
Correct or Inconclusive	67	2754	Correct or Inconclusive	48	1313
False Positive	1	19	False Negative	5	36
Cartridge case Evaluations in the First Round of the Study					
Nonmatching Sets			**Matching Sets**		
Evaluation	CMS Used	CMS Not Used	Evaluation	CMS Used	CMS Not Used
Correct or Inconclusive	4	2789	Correct or Inconclusive	7	1391
False Positive	1	25	False Negative	0	25

Of the 25 false positive evaluations of nonmatching cartridge case sets in which CMS was not used, 8 sets were produced with Beretta firearms and 17 sets were produced with Jimenez firearms. Of the 25 false negative evaluations of matching cartridge case sets in which CMS was not used, 17 sets were produced with Jimenez firearms and 8 were produced with Beretta firearms. In comparison, the ratio of total Beretta cartridge case examinations to total Jimenez examinations in Round 1 was exactly 2-to-1.

Relative Characteristics of Examiners who Made Errors

A total of 173 examiners participated in the study; hard errors (classifying a matching set as an Elimination or a nonmatching set as an Identification) were made by 34 examiners when examining bullet sets and 36 examiners when examining cartridge case sets in Round 1 (Table VII). It is of interest to know whether there are systematic patterns in examiner characteristics that might be useful in guiding education or quality improvement efforts.

Examiner Experience

Examiners were asked to report their years of training and of professional experience. The distributions of reported values for these quantities are similar for the examiners who did, and did not, make errors.

Figure 19 displays the distribution of stated experience for: (1) all examiners who responded to the question about experience, (2) those that did not make errors in evaluating bullets, (3) those that did make errors in evaluating bullets, (4) those that did not make errors in evaluating cartridge cases, and

(5) those that did make errors in evaluating cartridge cases. As is apparent from the figure, the distributions of stated experience were not materially different across these groups; Kolmogorov-Smirnov tests do not offer evidence for differences between these pairs of distributions (p = 0.978 for bullets, p = 0.751 for cartridge cases).

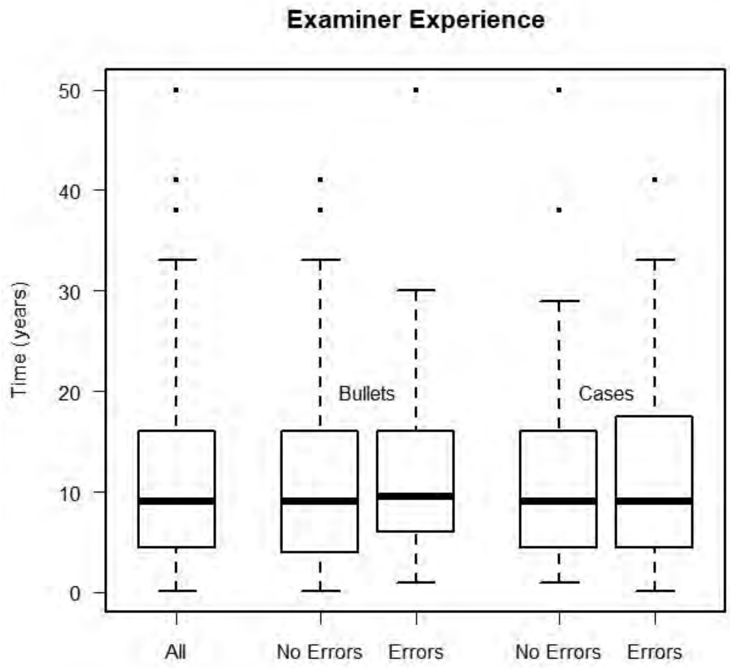


Figure 19: Stated experience (in years) for all examiners, examiners making no errors / errors in evaluating bullets, and examiners making no errors / errors in evaluating cartridge cases.

Examiners were also asked to state the extent of their training (in years). These values were separated into the five groups described for experience. In this case, the 25th percentile, median, and 75th percentile of the distributions for all five groups were all 2 years; that is, more than half of the examiners in each group responded with this value. This offers little evidence that there is any systematic difference between training and performance accuracy.

Overall Frequency of Evaluation Scores

While by definition examiners who make errors must have different frequencies of Identification and Exclusion determinations than examiners who don't, it is of interest to look at this more broadly to see if the frequency of Inconclusive determinations also differs between these groups. Table XXXIV separates the frequency-of-determination proportions from Round 1 evaluations (Table VI) by examiners who made errors and those who did not, by bullets and cartridge cases, and matching and nonmatching sets. Perhaps not surprisingly, examiners who made false positive errors used the Inconclusive-A category more often in examining nonmatching sets than did other examiners (p<.0001 for each of bullets and cartridge cases). The tabulated percentage of Inconclusive-C determinations made for matching sets was higher for examiners who made false negative determinations than for other examiners, but these differences are not statistically significant (p>.25 for each of bullets and cartridge cases).

Table XXXIV: First-Round Summary Percentages of Bullet and Cartridge case Evaluations by Set Type, and by Examiners who made no errors and Examiners who made errors.

<table>
<tr><th colspan="7">Bullet Evaluations -- Matching Sets</th></tr>
<tr><td>Examiners</td><td>ID</td><td>Inconclusive-A</td><td>Inconclusive-B</td><td>Inconclusive-C</td><td>Elimination</td><td>Total Sets</td></tr>
<tr><td><b>No F-Neg Errors</b></td><td>78.7%</td><td>9.59%</td><td>9.42%</td><td>2.33%</td><td><b>0%</b></td><td>1157</td></tr>
<tr><td><b>Made F-Neg Errors</b></td><td>66.9%</td><td>6.45%</td><td>6.45%</td><td>3.63%</td><td><b>16.5%</b></td><td>248</td></tr>
<tr><th colspan="7">Bullet Evaluations – Nonmatching Sets</th></tr>
<tr><td>Examiners</td><td>ID</td><td>Inconclusive-A</td><td>Inconclusive-B</td><td>Inconclusive-C</td><td>Elimination</td><td>Total Sets</td></tr>
<tr><td><b>No F-Pos Errors</b></td><td><b>0%</b></td><td>8.80%</td><td>30.5%</td><td>27.3%</td><td>33.4%</td><td>2659</td></tr>
<tr><td><b>Made F-Pos Errors</b></td><td><b>10.9%</b></td><td>18.6%</td><td>19.7%</td><td>10.4%</td><td>40.4%</td><td>183</td></tr>
<tr><th colspan="7">Cartridge case Evaluations – Matching Sets</th></tr>
<tr><td></td><td>ID</td><td>Inconclusive-A</td><td>Inconclusive-B</td><td>Inconclusive-C</td><td>Elimination</td><td>Total Sets</td></tr>
<tr><td><b>No F-Neg Errors</b></td><td>76.0%</td><td>12.5%</td><td>10.0%</td><td>1.43%</td><td><b>0%</b></td><td>1259</td></tr>
<tr><td><b>Made F-Neg Errors</b></td><td>61.5%</td><td>11.8%</td><td>8.70%</td><td>2.48%</td><td><b>15.5%</b></td><td>161</td></tr>
<tr><th colspan="7">Cartridge case Evaluations – Nonmatching Sets</th></tr>
<tr><td></td><td>ID</td><td>Inconclusive-A</td><td>Inconclusive-B</td><td>Inconclusive-C</td><td>Elimination</td><td>Total Sets</td></tr>
<tr><td><b>No F-Pos Errors</b></td><td><b>0%</b></td><td>5.04%</td><td>22.7%</td><td>22.6%</td><td>49.6%</td><td>2522</td></tr>
<tr><td><b>Made F-Pos Errors</b></td><td><b>8.31%</b></td><td>16.0%</td><td>20.4%</td><td>16.0%</td><td>39.3%</td><td>313</td></tr>
</table>

Examiner Repeatability

It is also of interest to know whether the repeatability characteristics of examiners who make errors are different from those who do not. Recall that repeatability is analyzed using data from Rounds 1 and 2 of the study, by comparing the two assessments of the same bullet and cartridge case sets by the same examiners. Overall proportions of intra-examiner agreement are summarized in Table XI; Table XXXV divides percent agreement further by examiners who did and did not make hard errors in Round 1. In each situation, but especially for nonmatching cartridge cases, the overall percentage of agreement is smaller for examiners who made hard errors.

Table XXXV: Proportion of Paired Classifications in Agreement by the Same Examiner (Repeatability), for Examiners who Made No Errors in Round 1, and for Examiners who Did Make Errors in Round 1.

Paired Bullet Classifications – Matching Sets		
Examiners	Agreement %	Number of Examiners
No F-Neg Errors in Round 1	80.4%	84
F-Neg Errors in Round 1	73.6%	21
Paired Bullet Classifications – Nonmatching Sets		
Examiners	Agreement %	Number of Examiners
No F-Pos Errors in Round 1	66.1%	95
F-Pos Errors in Round 1	51.2%	10
Paired Cartridge case Classifications – Matching Sets		
Examiners	Agreement %	Number of Examiners
No F-Neg Errors in Round 1	77.2%	93
F-Neg Errors in Round 1	63.3%	12
Paired Cartridge case Classifications – Nonmatching Sets		
Examiners	Agreement %	Number of Examiners
No F-Pos Errors in Round 1	64.6%	93
F-Pos Errors in Round 1	43.4%	12

Because the percentage values displayed in Table XXXV are pooled over examiners, they cannot be regarded as proportions of events from independent sampling; this makes direct statistical comparison difficult. Instead, examiner-specific observed agreement percentages (from repeated examinations of the same sets in Rounds 1 and 2) were compared for examiners who made errors in the first round and examiners who did not, using the nonparametric Kolmogorov-Smirnov Test. Resulting p-values are 0.1994 for matching bullet sets, 0.0886 for nonmatching bullet sets, 0.1321 for matching cartridge case sets, and 0.0093 for nonmatching cartridge case sets. While only the last of these would typically be regarded as strongly significant, the other three are all less than 0.2 and so are at least suggestive of possible differences between groups.

Examiners Making Multiple Errors

There is also interest in knowing whether, and to what extent, examiners made errors with both bullets and cartridge cases, or with multiple bullet or cartridge case sets. Table XXXVI displays more detail of this sort among the errors made in Round 1 evaluations by the 173 examiners.

Table XXXVI: Numbers of Examiners Making Hard Errors with Both Bullets and Cartridge cases in Round 1.

False Positives				
	No Case Errors	One Case Error	Two or More Case Errors	Total Examiners
No Bullet Errors	149	10	4	163
One Bullet Error	4	0	1	5
Two or More Bullet Errors	2	2	1	5
Total Examiners	155	12	6	173

False Negatives				
	No Case Errors	One Case Error	Two or More Case Errors	Total Examiners
No Bullet Errors	132	13	1	146
One Bullet Error	15	2	1	18
Two or More Bullet Errors	5	2	2	9
Total Examiners	152	17	4	173

Comparison Sets Evaluated Incorrectly by Multiple Examiners

Finally, by combining Rounds 1 and 3, and Rounds 2 and 3, comparison sets that were examined by two different examiners, and for which hard errors were made both times can be identified. Four matching bullet sets were found for which both evaluations were false negatives, and one nonmatching bullet set resulted in two false positive determinations. Two matching cartridge case sets were graded as false negatives by both examiners. These sets are (by group and pair codes):

Bullets:

- Org-Grp-112, 11-11 (false negatives)
- Org-Grp-35, N-N (false negatives)
- Org-Grp-81, K-K (false negatives)
- Org-Grp-22, 2-2 (false negatives)
- Org-Grp-305, 8-9 (false positives)

Cartridge cases:

- Org-Grp-43, 7-7 (false negatives)
- Org-Grp-250, 3-3 (false negatives)

An additional point should be made about the bullet set Org-Grp-81, K-K. This set was erroneously classified as Elimination in all three rounds; that is, one of the two examiners misclassified it in two different examinations.

Discussion

Accuracy

In discussing the results achieved comments will be directed initially in comparison to those obtained in the previous Baldwin [7] study, as this is the study judged by PCAST to be fundamentally sound and statistically valid. In doing so it is important to remember the distinct differences that exist between the two studies in terms of experimental parameters. These are summarized in Table XXXVII below.

Table XXXVII: Comparison of Baldwin et al. [7] vs this study.

	Baldwin Study	**Present Study**
Purpose	Accuracy	Accuracy Repeatability Reproducibility
Set Design	Open	Open
Sample Examined	Cartridge Cases	Cartridge Cases Bullets
Ammunition Used	Brass Jacketed Remington UMC 9-mm Pistol and Revolver Cartridges	Steel Jacketed Wolf Polyformance 9mm
Firearms Used	25 Ruger SR9 9mm (Radom)	Cases: 10 Jimenez JA9 9mm 27 Beretta M9A3 9mm Bullets: 10 Ruger SR9c 9mm 27 Beretta M9A3 9mm (23 of the 27 Beretta's were consecutively assembled)
Firing Sequence	Samples were fired in groups of 100 and all comparisons were within the group	Samples were fired in groups of 50. Samples divided into three Ranges (Early, Middle, Late) and comparisons were made for 9 EML combinations.
Number of Samples Collected per Firearm	Cartridge cases: 800	Cases: Jimenez 850 Beretta 700 Bullets: Ruger 850 Beretta 700
Number of Examiners	218	173
Comparisons in One Mailing	15 Cartridge case Comparisons (5 Same-Source Firearms, 10 Different-Source Firearms)	15 Case Comparisons (Variable-3 to 7 Same-Source Firearms, 8 to 12 Different-Source Firearms) 15 Bullet Comparisons (Variable-3 to 7 Same-Source Firearms, 8 to 12 Different-Source Firearms)
Single Comparison Set Makeup	3 Knowns to 1 Questioned	2 Knowns to 1 Questioned

Number of Case Comparisons	3270	10,110
Number of Bullet Comparisons	-	10,020
Error Rate (FP/FN): Cases	1.01% / 0.367%	0.933% / 1.87%
Error Rate (FP/FN): Bullets	-	0.656% / 2.87%
ID Decision Basis Information Collected	No	Yes

The overall error rates stated by Baldwin, namely, 1.01% for false positives and 0.367% for false negatives when examining cartridge cases [7] are consistent with the Accuracy rates found in this study. The most direct comparison possible is to consider cartridge case evaluations, where this study (Table VIII) found false positive rates of 0.933% and false negative error rates of 1.87% for cartridge cases. While the false positive rates match extremely well with the Baldwin study the somewhat higher false negatives recorded are possibly due to greater difficulties when faced with the steel Wolf Polyformance cartridge cases rather than brass Remington UMC since many examiners commented that they felt brass provides better marks for identification than steel. Anecdotally the Jimenez firearm is known to generate gross marks with high occurrences of subclass both for breech face marks and firing pin impressions compared to higher cost-point firearms such as the Berettas.

Sample sets were also created where there was a large number of firings of the firearm between the provided known bullets or cartridge cases and the corresponding questioned specimen. When this occurs wear of the barrel and/or lead deposition into the grooves of the rifling can be expected to be a factor, resulting in bullets with fewer characteristic marks in correspondence. Since the firearms were cleaned regularly in this study any change in characteristic markings that may have occurred are most likely related to wear. The data of Table XX, where much higher false negative error rates were seen for bullets when the comparisons were far removed from each other in firing order, would indicate that change has occurred due to wear. False negative rates for cartridge case comparisons were lower than bullets, yet still somewhat higher for greater differences in firing sequence. Thus, examinations of steel cartridge cases fired from a Jimenez firearm, with a considerable difference in firing order between the known and questioned specimens, can be expected to represent a “worst case scenario” for an examiner. Given these varying parameters (i.e. steel jacket / grossly marking firearm / time between firings), the higher false negative rates seen when compared to Baldwin are perhaps not surprising.

Various comments by examiners received during the course of the study attest to the difficulty of the comparisons and to the points discussed above. Some example comments are included to provide context. Concerning the difference in firing order:

“ … When we get test fires from a firearm, those test fires represent the condition of that gun at the time it was recovered. That becomes a baseline on which we can have absolute confidence. … One of the first and most critical questions in any test such as this is whether the “known” samples are being collected at the correct and relevant intervals with respect to the unknowns that are being generated. In normal case work TIME between the shooting event and the recovery of the firearm is a known.”

“ … when comparing these samples, even though the questioned item may have had absolutely no similarities with the known samples, I was very hesitant to eliminate it, even though I am very confident they were not fired in the same firearm. With a "single" unknown sample, there is no demonstrated repeatability of the patterns visible on this item. There is no way to determine or even estimate if there was 1 shot between the questioned and unknown or 5000 shots….”

Concerning the type of firearm possibly leading to difficulties:

“One of the major concerns is that they are presenting bullets with a high potential for subclass on some of those samples, and no mechanism for absolutely resolving that issue in the test, yet its resolution in a true laboratory setting would be as simple as doing a barrel cast and a quick visual exam…”

“In this research study, there were significant limitations on my ability to "evaluate the background" of the samples. Two questioned and one unknown sample. That's it. This study has really made me evaluate how "external information" influences my opinions; NOT regarding whether something is an IDENTIFICATION or not, but more so if two items should be ELIMINATED or INCONCLUSIVE.”

The Baldwin study did not examine bullets so a direct comparison is not possible. It is interesting that the overall error rates observed for bullet comparisons, being 0.656% and 2.87%, respectively, for false positives and false negatives from only first round comparisons (Table VIII), is statistically indistinguishable (*P*=0.19 and *P*=0.48, respectively) to the rates seen for the cartridge cases.

The concentration of the errors to a relatively small number of examiners, as was seen in Baldwin, was again noted. Examination of the data using Chi-square tests for independence show that the numbers cited above cannot be applied equally to all examiners; most examiners will perform better than the percentages cited above while a few will perform more poorly. Point estimates and confidence intervals were calculated under the assumption that examiners have different error probabilities, and that each can be represented by a beta distribution. The 95% confidence intervals presented in Table VIII represent the overall error rate that would be expected for a randomly selected examiner when asked to evaluate a randomly selected cartridge case or bullet set. The maximum likelihood estimates of error rates listed in this table should be interpreted the same way; they, rather than the overall simple proportions cited above in Table VI, should be regarded as the definitive error rate estimates from this study. For both bullets and cartridge cases the probability for a false positive is approximately ½ of what it is for a false negative, possibly reflecting examiner training that it is better to err on the side of caution.

Repeatability

When considering examiner repeatability the plots of Figures 11-13 show that examiners score high in repeatability, i.e. their observed performance generally exceeds the statistically expected agreement by a fairly wide margin. This is true whether all three inconclusive category are regarded separately, are pooled as a single category, or Identification and Inconclusive-A results are pooled and Elimination and Inconclusive-C are pooled. The greatest degree of less-than-expected agreement is seen when nonmatching sets are examined, especially for bullets. Examiner comments again indicate that this is somewhat expected (and even predicted) given the nature of the determination. As one examiner remarked**:**

“… I would be surprised if I did not "flip flop" on some of my "inconclusive-C" vs "elimination" conclusions, especially with the cartridge case comparisons.

That being said, I would be very surprised if I had "flip flopped" on any "Identification" conclusions. ... I would venture to guess that there is more variation within the Inconclusive A vs B; Inconclusive B vs C; and Inconclusive C vs Elimination.”

Only a limited number of examiners fall below the expected line. Some of these are undoubtedly due to chance, given the small number of sets re-examined by any one examiner.

Overall, this is evidence of the consistency that can be expected from examiners in evaluating cartridge case and bullet sets, even beyond what high accuracy alone would suggest. An examiner does not simply “take the same chance” (even a small one) in evaluating the same material twice, but presents findings in any one examination that can be taken as representative of what s/he would find in evaluating the same material again.

Reproducibility

Analysis of reproducibility was carried out in the same manner as for repeatability. The results displayed in Figures 14-16 are interesting in that while they show many of the same characteristics as for repeatability there are some obvious differences. For example, Figure 14 shows that in general the determinations made by different examiners are reproducible, i.e. the observed agreement falls above the expected agreement line. Just as for repeatability, the greatest variation is seen for nonmatching bullets; in this case the observed agreement approaches what is to be expected due to chance. As inconclusive ratings are pooled (Figure 15) or combined with ID or elimination categories (Figure 16) this trend continues, i.e. observed agreement generally matches expected agreement. It is not surprising that the trends shown in Figures 14-16 (Reproducibility) are not as dramatic as those seen in Figures 11-13 (Repeatability) since the reproducibility involves multiple examiners in the process rather than when a single examiner is involved for repeatability. Still, the general trend toward better observed agreement than expected agreement documents commonality in how the examination process is performed within the profession.

Inconclusive Ratings and Effect of Pooling

Concerning the ratings of Inconclusive, it is perhaps incumbent at this point to state that in the analyses conducted a clear difference exists between what are termed “hard errors” in this report and Inconclusive determinations. While Inconclusives were at times pooled with other determinations (e.g., Inconclusive-A and Identification, as in Figure 13 and Table XIII; Inconclusive C and Elimination as in Figure 16 and Table XVIII), this was not meant to imply in any way that these different ratings are equivalent.

Forensic examination must be regarded as (at least) a two-step process. The first step is an evaluation of the degree and quality of useful information associated with the material to be examined. At the second step, a conclusion of Identification or Elimination can only be justified after it is first determined that it can be supported based on the information available. For many reasons, fired bullets and cartridge cases simply do not always carry marks sufficient to support a definitive conclusion. While Inconclusive conclusions are not the most desirable outcomes, they certainly can be the most appropriate, and so should not be regarded as “errors” in the usual sense of the word.

In this context it should be clear that for examiners the Inconclusive rating is simply a reflection of the discrimination level they are imposing upon the data as they try to separate the “noise” that is present in the “system”. When confronted with a myriad of markings to be compared, a decision has to be made about whether the variations noted rise above a threshold level the examiner has unconsciously assigned for each examination. There may be numerous reasons for not obtaining this threshold, hence the different categories of uncertainty, but a declaration of uncertainty is not an error in this case any more than an instrument is in error when the set discrimination value says the incoming signal data cannot reliably be differentiated from background noise. The instrument is not broken, it is not in need of maintenance – it is functioning perfectly. It simply needs a higher signal in order to discern what is happening. In forensic examination, an Inconclusive determination is interpreted as an informed statement about the quantity of information available in the evidence – a positive statement that neither Identification nor Elimination determinations can be objectively justified.

The effect of pooling Inconclusive-A with Identification and Inconclusive-C with Elimination, as compared to addressing the data with all of the inconclusive categories simply considered together, provides some possible insights into examiner performance. When considering Repeatability of examiners, pooling of the Inconclusive A and C categories with Identification and Elimination, respectively, generally leads to greater agreement for both bullet and cartridge cases classifications for matching sets and less agreement for nonmatching sets. When considering bullets only a slightly improved proportion of paired agreements is seen when considering matching pairs. However, agreements go down considerably, over 10% difference, when nonmatching pairs are considered. A similar trend is seen for cartridge cases, although the drop seen in nonmatching pairs is somewhat less. This suggests that examiners may tend to be more certain of their Identifications than they are of Eliminations, or are more careful to err on the side of caution. In some cases examiners will not declare eliminations unless the class characteristics differ. As such, the shift seen supports the examiner comment quoted above (i.e. *" … I would be surprised if I did not “flip flop" on some of my "inconclusive-C" vs "elimination" conclusions…”*) that changes are more probable for evaluations on the Elimination end of the scale than on an Identification end.

The trends noted above are less apparent for pooling of categories when Reproducibility is considered. This is understandable given the discussion above - all examiners must establish for themselves a threshold value for evaluation – and the realization that examiners can not be treated as a homogeneous group, as shown by the error analysis conducted under the Accuracy portion of the results. Still, the paired classification results again show a greater degree of agreement for matching sets as opposed to nonmatching sets, with agreement again increasing slightly.

In summary, the value of the range of Inconclusive choices available is that it enables examiners of distinctly different personalities the ability to vary in their subjective evaluations but still render conclusions that are in agreement with one another to a point greater than what would be expected by random chance. This is true between examiners and especially evident when examiner repeatability is considered.

Manufacturing and Firing Sequence Considerations

The results obtained present clear evidence that firearm make and manufacturing effects play a role in examiner accuracy. Error proportions are relatively smaller, and correct conclusion proportions relatively larger, for the Beretta as opposed to the Ruger and Jimenez models. These results tally with examiner perceptions that more often rated matching bullet and cartridge case comparisons from the

Beretta firearms “Easy” compared to the corresponding bullets and cartridge cases from the Rugers and Jimenez firearms, respectively. For example, in the case of matching bullets the Beretta comparisons were almost four times as likely to be rated Easy as compared to the Rugers. This means that the performance of an examiner when confronted with a particular firearm will most likely vary from the average point estimate calculated in this study. However, it should be noted that the variations seen due to firearm type fall within the 95% confidence intervals (Table VIII) calculated in almost all cases, the exception being for false negatives in cartridge case comparisons where the variation is outside the 95% confidence interval by .03% on the high end.

Similarly, separation in firing order between knowns and unknowns can make a difference, especially if there is a large difference in firing order between the comparisons. While comparisons between samples fired relatively close together in sequence can easily be identified with errors within the bounds of the determined confidence intervals the same is not true for those widely separated. What is interesting is that the possibility for false positives can actually decrease while the possibility of a false negative can increase, depending on the particular make of firearm and whether a bullet or a cartridge case is being considered. This again might speak to a tendency of examiners to err on the “safe” side, i.e. being more likely to declare a false negative than a false positive.

Not knowing the difference in firing sequence was one of the complaints raised by the participating examiners. In actual practice examiners often would know the amount of time that had passed between the questioned use of a firearm in a crime and when it was made available to them for test fires. This might allow them to make some reasonable assumptions about the likelihood of there being a large number of firings between the comparison samples.

Other

Comments received from examiners during the study were generally to the effect that the study parameters did not allow them to follow normal procedures that would have better enabled them to arrive at either an Identification or Elimination. As seen above, not having access to the actual firearm and not knowing the number of firings between samples provided for examination, were common complaints that examiners stated would cause them to be more cautious than they ordinarily would have been when declaring eliminations.

When asked to comment on the difficulty level of any particular evaluation the most common response was “Average”. Examiners ranked Beretta examinations “Easy” as compared to Ruger or Jimenez, with the exception of nonmatching Jimenez cartridge case comparisons. “Easy” evaluations had lower error rates associated for them when compared to the overall average point estimates, a fact which perhaps is not surprising. It is interesting ,however, that examinations where the sample was rated as having “Extensive” markings did not always translate to a low error rate. For example, examiners rated 91.6% of the first round known bullet match samples as having “Extensive” markings, yet the false negative error rate was 3.34% as compared to 2.89% and 1.88% for the samples with “Some” or “Limited” markings. This suggests that possibly an overabundance of markings may cause confusion in the mind of an examiner. This trend did not hold true for nonmatching samples or for cartridge case samples, where the false positive and false negative errors varied. It would seem that the best conclusion that can be drawn is that the number of markings available is less important than the quality of the marks that are present, however many they may be.

When considering other effects that may contribute to the declaration of either an Identification or Elimination examiners clearly rely on Breech Face marks and Firing Pin impressions for cartridge case comparisons and Land impressions for bullets. For cartridge case comparisons, extractor and ejector marks are included more often in the examination if the comparison is nonmatching as opposed to matching. The majority of examiners spent 30 minutes or less for an examination, although times in excess of 90 minutes were reported by ≈ 10% of the examiners and some extremely long time (several hours) were reported in some cases (possibly erroneously). An analysis of Examiner Experience as compared to the tendency to commit an error (Fig. 19) showed no relationship between the two.

The absence of widespread use of the Consecutive Matching Striae (CMS) prevented any meaningful analysis of the effect on examiner performance of those using this technique. However, it was an interesting observation that a statistically significant increase in false negative determinations was made by those using CMS.

While the pilot study participants suggested no more than 10 comparison sets be provided per mailing the number of 30 proposed by the funding agency was felt necessary to obtain the required data. This resulted in the high number of dropouts seen in the early mailings of the study as many participants realized they did not have enough time to attend to their normal duties and complete the study examinations in a timely manner. This may have resulted in a potential bias toward examiners with a lower case load. Any arguments as to how this may have affected the determined error rates is mere speculation. The realization that examiners can not be treated as a homogenous entity, which led to the calculation of confidence intervals for examiner performance, effectively addresses the possibility of the effect this bias may have on the determined error rates. As the data acquired was obtained from a broad cross-section of qualified examiners practicing in various locations around the country the determined rates are believed to be as accurate as possible and any possible bias effect slight.

Summary and Conclusions

Based on the experimental results obtained in this study the following conclusions can be drawn:

1. The results observed in this study are consistent with the results of Baldwin, et al [7]. Using a beta-binomial model, maximum likelihood estimates for false positive and false negative error probabilities were calculated as 0.656% and 2.87% for bullets and 0.933 and 1.87%, for cartridges, respectively. The 95% confidence intervals for false positives and false negatives range from 1.42% – 0.305% and 4.26% - 1.89%, respectively, in bullets. Similarly, in cartridges the 95% confidence intervals are in the ranges 1.574% - 0.548% and 2.99% - 1.16% for false positives and false negatives, respectively.

2. Examiners scored high in repeatability with their observed performance generally exceeding expected agreement. Reproducibility between examiners was also generally above expected agreement. Both repeatability and reproducibility were less consistent for nonmatching sets.

3. The majority of errors were made by a limited number of examiners. For example, of the 173 examiners, 139 made no “hard errors” of either kind when examining bullets, and 3 made both kinds of errors. In the accuracy round of the study, “hard errors” were made by 34/36 of the 173 examiners

when examining bullets/cartridge cases. The six most error-prone examiners account for almost 30% (33 of the 112) of the total errors while thirteen examiners account for almost half of all the “hard errors” (54 of 112). These observations are consistent with the results of Baldwin.

4. The results of pooling Inconclusive-A results with Identification and Inconclusive-C results with Elimination to create a three factor scale rather than the AFTE five factor scale showed examiners tend to me more careful and sure of their Identification conclusions than of their Elimination ones. The five-factor system would appear to be a suitable way to account for variability in how different examiners view the amount of data (i.e. markings) available and their ability to arrive at a decision that satisfies their internal personal standards based on training and experience.

5. The Beretta firearm gave better results when considering both matching and nonmatching comparisons than the Ruger when considering bullets fired within a short temporal span. Similarly, the Beretta gave better results than the Jimenez when considering matching cartridge cases under the same time restrictions. This is in agreement with anecdotal statements about these firearms by examiners. While point estimates of errors for specific firearms are therefore predicted to vary, the values calculated generally fall within the 95% confidence intervals established in this study.

6. As the separation in firing order increased for both matching and non-matching comparisons, the results became more varied. The possibility of making a false-positive error decreased for all makes of firearms while the error rate for false negatives increased.

7. For both Matching and Nonmatching sets, the percentage of correct evaluations decreases substantially, and the percentage of each category of inconclusive evaluation generally increases, for comparisons rated as difficult by the examiners; this pattern is more pronounced for bullet than for cartridge case comparisons. Overall, the proportions of correct evaluations decrease, and those of the inconclusive determinations generally increase, when examiners rate the degree of individual characteristics as being more limited.

8. No definitive relationships were observed between an examiner’s length of experience or length of training and the propensity to make “hard errors” for either bullet or cartridge case comparisons.

9. Comparison sets that resulted in errors by more than one examiner have been identified in this report. It is suggested that these sets be examined by trained forensic examiners at the FBI to determine what may be the cause behind the errors. It is possible that lessons can be learned from these particular sets that can be used to increase examiner proficiency in the future.

References

1) President’s Council of Advisors on Science and Technology (PCAST), “Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods,” Office of Science and Technology Policy, 2016.

2) J. Gouwe, J. Hamby, and S. Norris, “Comparison of 10,000 Consecutively Fired Cartridge Cases from a Model 22 Glock .40 S&W Caliber Semiautomatic Pistol,” *AFTE Journal* **40** (1), 57-63 (2008).

3) Bunch, S. G., Murphy D., "A Comprehensive Validity Study for the Forensic Examination of Cartridge Cases," *AFTE Journal* **35** (2), 201-203 (2003).

4) James Hamby and James Thorpe, "The Examination, Evaluation and Identification of 9mm Cartridge Cases Fired from 617 Different Glock Model 17 & 19 Semiautomatic Pistols", *AFTE Journal* **41** (4) 310-324 (2009).

5) LaPorte, D., "An Empirical Validation Study of Breechface Marks on .380 ACP Caliber Cartridge Cases Fired from Ten Consecutively Finished Hi-Point Model C9 Pistols," *AFTE Journal* **43** (4), (2011).

6) "An Empirical Study to Improve the Scientific Foundation of Forensic Firearm and Tool Mark Identification Utilizing 10 Consecutively Manufactured Slides", Award Number 2009-DN-BX-K230, Final Report, Submitted By: Miami-Dade Police Department Crime Laboratory, 2011. https://www.ncjrs.gov/pdffiles1/nij/grants/237960.pdf

7) Baldwin, D.P., Bajic, S.J., Morris, M., and D. Zamzow (2014). "A study of false-positive and false-negative error rates in cartridge case comparisons," Ames Laboratory, USDOE, Technical Report #IS-5207.

8) Association of Firearms and Toolmark Examiners (AFTE), reproduced from https://afte.org/about-us/what-is-afte/afte-range-of-conclusions

9) Houck MA. 19th INTERPOL International Forensic Science Managers Symposium -- Review Papers 2019; 4-38, 68-107. Found at: https://www.interpol.int/en/content/download/14458/file/Interpol%20Review%20Papers%202019.pdf

10) Houck MA. 18th INTERPOL International Forensic Science Managers Symposium – Review Papers 2016; 4-36, 90-113. Found at: https://www.interpol.int/en/content/download/13471/file/Organising%20Committee%20Members%20IFSMS%202016.pdf

11) AFTE website, https://afte.org/resources/afte- position-documents

12) AFTE website, https://afte.org/resources/swggun-ark.

13) Smith TP, Smith GA, Snipes JB. A Validation Study of Bullet and Cartridge Case Comparisons Using Samples Representative of Actual Casework. Journal of Forensic Science 2016; 61(4): 939-946.

14) Keisler MA, Hartman S, Kilmon A, Oberg M, Templeton M. Isolated Pairs Research Study. AFTE Journal 2018; 50(1): 56-58.

15) Eric Hare, Heike Hoffman, and Alicia Carriquiry, "Automatic Matching of Bullet Land Impressions," Annals of Applied Statistics (2017) 11:2332-2356.

16) Eric Hare, Heike Hoffman, and Alicia Carriquiry, "Algorithmic Approaches to Match Degraded Land Impressions," Law, Probability and Risk, (2017) 16: 203-221

17) Pierre Duez, Todd Weller, Marcus Brubaker, Richard Hockensmith II, and Ryan Lilien, “Development and Validation of a Virtual Examination Tool for Firearms Forensics,” (2017) Journal of Forensic Sciences, Volume 62, Number 6.

18) Daniel Ott, Robert Thompson, and Junfeng Song, “Applying 3D Measurements and Computer Matching Algorithms to Two Firearms Examination Proficiency Tests,” Forensic Science International (2017) 271: 98-106

19) John Murdock, Nicholas D.K. Petraco, John Thornton, Michael Neel, Todd Weller, Robert Thompson, James Hamby, and Eric Collins, “The Development and Application of Random Match Probabilities to Firearm and Toolmark Identification,” Journal of Forensic Sciences (2017) May Vol 62, No. 3.

20) Xiao Tai, and William Eddy, “A Fully Automatic Method for Comparing Cartridge Case Images,” Journal of Forensic Sciences (2018) March Volume 63, Number 2.

21) Min Yang, Li Mou, Yi-Ming Fu, Yu Wang, and Jiang-Feng Wang, “Quantitative Statistics and Identification of Tool-Marks,” Journal of Forensic Sciences, (2019) Volume 64, Number 5.

22) Ganesh Krishnan, and Heike Hoffman, “Adapting the Chumbley Score to Match Striae on Land Engraved Areas (LEAs) of Bullets,” Journal of Forensic Sciences (2019) Volume 64, Number 3.

23) Danny Roberg, Alain Beauchamp, Serge Levesque, “Objective Identification of Bullets Based on 3D Pattern Matching and Line Counting Scores,” International Journal of Pattern Recognition and Artificial Intelligence (2019): Volume 33, Number 11.

24) James Hamby, David Brundage, Nicholas D. K. Petraco, and James Thorpe, “A Worldwide Study of Bullets Fired From 10 Consecutively Rifled 9MM RUGER Pistol Barrels – Analysis of Examiner Error Rates,” Journal of Forensic Sciences (2018) Volume 64, Number 2.

25) Erwin J.A.T. Mattijssen, Cilia L.M. Witteman, Charles E.H. Berger, Nicolaas W. Brand, Reinoud D. Stoel, “Validity and Reliability of Forensic Firearms Examiners,” Forensic Science International 307 (2020), Pages 1-14.

26) Dumville, J.C.; Torgerson, D.J.; Hewitt, C.E.; “Reporting attrition in randomised controlled trials”, Brit Med J 332(7547) (2006) 969-971.

27) Welch, A.K.; “History and Manufacturing Process of the Jennings / Bryco / Jimenez Arms Pistols,” AFTE J 45(3) (2013) 260-266.

28) Welch,A.K.; “Breech Face Subclass Characteristics of the Jimenez JA Nine Pistol,” AFTE J 45(4) (2013).

29) Monturo,C.; “Breech Face Marks Of The Bryco Arms Model Jennings Nine,” AFTE J 31(1) (1999).

30) Weller,T., Zheng, A., Thompson, R., Tulleners, F.; "Confocal Microscopy Analysis of Breech Face Marks on Fired Cartridge Cases from 10 Consecutively Manufactured Pistol Slides," J. Forensic Sci. 57(4) (2012) 912-917.

31) Nichols,R.; "Subclass characteristics: From origin to evaluation," AFTE J 50(2) (2018) 68-88.

32) Miller,J., Beach, G.; "Toolmarks: Examining the possibility of subclass characteristics," AFTE J 37(4) (2005) 296-34.

33) Lomoro, V.J.; "32 SWL Caliber, F.I.E. Corp., Titanic revolvers" AFTE Newsletter # 20 June 1972, AFTE Journal Vol. 6 No. 2, April 1974 and AFTE Journal Vol. 9 No. 2 July 1977.

34) Agresti, A. (2007). *An introduction to Categorical Data Analysis*, John Wiley and Sons, Hoboken, NJ.

35) Clopper, C. and E.S. Pearson (1934). "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika* **26** (4): 404-413.

36) R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

37) VGAM package: http://www2.uaem.mx/r-mirror/web/packages/VGAM/VGAM.pdf

38) Vardeman, S.B. and J.M. Jobe (2016). *Statistical Methods for Quality Assurance: Basics, Measurement, Control, Capability, and Improvement, 2^nd Edition*, Springer-Verlag, New York, NY.

Acknowledgments

This study was designed and supported by the Federal Bureau of Investigation (FBI) through contract with the Ames Laboratory, U.S. Department of Energy, under Interagency Agreement DJF-15-1200-0000667. The authors wish to acknowledge Dr. David Baldwin who was responsible for initiating an earlier related study. The authors also acknowledge useful discussions with Dr. Keith Monson, Dr. Michael Smith, Mr. Erich Smith, Dr. Gene Peters and Ms. Heather Seubert. Special thanks are due to Keith and Erich for obtaining the samples used in this study and for the people who assisted them, particularly Ms. Jennifer Stephens. This work was carried out at the Ames Laboratory. Ames Laboratory is operated for the U.S. Department of Energy by Iowa State University under Contract No. DE-AC02-07CH11358.

Appendices

Appendix A: Invitation and Informed Consent

AMES
LABORATORY
Creating Materials & Energy Solutions
U.S. DEPARTMENT OF ENERGY

March 7, 2018

Dear Firearm Examiner,

You are being invited to participate in a validation study of firearms forensic comparisons. This study assesses the accuracy, repeatability, and reproducibility of decisions involving forensic comparisons. Only Firearm Examiners who are currently conducting examinations and are members of AFTE or are employed in the firearms section of an accredited crime laboratory are being asked to participate. This study is sponsored by the Federal Bureau of Investigation (FBI) and will be carried out by the Ames Laboratory, USDOE.

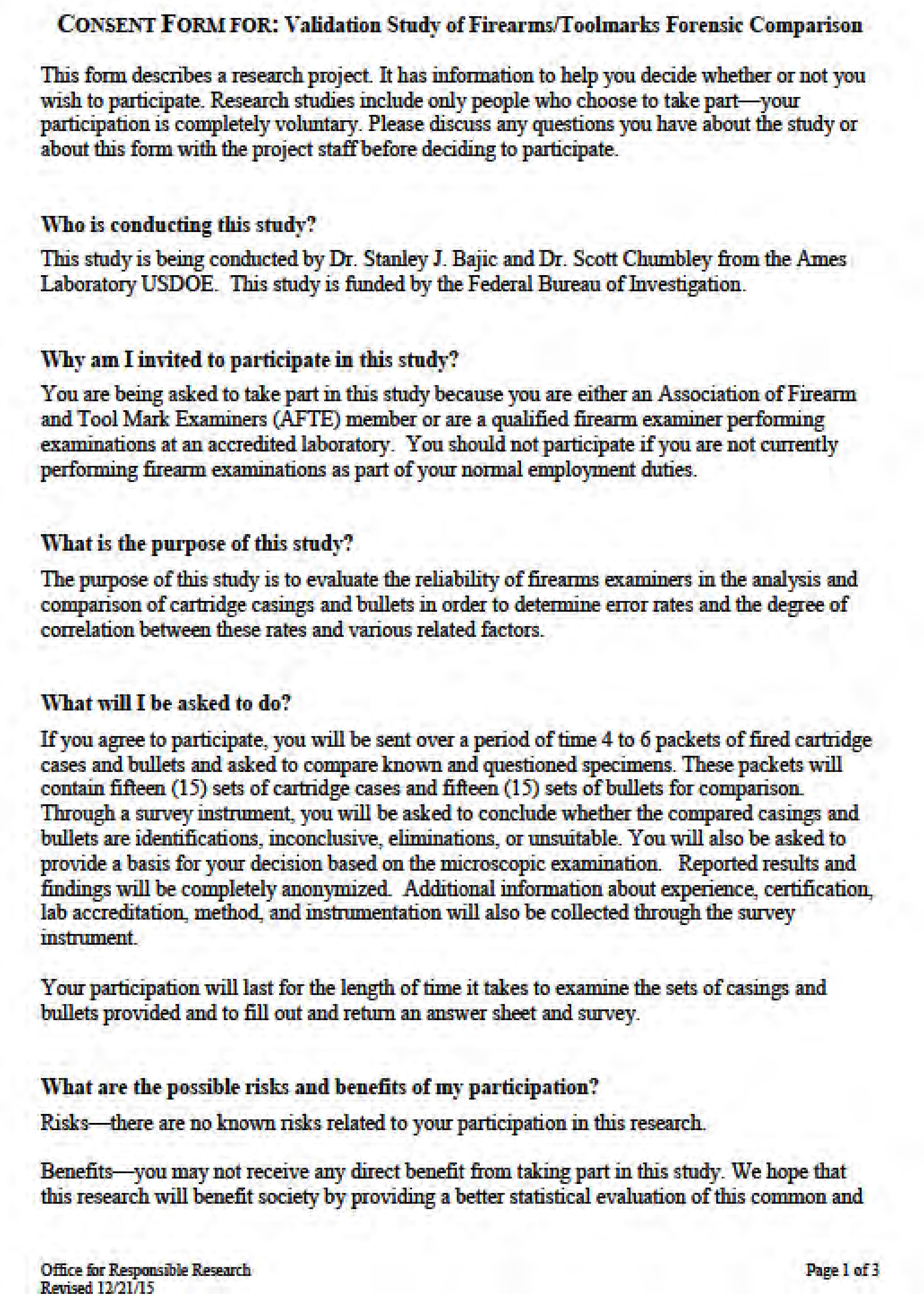
Participation is completely voluntary. There is no compensation for participating in this study. Participating examiners will be sent sets of cartridge cases and bullets and be asked to compare known and questioned specimens. The study will consist of 4 to 6 of these packets being sent to an examiner over an approximately two-year period. Through a survey instrument, you will be asked to conclude whether the compared samples are identifications, inconclusive, or eliminations. Reported results and findings will be completely anonymized. Individual results will not be disclosed to the subjects or their employers, even if requested by the examiners, unless compulsorily by law. Additional information about experience, certification, lab accreditation, method and instrumentation will also be collected through the survey instrument. The study findings will result in a peer-reviewed publication that will be relevant to the legal admissibility of such analyzes.

If you are interested in participating in this study, please complete the attached consent form and return it to the Ames Laboratory no later than March 30, 2018. If you need further information, please contact Dr. Scott Chumbley – (chumbley@ameslab.gov; 515-294- 1435).

Sincerely,

Scott Chumbley

L. Scott Chumbley
Faculty Scientist, Ames Laboratory
Professor, Iowa State University
Fellow ASM



CONSENT FORM FOR: Validation Study of Firearms/Toolmarks Forensic Comparison

This form describes a research project. It has information to help you decide whether or not you wish to participate. Research studies include only people who choose to take part—your participation is completely voluntary. Please discuss any questions you have about the study or about this form with the project staff before deciding to participate.

Who is conducting this study?

This study is being conducted by Dr. Stanley J. Bajic and Dr. Scott Chumbley from the Ames Laboratory USDOE. This study is funded by the Federal Bureau of Investigation.

Why am I invited to participate in this study?

You are being asked to take part in this study because you are either an Association of Firearm and Tool Mark Examiners (AFTE) member or are a qualified firearm examiner performing examinations at an accredited laboratory. You should not participate if you are not currently performing firearm examinations as part of your normal employment duties.

What is the purpose of this study?

The purpose of this study is to evaluate the reliability of firearms examiners in the analysis and comparison of cartridge casings and bullets in order to determine error rates and the degree of correlation between these rates and various related factors.

What will I be asked to do?

If you agree to participate, you will be sent over a period of time 4 to 6 packets of fired cartridge cases and bullets and asked to compare known and questioned specimens. These packets will contain fifteen (15) sets of cartridge cases and fifteen (15) sets of bullets for comparison. Through a survey instrument, you will be asked to conclude whether the compared casings and bullets are identifications, inconclusive, eliminations, or unsuitable. You will also be asked to provide a basis for your decision based on the microscopic examination. Reported results and findings will be completely anonymized. Additional information about experience, certification, lab accreditation, method, and instrumentation will also be collected through the survey instrument.

Your participation will last for the length of time it takes to examine the sets of casings and bullets provided and to fill out and return an answer sheet and survey.

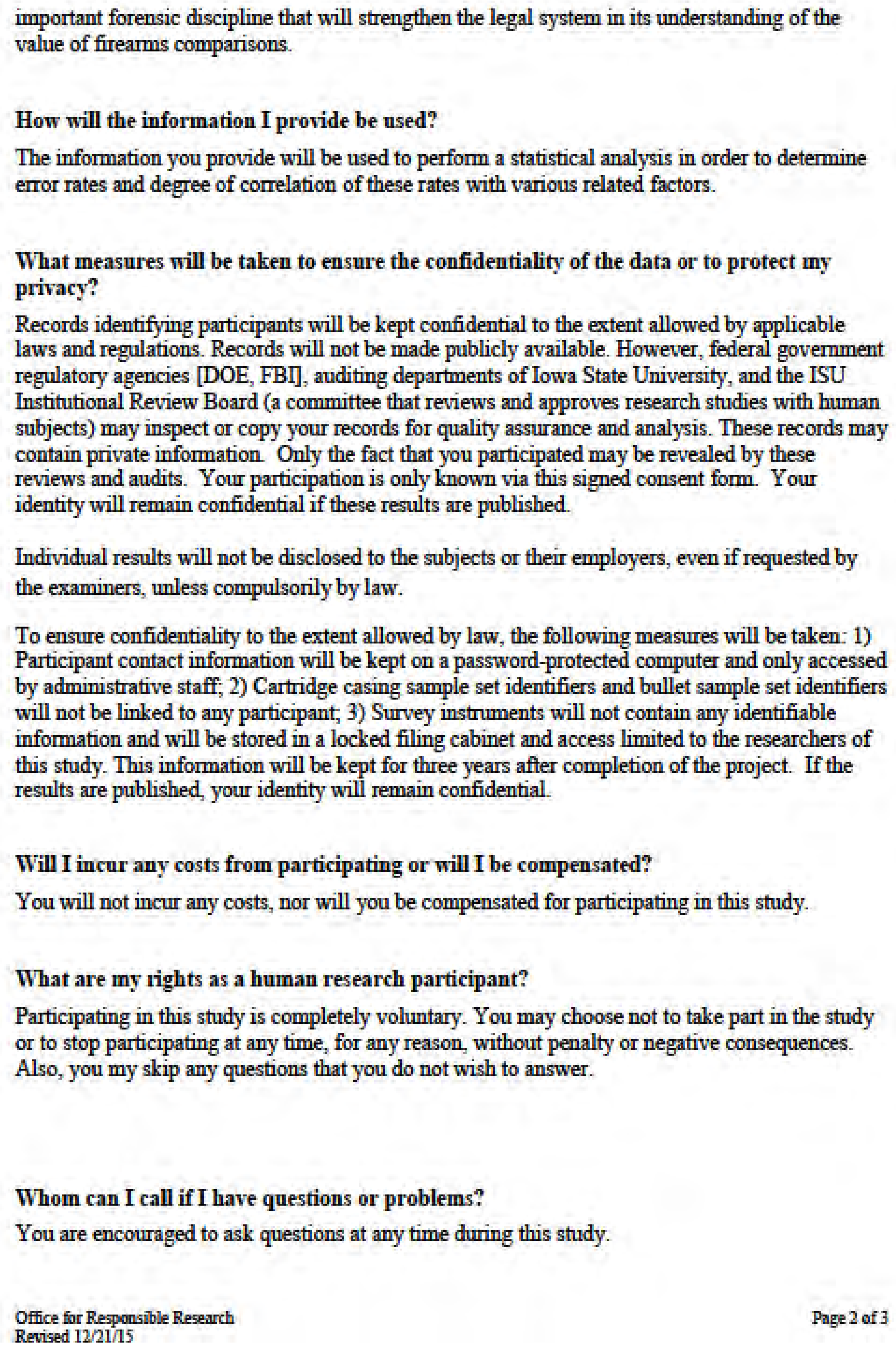
What are the possible risks and benefits of my participation?

Risks—there are no known risks related to your participation in this research.

Benefits—you may not receive any direct benefit from taking part in this study. We hope that this research will benefit society by providing a better statistical evaluation of this common and

Office for Responsible Research
Revised 12/21/15

Page 1 of 3



important forensic discipline that will strengthen the legal system in its understanding of the value of firearms comparisons.

How will the information I provide be used?

The information you provide will be used to perform a statistical analysis in order to determine error rates and degree of correlation of these rates with various related factors.

What measures will be taken to ensure the confidentiality of the data or to protect my privacy?

Records identifying participants will be kept confidential to the extent allowed by applicable laws and regulations. Records will not be made publicly available. However, federal government regulatory agencies [DOE, FBI], auditing departments of Iowa State University, and the ISU Institutional Review Board (a committee that reviews and approves research studies with human subjects) may inspect or copy your records for quality assurance and analysis. These records may contain private information. Only the fact that you participated may be revealed by these reviews and audits. Your participation is only known via this signed consent form. Your identity will remain confidential if these results are published.

Individual results will not be disclosed to the subjects or their employers, even if requested by the examiners, unless compulsorily by law.

To ensure confidentiality to the extent allowed by law, the following measures will be taken: 1) Participant contact information will be kept on a password-protected computer and only accessed by administrative staff; 2) Cartridge casing sample set identifiers and bullet sample set identifiers will not be linked to any participant; 3) Survey instruments will not contain any identifiable information and will be stored in a locked filing cabinet and access limited to the researchers of this study. This information will be kept for three years after completion of the project. If the results are published, your identity will remain confidential.

Will I incur any costs from participating or will I be compensated?

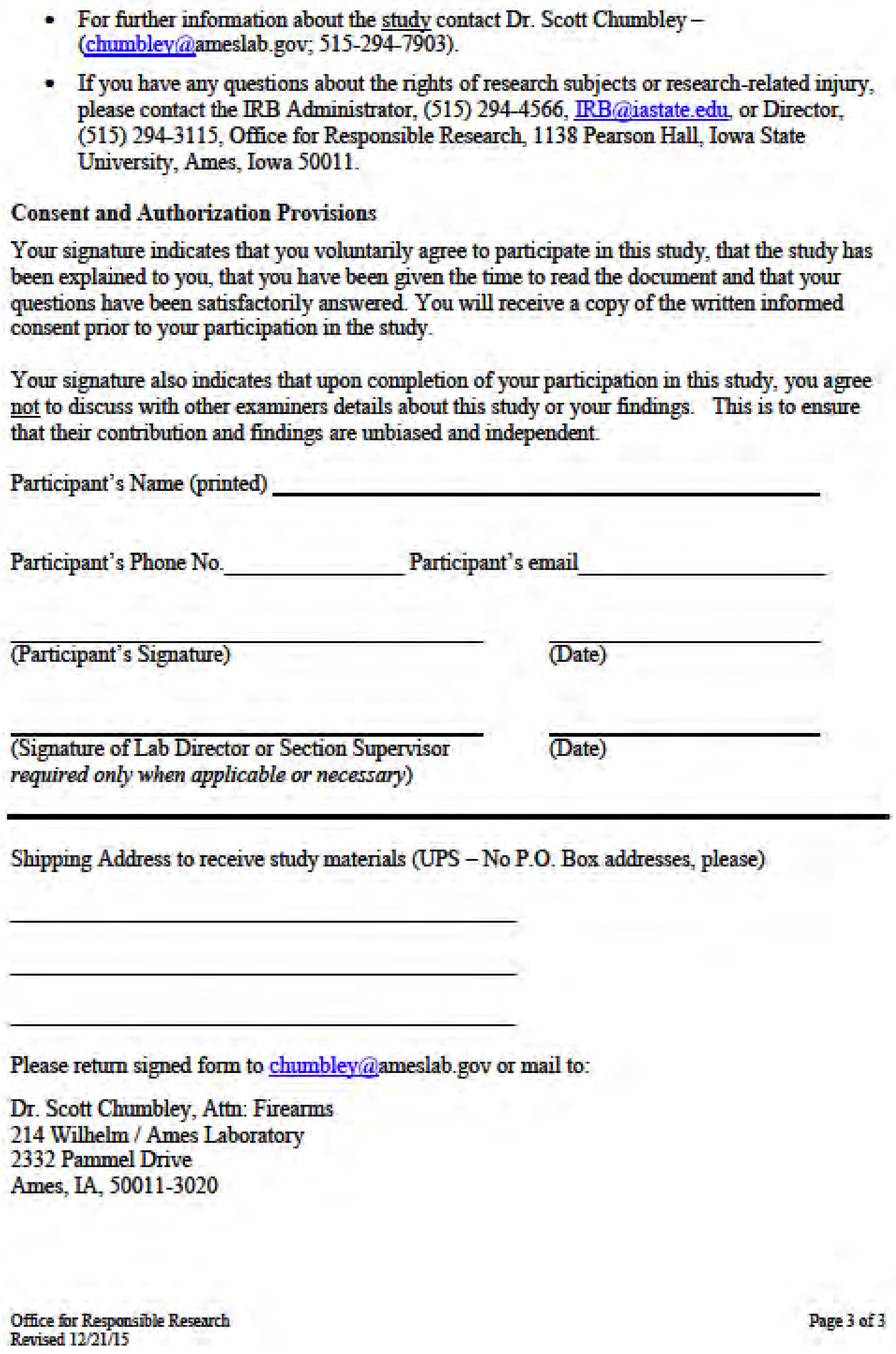
You will not incur any costs, nor will you be compensated for participating in this study.

What are my rights as a human research participant?

Participating in this study is completely voluntary. You may choose not to take part in the study or to stop participating at any time, for any reason, without penalty or negative consequences. Also, you my skip any questions that you do not wish to answer.

Whom can I call if I have questions or problems?

You are encouraged to ask questions at any time during this study.



- For further information about the study contact Dr. Scott Chumbley – (chumbley@ameslab.gov; 515-294-7903).
- If you have any questions about the rights of research subjects or research-related injury, please contact the IRB Administrator, (515) 294-4566, IRB@iastate.edu, or Director, (515) 294-3115, Office for Responsible Research, 1138 Pearson Hall, Iowa State University, Ames, Iowa 50011.

Consent and Authorization Provisions

Your signature indicates that you voluntarily agree to participate in this study, that the study has been explained to you, that you have been given the time to read the document and that your questions have been satisfactorily answered. You will receive a copy of the written informed consent prior to your participation in the study.

Your signature also indicates that upon completion of your participation in this study, you agree not to discuss with other examiners details about this study or your findings. This is to ensure that their contribution and findings are unbiased and independent.

Participant's Name (printed) ______________________________

Participant's Phone No.________________ Participant's email______________________

______________________________ ______________
(Participant's Signature) (Date)

______________________________ ______________
(Signature of Lab Director or Section Supervisor (Date)
required only when applicable or necessary)

Shipping Address to receive study materials (UPS – No P.O. Box addresses, please)

____

____

____

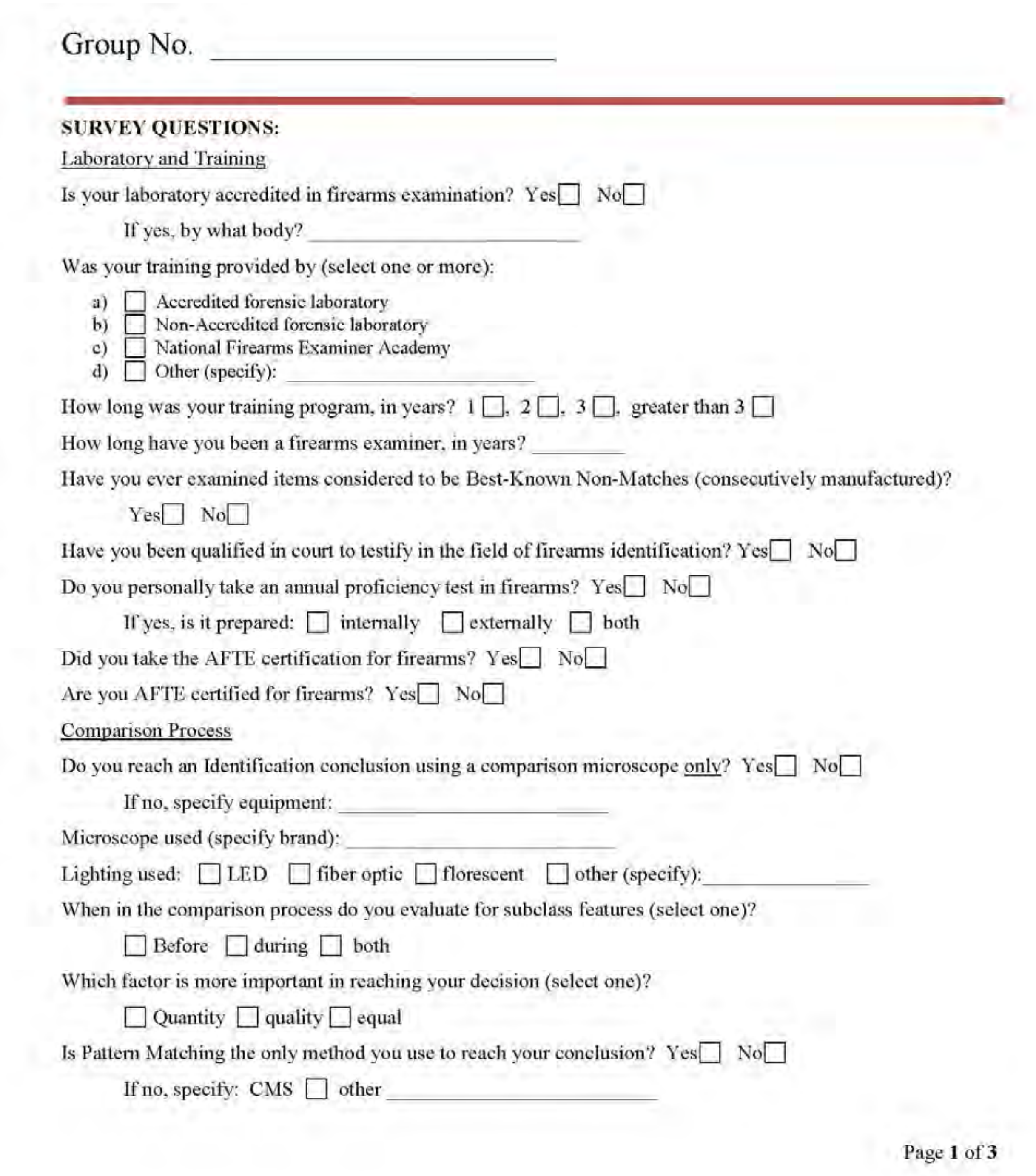
Please return signed form to chumbley@ameslab.gov or mail to:

Dr. Scott Chumbley, Attn: Firearms
214 Wilhelm / Ames Laboratory
2332 Pammel Drive
Ames, IA, 50011-3020

Appendix B: Participant Survey and Response Summary

Form B1: Blank Participant Survey Form

Participant Survey



Group No. ______________________

SURVEY QUESTIONS:

Laboratory and Training

Is your laboratory accredited in firearms examination? Yes□ No□

If yes, by what body? ______________________

Was your training provided by (select one or more):

a) □ Accredited forensic laboratory
b) □ Non-Accredited forensic laboratory
c) □ National Firearms Examiner Academy
d) □ Other (specify): ______________________

How long was your training program, in years? 1 □, 2 □, 3 □, greater than 3 □

How long have you been a firearms examiner, in years? ________

Have you ever examined items considered to be Best-Known Non-Matches (consecutively manufactured)?

Yes□ No□

Have you been qualified in court to testify in the field of firearms identification? Yes□ No□

Do you personally take an annual proficiency test in firearms? Yes□ No□

If yes, is it prepared: □ internally □ externally □ both

Did you take the AFTE certification for firearms? Yes□ No□

Are you AFTE certified for firearms? Yes□ No□

Comparison Process

Do you reach an Identification conclusion using a comparison microscope only? Yes□ No□

If no, specify equipment: ______________________

Microscope used (specify brand): ______________________

Lighting used: □ LED □ fiber optic □ florescent □ other (specify):______________

When in the comparison process do you evaluate for subclass features (select one)?

□ Before □ during □ both

Which factor is more important in reaching your decision (select one)?

□ Quantity □ quality □ equal

Is Pattern Matching the only method you use to reach your conclusion? Yes□ No□

If no, specify: CMS □ other ______________________

Page **1** of **3**

Do you use CMS in your decision making process? Yes□ No□

If no, skip to the section, *Documenting and Reporting Eliminations and Inconclusives*

CMS (applies to striated marks only)

Do you use CMS for documentation only? Yes□ No□

Is CMS recognized by your laboratory protocols as a method for documentation? Yes□ No□

Is CMS recognized by your laboratory protocols for determination of an Identification? Yes□ No□

If yes, is it optional? Yes□ No□

Do you use a comparison microscope to reach your conclusion by CMS? Yes□ No□

Do you use a photograph to reach your conclusion by CMS? Yes□ No□

If your Pattern Matching (PM) decision is Inconclusive do you attempt CMS? Yes□ No□ N/A□

If your PM decision is Inconclusive, could the results of CMS change the final conclusion to Identification?

Yes□ No□ N/A□

If your PM decision is Identification, could the results of CMS change the final conclusion to Inconclusive?

Yes□ No□ N/A□

What is your count threshold for CMS? ___________________________

During your evaluation and use of CMS, have you ever considered that your threshold count should change?

Yes□ No□ If yes, how? ___________________________

Could a conclusion of Identification be made if the CMS count threshold was not met? Yes□ No□

Documenting and Reporting Eliminations and Inconclusives

Do your laboratory protocols permit an Elimination decision based on differences in individual characteristics?

Yes□ No□

Do your laboratory protocols permit an Inconclusive decision? Yes□ No□ If no:

Is a comparison that is not an Identification considered an Elimination? Yes□ No□

If you use some other term to report such a conclusion, what is it? ___________________________

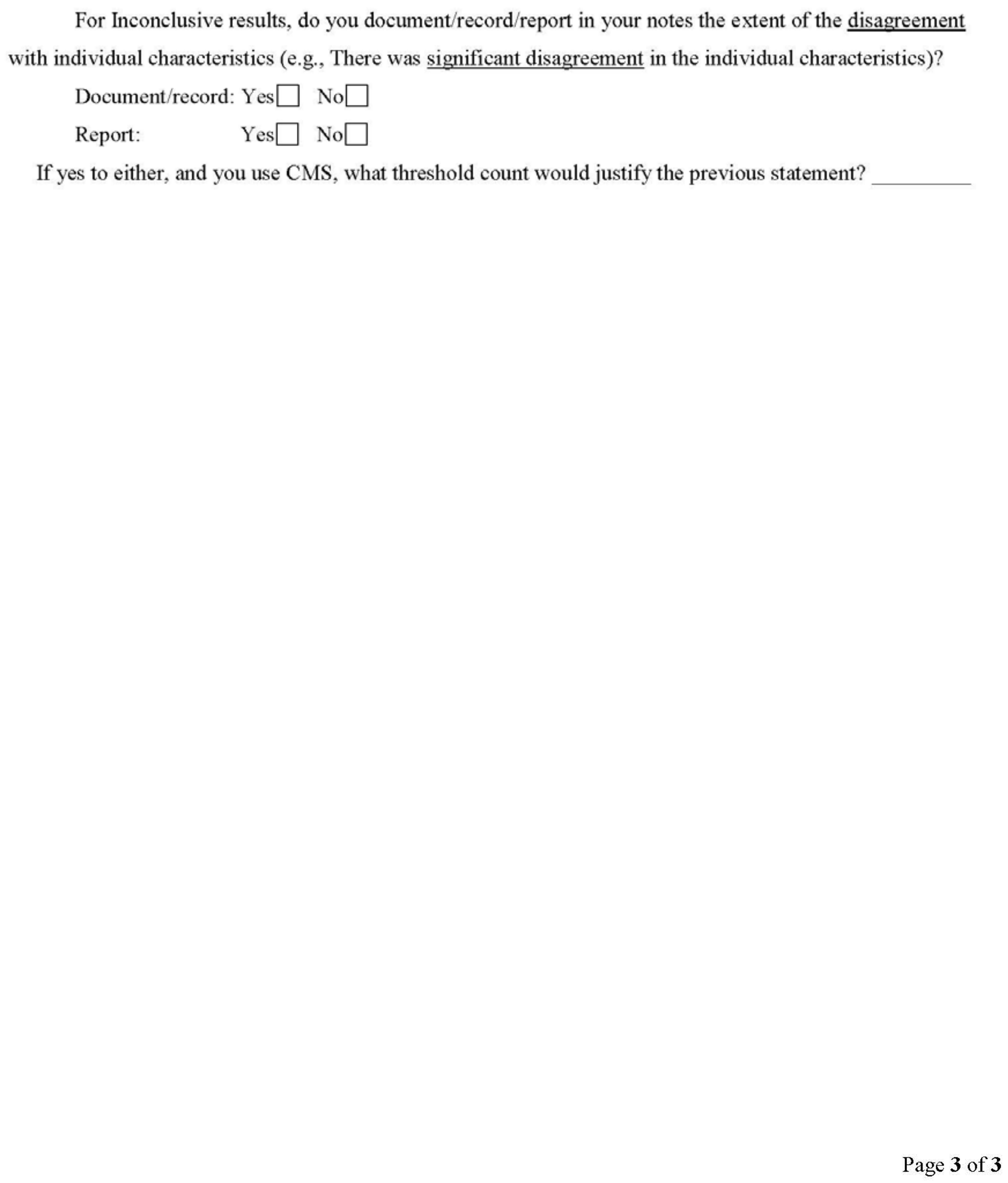
Is this policy explained in your report? Yes□ No□

For Inconclusive results, do you document/record/report in your notes the extent of the agreement with individual characteristics (e.g., There was significant agreement in the individual characteristics)?

Document/record: Yes□ No□

Report: Yes□ No□

If yes to either, and you use CMS, what threshold count would justify the previous statement?_________



For Inconclusive results, do you document/record/report in your notes the extent of the disagreement with individual characteristics (e.g., There was significant disagreement in the individual characteristics)?

Document/record: Yes□ No□

Report: Yes□ No□

If yes to either, and you use CMS, what threshold count would justify the previous statement? _________

Page **3** of **3**

Table B1: Summary of Survey Answers

Laboratory and Training

	Count	Percent
Is your laboratory accredited in firearms examination?		
Yes	170	98.3%
No	2	1.2%
No response	1	0.6%
If yes, by what body?		
A2LA	1	0.6%
ANAB	48	27.7%
ANAB-ISO 17020	7	4.0%
ASCLD	6	3.5%
ASCLD LAB - International	12	6.9%
ASCLD/ANAB	6	3.5%
ASCLD-ISO 17025	1	0.6%
ASCLD-LAB	39	22.5%
ASCLD-LAB ISO 17025	12	6.9%
ASCLD-LAB ISO 17025/ANAB	1	0.6%
ASCLD-LAB/ANAB	15	8.7%
ASCLD-LAB/other	3	1.7%
FQSI/ISO	1	0.6%
ISO 17025	2	1.2%
ISO 17025 / ANAB	4	2.3%
ISO/IEC 17025 / ANAB	3	1.7%
No Response	12	6.9%
Was your training provided by (select one or more):		
Accredited forensic laboratory	138	70.4%
Non-Accredited forensic laboratory	18	9.2%
National Firearms Examiner Academy	26	13.3%
Other (specify):		
California Criminalistic Institute (CCI)	5	2.6%
NFSTC	3	1.5%
Other	6	3.1%
Training by single organization	148	85.5%
Training by more than one organization	24	13.9%
No response	1	0.6%

How long was your training program, in years?		
1	15	8.7%
1.5	8	4.6%
2	128	74.0%
2.5	1	0.6%
3	13	7.5%
greater than 3	6	3.5%
No response	2	1.2%
How long have you been a firearms examiner, in years?		
Less than 5	45	26.0%
5-9	50	28.9%
10-19	59	34.1%
20-29	13	7.5%
30-39	3	1.7%
40+	2	1.2%
No response	1	0.6%
Minimum: 0.1; Median: 9; Maximum: 50; Mean: 10.7; St. Dev.: 8.1		
Have you ever examined items considered to be Best-Known Non-Matches (consecutively manufactured)?		
Yes	163	94.2%
No	8	4.6%
No response	2	1.2%
Have you been qualified to testify on court in the field of firearms identification?		
Yes	157	90.8%
No	15	8.7%
No response	1	0.6%
Do you personally take an annual proficiency test in firearms?		
Yes, internal test	3	1.7%
Yes, external test	117	67.6%
Yes, both internal and external tests	51	29.5%
No	1	0.6%
No response	1	0.6%
Did you take the AFTE certification for firearms?		
Yes	58	33.5%
No	114	65.9%
No response	1	0.6%
Are you AFTE certified for firearms?		
Yes	54	31.2%
No	118	68.2%
No response	1	0.6%

Comparison Process

	Count	Percent
Do you reach an Identification conclusion using a comparison microscope only?		
Yes	166	96.0%
No	6	3.5%
No response	1	0.6%
If no, specify equipment:		
Stereomicroscope	3	
Caliper and balance	3	
Microscope used (specify brand):		
Leica	103	59.5%
Leeds	52	30.1%
Leica / Leeds	5	2.9%
Leeds/Olympus	4	2.3%
Reichert	3	1.7%
No response	6	3.5%
Lighting used:		
Fiber optic	15	8.7%
Fluorescent	95	54.9%
Fiber optic-Halogen	1	0.6%
Fluorescent - Fiber optic	6	3.5%
Fluorescent - LED	8	4.6%
Fluorescent - LED - Fiber optic	5	2.9%
Halogen	3	1.7%
LED	30	17.3%
LED - Fiber optic	6	3.5%
No response	4	2.3%
When in the comparison process do you evaluate for subclass features?		
Before	113	65.3%
During	57	32.9%
Both	1	0.6%
No response	2	1.2%
Which factor is more important in reaching your decision?		
Quantity	4	2.3%
Quality	31	17.9%
Equal	134	77.5%
No response	4	2.3%

Is Pattern Matching the only method you use to reach your conclusion?		
Yes	161	93.1%
No, CMS	10	5.8%
No response	2	1.2%
Do you use CMS in your decision making process?		
Yes	12	6.9%
No	159	91.9%
No response	2	1.2%

<u>Documenting and Reporting Eliminations and Inconclusives</u>

	Count	Percent
Do your laboratory protocols permit an Elimination decision based on differences in individual characteristics?		
Yes	159	91.9%
No	11	6.4%
No response	2	1.2%
Yes-only if verified	1	0.6%
Do your laboratory protocols permit an Inconclusive decision?		
Yes	171	98.8%
No	0	0.0%
No response	2	1.2%
If no, Is a comparison that is not an Identification considered an Elimination?		
Yes	1	0.6%
No	52	30.1%
N/A	120	69.4%
For Inconclusive results, do you document/record/report in your notes the extent of the agreement with individual characteristics (e.g., There was significant agreement in the individual characteristics)?		
Yes, document/record	104	60.1%
No	67	38.7%
No response	2	1.2%
Yes, report	49	28.3%
No	120	69.4%
No response	4	2.3%
If yes to either, and you use CMS, what threshold count would justify the previous statement?		

For Inconclusive results, do you document/record/report in your notes the extent of the disagreement with individual characteristics (e.g., There was significant disagreement in the individual characteristics)?		
Yes, document/record	97	56.4%
No	71	41.3%
No response	4	2.3%
Yes, report	45	26.0%
No	123	71.1%
No response	5	2.9%
If yes to either, and you use CMS, what threshold count would justify the previous statement?		

CMS (applies to striated marks only)

	Count	Percent
Do you use CMS for documentation only?		
Yes	2	1.2%
No	14	8.1%
No Response	156	90.2%
Yes and No. Mostly used for documentation and borderline comparisons	1	0.6%
Is CMS recognized by your laboratory protocols as a method for documentation?		
Yes	6	3.5%
No	10	5.8%
No Response	157	90.8%
Is CMS recognized by your laboratory protocols for determination of an Identification?		
Yes	8	4.6%
No	9	5.2%
No Response	156	90.2%
If yes, is it optional?		
Yes	6	3.5%
No	2	1.2%
No Response	165	95.4%

Do you use a comparison microscope to reach your conclusion by CMS?		
Yes	9	5.2%
No	3	1.7%
No Response	161	93.1
Do you use a photograph to reach your conclusion by CMS?		
Yes	2	1.2%
No	10	5.8%
No Response	161	93.1%
If your Pattern Matching (PM) decision is Inconclusive do you attempt CMS?		
Yes	6	3.5%
No	7	4.0%
No Response	159	91.9%
N/A	1	0.6%
If your PM decision is Inconclusive, could the results of CMS change the final conclusion to Identification?		
Yes	5	2.9%
No	7	4.0%
No Response	158	91.3%
N/A	3	1.7%
If your PM decision is Identification, could the results of CMS change the final conclusion to Inconclusive?		
Yes	1	0.6%
No	11	6.4%
No Response	158	91.3%
N/A	3	1.7%
What is your count threshold for CMS?		
During your evaluation and use of CMS, have you ever considered that your threshold count should change?		
Yes	2	1.2%
No	8	4.6%
No Response	163	94.2%
Could a conclusion of Identification be made if the CMS count threshold was not met?		
Yes	6	3.5%
No	5	2.9%
No Response	162	93.6%

Note that essay responses are not tabulated above for fear of possibly revealing the identification of the writer through examination of their writing style, references to methods used, or specific comments made in their answer.

Appendix C: Instruction Sheet and Reporting Form

INSTRUCTION SHEET

ITEMS CONTAINED IN THIS PACKAGE:

-This instruction sheet.

-A sealed and marked envelope containing the cartridge cases and bullets for comparison, and a recording sheet to record your findings.

-An unsealed and labeled UPS Express shipping box and an unsealed and marked Tyvek envelope to return the bullets and cartridge cases, and recording sheet.

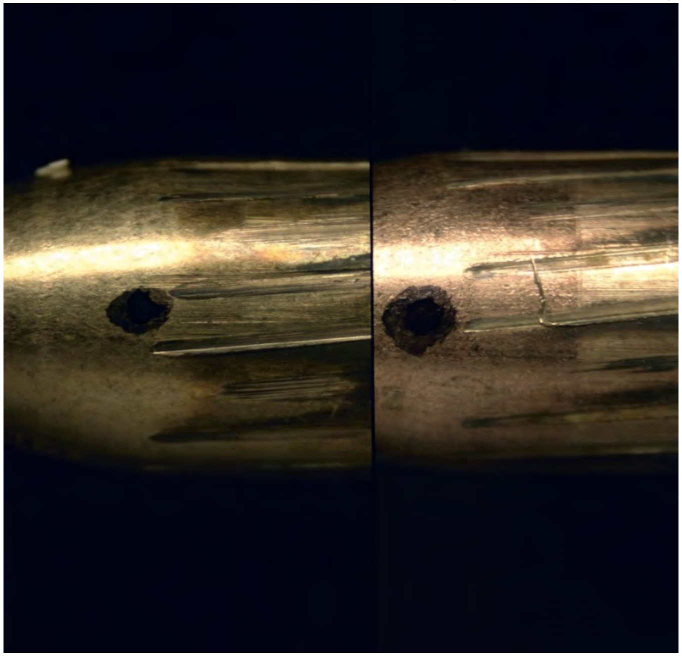
HOW TO CONDUCT THE COMPARISONS:

The cartridge cases and bullets are divided into sets of two known's (k's –produced from the same firearm) and one questioned (q). Each *set* comes in its own numbered plastic bag. This number corresponds to the set number on the recording sheet. Each *set* contains two smaller plastic bags that contain the k's and q respectively.

Work with one set at a time. There are a total of 30 sets consisting of 15 cartridge case sets and 15 bullet sets for examination. Return specimens to their respective bags after comparison is complete in order to minimize mixing up the samples. Record findings for each set on the provided sheet.

1. You are being asked to compare the k's to the q and render a finding of either "Identification, Elimination, Inconclusive or Unsuitable" as defined in the AFTE Glossary "Range of Conclusions Possible when Comparing Toolmarks." Please provide a basis for your conclusion using the check boxes on the answer sheet. Please further qualify any inconclusive decision within the three options indicated, *even if this is not your laboratory's practice.*

2. If an Identification conclusion is rendered for bullets, place a single corresponding mark indicating the land or groove impression used for identification as shown in the photo. Mark the best one if more than one land or groove impression was used to arrive at the decision. **Mark bullet to indicate land or groove decision area used for identification decision with a Sharpie®. Please do not make any permanent marks**



3. If your finding for a particular set is Inconclusive, please select a basis for this finding. The three choices come from the AFTE Glossary "Range of Conclusions Possible when Comparing Toolmarks."

4. Indicate on the recording sheet the number of k's (none, one or two) that have satisfactorily reproduced marks for comparison and are sufficient to substantiate an Identification between the k's from the known same source.

5. Please complete all analyzes and return materials within 30 days after receiving.

NOTICE:

Please do not peer-review or confirm results of your examination.

AFTER COMPARISONS ARE COMPLETE:

Place the recording sheet and the cartridge case and bullet sets in the provided marked Tyvek envelope and seal. Please do not include any identifying marks or numbers that would indicate your identity on either the recording sheet or Tyvek envelope.

Place the sealed envelope (containing materials) into the provided UPS Express shipping box and seal the box.

Return materials to the Ames Laboratory/Iowa State University.

Please do *not* discuss your findings or this study with other examiners who may be participating, so that their contribution and findings may be unbiased and independent.

++

If you have any questions, please contact Dr. Scott Chumbley (515-294-7903, chumbley@ameslab.gov) for assistance.

Thank you for your participation!

Group No. ________________________

Comparison Set No. 1

Finding for Comparison set:

□ IDENTIFICATION

Indicate all areas that you used to arrive at your identification decision:

If a cartridge case:

□ breech face marks

□ firing pin impression

□ chamber marks

□ extractor marks

□ ejector marks

If a bullet:

□ land impression(s)

□ groove impression(s)

□ INCONCLUSIVE*

***Please further qualify your decision, even if this is not your laboratory's practice**

□ a) Some agreement of individual characteristics and all discernible class characteristics, but insufficient for an identification.

□ b) Agreement of all discernible class characteristics without agreement or disagreement of individual characteristics due to an absence, insufficiency, or lack of reproducibility.

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

□ ELIMINATION – (Please provide basis for decision)

□ Using difference in class characteristics only

□ Using difference in individual characteristics

□ UNSUITABLE – (Please provide basis for decision – one or more)

□ a) Lack of sufficient marks on the *Questioned* sample

□ b) Lack of sufficient marks on either of the *Known* samples

□ c) Missing or damaged sample: *Questioned* □ *Known*□

Was CMS used to arrive at your decision? Yes □ No □

Number of knowns with sufficient reproduced detail for comparison: 0 □ 1□ 2□

How long did it take to conduct the comparison? ________________________________

What was the relative difficulty of this comparison? Easy □ Average □ Hard □

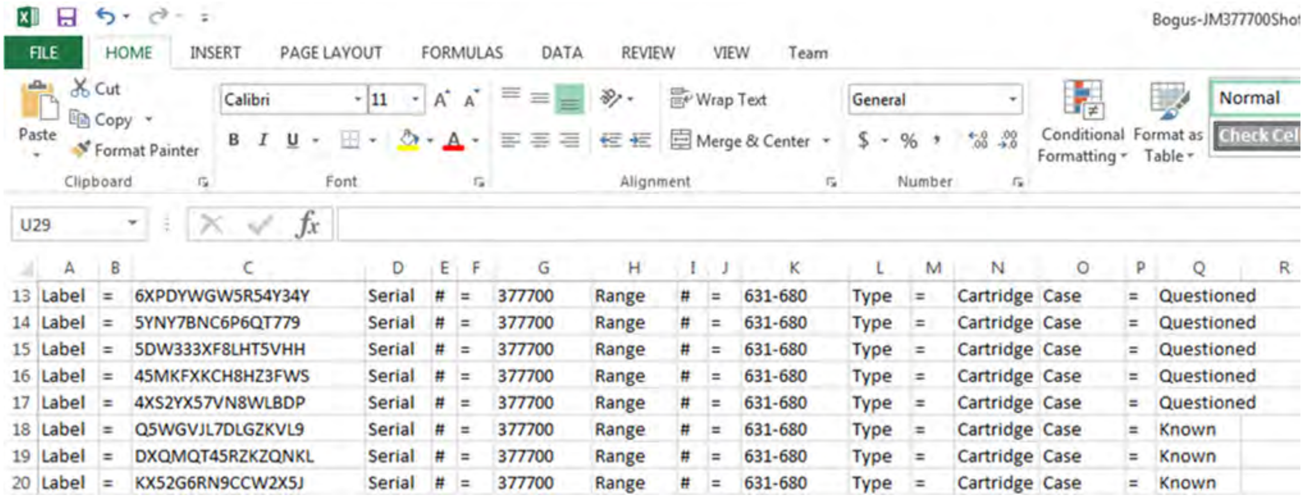
When evaluating/comparing the samples, what level of individual characteristics (quality/ quantity) were available? Limited □ Some □ Extensive □

Appendix D: Barcode Reader, Interface, and Sample Verification

Barcode Reader

The barcode reader is a Cognex DataMan 260 equipped with a 16mm lens with autofocus capability and integrated LEDs for illumination. A barcode-reading application was developed using the Cognex software development kit, DataMan DMCC.NET SDK v5.6.0, providing a .NET interface for using DMCC commands, image and data transfer, device discovery, and loading and saving files. Code editing and debugging was performed using Microsoft Visual Studio Express 2012 for Windows Desktop-Microsoft .NET framework.

The initial GUI provided user-fillable fields to inventory all of the labeled samples. The information saved during the inventorying phase included the individual identifying label on the sample, the serial number of the firearm, the test firing-order range, the sample type, and whether the sample was a questioned or known sample. The information was first saved into a conventional text file from the GUI and then imported into Microsoft Excel for further processing. Separate files were generated for each group of 50 cartridge case or bullets samples, as described in the main body of the report. An example of a small section of an Excel file with inventory information is shown below.

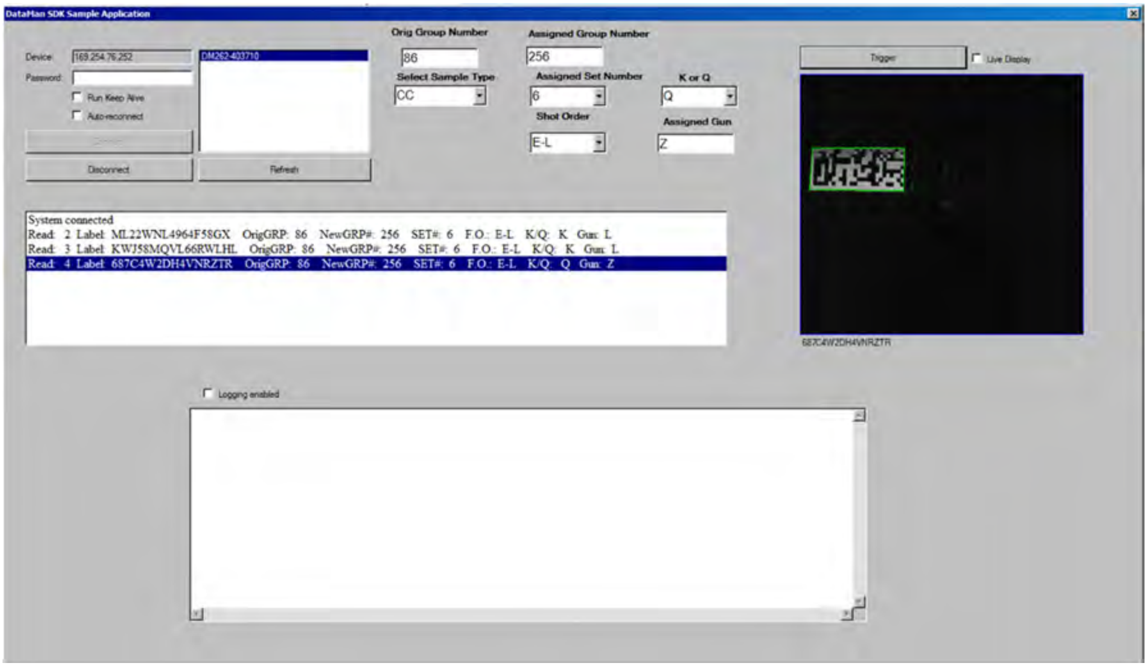


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
13	Label	=	6XPDYWGW5R54Y34Y	Serial	#	=	377700	Range	#	=	631-680	Type	=	Cartridge	Case	=	Questioned	
14	Label	=	5YNY7BNC6P6QT779	Serial	#	=	377700	Range	#	=	631-680	Type	=	Cartridge	Case	=	Questioned	
15	Label	=	5DW333XF8LHT5VHH	Serial	#	=	377700	Range	#	=	631-680	Type	=	Cartridge	Case	=	Questioned	
16	Label	=	45MKFXKCH8HZ3FWS	Serial	#	=	377700	Range	#	=	631-680	Type	=	Cartridge	Case	=	Questioned	
17	Label	=	4XS2YX57VN8WLBDP	Serial	#	=	377700	Range	#	=	631-680	Type	=	Cartridge	Case	=	Questioned	
18	Label	=	Q5WGVJL7DLGZKVL9	Serial	#	=	377700	Range	#	=	631-680	Type	=	Cartridge	Case	=	Known	
19	Label	=	DXQMQT45RZKZQNKL	Serial	#	=	377700	Range	#	=	631-680	Type	=	Cartridge	Case	=	Known	
20	Label	=	KX52G6RN9CCW2X5J	Serial	#	=	377700	Range	#	=	631-680	Type	=	Cartridge	Case	=	Known	

After all of the samples were inventoried, the individual Excel files were concatenated into larger "master" files for each type of sample, cartridge cases and bullets, respectively. Each master file was checked to ensure that there were no duplicate sample-identifying labels either intra- or inter-sample. The master files were used with Microsoft Access Database to verify the contents of the assembled test packages sent to examiners for analysis against the generated distribution tables. They were also used to check and verify comparison sets that were presumptively scored as errors were in fact incorrectly analyzed by the examiners. Other information added to the master files included lot numbers (where known), firing-order sequence number, and the appropriate early, middle, or late (EML) designation.

Interface

The developed GUI used when assembling the accuracy test-packets is shown in the figure below. User-fillable fields were provided that allowed the original identifiers (original group number, K-gun and Q-gun, and firing order) of the samples to be captured, as well as the randomly-assigned group number, comparison set numbers and the sample-identifying labels. (Note: The original group number was used to track the assembled test-packet (with its group of comparison sets) through all phases of the project.) The GUI also provided for an image capture of what the reader was viewing when triggered. The image was temporarily displayed within the GUI but not stored in the text file. Separate barcode-read files were generated for cartridge cases and bullets for each test-packet group.



Device: 169.254.76.252

Password:

□ Run Keep Alive

□ Auto-reconnect

DM262-403710

Orig Group Number: 86

Select Sample Type: CC

Assigned Group Number: 256

Assigned Set Number: 6

Shot Order: E-L

K or Q: Q

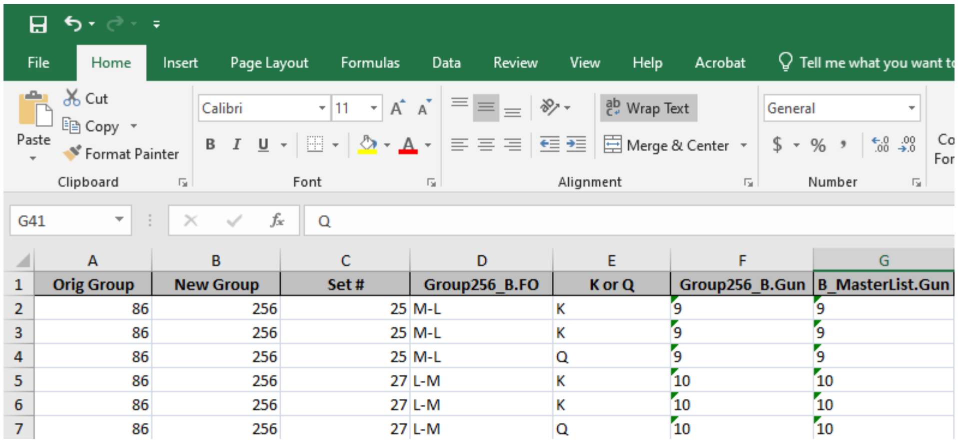
Assigned Gun: Z

```
System connected
Read  2 Label ML22WNL4964F58GX   OrigGRP: 86   NewGRP#: 256   SET#: 6   F.O.: E-L   K/Q: K   Gun: L
Read  3 Label KWJ58MQVL66RWLHL   OrigGRP: 86   NewGRP#: 256   SET#: 6   F.O.: E-L   K/Q: K   Gun: L
Read  4 Label 687C4W2DH4VNRZTR   OrigGRP: 86   NewGRP#: 256   SET#: 6   F.O.: E-L   K/Q: Q   Gun: Z
```

687C4W2DH4VNRZTR

□ Logging enabled

These files were likewise converted into Excel spreadsheet files, and imported into Access to verify the packet contents. Packet contents were verified in Access by executing a query between the packet read-file and the master file list of samples, linked by the identifying-sample labels. The resulting Access query file was exported to Excel and checked for accuracy. The figure below shows a small section of the output of a bullet query file. The last two columns show the values (or guns) that were inputted in the packet read-file (Group256_B.Gun) and the corresponding values from the master list file (B.MasterList.Gun), linked by the identifying-sample label. If the values for all the sets agreed, then the sets (and hence test packets) were assembled correctly, and the contents verified.

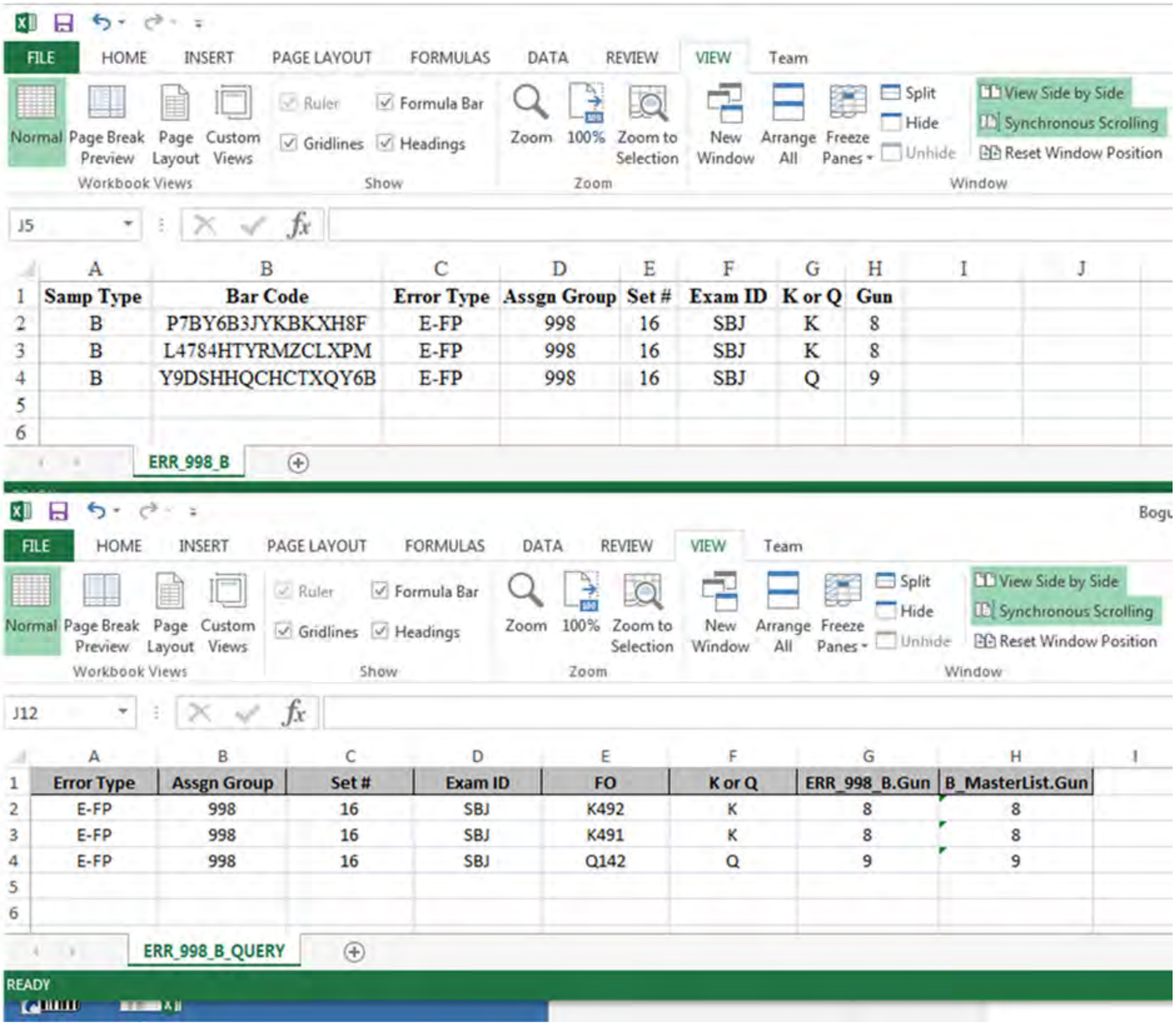


G41 Q

	A	B	C	D	E	F	G
1	Orig Group	New Group	Set #	Group256_B.FO	K or Q	Group256_B.Gun	B_MasterList.Gun
2	86	256	25	M-L	K	9	9
3	86	256	25	M-L	K	9	9
4	86	256	25	M-L	Q	9	9
5	86	256	27	L-M	K	10	10
6	86	256	27	L-M	K	10	10
7	86	256	27	L-M	Q	10	10

Error Verification

When test packets were returned, the examiners' analyses were scored and tabulated. Any comparison set analysis that was scored as an error was barcode-read again to confirm that the set samples (K's and Q) came from the guns listed in the distribution table. This was done by slightly modifying the original GUI to include two additional user-fillable fields: type of error and the examiner three letter identification code. Other information such as sample type, the individual identifying-sample label, assigned group and set numbers, whether the scanned sample was a questioned or known sample, and the presumptive gun the sample came from was included and saved in the text file. As before, the text file was imported into an Excel spreadsheet file for further processing. The top portion of the figure below shows an example of the GUI read-file output (in Excel) from a comparison set scored as an error. In this example, the examiner rendered a decision of "Identification" for this specific comparison set. According to the distribution table this set is a known non-match (two Gun 8 K's and one Gun 9 Q) and should have been marked as an "Elimination." The examiner committed a false-positive error (E-FP).



	A	B	C	D	E	F	G	H	I	J
1	Samp Type	Bar Code	Error Type	Assgn Group	Set #	Exam ID	K or Q	Gun		
2	B	P7BY6B3JYKBKXH8F	E-FP	998	16	SBJ	K	8		
3	B	L4784HTYRMZCLXPM	E-FP	998	16	SBJ	K	8		
4	B	Y9DSHHQCHCTXQY6B	E-FP	998	16	SBJ	Q	9		
5										
6										

	A	B	C	D	E	F	G	H	I
1	Error Type	Assgn Group	Set #	Exam ID	FO	K or Q	ERR_998_B.Gun	B_MasterList.Gun	
2	E-FP	998	16	SBJ	K492	K	8	8	
3	E-FP	998	16	SBJ	K491	K	8	8	
4	E-FP	998	16	SBJ	Q142	Q	9	9	
5									
6									

The comparison set contents were verified, as before, in Access by executing a query between the error comparison set read-file and the master-file list of samples, linked by the identifying-sample labels. The resulting Access query file was exported to Excel and checked for accuracy. The bottom portion of the above figure shows the output of an error bullet query file. The last two columns show the values (or guns) that were inputted in the error set read-file (ERR_998_B.Gun) and the corresponding value from the master list file (B.MasterList.Gun), linked by the identifying-sample labels. Since the values agree, then the set was assembled as a known non-match, and the false-positive error verified.

Repackaging for Reproducibility and Repeatability

When a returned test packet was needed for a repeatability or reproducibility analysis, each specimen in each comparison set was visually examined and gently cleaned of any debris or marks. The K's and Q for a particular set remained together, but assigned new, random set and group numbers, and readied for the next mailing round. The same verification procedure for assembling the initial test packets described above was followed for the now-repackaged test packet with slight alterations in the collected information. The original group number user-fillable field was filled with the "previous" group number (i.e., the sample-set group number prior to repackaging) and the firing-order user-fillable field was used to indicate whether it was the first or second repackaging of the samples (with the original group number for tracking purposes included). This field would be filled with, for example, *RPKxxx* for the first repackaging or *RPKRPTxxx* which would signify the second repackaging, where xxx = the original group number. This information was saved in the corresponding text and Excel files.

* The exact sequence number was not known. The assigned sequence number was estimated within a collected group of 50 specimens of the firing order.

Appendix E: Sample Firing, Firearm, and Pairing Information

Table E1: Firing order EML ranges for bullet and cartridge case samples

Firing-Order EML Ranges for Bullet and Cartridge case Samples

	Bullet		**Case**	
	Beretta	**Ruger**	**Beretta**	**Jimenez**
Early	51-100	61-110	51-100	31-80
	101-150	111-160	101-150	81-130
	151-200	161-210	151-200	131-180
	201-250	211-260	201-250	181-230
	251-300	261-310	251-300	231-280
		311-360		281-330
Middle	301-350	361-410	301-350	331-380
	351-400	411-460	351-400	381-430
	401-450	461-510	401-450	431-480
	451-500	511-560	451-500	481-530
		561-610		531-580
Late	501-550	611-660	501-550	581-630
	551-600	661-710	551-600	631-680
	601-650	711-760	601-650	681-730
	651-700	761-810	651-700	731-780
	701-750	811-860	701-750	781-830
		861-910		831-880

Table E2: Barrel serial numbers and letter/number designations for tracking of bullets.

Beretta barrel manufacturing sequencing is indicated by shading.

Ruger

Barrel Serial Number	Assigned Number Designation
334-57888	7
CB1	6
CB2	11
CB3	2
CB33	5
CB4	10
CB5	4
CB6	9
CB7	3
CB8	8
CBO	1

Note: 334-57888 added from FBI gun collection. Other barrels consecutively manufactured.

Beretta

Barrel Production Sequence	Barrel Serial Number	Assigned Letter Designation
1	Ber733320	I
2	Ber733379	B
3	Ber733380	K
4	Ber733381	T
5	Ber733382	AA
16	Ber733383	R
17	Ber733384	D
18	Ber733385	M
19	Ber733386	V
31	Ber733387	W
32	Ber733388	N
33	Ber733389	E
35	Ber733390	G
46	Ber733391	L
47	Ber733392	C
48	Ber733393	U
49	Ber733394	O
50	Ber733395	X
62	Ber733396	A
63	Ber733397	J
64	Ber733398	S
65	Ber733399	Z
66	Ber733400	F
x1	Ber069589	H
x2	Ber722074	P
x3	Ber133523z	Q
x4	BerC92652z	Y

Note: Barrels x1, x2, x3, and x4 added from FBI gun collection.

Table E3: Slide serial numbers and letter/number designations used for tracking of cartridge case samples.

Beretta slide manufacturing sequencing is indicated by shading.

Jimenez

Slide Serial Number	Assigned Number Designation
377700	1
377701	6
377702	11
377703	5
377704	10
377705	4
377706	9
377707	3
377708	8
377709	2
BR1351894	7

Note: Slide BR1351894 added from FBI gun collection. Other slides consecutively manufactured.

Beretta

Slide Production Sequence	Slide Serial Number	Assigned Letter Designation
1	Ber733320	D
2	Ber733379	M
3	Ber733380	V
4	Ber733381	C
5	Ber733382	L
16	Ber733383	U
17	Ber733384	W
18	Ber733385	B
19	Ber733386	K
31	Ber733387	T
32	Ber733388	A
33	Ber733389	J
35	Ber733390	S
46	Ber733391	AA
47	Ber733392	I
48	Ber733393	R
49	Ber733394	Z
50	Ber733395	H
62	Ber733396	Q
63	Ber733397	Y
64	Ber733398	G
65	Ber733399	P
66	Ber733400	X
x1	BerC92652z	O
x2	Ber069589	E
x3	Ber133523z	N
x4	Ber722074	F

Note: Slides x1, x2, x3, and x4 added from FBI gun collection.

Table E4: Possible pairings for each firearm

Matrix showing the nonmatch pairings for each firearm. Numbers 1-11 designate Ruger barrels and Jimenez slides, letters A-AA designate Beretta barrels and slides. For example, when gun A is used as a Known (left side of table), it is paired with guns F, G, H, I, and J (top of table) as known nonmatched sets designated by the "X". When it's paired with itself, it is a matched set, designated by the "M" in the table.

Questioned

Known	1	2	3	4	5	6	7	8	9	10	11
1	m	x	x	x	x	x					
2		m	x	x	x	x	x				
3			m	x	x	x	x	x			
4				m	x	x	x	x	x		
5					m	x	x	x	x	x	
6						m	x	x	x	x	x
7	x						m	x	x	x	x
8	x	x						m	x	x	x
9	x	x	x						m	x	x
10	x	x	x	x						m	x
11	x	x	x	x	x						m

Questioned

Known	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
A	m					x	x	x	x	x																	
B		m					x	x	x	x	x																
C			m					x	x	x	x	x															
D				m					x	x	x	x	x														
E					m					x	x	x	x	x													
F						m					x	x	x	x	x												
G							m					x	x	x	x	x											
H								m					x	x	x	x	x										
I									m					x	x	x	x	x									
J										m					x	x	x	x	x								
K											m					x	x	x	x	x							
L												m					x	x	x	x	x						
M													m					x	x	x	x	x					
N														m					x	x	x	x	x				
O															m					x	x	x	x	x			
P																m					x	x	x	x	x		
Q																	m					x	x	x	x	x	
R																		m					x	x	x	x	x
S	x																		m					x	x	x	x
T	x	x																		m					x	x	x
U	x	x	x																		m					x	x
V	x	x	x	x																		m					x
W	x	x	x	x	x																		m				
X		x	x	x	x	x																		m			
Y			x	x	x	x	x																		m		
Z				x	x	x	x	x																		m	
AA					x	x	x	x	x																		m

Table E5: Lot numbers for sample bullets and cartridge cases

Beretta Bullets and Cases

Serial Number	Fired Rounds	Lot Number		
BER733379	700	**π 75**	**96**	**2016-04**
BER733380	700	**π 75**	**96**	**2016-04**
BER733381	700	π 78	35	2016-04
BER733382	700	π 02	81	2016-01
BER733383	700	π 78	35	2016-04
BER733384	700	π 78	35	2016-04
BER733385	700	π 78	35	2016-04
BER733386	700	π 78	35	2016-04
BER733387	700	π 78	35	2016-04
BER733388	700	π 78	35	2016-04
BER733389	700	π 78	35	2016-04
BER733390	700	**π 78**	**188**	**2016-04**
BER733391	700	**π 75**	**96**	**2016-04**
BER733392	700	**π 02**	**94**	**2016-01**
BER733393	700	π 116	179	2016-06
BER733394	700	**π 116**	**125**	**2016-06**
BER733395	700	π 154	35	2016-08
BER733396	700	π 154	35	2016-08
BER733397	700	**π 154**	**125**	**2016-08**
BER733398	550	π 154	35	2016-08
	150	**π 154**	**125**	**2016-08**
BER733399	600	π 116	119	2016-06
	100	π 116	35	2016-06
BER733400	300	**π 78**	**188**	**2016-04**
	250	**π 165**	**91**	**2016-08**
	150	**π 165**	**96**	**2016-08**
BER733320	700	π 78	35	2016-04
BER722074	700	π 154	96	2016-08
BER133523Z	400	π 02	90	2016-01
	300	π 116	179	2016-06
C92652Z	700	π 111	21	2014-06
BER069589	200	**π 02**	**94**	**2016-01**
	200	**π 165**	**96**	**2016-08**
	100	π 116	35	2016-06
	150	**π 154**	**125**	**2016-08**
	50	**π 154**	**188**	**2016-08**

Jimenez Cases

Serial Number	Fired Rounds	Lot Number		
377700	850	π 02	90	2016-01
377701	750	π 02	90	2016-01
	100	**π 02**	**94**	**2016-01**
377702	750	π 02	90	2016-01
	100	**π 02**	**94**	**2016-01**
377703	750	π 02	90	2016-01
	100	**π 02**	**94**	**2016-01**
377704	850	π 02	81	2016-01
377705	850	π 02	81	2016-01
377706	750	π 02	90	2016-01
	100	**π 02**	**94**	**2016-01**
377707	850	**π 02**	**94**	**2016-01**
377708	850	π 02	90	2016-01
377709	850	π 02	90	2016-01
BR1351894	500	**π 154**	**188**	**2016-08**
	200	**π 154**	**96**	**2016-08**
	150	π 02	90	2016-01

Note: 1) Color is indicative of the printed ink color observed in ammo box. 2) Lot numbers not available for Ruger bullets.

Appendix F: Tabulation of Data Gathered

Table F1: Listing of Examiner Degree of Difficulty Ratings by Firearm for Round 1

Numbers are reported as fractional values where the ratings for an individual firearm total 1 (easy + average + hard) for either match sets or nonmatched sets.

Degree of Difficulty

Bullets								Cases							
	Known Match				Known Nonmatch				Known Match				Known Nonmatch		
	Easy	Average	Hard		Easy	Average	Hard		Easy	Average	Hard		Easy	Average	Hard
Ruger								**Jimenez**							
1	0.17	0.58	0.25	1	-	0.67	0.33	1	0.17	0.60	0.23	1	0.22	0.65	0.13
2	0.04	0.50	0.46	2	0.07	0.53	0.40	2	0.26	0.49	0.26	2	0.36	0.52	0.12
3	0.11	0.49	0.40	3	0.03	0.54	0.43	3	0.16	0.55	0.29	3	0.27	0.55	0.18
4	0.12	0.46	0.42	4	0.07	0.48	0.45	4	0.18	0.53	0.29	4	0.29	0.43	0.28
5	0.04	0.57	0.40	5	0.01	0.61	0.37	5	0.21	0.46	0.33	5	0.17	0.62	0.21
6	0.08	0.65	0.27	6	0.04	0.73	0.23	6	0.02	0.62	0.36	6	0.18	0.50	0.32
7	0.32	0.45	0.23	7	0.09	0.78	0.13	7	0.24	0.63	0.12	7	0.75	0.25	-
8	0.12	0.46	0.42	8	0.01	0.55	0.43	8	0.06	0.48	0.46	8	0.08	0.69	0.23
9	0.05	0.52	0.43	9	0.01	0.51	0.47	9	0.20	0.49	0.31	9	0.14	0.70	0.16
10	0.17	0.54	0.29	10	0.05	0.69	0.26	10	0.25	0.43	0.32	10	0.09	0.74	0.17
11	0.07	0.69	0.24	11	0.05	0.63	0.33	11	0.20	0.54	0.26	11	0.11	0.68	0.20
All	0.12	0.54	0.35	**All**	0.04	0.62	0.34	**All**	0.18	0.53	0.30	**All**	0.24	0.58	0.18
Beretta								**Beretta**							
A	0.41	0.46	0.13	A	0.06	0.83	0.11	A	0.32	0.56	0.12	A	0.14	0.62	0.23
B	0.46	0.44	0.10	B	0.16	0.70	0.14	B	0.39	0.50	0.11	B	0.06	0.69	0.25
C	0.29	0.63	0.09	C	0.11	0.73	0.16	C	0.26	0.66	0.08	C	0.09	0.69	0.22
D	0.46	0.54	-	D	0.05	0.81	0.14	D	0.09	0.53	0.38	D	0.04	0.58	0.38
E	0.48	0.44	0.07	E	0.10	0.77	0.13	E	0.07	0.31	0.62	E	0.09	0.62	0.29
F	0.27	0.63	0.10	F	0.10	0.73	0.16	F	0.53	0.47	-	F	0.29	0.63	0.08
G	0.50	0.47	0.03	G	0.12	0.72	0.16	G	0.32	0.61	0.08	G	0.23	0.60	0.17
H	0.39	0.54	0.07	H	0.12	0.78	0.09	H	0.13	0.55	0.32	H	0.23	0.56	0.22
I	0.37	0.57	0.06	I	0.14	0.74	0.12	I	0.23	0.58	0.19	I	0.24	0.56	0.20
J	0.45	0.55	-	J	0.10	0.81	0.09	J	0.12	0.77	0.12	J	0.17	0.56	0.27
K	0.35	0.60	0.05	K	0.09	0.80	0.11	K	0.11	0.57	0.32	K	0.10	0.68	0.22
L	0.38	0.63	-	L	0.08	0.82	0.10	L	0.07	0.63	0.30	L	0.07	0.64	0.30
M	0.67	0.26	0.07	M	0.15	0.77	0.08	M	0.17	0.43	0.40	M	0.15	0.58	0.27
N	0.52	0.48	-	N	0.04	0.85	0.11	N	0.53	0.47	-	N	0.26	0.64	0.10
O	0.51	0.49	-	O	0.06	0.79	0.15	O	0.65	0.32	0.03	O	0.57	0.41	0.01
P	0.43	0.54	0.03	P	0.07	0.83	0.10	P	0.22	0.56	0.22	P	0.05	0.53	0.42
Q	0.39	0.54	0.07	Q	0.15	0.75	0.10	Q	0.28	0.48	0.24	Q	0.10	0.61	0.29
R	0.38	0.54	0.08	R	0.08	0.74	0.18	R	0.38	0.47	0.16	R	0.10	0.76	0.14
S	0.38	0.46	0.15	S	0.08	0.79	0.13	S	0.45	0.45	0.10	S	0.31	0.58	0.11
T	0.69	0.27	0.04	T	0.12	0.74	0.14	T	0.32	0.58	0.10	T	0.06	0.74	0.21
U	0.42	0.52	0.06	U	0.06	0.77	0.17	U	0.16	0.69	0.16	U	0.07	0.69	0.24
V	0.51	0.46	0.03	V	0.10	0.81	0.10	V	0.21	0.58	0.21	V	0.12	0.62	0.26
W	0.52	0.45	0.03	W	0.10	0.78	0.13	W	0.14	0.55	0.31	W	0.08	0.71	0.21
X	0.38	0.52	0.10	X	0.05	0.76	0.19	X	0.23	0.50	0.27	X	0.14	0.58	0.28
Y	0.37	0.52	0.11	Y	0.05	0.80	0.15	Y	0.28	0.55	0.17	Y	0.10	0.58	0.32
Z	0.50	0.46	0.04	Z	0.11	0.78	0.11	Z	0.03	0.60	0.37	Z	0.14	0.42	0.43
AA	0.33	0.58	0.09	AA	-	0.79	0.21	AA	0.31	0.53	0.16	AA	0.08	0.69	0.23
All	0.44	0.50	0.06	**All**	0.09	0.78	0.13	**All**	0.27	0.54	0.20	**All**	0.15	0.61	0.23

Table F2: Listing of Examiner Individual Characteristics ratings by specific firearm for Round 1.

Numbers are reported as fractional values where the ratings for an individual firearm total 1 (extensive + some + limited) for either match sets or nonmatched sets.

Individual Characteristics

Bullets								Cases							
	Known Match				Known Nonmatch				Known Match				Known Nonmatch		
	Extensive	Some	Limited		Extensive	Some	Limited		Extensive	Some	Limited		Extensive	Some	Limited
Ruger								**Jimenez**							
1	0.20	0.53	0.27	1	0.11	0.48	0.41	1	0.35	0.58	0.08	1	0.34	0.50	0.16
2	0.12	0.50	0.38	2	0.14	0.36	0.50	2	0.43	0.51	0.06	2	0.35	0.54	0.12
3	0.13	0.57	0.30	3	0.12	0.41	0.47	3	0.29	0.59	0.12	3	0.37	0.53	0.10
4	0.08	0.51	0.41	4	0.09	0.41	0.51	4	0.20	0.69	0.10	4	0.22	0.61	0.18
5	0.06	0.47	0.47	5	0.09	0.49	0.42	5	0.33	0.48	0.19	5	0.28	0.57	0.16
6	0.15	0.44	0.40	6	0.11	0.49	0.41	6	0.18	0.49	0.33	6	0.19	0.53	0.28
7	0.36	0.51	0.13	7	0.31	0.58	0.12	7	0.33	0.59	0.08	7	0.62	0.34	0.04
8	0.14	0.51	0.35	8	0.09	0.39	0.52	8	0.22	0.44	0.34	8	0.32	0.51	0.18
9	0.05	0.34	0.61	9	0.07	0.32	0.61	9	0.24	0.49	0.27	9	0.29	0.53	0.18
10	0.10	0.52	0.38	10	0.16	0.44	0.40	10	0.32	0.59	0.09	10	0.32	0.55	0.13
11	0.07	0.53	0.40	11	0.08	0.43	0.50	11	0.28	0.62	0.10	11	0.30	0.58	0.11
All	0.13	0.50	0.37	**All**	0.12	0.44	0.44	**All**	0.29	0.55	0.16	**All**	0.33	0.53	0.15
Beretta								**Beretta**							
A	0.59	0.38	0.03	A	0.40	0.57	0.03	A	0.62	0.35	0.03	A	0.26	0.59	0.14
B	0.62	0.38	-	B	0.59	0.41	-	B	0.53	0.39	0.08	B	0.31	0.51	0.18
C	0.57	0.40	0.03	C	0.42	0.55	0.03	C	0.50	0.42	0.08	C	0.32	0.56	0.12
D	0.86	0.14	-	D	0.61	0.39	-	D	0.25	0.44	0.31	D	0.14	0.55	0.32
E	0.70	0.30	-	E	0.57	0.40	0.03	E	0.03	0.34	0.62	E	0.05	0.48	0.48
F	0.50	0.47	0.03	F	0.59	0.38	0.04	F	0.58	0.39	0.03	F	0.32	0.65	0.03
G	0.74	0.26	-	G	0.55	0.43	0.01	G	0.37	0.53	0.11	G	0.24	0.63	0.13
H	0.53	0.45	0.03	H	0.62	0.34	0.04	H	0.16	0.48	0.35	H	0.18	0.54	0.28
I	0.51	0.49	-	I	0.51	0.49	-	I	0.30	0.48	0.22	I	0.25	0.58	0.16
J	0.83	0.17	-	J	0.56	0.43	0.01	J	0.23	0.54	0.23	J	0.17	0.67	0.17
K	0.75	0.25	-	K	0.63	0.38	-	K	0.25	0.50	0.25	K	0.19	0.66	0.14
L	0.54	0.46	-	L	0.49	0.44	0.06	L	0.33	0.44	0.22	L	0.11	0.57	0.32
M	0.81	0.19	-	M	0.54	0.46	-	M	0.10	0.47	0.43	M	0.14	0.59	0.27
N	0.67	0.33	-	N	0.45	0.54	0.01	N	0.53	0.41	0.06	N	0.43	0.49	0.08
O	0.67	0.33	-	O	0.50	0.50	-	O	0.71	0.29	-	O	0.47	0.49	0.04
P	0.69	0.31	-	P	0.49	0.48	0.03	P	0.25	0.53	0.22	P	0.11	0.50	0.39
Q	0.79	0.14	0.07	Q	0.70	0.29	0.01	Q	0.31	0.48	0.21	Q	0.17	0.56	0.27
R	0.65	0.35	-	R	0.50	0.47	0.03	R	0.16	0.69	0.16	R	0.24	0.55	0.21
S	0.54	0.46	-	S	0.63	0.31	0.07	S	0.39	0.61	-	S	0.31	0.55	0.14
T	0.81	0.19	-	T	0.56	0.40	0.04	T	0.32	0.68	-	T	0.23	0.61	0.17
U	0.58	0.42	-	U	0.49	0.46	0.05	U	0.34	0.44	0.22	U	0.12	0.67	0.21
V	0.62	0.35	0.03	V	0.51	0.44	0.04	V	0.45	0.39	0.15	V	0.12	0.62	0.26
W	0.76	0.24	-	W	0.51	0.47	0.01	W	0.24	0.55	0.21	W	0.07	0.71	0.22
X	0.69	0.24	0.07	X	0.51	0.45	0.04	X	0.37	0.43	0.20	X	0.25	0.49	0.26
Y	0.67	0.33	-	Y	0.44	0.55	0.01	Y	0.48	0.45	0.07	Y	0.24	0.56	0.21
Z	0.82	0.18	-	Z	0.54	0.41	0.05	Z	0.20	0.50	0.30	Z	0.04	0.49	0.46
AA	0.45	0.55	-	AA	0.47	0.49	0.04	AA	0.28	0.59	0.13	AA	0.24	0.56	0.20
All	0.66	0.33	0.01	**All**	0.53	0.44	0.03	**All**	0.35	0.47	0.17	**All**	0.21	0.57	0.22

Table F3: Listing of Examiner evaluations for both bullet and cartridge case comparisons by firearm for Round 1.

Numbers are reported as fractional values where evaluations for an individual firearm total 1 for either match sets or nonmatched sets. Overall values are reported as percentages

Bullets

Beretta

Matching	ID	Inc-A	Inc-B	Inc-C	Elim	Sets
A	0.92	0.03	0.03	0.03	-	39
B	0.88	0.10	-	0.03	-	40
C	0.92	0.06	0.03	-	-	36
D	0.90	0.03	0.03	0.03	-	29
E	0.85	0.04	-	0.04	0.07	27
F	0.93	0.03	-	-	0.03	30
G	0.95	-	-	0.03	0.03	38
H	0.73	0.05	0.12	0.05	0.05	41
I	0.91	0.06	-	-	0.03	35
J	0.97	-	-	-	0.03	29
K	0.85	0.05	0.05	-	0.05	20
L	0.88	0.08	0.04	-	-	24
M	0.89	0.07	-	-	0.04	27
N	0.91	-	0.03	0.03	0.03	33
O	0.92	0.08	-	-	-	39
P	0.91	0.06	-	0.03	-	35
Q	0.93	-	-	0.04	0.04	28
R	0.96	-	-	-	0.04	26
S	0.85	0.04	0.12	-	-	26
T	1.00	-	0.00	-	-	26
U	0.88	-	0.12	-	-	33
V	0.95	-	-	0.03	0.03	37
W	0.91	-	0.06	-	0.03	33
X	0.72	0.14	0.03	-	0.10	29
Y	0.89	0.07	0.04	-	-	27
Z	0.96	0.04	-	-	-	28
AA	0.88	0.09	-	-	0.03	33
All	89.70%	4.13%	2.59%	1.30%	2.24%	848

Beretta

Nonmatching	ID	Inc-A	Inc-B	Inc-C	Elim	Sets
A	0.01	0.14	0.23	0.36	0.26	70
B	0.01	0.09	0.29	0.32	0.29	69
C	0.01	0.07	0.28	0.35	0.29	75
D	-	0.13	0.18	0.26	0.43	77
E	0.03	0.09	0.22	0.31	0.36	78
F	-	0.07	0.20	0.30	0.43	81
G	-	0.05	0.23	0.26	0.45	77
H	0.01	0.12	0.20	0.25	0.41	75
I	-	0.12	0.22	0.22	0.45	78
J	-	0.14	0.19	0.24	0.44	80
K	-	0.07	0.21	0.25	0.47	81
L	-	0.10	0.18	0.23	0.49	79
M	-	0.05	0.29	0.22	0.43	76
N	-	0.09	0.23	0.31	0.36	74
O	-	0.07	0.29	0.23	0.41	69
P	-	0.10	0.28	0.26	0.36	69
Q	-	0.12	0.19	0.27	0.41	73
R	-	0.13	0.26	0.22	0.39	72
S	-	0.16	0.21	0.23	0.40	75
T	-	0.06	0.19	0.29	0.45	78
U	0.01	0.13	0.26	0.32	0.28	78
V	0.01	0.06	0.25	0.28	0.40	72
W	-	0.13	0.17	0.29	0.42	72
X	0.01	0.07	0.23	0.36	0.33	75
Y	0.01	0.07	0.21	0.36	0.35	75
Z	-	0.07	0.18	0.39	0.36	74
AA	0.01	0.10	0.36	0.26	0.27	70
All	0.54%	9.59%	22.90%	28.20%	38.70%	2022

Ruger

Matching	ID	Inc-A	Inc-B	Inc-C	Elim	Sets
1	0.75	0.12	0.12	0.02	-	52
2	0.48	0.14	0.18	0.14	0.06	50
3	0.60	0.15	0.13	-	0.11	53
4	0.47	0.22	0.18	0.08	0.06	51
5	0.60	0.17	0.19	0.02	0.02	53
6	0.58	0.23	0.12	0.04	0.04	52
7	0.79	0.13	0.06	0.02	-	53
8	0.49	0.08	0.34	0.06	0.04	53
9	0.40	0.24	0.31	0.02	0.02	45
10	0.59	0.22	0.14	0.02	0.02	49
11	0.41	0.13	0.30	0.09	0.07	46
All	56.60%	16.50%	18.50%	4.49%	3.95%	557

Ruger

Nonmatching	ID	Inc-A	Inc-B	Inc-C	Elim	Sets
1	-	0.08	0.44	0.25	0.24	80
2	0.03	0.04	0.46	0.25	0.22	68
3	-	0.13	0.47	0.24	0.16	68
4	-	0.07	0.59	0.23	0.10	69
5	0.01	0.13	0.49	0.15	0.21	67
6	0.03	0.09	0.45	0.17	0.25	75
7	-	0.05	0.19	0.35	0.41	78
8	-	0.13	0.56	0.21	0.10	78
9	0.03	0.11	0.63	0.12	0.12	75
10	-	0.07	0.45	0.17	0.30	82
11	0.03	0.09	0.45	0.20	0.24	80
All	1.10%	9.02%	47.00%	21.20%	21.70%	820

Cartridge cases

Beretta

Matching	ID	Inc-A	Inc-B	Inc-C	Elim	Sets
A	0.97	0.03	-	-	-	34
B	0.86	0.11	0.03	-	-	37
C	0.85	0.10	0.03	0.03	-	39
D	0.70	0.12	0.18	-	-	33
E	0.23	0.17	0.53	-	0.07	30
F	0.95	0.05	-	-	-	38
G	0.79	0.05	0.11	0.03	0.03	38
H	0.68	0.10	0.23	-	-	31
I	0.74	0.07	0.15	-	0.04	27
J	0.77	0.08	0.12	0.04	-	26
K	0.79	0.11	0.07	-	0.04	28
L	0.81	0.04	0.15	-	-	27
M	0.67	0.23	0.10	-	-	30
N	0.94	0.03	-	-	0.03	33
O	0.97	0.03	-	-	-	34
P	0.84	0.13	0.03	-	-	32
Q	0.83	0.13	0.03	-	-	30
R	0.81	0.16	0.03	-	-	32
S	0.90	0.06	-	-	0.03	31
T	0.97	-	0.03	-	-	31
U	0.84	0.13	0.03	-	-	32
V	0.85	0.12	0.03	-	-	33
W	0.76	0.10	0.14	-	-	29
X	0.83	0.03	0.13	-	-	30
Y	0.90	0.07	0.03	-	-	29
Z	0.60	0.17	0.20	-	0.03	30
AA	0.82	0.18	-	-	-	33
All	80.70%	9.57%	8.40%	0.35%	0.93%	857

Beretta

Nonmatching	ID	Inc-A	Inc-B	Inc-C	Elim	Sets
A	0.01	0.09	0.23	0.28	0.39	69
B	-	0.06	0.31	0.24	0.40	72
C	-	0.12	0.28	0.22	0.38	74
D	0.01	0.09	0.34	0.25	0.30	76
E	-	-	0.32	0.26	0.42	66
F	-	0.03	0.15	0.21	0.62	73
G	-	0.03	0.20	0.24	0.54	76
H	-	0.04	0.24	0.13	0.60	80
I	0.01	0.08	0.16	0.13	0.63	80
J	0.01	0.06	0.16	0.27	0.49	79
K	-	0.06	0.21	0.31	0.42	78
L	0.01	0.09	0.23	0.31	0.35	74
M	-	0.03	0.21	0.16	0.60	67
N	-	0.04	0.10	0.17	0.69	72
O	-	-	0.09	0.06	0.86	70
P	0.02	0.12	0.41	0.20	0.26	66
Q	0.01	0.06	0.27	0.23	0.43	70
R	0.01	0.08	0.25	0.22	0.43	72
S	-	0.10	0.21	0.14	0.55	71
T	0.01	0.13	0.19	0.28	0.39	72
U	0.01	0.07	0.28	0.31	0.33	75
V	0.01	0.06	0.29	0.36	0.27	78
W	0.01	0.08	0.28	0.22	0.41	76
X	0.01	0.07	0.26	0.22	0.43	72
Y	0.03	0.14	0.22	0.25	0.36	72
Z	0.03	0.01	0.35	0.22	0.39	69
AA	-	0.06	0.28	0.28	0.39	72
All	0.86%	6.65%	24.00%	22.70%	45.70%	1971

Jimenez

Matching	ID	Inc-A	Inc-B	Inc-C	Elim	Sets
1	0.75	0.13	0.04	0.04	0.04	53
2	0.72	0.13	0.06	0.04	0.04	47
3	0.65	0.14	0.12	0.06	0.04	51
4	0.69	0.12	0.16	0.02	-	49
5	0.63	0.17	0.15	0.04	0.02	48
6	0.47	0.20	0.27	0.02	0.04	55
7	0.76	0.06	0.08	0.04	0.06	49
8	0.49	0.22	0.24	0.04	0.02	51
9	0.60	0.27	0.08	0.02	0.04	52
10	0.72	0.16	0.09	0.02	0.02	57
11	0.65	0.25	0.04	0.04	0.02	51
All	64.70%	16.90%	12.10%	3.37%	3.02%	563

Jimenez

Nonmatching	ID	Inc-A	Inc-B	Inc-C	Elim	Sets
1	-	0.02	0.20	0.22	0.55	83
2	-	0.04	0.17	0.18	0.62	78
3	0.01	0.05	0.22	0.15	0.57	79
4	0.01	0.07	0.19	0.23	0.49	81
5	0.01	0.09	0.27	0.23	0.39	77
6	0.04	0.08	0.24	0.25	0.39	72
7	-	-	-	0.03	0.97	79
8	0.01	0.07	0.24	0.24	0.43	82
9	0.01	0.05	0.19	0.23	0.51	77
10	0.01	0.06	0.14	0.26	0.52	77
11	-	0.04	0.22	0.16	0.58	79
All	1.04%	5.32%	18.90%	19.90%	54.90%	864

Table F4: Tabulation of bullet and cartridge case comparison sets deemed unsuitable by Examiners.

The columns relate to the responses available to examiners on the answer sheet shown in Appendix C. K refers to an unsuitable known or knowns in the set; Q refers to an unsuitable questioned sample; K,Q refers to both; Other refers to instances where examiners did not specify a reason for the unsuitability.

Unsuitable Bullets by Firearm

Beretta

Match	K	Q	K, Q	Other	Sets	Nonmatch	K	Q	K, Q	Other	Sets
A	-	-	-	-	93	A	-	-	-	-	166
B	-	-	-	-	93	B	-	-	-	-	163
C	-	-	-	-	87	C	-	-	-	-	175
D	-	-	-	-	69	D	-	-	-	-	180
E	-	-	-	-	62	E	-	-	-	-	182
F	-	-	-	-	75	F	-	-	-	-	186
G	-	-	-	-	93	G	-	-	-	-	178
H	-	-	-	-	91	H	-	-	-	-	180
I	-	-	-	-	80	I	-	-	-	-	182
J	-	-	-	-	64	J	-	-	-	-	184
K	-	-	-	-	43	K	-	-	-	-	184
L	-	-	-	-	51	L	-	-	-	-	179
M	-	-	-	-	60	M	-	-	-	-	172
N	-	-	-	-	80	N	-	-	-	-	166
O	-	-	-	-	94	O	-	-	-	-	156
P	-	-	-	-	87	P	-	-	-	-	152
Q	-	-	-	-	68	Q	-	-	-	-	161
R	-	-	-	-	58	R	-	-	-	-	169
S	-	-	-	-	58	S	-	-	-	-	179
T	-	-	-	-	61	T	-	-	-	-	183
U	-	-	-	-	77	U	1	-	-	-	184
V	-	-	-	-	88	V	-	-	-	-	172
W	-	-	-	-	80	W	-	-	-	-	173
X	-	-	-	-	72	X	-	-	-	-	171
Y	-	-	-	-	63	Y	-	-	-	-	174
Z	-	-	-	-	65	Z	-	-	-	-	175
AA	-	-	-	-	75	AA	-	-	-	-	167
Sum	**0**	**0**	**0**	**0**	**1987**	Sum	**1**	**0**	**0**	**0**	**4693**

Ruger

Match	K	Q	K, Q	Other	Sets	Nonmatch	K	Q	K, Q	Other	Sets
1	-	-	-	-	128	1	1	2	-	-	176
2	1	-	4	-	130	2	5	1	-	1	164
3	5	-	1	-	137	3	4	-	2	-	172
4	2	-	-	-	125	4	8	1	1	-	183
5	-	-	2	-	124	5	4	1	-	-	173
6	2	1	-	-	127	6	1	-	1	-	184
7	-	-	-	-	124	7	-	2	-	-	184
8	-	1	2	-	121	8	5	2	-	1	188
9	4	2	1	-	111	9	12	4	-	-	184
10	2	-	-	-	114	10	2	1	2	-	190
11	3	1	1	-	124	11	1	4	-	-	177
Sum	**19**	**5**	**11**	**0**	**1365**	Sum	**43**	**18**	**6**	**2**	**1975**

Unsuitable Cartridge Cases by Firearm

Beretta

Match	K	Q	K, Q	Other	Sets	Nonmatch	K	Q	K, Q	Other	Sets
A	-	1	-	-	80	A	1	-	-	-	171
B	-	-	-	-	88	B	-	-	-	-	172
C	-	-	-	-	93	C	-	-	-	-	171
D	-	1	-	-	75	D	2	-	-	-	179
E	2	-	7	-	77	E	19	-	-	-	171
F	-	-	-	-	88	F	-	1	-	-	177
G	-	-	-	-	84	G	-	-	-	-	181
H	3	-	-	-	70	H	1	-	-	-	188
I	-	-	-	-	64	I	1	-	-	-	184
J	1	-	-	-	60	J	1	1	-	-	182
K	-	-	-	-	60	K	-	1	-	-	179
L	-	-	1	-	64	L	1	-	1	-	172
M	1	1	-	-	69	M	2	1	-	-	160
N	-	-	-	-	77	N	-	-	-	-	165
O	-	-	-	-	84	O	-	-	-	-	157
P	-	1	-	-	82	P	-	2	1	-	149
Q	-	-	-	-	72	Q	-	-	-	-	162
R	-	-	1	-	81	R	-	-	-	-	167
S	-	-	1	-	76	S	-	-	1	-	166
T	-	-	-	-	80	T	1	2	-	-	173
U	-	-	-	-	83	U	1	-	-	-	177
V	-	-	-	-	83	V	1	1	-	-	186
W	1	-	-	-	75	W	2	-	2	-	186
X	2	-	-	-	70	X	1	-	1	-	180
Y	-	-	-	-	64	Y	-	3	-	-	180
Z	3	-	1	-	75	Z	5	2	1	-	178
AA	-	-	-	-	76	AA	2	-	1	-	177
Sum	**13**	**4**	**11**	**0**	**2050**	Sum	**41**	**14**	**8**	**0**	**4690**

Jimenez

Match	K	Q	K, Q	Other	Sets	Nonmatch	K	Q	K, Q	Other	Sets
1	-	-	-	-	122	1	1	-	-	-	196
2	2	-	2	-	124	2	2	1	-	-	182
3	1	-	-	-	124	3	-	-	-	-	187
4	1	-	-	-	119	4	-	-	-	-	193
5	-	-	-	-	108	5	1	-	1	-	198
6	2	-	1	-	121	6	1	-	1	-	183
7	1	-	-	-	105	7	-	-	-	-	183
8	-	-	1	-	119	8	2	-	-	-	188
9	1	-	-	-	121	9	-	1	1	-	178
10	1	-	-	-	129	10	-	2	-	-	182
11	2	-	1	-	117	11	2	-	2	-	191
Sum	**11**	**0**	**5**	**0**	**1309**	Sum	**9**	**4**	**5**	**0**	**2061**

Table F5: Proportion of useable and unusable known’s by individual Firearm reported in Round 1.

Beretta				
		Reported	Proportion	Proportion
Gun	Available K's	Usable	Usable	Unusable
A	218	215	98.6%	1.4%
B	216	216	100.0%	0.0%
C	218	217	99.5%	0.5%
D	210	210	100.0%	0.0%
E	208	207	99.5%	0.5%
F	220	218	99.1%	0.9%
G	228	228	100.0%	0.0%
H	228	228	100.0%	0.0%
I	224	221	98.7%	1.3%
J	216	215	99.5%	0.5%
K	200	200	100.0%	0.0%
L	206	205	99.5%	0.5%
M	204	203	99.5%	0.5%
N	214	214	100.0%	0.0%
O	216	216	100.0%	0.0%
P	208	205	98.6%	1.4%
Q	202	201	99.5%	0.5%
R	196	193	98.5%	1.5%
S	202	197	97.5%	2.5%
T	208	206	99.0%	1.0%
U	222	218	98.2%	1.8%
V	218	215	98.6%	1.4%
W	210	209	99.5%	0.5%
X	208	205	98.6%	1.4%
Y	204	201	98.5%	1.5%
Z	204	201	98.5%	1.5%
AA	206	203	98.5%	1.5%
Total	5714	5667	99.2%	0.8%

Unusable

Minimum	0.0%	Mean	0.8%
Median	0.7%	St. Dev.	0.7%
Maximum	2.5%		

Ruger				
		Reported	Proportion	Proportion
Gun	Available K's	Usable	Usable	Unusable
1	262	241	92.0%	8.0%
2	234	205	87.6%	12.4%
3	240	205	85.4%	14.6%
4	240	205	85.4%	14.6%
5	240	220	91.7%	8.3%
6	254	233	91.7%	8.3%
7	260	255	98.1%	1.9%
8	258	226	87.6%	12.4%
9	238	170	71.4%	28.6%
10	256	221	86.3%	13.7%
11	250	204	81.6%	18.4%
Total	2732	2385	87.3%	12.7%

Unusable

Minimum	1.9%	Mean	12.8%
Median	12.6%	St. Dev.	6.5%
Maximum	28.6%		

Table F6. Proportion of usable and unusable cartridge case knowns by individual firearm reported in Round 1

Beretta				
		Reported	Proportion	Proportion
Gun	Available K's	Usable	Usable	Unusable
A	206	201	97.6%	2.4%
B	216	205	94.9%	5.1%
C	222	215	96.8%	3.2%
D	212	189	89.2%	10.8%
E	190	127	66.8%	33.2%
F	218	214	98.2%	1.8%
G	226	219	96.9%	3.1%
H	218	186	85.3%	14.7%
I	212	194	91.5%	8.5%
J	208	195	93.8%	6.3%
K	208	191	91.8%	8.2%
L	202	181	89.6%	10.4%
M	194	173	89.2%	10.8%
N	210	202	96.2%	3.8%
O	208	205	98.6%	1.4%
P	196	165	84.2%	15.8%
Q	198	178	89.9%	10.1%
R	208	195	93.8%	6.3%
S	204	198	97.1%	2.9%
T	206	193	93.7%	6.3%
U	214	198	92.5%	7.5%
V	218	197	90.4%	9.6%
W	210	194	92.4%	7.6%
X	204	195	95.6%	4.4%
Y	202	192	95.0%	5.0%
Z	196	159	81.1%	18.9%
AA	204	195	95.6%	4.4%
Total	5610	5156	91.9%	8.1%

Unusable

Minimum	1.4%	Mean	8.2%
Median	6.9%	St. Dev.	6.5%
Maximum	33.2%		

Jimenez				
		Reported	Proportion	Proportion
Gun	Available K's	Usable	Usable	Unusable
1	270	256	94.8%	5.2%
2	250	245	98.0%	2.0%
3	260	246	94.6%	5.4%
4	256	235	91.8%	8.2%
5	250	237	94.8%	5.2%
6	254	214	84.3%	15.7%
7	254	243	95.7%	4.3%
8	262	222	84.7%	15.3%
9	252	240	95.2%	4.8%
10	264	255	96.6%	3.4%
11	258	245	95.0%	5.0%
Total	2830	2638	93.2%	6.8%

Unusable

Minimum	2.0%	Mean	6.8%
Median	5.2%	St. Dev.	4.4%
Maximum	15.7%		

Appendix G: Error Probability Confidence Intervals

Estimation of Confidence Interval - False Positive

For specificity, this discussion will address the estimation of the probability of false positive determinations. As noted in the text, this is regarded as the probability that a true non-match is classified as an Identification, given that it is classified as one of Identification, Inconclusive-A, Inconclusive-B, Inconclusive-C, or Elimination.

The simplest, and probably most intuitive statistic that might be considered to estimate this probability is the number of false Identification determinations divided by the total number of the five determinations listed above, from among those comparison sets that are non-matching, i.e. the proportion of “hard errors” made over all examiners and non-matching comparison sets. Because this ratio is a simple proportion, it is then tempting to use a standard statistical method for computing a confidence interval based on an assumption that all examiner determinations are independent and have the same error probability; the Clopper-Pearson (1934) method is perhaps the most often-used. As strongly suggested by the analysis presented in section on Accuracy, the assumption that the same hard error probabilities apply to each examiner is suspect. As a result, the ``simple estimate’’ described above is actually an estimate of a composite error probability arbitrarily weighted toward the characteristics of examiners who evaluated more comparison sets, and any confidence interval methodology that regards all errors as equally probable is based on an incorrect mathematical model and so cannot be trusted.

In place of this, the confidence intervals are based on methodology that allows for different error probabilities for each examiner. This approach is based on two different families of probability distributions:

1. **Beta distribution**: The beta distribution is a continuous probability distribution over the interval [0,1], which has a flexible shape that is governed by two parameters (e.g. Evans et al, 2000). The analysis regards each examiner’s error probability as an independent ``draw’’ from this distribution, allowing them to be different. The analysis does not require that these (true) error probabilities are actually observed, but incorporates indirect information based on the number of errors made by each examiner.

2. **Binomial distribution**: The binomial distribution is a discrete probability distribution over non-negative integer values up to a specified value, often called *n* - one of the parameters of the distribution (e.g. Evans et al, 2000). The other parameter is a probability, and the modeled (random) variable is the number of errors that are made if each of the *n* calls is subject to this same probability of error. The binomial distribution (alone) is the basis of the most commonly used confidence intervals associated with proportions. In the method used, the number of errors made by each examiner is modeled as a ``draw’’ from an individual binomial distribution characterized by the number of comparison sets examined by that examiner, and that examiner’s specific and unknown error probability.

The beta distribution has a mathematical form that is often called *conjugate*, with respect to the binomial. In essence, this means that the two distributions can be combined to form a new distribution appropriate for modeling the number of errors to be made by an *unspecified* examiner; that is, a distribution that models the number of errors made by an examiner drawn randomly from among the relevant population of examiners. This property allows for the construction of what statisticians call a *likelihood function* – the mathematical form central to computing confidence intervals for the parameters of the beta, based on the numbers of errors made by each examiner (which must be kept separate in the analysis to account for the fact that the error probabilities are not the same for each). The beta parameters, in turn, characterize the distribution of ``true'' examiner-specific error probabilities. *The maximum likelihood estimates and confidence intervals cited in this report are estimates of the mean of the examiner-specific error probabilities.*

Note that, given this situation, the confidence interval should not be interpreted as bounding the error probability of *any one examiner*. Again, it is not assumed that these probabilities are the same, and the data available for any one examiner is quite limited. A valid, if artificial, alternative explanation of the interval we've offered is the following: If many examiners are randomly selected from the population and each asked to make a single determination for a (different) comparison set known to be a nonmatch, the intervals given bound, with stated confidence, the overall proportion of errors made in this process.

It should also be noted that this method is not completely assumption-free (even though the assumptions are weaker than those on which the Clopper-Pearson intervals are based). Specifically, it is assumed without formal evidence that the beta distribution is appropriate for modeling the population of examiner-specific errors probabilities. The flexibility of the beta distribution family (i.e. the variety of shapes the distribution can take, controlled by its parameters) ensures that the methodology can be appropriate for a wide variety of situations. Because the examiner-specific error probabilities are not directly observable, and there is relatively limited information available on the accuracy of each examiner's determinations, it would be difficult to build a convincing case for a more appropriate distribution. (And even if a different distribution really should be used, the beta distribution is certainly a more appropriate approximation than the single-value distribution assumed by the Clopper-Pearson approach.)

Calculation:

As noted in the text, the VGAM package (in R; web address below) was used to compute likelihood-based estimates. As should be expected from the discussion above, the data required for this calculation are the number of examinations made and (of these) the number of hard errors made *for each examiner individually* – i.e. not combined. (Further information on VGAM is available at the reference cited below.) Given the maximum likelihood estimates of the parameters, calculation of the confidence interval was accomplished via evaluation of the likelihood function over a grid of the parameter values.

1.) <u>Maximum likelihood estimates:</u> Estimation of the beta parameters requires data on the number of examinations and errors made by each examiner. Given a vector *y* of error counts and a vector *n* of examination counts (each of length 173 for data taken from round 1), the specific commands used (for false positive cartridge case evaluations) are:

```
fit <- vglm(cbind(y,n-y) ~ 1, betabinomial, irho=.9)
coef(fit)
      mu          rho
0.00933045 0.05026947
```

These are the maximum likelihood estimates for the two parameters that characterize the distribution; the first (“mu”) is the mean of the beta distribution being used to model the examiner-specific error probabilities, and is the quantity of interest here.

2.) Confidence intervals for mu*:* While fitting maximum likelihood estimates for this model is a bit tricky (and so accomplished with the special program VGAM), the likelihood function itself is relatively easy to calculate using standard R commands. Specifically:

```
ll <- 0
for(i in 1:173){
ll <- ll + dbetabinom(y[i], n[i], mu, rho, log = TRUE)
}
```

can be used to compute the beta-binomial log-likelihood (ll) for specified parameter values mu and rho. (The likelihood function is computed on a log scale here because this is the quantity on which the confidence interval is based.) Because rho is a ``nuisance parameter’’ in this application, the profile-log-likelihood – i.e. for a given value of mu, the maximum of the log-likelihood function over values of rho – is the basis for a confidence interval for mu alone. The log-likelihood function was calculated over a grid of both parameter values, and this was reduced to a one-dimensional grid (over mu) by identifying the mu-specific value of rho for which the log-likelihood is maximized. Using large-sample likelihood theory, the upper and lower 95% confidence limits for mu were identified as the two values for which the profile-log-likelihood is 3.8410/2 = 1.9205 less than the log-likelihood value at the maximum likelihood estimates. (The value of 3.8410 is the 95th quantile of the chi-square distribution with one degree of freedom.)

A more detailed description of the construction of confidence intervals using profile likelihood can be found, for example, in Cox and Snell (1989).

Appendix G References:

Clopper, C. and E.S. Pearson (1934). “The use of confidence or fiducial limits illustrated in the case of the binomial,” *Biometrika* **26** (4): 404-413.

Cox, D.R. and E.J. Snell (1989). *Analysis of Binary Data, 2nd edition*, ISBN 978-0412306204, Chapman & Hall/CRC Press, London.

Evans, M., N. Hastings and B. Peacock (2000). *Statistical Distributions, 3rd edition*, ISBN 0-471-37124-6, John Wiley and Sons, New York.

VGAM package: http://www2.uaem.mx/r-mirror/web/packages/VGAM/VGAM.pdf

Appendix H: Contingency Tables

Observed and Expected Agreement

The observed and expected proportions of agreement depicted in Figure 11 (repeatability) and Figure 14 (reproducibility) are computed from 6x6 contingency tables of counts, formatted as in Tables X and XI (repeatability) and Tables XIV and XV (reproducibility), but with each including only data for repeated measurements made by one examiner (repeatability) or pair of examiners (reproducibility). An example of this calculation is offered here for a hypothetical case in which one examiner evaluates the same 10 non-matching bullet sets in both of Rounds 1 and 2 of the study:

Table H1: Example Data Table for One Examiner and Ten Bullet or Cartridge case Sets in Rounds 1 and 2

			Nonmatching Sets			
Classification on First Evaluation	Classification on Second Evaluation					
	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Unsuitable
ID	0	0	0	0	0	0
Inconclusive-A	0	0	0	0	0	0
Inconclusive-B	0	0	2	0	1	0
Inconclusive-C	0	0	0	0	0	0
Elimination	0	0	0	1	6	0
Unsuitable	0	0	0	0	0	0

The number of paired examinations in which agreement occurred is the sum of entries in the upper-left to lower-right diagonal cells in the table – that is, the number of times the examiner made the same determination in round 2 as in round 1. In the example, this is 0+0+2+0+6+0=8, or 80% of the 10 replicated evaluations of the same material.

The proportion of expected agreement is computed using the marginal proportions in each column and row:

Table H2: Example Data Table for One Examiner and Ten Bullet or Cartridge case Sets in Rounds 1 and 2, with Marginal Row and Column Proportions

				Nonmatching Sets			
Classification on First Evaluation	Classification on Second Evaluation						
	ID	Inc-A	Inc-B	Inc-C	Elim	Unsuitable	**Marginal Proportion**
ID	0	0	0	0	0	0	**0/10 = 0**
Inconclusive-A	0	0	0	0	0	0	**0/10 = 0**
Inconclusive-B	0	0	2	0	1	0	**3/10=0.3**
Inconclusive-C	0	0	0	0	0	0	**0/10 = 0**
Elimination	0	0	0	1	6	0	**7/10=0.7**
Unsuitable	0	0	0	0	0	0	**0/10 = 0**
Marginal Proportion	**0/10=0**	**0/10=0**	**2/10=0.2**	**1/10=0.1**	**7/10=0.7**	**0/10=0**	

The expected proportion is computed as the sum of products of the corresponding marginal proportions – in this case:

$$0x0 + 0x0 + 0.2x0.3 + 0.1x0 + 0.7x0.7 + 0x0 = 0.55 = 55\%$$

So in this case, the observed agreement exceeds the expected agreement (as it does in most of the repeated measurement sets displayed in Figure 11). Observed and expected proportions of agreement are computed similarly for the two alternative scoring schemes considered (pooled Inconclusives; and pooled ID's and Inconclusive-A's, and Eliminations and Inconclusive-C's) after combining/collapsing rows and columns of counts in the table to reflect the pooling of scores. Observed and expected proportions of agreement for reproducibility are also computed in this manner from tables in which rows correspond to the evaluations of one examiner and columns to the evaluations of the other.

IN THE CIRCUIT COURT FOR PRINCE GEORGE'S COUNTY, MARYLAND
(Criminal Division)

STATE OF MARYLAND :
:
v. : Criminal Number: CT121375X
:
KOBINA EBO ABRUQUAH :
:
Defendant. :

AFFIDAVIT OF DAVID L. FAIGMAN

I, David L. Faigman, am over the age of twenty-one and am competent to make this affidavit. My date of birth is September 12, 1957. My address is 20 Saint Jude Road, Mill Valley, CA 94941. I declare as follows:

I. Relevant Education and Experience

1. Affiant is the Chancellor & Dean and John F. Digardi Distinguished Professor of Law at University of California Hastings College of the Law in San Francisco. Affiant also holds a position as Professor in the School of Medicine (Dept. of Psychiatry), University of California, San Francisco.

2. Affiant received his BA in Psychology and History from the State University of New York, College at Oswego, his MA in Psychology from the University of Virginia, and his JD from the University of Virginia, School of Law.

3. Affiant is one of the leading scholars in the United States on the subject of the use of scientific research in legal decision making. He was recently identified as the second most-cited evidence scholar in the nation. He has written and taught extensively in the area of forensic science and issues surrounding proper scientific methodology for scientific evidence offered in applied settings, in particular including courtrooms.

Exhibit E

4. Affiant served on the National Research Council's Committee examining the scientific validity of polygraphs,[1] which principally considered the use of polygraph tests at the nation's nuclear labs for national security purposes. Affiant was a Senior Advisor for the President's Council of Advisors on Science and Technology's (PCAST) Report on forensic science.[2] Additionally, he was a member of the MacArthur Foundation's two Networks (Phase I and Phase II) on Law and Neuroscience.

5. In his role as Senior Advisor to the PCAST Report, Affiant reviewed several drafts of that Report, asking questions, providing comments and offering suggestions—both as to content and form. In addition, Affiant participated in several phone conference calls with members of PCAST and other advisors. In effect, Affiant acted as a consultant and peer reviewer for PCAST as it finalized its Report.

6. Affiant regularly presents on the subject of forensic science to both federal and state judges. He has participated in programs organized by the Federal Judicial Center and judicial education conferences in numerous states, including California, Texas, Illinois, North Dakota, Virginia, and Florida. In addition, over the last nineteen years, Affiant has presented lectures to judges regarding the scientific literatures relevant to their work, including the forensic sciences, at the National Judicial College in Reno, Nevada.

7. In his presentations and lectures to federal and state judges, Affiant focuses on the statistical and methodological bases for proffered scientific evidence. In particular, under virtually all evidence codes, federal or state, judges have gatekeeping responsibilities to ensure that expert opinion is based on good grounds.

[1] THE POLYGRAPH AND LIE DETECTION (National Research Council (NAS) 2003).

[2] President's Council of Advisors on Science and Technology, Report to the President, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* (2016) (hereinafter "PCAST Report").

8. The Maryland Court of Appeals recently adopted the *Daubert* test under Maryland Rule 5-702 in *Rochkind v. Stevenson*.[3] Under *Rochkind*, the trial court is charged with assessing reliability, which has as its focus the principles and methodology underlying proffered expert opinion, "'not on the conclusions that they generate.'"[4] Importantly, however, as the *Rochkind* Court emphasized:

> *Joiner* clarified that "conclusions and methodology are not entirely distinct from one another." A trial court must also consider the relationship between the methodology applied and conclusion reached. Indeed, "[t]rained experts commonly extrapolate from existing data. But nothing in either *Daubert* or the Federal Rules of Evidence requires a [trial] court to admit opinion evidence that is connected to existing data only by the ipse dixit of the expert." A court may conclude that there is simply too great an analytical gap between the data and the opinion proffered.[5]

The *Rochkind* Court also indicated that "general acceptance," which had been the touchstone under the previous standard, "remains an important consideration in the reliability analysis, but it cannot remain the *sole* consideration."[6]

9. The above paragraph is not intended as an assertion regarding applicable law in this case or under the Maryland Evidence Code more generally. However, since applicable law establishes the framework for the scope and content of this Affidavit, the above paragraph sets forth the Affiant's understanding of the applicable tests against which the remainder of this Affidavit is written. In short, that is, this Affidavit contemplates the state of firearms examination under the test set forth in *Rochkind*, which calls upon trial courts to evaluate the evidentiary reliability (i.e., scientific validity) of the methods and principles underlying firearms identification expert testimony, with general acceptance in the pertinent field as a factor in that determination.

[3] 236 A.3d 630 (Md. 2020)

[4] *Rochkind*, 236 A.3d at 651 (quoting Daubert, 113 S.Ct. at 595).

[5] *Rochkind*, 236 A.3d at 651 (internal citations omitted).

[6] *Id*. at 647 (emphasis in original).

10. For well-over thirty years in the profession, Affiant has dedicated his scholarship and teaching to the use of scientific research in legal decision making. He has written over sixty articles, published in the leading law reviews, including the Chicago, Virginia, Pennsylvania, Texas, and Northwestern law reviews, and peer-reviewed science journals, including *Science*, *Sociological Methods and Research*, *Nature Reviews Neuroscience*, *Law & Human Behavior*, and *Current Biology*. He has written three books on subjects related to the law's use of scientific research.[7] In addition to courtroom use of applied science, these books have considered the use of science by administrative agencies, legislatures, and in constitutional cases. In addition, Affiant is the managing author/editor of the leading treatise on scientific evidence, *Modern Scientific Evidence: The Law and Science of Expert Testimony*,[8] which has been cited several times by the US Supreme Court. Affiant's treatise, books and articles have been cited numerous times by federal and state courts.

11. Affiant's CV is attached to this Affidavit in Appendix A.

II. REFERRAL

12. Counsel for Kobina Ebo Abruquah asked Affiant to provide his review and analysis of several issues pertinent to the admission of firearms identification expert testimony in Mr. Abruquah's trial.

[7] LEGAL ALCHEMY: THE USE AND MISUSE OF SCIENCE IN THE LAW (W.H. Freeman 1999); LABORATORY OF JUSTICE: THE SUPREME COURT'S 200-YEAR STRUGGLE TO INTEGRATE SCIENCE AND THE LAW (Times Books 2004); CONSTITUTIONAL FICTIONS: A UNIFIED THEORY OF CONSTITUTIONAL FACTS (Oxford Univ. Press 2008).

[8] DAVID L. FAIGMAN, EDWARD K. CHENG, JENNIFER L. MNOOKIN, ERIN E. MURPHY, JOSEPH SANDERS & CHRISTOPHER SLOBOGIN, MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (Thomson Reuters (WestLaw) 2019-20 edition (5 volumes)) (hereinafter MODERN SCIENTIFIC EVIDENCE).

III. Findings/Observations

A. Summary of Relevance and Conclusions

13. The core concern of this Affidavit primarily involves the question whether firearms identification expertise—a subject matter long assumed to be valid for courtroom use—possesses a methodological and statistical foundation adequate to support the opinion of expert witnesses in court.

14. Courts historically did not closely examine these matters and largely relied on the acceptance of the specialty among those who practiced it, and judicial precedents that had not examined it in detail. Indeed, for the most part, non-DNA forensic identification disciplines went largely unchallenged until the *Daubert* decision in 1993. *Daubert* called upon courts to examine the research underlying expert forensic claims and, as it turned out, many acclaimed areas of supposed expertise had little or no data underlying them. These included celebrated areas of forensic identification, such as fingerprinting, and more suspect applications, such as bitemarks. Firearms identification—a subfield of toolmark identification—also came under scrutiny and has been found to lack conventional empirical support by mainstream academic scientists who have examined the research literature.

15. Although this Affidavit is primarily not case-specific, in that it is directed generally at the field of firearms identification, its relevance is specific to the testimony offered in Mr. Abruquah's matter. The firearms expert claims to be able to identify the source of bullets and cartridge cases fired in this case. This opinion testimony is not supported by scientific studies and the underlying scientific theory and technique are not accepted as valid by the relevant scientific community.

16. The expert’s claim that bullets and cartridge cases can be linked to one another or specifically to a particular gun is not supported by the scientific literature or by scientists who have evaluated that literature.

17. Nonetheless, while the science does not support an individualized identification, the literature would appear to support a firearms expert identifying the class of gun used. As discussed *infra*, the difference is akin to saying that the perpetrator drove a red mustang versus saying that the perpetrator drove a particular red mustang.

18. In summary, therefore, based on the existing literature, firearms examiners should not be permitted to offer an opinion that a particular bullet or cartridge case came from a particular firearm. A firearms examiner should be limited to testifying only that a particular bullet or cartridge case came from a general type or class of firearms.

B. Scientific Assessment of Firearms

19. The reviews and recommendations of the scientific community are to be found principally in two government reports, the 2009 NRC and PCAST Reports, as discussed below. In short, whereas the technicians that use firearms identification methods believe in their value, academic scientists who have been asked to review those methods have uniformly questioned their validity.

20. It must be emphasized at the outset that the NRC and PCAST Reports do not stand alone in their criticism of non-DNA forensic identification methods. The criticism of non-DNA identification evidence is longstanding and was largely initiated by legal scholars with scientific training who began raising substantial concerns with the methodologies employed by a variety of

specialties, including fingerprints, bitemarks, firearms and toolmarks, hair, and handwriting.[9] Among scientists, an important entry in the critical assessment of many forensic science fields was a 2003 editorial in the prestigious journal *Science* by the then editor Donald Kennedy, entitled, *Forensic Science: Oxymoron?* Beyond Kennedy's early indictment of the discipline, increasing numbers of scientists have joined the chorus of concern over the validity and reliability of forensic science.[10]

21. The most substantial challenges to forensics, however, have come from reports whose findings are enormously significant and carry considerable weight. In 2008, the National Research Council published its Report, *Ballistic Imaging*.[11] This was followed in 2009 by the National Research Council Report, *Strengthening Forensic Science in the United States: A Path Forward*.[12] And PCAST delivered its highly critical Report in 2016. These reports, and the scientific community's critical commentary on the state of the art of forensic firearms identification are marked by objectivity and neutrality. Unlike the firearms examiner community, which very much has a horse in this race, these independent scientists bring both skill and detachment to their evaluations.

1. The National Academies of Science

22. The National Research Council (NRC) of the National Academies of Science (NAS) has considered firearms analysis in two reports. The NAS is the most prestigious scientific

[9] *See, e.g.*, D. Michael Risinger, Mark P. Denbeaux & Michael J. Saks, *Exorcism of Ignorance as a Proxy for Rational Knowledge: The Lessons of Handwriting Identification Expertise*, 137 U. PA. L. REV. 731 (1989).

[10] *See, e.g.*, Stephen E. Fienberg, *Editorial: Statistics and Forensic Science*, 1 ANNALS OF APPLIED STATISTICS 285 (2007) ("At the U.S. National Academies of Science/National Research Council, there have been symposia, reports and other publications on various forensic scientific methods, all of which have raised serious questions about how virtually every form of forensic evidence except DNA comparisons has been used.") (citations omitted).

[11] NATIONAL RESEARCH COUNCIL, BALLISTICS IMAGING, Wash. DC: The National Academies Press (2008) (hereinafter "BALLISTICS IMAGING").

[12] NATIONAL RESEARCH COUNCIL, STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES: A PATH FORWARD, Wash. DC: The National Academies Press (2009) (hereinafter "STRENGTHENING FORENSIC SCIENCE" or "2009 NRC Report").

organization in the United States. Its history and renown are unmatched. In 1863, President Abraham Lincoln signed a congressional charter creating the NAS to "investigate, examine, experiment, and report upon any subject of science."[13] Its membership includes the most distinguished scientists, engineers, physicians, and researchers, including more than 300 Nobel laureates. The NAS's two reports involving firearms raised considerable alarm about the validity of the techniques used by specialists in this area of study.

a. 2008 NRC Report: *Ballistics Imaging*

23. The 2008 NRC Report, *Ballistic Imaging*, was commissioned to consider the creation of "a national reference ballistic image database (RBID) that would house images from firings of all newly manufactured or imported firearms."[14] The Report expressly stated that its study was "neither a verdict on the uniqueness of firearms-related toolmarks generally nor an assessment of the validity of firearms identification as a discipline."[15]

24. The fact that the 2008 Committee had a limited charge is not controversial and is largely beside the point. No one has argued otherwise. This does not mean that certain findings of that Report are not relevant to one of the core empirical premises of firearms identification when used for forensic purposes.

25. Because its charge was to consider the value, for purposes of identification or investigative purposes of a database with firearms markings, the issue of the relevance of the markings themselves was inevitably presented. The Report explained: "Underlying the specific tasks with which the committee was charged is the question of whether firearms-related toolmarks are unique: that is, whether a particular set of toolmarks can be shown to come from one weapon

[13] See http://www.nationalacademies.org/about/whoweare/index.html.
[14] BALLISTICS IMAGING, *supra* note 11, at 1. It should be noted that the Affiant was an invited Reviewer for the 2008 NRC Report.
[15] *Id*. at 1-2.

to the exclusion of all others."[16] The Committee found as follows: "The validity of the fundamental assumptions of uniqueness and reproducibility of firearms-related toolmarks has not yet been fully demonstrated."[17]

26. Although the 2008 NRC Report was not directed at the identification techniques used by courtroom experts, the assumptions of uniqueness and reproducibility are central to the operating theory of those experts. For instance, suppose that the NRC Committee had been charged with determining whether it was possible to drive from New York to San Francisco in a single day, but in concluding that it was not, pointed out that the car originally intended for this trek had no engine. The Committee's focus would have been on the difficulty of the trip given the distance, but the finding that the car was non-functional would be relevant to the question of the usefulness of the car for other purposes. The NRC's conclusion that those key assumptions had yet to be demonstrated raised serious doubts about the soundness of a field that proceeded on the basis of those assumptions. Later scientific committees were charged with taking up that issue more directly.

b. 2009 NRC Report: *Strengthening Forensic Science*

27. The 2009 NRC Report considered a wide number of forensic identification disciplines, from bitemarks to DNA profiling, and including firearms and toolmarks. It was highly critical of all of these disciplines, with the notable exception of DNA analysis.[18] Significantly, the

[16] *Id*. at 3.

[17] *Id*.

[18] STRENGTHENING FORENSIC SCIENCE, *supra* note 12, at 130 ("DNA typing is now universally recognized as the standard against which many other forensic individualization techniques are judged. DNA enjoys this preeminent position because of its reliability and the fact that, absent fraud or an error in labeling or handling, the probabilities of a false positive are quantifiable and often miniscule."). It is worth noting that DNA profiling is the only forensic identification discipline studied in the Report that originated in basic academic science. Forensic applications in areas such as firearms, handwriting, bitemarks, fingerprints, and so forth, were products of police laboratories, not academic departments. *See* Paul C. Giannelli, *Forensic Science: Why No Research?,* 38 FORDHAM URB. L.J. 503 (2010). *See also* Michael J. Saks & David L. Faigman, *Failed Forensics: How Forensic Science Lost Its Way and*

2009 Report quoted the 2008 NRC Report finding that "the fundamental assumptions of uniqueness and reproducibility of firearms-related toolmarks has not yet been fully demonstrated."[19]

28. The 2009 NRC Report primarily criticized the firearms field on the basis that it lacked a "precisely defined process."[20] In particular, it found that the guidelines employed by the Association of Firearm and Toolmark Examiners (AFTE) failed to "provide a specific protocol."[21] These guidelines were described in AFTE's 1992 comprehensive guide for toolmark identification.[22] That Report stated as follows:

> *Theory of Identification as it Relates to Toolmarks*
>
> a) The theory of identification as it pertains to the comparison of toolmarks enables opinions of common origin to be made when the unique surface contours of two toolmarks are in "sufficient agreement."
>
> b) 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 it exceeds the best agreement demonstrated between toolmarks known to have been produced by different tools and is consistent with the agreement demonstrated by toolmarks known to have been produced by the same tool. The statement that "sufficient agreement" exists between two toolmarks means that the agreement 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.[23]

How It Might Yet Find It, 4 ANN. REV. L. & SOC. SCI. 149 (2008); SIMON A. COLE, SUSPECT IDENTITIES: A HISTORY OF FINGERPRINTING AND CRIMINAL IDENTIFICATION (2002).

[19] STRENGTHENING FORENSIC SCIENCE, *supra* note 12, at 154 (*quoting* BALLISTICS IMAGING, *supra* note 11, at 3).

[20] *Id*. at 155.

[21] This was a central criticism of the PCAST Report as well, as discussed in the next section.

[22] *Theory of Identification, Range of Striae Comparison Reports, and Modified Glossary Definitions – An AFTE Criteria for Identification Committee Report*, 24 ASS'N. FIREARM & TOOLMARK EXAMINERS J. 336 (1992).

[23] *Id*.

29. As regards the important construct of "sufficient agreement," AFTE adopted the following non-quantitative and non-objective standard:

> c) Currently the interpretation of individualization/identification is subjective in nature, founded on scientific principles and based on the examiner's training and experience.[24]

30. The 2009 Report explained the fatal flaw in the AFTE defined procedure as follows:

> 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.[25]

31. Also of note is the assertion in the AFTE guidelines that the forensic examination should lead to a finding "that the likelihood another tool could have made the mark is so remote as to be considered a practical impossibility." This notion has led to courtroom claims of the ability to identify the suspect weapon to the exclusion (or practical exclusion) of all guns in the world, or of having a "zero error rate." These claims are patently absurd. As to the former claim, individualization in applied science is impossible and even the gold standard of DNA analysis provides probabilistic statements of the likelihood of randomly finding a match in some relevant population (i.e., "random match probability").[26] As to the latter claim of zero error rate, such perfection will never be met so long as humans have anything to do with the procedure. Given the

[24] *Id*.

[25] *Id*. at 155.

[26] *See* Michael J. Saks & Jonathan J. Koehler, *The Individualization Fallacy in Forensic Science*, 61 VAND. L. REV. 199, 208-09 (2008).

admittedly subjective nature of firearms identification, error rates of zero are, to say the least, far-fetched.

32. On the issue of what may be said about an "identification," the December 2014 publication, *Approved Standards for Scientific Testimony and Report Language for the Firearms/Toolmarks Discipline* (ASSTR) provides that an examiner cannot "state or imply that a toolmark was created by a specific tool to absolute certainty or to the exclusion of all other tools in the world"; and "an examiner cannot assign a numerical degree of certainty nor provide a precise error rate to a toolmarks identification."[27] On these points, no one should disagree. Of course, the operative question is what an expert should be allowed to say, given the state of the art of the technique.

33. ASSTR embraces the AFTE notion of "practical impossibility," stating as follows:

> When sufficient agreement exists between two toolmarks, the agreement of the microscopic marks is of a quantity and quality that the likelihood another tool could have produced the same mark is so remote as to be considered a practical impossibility.[28]

34. From a fact-finder's perspective, the difference between saying that a toolmark was created by a specific tool "to the exclusion of all other tools in the world" and saying that it was created by a specific tool "to the practical exclusion of all other tools in the world" is likely to be non-existent. An expert's assertion that a bullet or shell casing was fired by a particular gun is very likely to be heard by the jury as a statement of implied certainty. The research literature, however, does not support an examiner's ability to do the task of linking a toolmark to a particular tool with any known measure of accuracy. A witness, for example, might be highly confident, even nearly certain, that the getaway car was a late-model red Ford Mustang with a broken tail-light; however,

[27] *See Department Of Justice Uniform Language For Testimony And Reports For The Forensic Firearms/Toolmarks Discipline – Pattern Match Examination*, available at https://www.justice.gov/olp/page/file/1083671/download.
[28] *Id*.

this is a quite different assertion from a witness expressing an opinion that *the defendant's* late model red Ford Mustang with a broken tail-light was *the* getaway car. Accordingly, while a firearms expert should be allowed to testify that the bullet or shell casing was fired by a *particular* ***type of*** *gun*, she should not be permitted to testify that the bullet or shell was fired by a *particular gun*.

2. 2016 PCAST Report

35. The PCAST Report began where the two NRC Reports left off.[29] It cited and quoted the 2008 NRC Report's finding "that 'the validity of the fundamental assumptions of uniqueness and reproducibility of firearms-related toolmarks' had not yet been demonstrated."[30]

36. Like the 2009 NRC Report, PCAST raised serious doubts regarding the central operating premise of AFTE's theory of identification as it relates to toolmarks. The Report explained: "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."[31] As PCAST pointed out, this core operating assumption is hopelessly circular.

37. Going beyond the NRC Reports, PCAST provided an extensive review of the literature. It would be redundant to recite PCAST's analysis of the firearms research literature here, since it is clearly set forth in the Report. For ease of reference, however, the pertinent pages are included with this Affidavit in Appendix B.[32] The basic insight that PCAST used to evaluate

[29] PCAST is the leading scientific and technological advisory body to the executive branch, originally chartered by President Eisenhower in the weeks after the launch of Sputnik. See Celebrating the Contributions of the President's Council of Advisors on Science and Technology, White House (Jan. 9, 2017, 2:30 PM), https://obamawhitehouse.archives.gov/blog/2017/01/09/celebrating-contributions-presidents-council-advisors-science-and-technology.

[30] *PCAST Report*, *supra* note 2, at 105 (*quoting* BALLISTICS IMAGING, *supra* note 11, at 3).

[31] *Id.* at 104.

[32] Appendix B includes pp. 104-114 of the PCAST Report. Of course, other sections of the Report are relevant to the question presented here and are cited and quoted throughout this Affidavit.

the research literature was sound. As the Report describes, many of the studies advanced in support of forensic firearms analysis were not well-designed to test the practice employed by courtroom experts. Black box studies[33] would provide such support, but, at the time of the PCAST Report, only one such study arguably came close to providing the needed test.

38. For purposes of emphasis, PCAST's conclusions regarding this literature are worth highlighting. The PCAST Report observed as follows regarding the literature on firearms identification:

> 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.[34]

39. As the Report goes on to detail, there is only one study—the Ames Laboratory study—that generally met minimum methodological criteria for research of this type, and it had yet to be published in a peer reviewed journal. It still has yet to be published in a peer-reviewed journal.[35]

40. The PCAST Report concluded as follows:

> The early studies indicate that examiners can, under some circumstances, associate ammunition with the gun from which it was fired. However, … most of these studies involved designs that are not appropriate for assessing the scientific validity

[33] Black box studies are studies that are measured exclusively by their inputs and outputs. In the context of firearms, this means that the experimenter would know "ground truth" regarding whether a comparison sample is a match or not a match. The subject (i.e., forensic examiner) effectively operates as a "black box" in which a subjective assessment of identification is made but which is not accessible to a third-party. Her success at making the identification would be the measured output.

[34] *PCAST Report, supra* note 2, at 106.

[35] A search of the PubMed database did not reveal any published articles resulting from the Ames Laboratory Study, nor did a search of the Ames Laboratory website. The study itself is included on the AFTE website. See David P. Baldwin, Stanley J. Bajic, Max Morris & Daniel Zamzow, *A Study of False-positive and False-negative Error Rates in Cartridge Case Comparisons, Ames Laboratory*, USDOE, Technical Report #IS-5207 (2014), at https://afte.org/uploads/documents/swggun-false-postive-false-negative-usdoe.pdf.

> 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) ….

41. 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.[36] PCAST concluded that there was a need for additional, appropriately designed black-box studies to provide estimates of reliability.[37]

[36] *Id*. at 111 (emphasis added). PCAST explained that the basic scientific criteria of foundational validity require two key elements:

> (1) a reproducible and consistent procedure for (a) identifying features within evidence samples; (b) comparing the features in two samples; and (c) determining, based on the similarity between the features in two samples, whether the samples should be declared to be a proposed identification ("matching rule").
>
> (2) empirical measurements, from multiple independent studies, of (a) the method's false positive rate—that is, the probability it declares a proposed identification between samples that actually come from different sources and (b) the method's sensitivity—that is, probability that it declares a proposed identification between samples that actually come from the same source.

Id. at 48.

[37] Two firearms associated professional organizations published formal replies to the PCAST Report. These were: (1) Organization of Scientific Area Committees (OSAC) Firearms and Toolmarks Subcommittee, Response to the President's Council of Advisors on Science and Technology (PCAST) Call for Additional References Regarding its Report "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods" (December 14, 2016); and (2) Association of Firearms & Toolmarks Examiners (AFTE), Response to PCAST Report on Forensic Science (October 31, 2016). On January 6, 2017, PCAST responded to these two replies in An Addendum to the PCAST Report on Forensic Science in Criminal Courts.

The firearms associations disagreed with the findings of PCAST and provided detailed rebuttals to PCAST's findings. For instance, OSAC particularly disagreed "with PCAST's conclusion 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.'" OSAC Response, at 1. PCAST did not find OSAC's arguments persuasive, stating as follows:

> 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).

PCAST Addendum, at 7.

In the first-half of January 2021, the Department of Justice published a response critical of the PCAST Report. *See* United States Department of Justice Statement on the PCAST Report: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, January 13, 2021, available at https://www.justice.gov/olp/page/file/1352496/download. The DOJ Response asserts that the PCAST Report made "several fundamentally incorrect claims," including:

> 1) that traditional forensic pattern comparison disciplines, as currently practiced, are part of the scientific field of metrology; 2) that the validation of pattern comparison methods can only be accomplished by strict

C. The Firearms Community

42. The firearms community disagrees with the findings, and the implications for their work, of the scientific committees charged with evaluating the foundational validity of their field. They assert that the extant research literature, contrary to the 2009 NRC and PCAST Reports, is sufficient to permit such testimony and, moreover, that research completed after 2016 further supports their position. They contend that this literature demonstrates that a firearms examiner can identify a cartridge case or bullet to a specific firearm with a very high degree of accuracy.[38]

1. Methodological Limitations of the Firearms Literature

43. Forensic examiners' principal point of disagreement with PCAST is their belief that the Report failed "to recognize the importance of non-black box studies."[39] This, of course, was a point extensively considered and explained by PCAST.[40] Simply put, the PCAST scientists believed that the methodologies employed in the firearms literature did not test the task-at-hand presented in the courtroom. This is a classic consideration in science—indeed, one expressly noted in *Kumho Tire*[41]—and a fundamental matter in assessing the probative value of scientific findings. This criticism is akin to faulting the FDA for requiring clinical trials with humans, beyond animal studies, before approving a new drug. The firearms research literature amounts to mouse studies, and PCAST called for a more sophisticated research paradigm.

> adherence to a non-severable set of experimental design criteria; and 3) that error rates for forensic pattern comparison methods can only be established through "appropriately designed" black box studies.

Id. at 1. These claimed mistakes, however, are largely quibbles about phraseology and emphasis over appropriate research designs. As described herein, whether firearms is, or should be, part of the field of metrology, or other experimental designs in addition to those specified by PCAST might be sometimes appropriate, or study designs in addition to black box studies might be relevant, are all rather beside the point. The experimental work done to date indicates that firearms examiners cannot validly do what they purport to do.

[38] *OSAC Response*. at 3.

[39] *Id*. at 13.

[40] PCAST Report, *supra* note 2, at 106-111.

[41] *Kumho Tire v. Carmichael*, 526 U.S. 137, 141 (1999)("[T]he Federal Rules of Evidence 'assign to the trial judge the task of ensuring that an expert's testimony both rests on a reliable foundation and is relevant to the task at hand.'" (*quoting Daubert v. Merrell Dow Pharm. Inc*., 509 U.S. 579, 597 (1993)).

44. Despite the multitude of cogent and significant challenges to the research basis for firearms examination, the field continues to insist that firearms examiners are highly accurate. They have yet to provide a sound basis for this claim. The first principle of science is that other researchers can check empirical claims. The assertion that examiners are "highly accurate" is vague and ambiguous. Without quantitative proof, this is merely unverified subjective belief. Their assertions regarding the accuracy of courtroom firearms testimony, based on an inapposite research literature, is the worst form of *ipse dixit*. As the Court in *Joiner* pointed out: "A court may conclude there is simply too great an analytical gap between the data and the opinion proffered."[42]

45. Black-box studies, woefully absent in firearms research, would provide general parameters regarding the accuracy rates of forensic firearms identification. The value of such studies, of course, depends on their verisimilitude to the actual tasks of forensic firearms cases. Research in which subjects achieve near perfect results might indicate either that the analytical techniques are highly valuable or that the tests were too easy. Only close evaluation of the methods employed in the research can reveal which is the case. Also, replication in science serves this purpose. In most scientific settings empirical research seeks to test the limits of a hypothesis. In short, research should attempt to falsify hypotheses—i.e., truly test them—not simply seek to corroborate received wisdom. Only when research fully subjects a hypothesis or technique to rigorous test should the mainstream scientific community—or the courts—come to regard it as valid.

[42] *General Electric Co. v. Joiner*, 522 U.S. 136, 137 (1997).

2. Measuring Error Rates

46. Since the PCAST Report was published, the question of how to measure "error" in firearms research has become a central issue. Indeed, for courts, it might be *the* central issue for determining admissibility.[43] In particular, the question presented is how to handle the category of "inconclusive" in research studies and how this determination relates to casework.

47. The task presented in firearms identification is straightforward, whether in casework or in research. An examiner is asked to compare known items—i.e., bullets or cartridge cases known to have come from a particular source—to items that might or might not have come from that source—the unknown items. The issue is whether the unknown item came from the same source as the known items. Hence, there are two basic possibilities in each comparison, either that they came from the same source (i.e., "match" or "identification") or they did not (i.e., "no match," "exclusion," or "elimination"). However, at least in casework, there will be times when there is too little information, and a third category is possible: "inconclusive."

48. Given that there are two fundamentally correct answers for firearms comparisons, there are also two fundamental mistakes that examiners might make. These are "false positives"—i.e., finding a match when there is no match—and "false negatives"—finding no match when there is a match.

49. As noted, in casework at least, there is a third possible answer, "inconclusive." Hence, an examiner who correctly categorizes an item as an inconclusive would be "correct," but who labels a match or a non-match as inconclusive would have made a mistake.

[43] For an excellent overview of the error rate problem in forensic sciences, such as firearms, see Itiel E. Dror & Nicholas Scurich, *(Mis)Use of Scientific Measurements in Forensic Science*, 2 FORENSIC SCIENCE INT'L: SYNERGY 333 (2020). For a somewhat contrary view, see Alex Biedermann & Kyriakos N. Kotsoglou, *Forensic Science and the Principle of Excluded Middle: "Inconclusive" Decisions and the Structure of Error Rate Studies*, 3 FORENSIC SCIENCE INT'L.: SYNERGY 100147 (2021).

50. However, what makes sense in casework makes significantly less sense in the research. In research, samples are created so that the experimenters know "ground truth"—i.e., whether the samples match or do not match. Hence, whereas in fieldwork an "inconclusive" could be the "right answer," it is rarely, if ever, the correct answer in the studies.

51. Yet, when the field touts low error rates in the research, they do not include inconclusives as errors. While technically an inconclusive may not be a "false positive," it is certainly an error when ground truth is known and the correct answer is either match or no-match.[44]

52. In many testing contexts, of course, it is common practice to fashion a test that requires subjects to identify true cases, false cases, and indeterminate cases, where additional investigation might be needed. Consider, for example, a medical school exam in which the student is asked to review blood work to determine whether the patient is suffering from diabetes. The test might manipulate the information to make the correct answer positive, negative, or indeterminate (i.e., inconclusive). Test subjects would be tasked with correctly identifying whether the blood work demonstrates a positive result, a negative result, or an inconclusive result. "Inconclusive" thus could be a correct answer, and in medicine would indicate the need for further testing, which likely would be more expensive and more intrusive. A medical student who mistakes a positive

[44] This issue is the crux of the disagreement between two recent entries on the status of inconclusives in forensic identification studies. Dror and Scurich argue that inconclusives are properly a category in forensic casework, since samples may not have sufficient identifying marks. In the research, although the experimenters know the source of known comparators, it's possible, though presumably unusual, that the marks are insufficient such that the reasonable (or "correct") answer is inconclusive." Dror & Scurich, *supra* note 43, at 334. Bierdermann and Kotoglou reject this view, arguing from metaphysics, that in the research, since ground truth is known, there are only two possible answers, either same source or different source; like pregnancy, they argue, the items match or do not match—there is no middle ground. Although Bierdermann and Kotoglou are correct as a philosophical matter, Dror and Scurich have the better of the argument as an empirical and methodological matter. It may be that a woman is either pregnant or not pregnant, but diagnostic tests designed to assess pregnancy may sometimes give ambiguous—or inconclusive—results. A doctor might be correct then in saying such a result is "inconclusive" and order a more expensive (and possibly more invasive) test. If that diagnostic test resulted in overwhelming numbers of inconclusives, it would, and should, be abandoned. In a nutshell, that is the issue in firearms research, where the number of inconclusives are overwhelming when ground truth is actually known.

for an inconclusive, or a negative for an inconclusive, would thus be subjecting the patient to unneeded testing and potentially deleterious health outcomes.

53. The consequences of labeling an evaluation in firearms identification an inconclusive when it is, in fact, a match or not a match could have similarly significant outcomes. Labeling a match as inconclusive might lead to a guilty person being freed; and labeling a non-match as inconclusive might lead to an innocent person being further detained and possibly wrongly prosecuted. Hence, in research, where the ground truth is known, and the correct answers are match or no-match, a subject choosing "inconclusive" has made a mistake. The field should count those mistakes as errors.

54. When one considers the research literature, counting inconclusives as errors makes the error rate balloon. For example, in the Ames Laboratory study considered by PCAST, David Baldwin and colleagues found that of the 2,178 different-source comparisons conducted, there were 22 identifications, and 705 inconclusive responses.[45] He reported a false positive error rate of 1.01% based on 22/2,178.[46] But counting any response other than an 'exclusion' as an error, the error rate exploded to 33% ((22+705)/2,178)).

55. This explosion of inconclusives in black box studies appears to be a function of the testing methods used. As anticipated by PCAST, black box studies are more challenging for examiners than designs employed in earlier research. Because of this difficulty, and more so the fact that inconclusives are not counted as errors by friendly researchers, examiners default to "inconclusive" when they are not confident of the answer. In effect, examiners are able to skip the

[45] David P. Baldwin, et al., *A Study of False-Positive And False-Negative Error Rates In Cartridge Case Comparisons*, Ames Laboratory, USDOE, Technical Report #IS-5207 (2014)
[46] *Id*. at 16.

hard questions by labeling them inconclusive, answer the easy ones, and get an A+ for a score. This is a method of testing seemingly found through the looking glass.[47]

56. This method is akin to a State Bar allowing examinees to avoid answering any of the two hundred MBE questions that they found to be too hard, too ambiguous, or too "inconclusive," and then calculating percentage correct only on the basis of the questions they did answer—or worse, including those questions not answered as correct! An examinee who gets to choose which questions to answer is likely to do very well indeed on the test. Such a testing protocol, of course, would be absurd. It is similarly absurd as a research design.

57. Since the PCAST Report was published, the field has engaged in additional research, most notably including two black-box studies, as had been recommended by PCAST. Although the field should be applauded for its commitment to carry out research, particularly taking into account the need for black-box studies, these studies do little to buttress the inadequate literature reviewed by PCAST. Indeed, as discussed below, at least one of these studies seems to affirmatively demonstrate that examiners cannot do what they claim to be able to do.

3. Black-box Studies Completed After PCAST

58. This section considers the two notable black-box studies completed after the 2016 PCAST Report was published. The most recent is "Report: Validation Study of the Accuracy, Repeatability, and Reproducibility of Firearm Comparisons" ("Ames II").[48] I consider this Report first because it is the latest and seemingly most substantial to have been done since 2016. The second is research carried out by Mark Keisler and his colleagues.[49]

[47] *See* LEWIS CARROLL, ALICE'S ADVENTURES IN WONDERLAND & THROUGH THE LOOKING GLASS (Bantam Classics 1984) ("Well, I never heard it before, but it sounds uncommon nonsense.").

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

[49] Mark A. Keisler, *Isolated Pairs Study*, 50 AFTE JOURNAL 56 (Winter 2018).

a. Ames II (2020)

59. Dated October 7, 2020, the Ames Laboratory of the U.S. Department of Energy made available its findings of a second study on firearms identification, Ames II. The research was prepared in consultation with and for the Federal Bureau of Investigation. It has not yet been published in a peer-reviewed journal.

60. Ames II appears to aspire to be the definitive black box study, a design expressly called for by PCAST.[50] Because of the length and considerable detail contained in the Ames II Report, a point-by-point assessment of this work is beyond the scope of this Affidavit. (The Report is 127 pages long, inclusive.) Instead, I consider several key methodological concerns, but focus on the single most significant finding of the study, that is, the data reported indicate unambiguously that the examiners in this study were unable to accurately carry out firearms comparisons.[51]

61. It is important to begin with the task given to the subjects to complete. Subjects received packets each containing three samples of either cartridge cases or bullets. These consisted of two known samples (i.e., fired from the same gun) and one unknown sample (i.e., either fired from the same gun or a different gun of the same make and model).

62. Ground truth was known for all of the sample packets, meaning that there were only two possible correct answers for each task, "match" (also referred to as "identification") or "no-match" (also referred to as "elimination"). However, based on their examination of the items, subjects were asked to provide responses of "match," "no-match," or "inconclusive," with this

[50] Ames II, *supra* note 48, at 12 ("The PCAST report further stated that in order to establish foundational validity the principle of reproducibility needed to be satisfied by an additional study. The investigative work planned and discussed below was designed to provide that necessary information.").

[51] I focus on the first round of the study, because of its relevance to the question presented in this Affidavit. This is consistent with the Report's approach itself, which used accuracy data for "only those evaluations made in the first round of the study." Ames II, *supra* note 48, at 33.

latter category divided into three further classifications, based on the reasons for finding insufficiency to make an identification or an elimination.

63. In reporting on accuracy, Ames II explained that:

> In the first round of the study each of 173 examiners evaluated sets of bullets and cartridge cases, each consisting of 2 known items and 1 questioned item. Individual examiners evaluated 15, 30, or (in one instance) 45 sets in the first round. A total of 4320 bullet set examinations and 4320 cartridge case set examinations were performed.

The following table, reprinted from Ames II, provides case summary counts (bold in original):

Table V: First-Round Bullet and Cartridge case Summary Counts.

Bullet Evaluations by Set Type

	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Other
Matching	1076	127	125	36	**41**	24
Nonmatching	**20**	268	848	745	961	49

Cartridge case Evaluations by Set Type

	ID	Inconclusive-A	Inconclusive-B	Inconclusive-C	Elimination	Other
Matching	1056	177	140	22	**25**	25
Nonmatching	**26**	177	637	620	1375	40

64. Ames II then goes on to calculate error rates based on these data. Tellingly, the Report focuses on what it refers to as "hard errors," which it defines as either false positives (i.e., finding a match when they do not match) or false negatives (i.e., finding no-match when they match). The Report does not define what would qualify as "soft errors," though presumably, since ground truth is known, they refer to inconclusives.

65. The Report calculates the false positive rate by including in the numerator the number of correct identifications, with all of the cells[52] included in the denominator. For bullet sets, the False Positive rate was thus calculated as 0.704% and for cartridge cases it was 0.92%.

[52] The Report excludes from its calculations the "other" category, which refers to "records from which an evaluation was not coded or was recorded an inconclusive without a level designation (A, B, or C)." Ames II, *supra* note 48, at 34.

66. The Report calculates the false negative rate by including correct eliminations in the numerator, with all of the cells included in the denominator, which results in a False Negative rate of 2.92% and for cartridge cases a rate of 1.76%.

67. These statistics for hard errors would appear to demonstrate the truly remarkable—almost unbelievable—accuracy of firearms examiners. In fact, however, these statistics are not believable, because they leave out the vast number of mistakes made by the subject examiners.

68. As noted above, a pivotal—perhaps the pivotal—consideration in this area of expert evidence is how to treat inconclusives in the research. To treat them as anything but errors is a profound mistake.

69. In effect, the Ames II study permits subjects to choose only the questions (sets) that they are confident in answering. Since all of the firearms-sponsored research in this area treat inconclusives as not incorrect, subjects know that choosing "inconclusive," even when the only answer must be match or no-match, will not count against calculated error rates.[53]

70. The researchers in this study expressly created packets in which ground truth was known, either "match" or "no-match." Answering "inconclusive," therefore, was an error.

71. The Ames II researchers implicitly concede that inconclusives are errors, since they refer to false positives and false negatives as "hard errors." Presumably, a comparison labeled inconclusive when it is, in fact, a match or non-match, is a "soft error." Whether it is indeed "soft" is a matter of interpretation; that it is an error is beyond question.

72. Therefore, it is appropriate—indeed compelled by fundamentals of research design and common sense—to calculate error rates for Ames II by including both "hard" and "soft" errors.

[53] *See* Dror & Scurich, *supra* note 43, at 336 ("Examiners resort to making more inconclusive decisions during error rate studies than they do in casework.").

73. Calculating actual error rates is a relatively simple mathematical exercise. For the bullet comparison, it requires calculating total examinations (4,320), the number of correct identifications/eliminations (true positives (1076) + true negatives (961)), which equals 2,037. It then requires calculating total mistakes made (which includes false positives (41), false negatives (20), inconclusives when the correct answer was "identification" (288), inconclusives when the correct answer was "elimination" (1,861), and all others (24 + 49 = 73)), which equals 2,283.

74. The error rate for comparing bullets in the Ames II study, therefore, was as much as a whopping 53%.[54] (2,283/4,320=0.528)

75. Error rates for cartridge case comparisons can be calculated similarly.[55]

76. The error rate for comparing cartridge cases in the Ames II study was as much as a similarly eye-popping 44%. (1,889/4320=0.437)

77. The Ames II study thus indicates that in a controlled black-box study, where ground truth is known, examiners are worse than flipping a coin in making bullet comparisons and only slightly better than flipping a coin in making cartridge case comparisons.

b. Keisler et al. (2018)

78. The Keisler et al. study, though an improvement over previous work, suffers fundamental flaws that undermine its value as support for the claim that "a trained, qualified

[54] It should be noted that it is possible that within the design of Ames II, "inconclusive" could have been the right answer sometimes, if the exemplars provided too little information to make an identification, despite the experimenters knowing the source of the item. *See* Dror & Scurich, *supra* note 43, at 334. Unfortunately, their research design did not control for that possibility.

[55] The calculation of error for cartridge cases was as follows: Total examinations=4,320; Total accurate answers (true positives + true negatives) = 2,431; Total mistakes (false positives + false negatives + inconclusives + others) = 1,838. Error rate = Total Errors/Total Examinations = 1, 8889/4320 = 43.73%.

firearm examiner is able to correctly identify and exclude cartridge[s] as having been fired from the same firearm."[56] The flaws are multifold.

79. Keisler et al. followed convention in maintaining that "[i]nconclusive answers are not considered incorrect."[57] This method was employed despite the fact that the samples were specifically selected to allow only two possible answers, match or no-match.[58] In this study, this flaw was particularly evident in cases in which there was a true-exclusion and examiners defaulted to inconclusive. The study states that of 1008 "true exclusions" possible, there were 805 reported exclusions. This is a 20% error rate.[59] Other studies suggest that this flaw impacts reported error rates for both identifications and exclusions.

80. Keisler et al., perhaps most problematically, conducted no pre-testing to determine the difficulty of the task. This should be a basic and integral component of any sort of proficiency testing. After all, getting 100% correct on a test that could be successfully completed by a ninth grader is no test at all.

81. Keisler et al., themselves, recognized additional weaknesses in their methodology, including whether subjects completed the test employing their respective laboratory policies or, even, whether the subjects completed the assignment alone or received feedback or assistance from colleagues.[60]

[56] *Id*. at 58.
[57] *Id*. at 56.
[58] *Id*. at 56-57.
[59] In fairness, it should be noted that the examiners in the study did considerably better in identifying true identifications. Out of 1512 possible, they identified 1508. Of course, as noted in the text, other problems raise doubts about this rate of accuracy, including the fact that there is no indication regarding the difficulty of the task.
[60] *Id*. at 58 ("[I]t is unclear if participants used quality assurance measures, such as verifications, when conducting the research.")

IV. APPLICABLE STANDARDS FOR APPLIED FIREARMS IDENTIFICATION

82. It is not the purpose of this Affidavit to offer views on the correct interpretation of Maryland law. Nonetheless, expert testimony's relevance necessarily depends on the evidentiary rules within which it is offered. Thus, this section is meant only to provide a framework for the opinions offered in this Affidavit. This section considers factors for evaluating firearms under Maryland's *Rochkind* test, and includes a section on evaluating "general acceptance" in the field of firearms identification.

A. Firearms Under the *Rochkind* Test

83. Maryland Rule 5-702 provides as follows:

> Expert testimony may be admitted, in the form of an opinion or otherwise, if the court determines that the testimony will assist the trier of fact to understand the evidence or to determine a fact in issue. In making that determination, the court shall determine (1) whether the witness is qualified as an expert by knowledge, skill, experience, training, or education, (2) the appropriateness of the expert testimony on the particular subject, and (3) whether a sufficient factual basis exists to support the expert testimony.[61]

84. For purposes of the present Affidavit, the latter two criteria are pertinent—the appropriateness of the expert testimony on the particular subject and whether a sufficient factual basis exists to support the expert testimony. And these factors are to be considered though the lens of *Rochkind* and the criteria set forth therein from the *Daubert* decision and its progeny. Because of its centrality to the issues presented, I begin with the expectations of *Daubert*.

1. The *Daubert* Factors

85. In *Rochkind*, the Court outlined the five-principal factors of the *Daubert* test and, additionally, listed another five factors that have been identified by other courts. This section very briefly summarizes these factors in light of the discussion above.

[61] *Rochkind*, 236 A.3d at 642 (quoting Maryland Rule 5-702)..

86. (1) *Whether a theory or technique can be (and has been) tested*. Firearms identification techniques are testable, though as the PCAST Report found, the lack of black box studies means that they had yet to be tested appropriately or adequately. As discussed herein, since 2016, additional black box studies have been completed, but they suffer significant methodological flaws and, in any case, indicate huge error rates.

87. (2) *Whether a theory or technique has been subjected to peer review and publication*. As noted above, PCAST largely dismissed non-black box studies as employing inappropriate research designs for demonstrating foundational validity. The two principal black box studies advanced to support firearms identification—Ames I and Ames II—have yet to be published in peer reviewed journals. The Keisler study considered above was published in a peer reviewed journal, but nonetheless suffers from significant and fatal methodological flaws. Moreover, to the extent that this factor includes post-publication peer review, the 2009 NRC Report and the PCAST Report arguably are damning indictments of the "theory" and "technique" by two separate groups of reviewing scientists.

88. (3) *Whether a particular scientific technique has a known or potential rate of error*. As discussed in detail above, this is a pivotal factor for consideration by anyone considering the foundational validity of the field of firearms identification. If one accepts the friendly researchers' definition of error—that is, counting only the mistakes made for the set of questions that the examiners decided to answer—examiners display vanishingly small error rates. However, in the annals of scientific research or of proficiency testing, it would be difficult to find a more risible manner of measuring error. As is the case in this research, when ground truth is known, and the only correct answers are either "match" or "no-match," answering "inconclusive" is an error. When the actual error rate is calculated, examiners' performance hovers around chance.

89. (4) *The existence and maintenance of standards and controls*. Both the 2009 NRC Report and the PCAST Report criticized the field for its lack of standards for determining when to declare a match. As champions of the field readily admit, the judgment employed by examiners is inherently subjective. This, alone, is not fatal to a technique. But some set of standards need to guide decisions so that others might be able to assess the reliability of the technique. Indeed, this very fact is the reason that black box studies are imperative. If practitioners in the field cannot specify objective standards by which they employ their technique, black box studies would allow them to demonstrate the validity of their subjective methods. To date, black box studies appear to demonstrate just the opposite, that the subjective standards employed cannot reliably identify the specific source of marks on bullets and cartridge cases.

90. (5) *Whether a theory or technique is generally accepted*. See paragraphs 114-120, infra.

91. (6) *Whether experts are proposing to testify about matters growing naturally and directly out of research they have conducted independent of the litigation, or whether they have developed their opinions expressly for purposes of testifying*. There is little or no market for firearms identification outside of law enforcement and, ultimately, to provide "opinions expressly for purposes of testifying." Much of the research carried out in this area appears to be done in order to demonstrate that the claims made all along by firearms examiners are supported. For example, in Ames II, the Report asserts: "The PCAST report further stated that in order to establish foundational validity the principle of reproducibility needed to be satisfied by an additional study. The investigative work planned and discussed below was designed to provide that necessary information."[62] This is not how research is done. Good empirical work sets out to "falsify"

[62] Ames II, supra note 48, at 12.

hypotheses and it is in this rigorous testing that we can be confident that hypotheses that survive are worth relying upon.

92. (7) *Whether the expert has unjustifiably extrapolated from an accepted premise to an unfounded conclusion.* Without question, there is information embedded in the toolmarks left by firearms. In particular, class characteristics provide relevant information that examiners can identify. Contrary to their claims, however, there is little support in the research literature that examiners can extrapolate from the premise that guns leave marks on bullets and cartridge cases to the conclusion that particular marks can be matched to a specific gun.

93. (8) *Whether the expert has adequately accounted for obvious alternative explanations*. Although this factor most often arises in medical causation cases, it is relevant to firearms examiners to the extent that they believe that a finite set of marks can be linked to a particular tool that made them, despite not knowing the base rate for those marks or otherwise specifying the objective basis for relying on them. Fundamentally, the process of identification presumes to identify the source of the marks, effectively ruling out all alternative sources. Firearms examiners have yet to demonstrate their capacity to successfully carry out this task.

94. (9) *Whether the expert is being as careful as he [or she] would be in his [or her] regular professional work outside his [or her] paid litigation consulting*. One can presume that firearms experts are acting in good faith and acting professionally. Their good faith, however, cannot assure accuracy. Of course, as regards this factor, there is no space between a firearms expert's "professional work" and his or her "paid litigation consulting." Firearms examiners are, by their trade, always acting in a litigation capacity.

95. (10) *Whether the field of expertise claimed by the expert is known to reach reliable results for the type of opinion the expert would give*. Although firearms experts have long been

believed to provide reliable opinions, this assumption did not come under close scrutiny until after *Daubert* called upon courts to examine the underlying bases—the methods and principles—for their claimed expertise. Like other long-believed areas of forensic identification expertise—including non-DNA hair, certain arson indicators, comparative bullet-lead analysis, bitemarks, and handwriting—this scrutiny of many forensic techniques revealed that there was no "there there." Both the 2009 NRC Report and the PCAST Report cast a wide and critical net over many of these failed forensic specialties. Firearms identification simply fails to pass muster when evaluated under conventional scientific scrutiny.

2. The Appropriateness of the Expert Testimony on the Particular Subject

96. As *Daubert* makes clear, the pivotal test of proffered expert opinion concerns the empirical basis for that opinion. If that basis is not supported by good grounds, that is, the methods and principles underlying that opinion, it should not be relied upon. As demonstrated by the substantive analysis of the 2009 NRC Report and the PCAST Report, the field of firearms identification has not employed acceptable methods and lacks adequate principles to support their courtroom testimony. Indeed, to the extent that research has truly tested their competence at making identifications, it demonstrates their inability to do what they purport to do.

97. As discussed *infra*, all three scientific reviews of firearms noted that "[t]he validity of the fundamental assumptions of uniqueness are reproducibility of firearms-related toolmarks has not yet been fully demonstrated."

98. However, even if one grants the plausible assumption that tools, such as firearms, leave marks that are relevant to making individual identifications with some level of confidence, this does not mean that examiners have the capacity to state with any level of certainty that unknown samples "matched" a known tool. Unlike DNA profiling, there has been no suggestion

in the literature that the base-rates for marks left by tools can be quantified. Such base-rates in DNA permit the calculation of random match probabilities. There is no corresponding quantification measure in firearms identification.

99. Instead of objective quantification, firearms experts revert to seeking "sufficient agreement" of two toolmarks, which, according to AFTE, "means that the agreement 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."

100. As found by the 2009 NRC Report and the PCAST Report, this theory of identification is subjective and hopelessly circular. The standards of agreement for "quality" and "quantity" of marks are undefined, and terms such as "likelihood" and "practical impossibility" are vague and left unspecified.

101. The underlying scientific theory of firearms identification has not only not been demonstrated to be valid by a preponderance of the evidence, it is, for all intents and purposes, not a scientific theory at all.

3. Whether a Sufficient Factual Basis Exists to Support the Testimony

102. As the PCAST Report contemplated in its review of firearms, applied science does not necessarily require a valid scientific theory to be useful and, arguably, "valid." Consider, for instance, the example of aspirin. Dr. Lawrence Craven first recommended in 1948 that an aspirin a-day would reduce heart attack risk. Only years later was the reason discovered, that it inhibits production of hormones called prostaglandins, which are responsible for forming clots that lead to heart attacks.[63]

[63] DIAMUID JEFFREYS, ASPIRIN: THE REMARKABLE STORY OF A WONDER DRUG (2008).

103. However, if we do not have a scientific theory to explain aspirin's effectiveness against heart disease, how is its validity established? The answer, of course, is the standard fare of clinical studies. In effect, clinical studies are the equivalent of the black box studies prescribed by the PCAST scientists. Adequate black box studies would give us considerable confidence that, though we might not know why the applied technique (whether aspirin or AFTE's subjective method) works, it does work. Moreover, they would give us insights about how well the technique works. Aspirin, after all, does not eliminate the risk of heart attacks; and firearms examiners have error rates far greater than zero. Good quality research provides the answer to the effectiveness of aspirin in risk reduction; similarly, good quality research would provide answers to the error rates of forensic examiners.

104. As the PCAST Report concluded, the current state of the art of the scientific literature does not support a finding that the techniques applied by firearms examiners are valid.

4. Does Failed Science Qualify as "Technical" or "Specialized" Knowledge?

105. The Maryland rule applies not only to scientific knowledge, but also to technical or other specialized knowledge.[64] Forensic examiners have sometimes sought to redefine their knowledge from scientific to specialized, typically invoking their years of experience as a basis for their expertise.

106. Justice Antonin Scalia anticipated the possibility that proponents of expert evidence might seek to use the backdoor of technical or specialized knowledge to gain admission of failed science. Concurring in *Kumho Tire*, joined by Justices O'Connor and Thomas, he warned that the discretion afforded to trial courts under Rule 702 in "choosing the manner of testing reliability" should not be understood as "discretion to abandon the gatekeeping function." He continued:

[64] *See Rochkind*, 236 A.3d at 638.

> Rather, it is discretion to choose among reasonable means of excluding expertise that is fausse and science that is junky. Though, as the Court makes clear today, the *Daubert* factors are not holy writ, in a particular case the failure to apply one or another of them may be unreasonable, and hence an abuse of discretion.[65]

107. The operative question under Rule 702 is, simply, what is the basis for the expert's proffered opinion? Stated another way, how does the expert know what he thinks he knows? The methods of scientific inquiry, of course, provide the basis on which experts testifying to scientific knowledge rely; and these methods were the focus of the *Daubert* factors. In contrast, technical and specialized expertise often relies on experience and judgment, factors that forensic examiners trumpet.

108. The PCAST Report did not ignore the value of experience. It highlights the value of experience in the following passage:

> In some settings, an expert may be scientifically capable of rendering judgments based primarily on his or her 'experience' and 'judgment.' Based on experience, a surgeon might be scientifically qualified to offer a judgment about whether another doctor acted appropriately in the operating theater or a psychiatrist might be scientifically qualified to offer a judgment about whether a defendant is mentally competent to assist in his or her defense.[66]

109. The Report, however, goes on to distinguish forensic contexts:

> By contrast, 'experience' or 'judgment' cannot be used to establish the scientific validity and reliability of a metrological method, such as a forensic feature-comparison method. The frequency with which a particular pattern or set of features will be observed in different samples, which is an essential element in drawing conclusions, is not a matter of 'judgment.' It is an empirical matter for which only empirical evidence is relevant. Moreover, a forensic examiner's 'experience' from extensive casework is not informative—because the 'right answers' are not typically known in casework and thus examiners cannot accurately know how often they erroneously declare matches and cannot readily hone their accuracy by learning from the mistakes in the course of casework.[67]

110. The logic of PCAST is fairly plain. There are actually two principles operating implicitly in the two paragraphs quoted. One is the benefit of experience to assess common

[65] *Kumho Tire v. Carmichael*, 526 U.S. 137, 158-59 (1999) (Scalia, J., concurring).

[66] *PCAST Report*, *supra* note 2,at 55.

[67] *Id.*

practices and the other is the benefit of experience when the beneficiary of that experience receives quality feedback.

111. In the first quotation, the surgeon can speak to the proper manner of acting in an operating room because he has witnessed such practices multiple times over his career. For example, if the relevant issue in court was whether it was common practice in forensic labs to wash machinery every evening, the forensic examiner would have sufficient expertise based on experience (and possibly judgment too) to testify. Of course, that is not the relevant issue in most firearms cases. Rather, the frequency of particular patterns and feedback about performance in actual casework are the relevant concerns, both of which are empirical matters.

112. The second quoted paragraph points to a well-researched subject in psychology, the question of how expertise develops. Outside of reliance on empirical tests, as PCAST finds necessary for forensic pattern recognition, researchers have identified "good feedback" as essential.[68] A good feedback loop provides information to the practitioner regarding the success or failure of some procedure, technique, or belief. But not all feedback loops are equally valuable. Doctors bled patients for centuries, because the procedure conformed to the medical theory of the time and some of their patients improved following this therapy. In those cases, of course, the patients would have recovered anyway, and likely sooner, without having been bled.

113. This concept of good feedback explains why harbor pilots should be allowed to testify and firearms specialists should not—or, at least, the latter should be limited in what they are allowed to say. If the issue concerns safe conditions in a particular harbor, a harbor pilot with extensive experience in that waterway is likely to have received considerable feedback on the

[68] *See* Daniel Kahneman & Gary Klein, *Conditions for Intuitive Expertise: A Failure to Disagree*, 64 AMER. PSYCHOLOGIST 515, 522 (2009); *see also* J. Shanteau, *Competence in Experts: The Role of Task Characteristics, 53* ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES, 252 (1992).

factual question in issue. A forensics examiner, in contrast, will not receive any comparable feedback from even extensive casework, since he receives little, if any, information on his accuracy in each case. Indeed, even the judgment about whether he was correct in distinguishing class and sub-class characteristics from individual characteristics is not available to improve his future work.

B. Is Firearms Expertise Generally Accepted?

114. A frequently used factor for determining whether a proponent meets the burden of showing scientific validity under *Daubert* is the extent to which "the theory or technique has general acceptance within a relevant scientific community." Given the paucity of adequately designed research studies available to demonstrate the validity of firearms identification, this is an often-cited basis for admitting such expert testimony.

115. The forensic field's claim that firearms identification evidence is generally accepted suffers from a fundamental error in measuring this criterion. To be sure, they are correct that firearms evidence is accepted among those who practice this art for a living. One would imagine that tea-leaf reading is generally accepted among tea-leaf readers.

116. As Upton Sinclair put it, "It is difficult to get a man to understand something, when his salary depends on his not understanding it."[69]

117. What the 2009 NRC and PCAST Reports make clear is that firearms identification opinion evidence is not generally accepted among scientists. The general acceptance criterion does not mean very much if it is limited to surveying only those who depend on the technique to make a living. Although it is admittedly a difficult judgment to make regarding how widely to define a field for measuring acceptance, the relevant field must extend beyond true believers. In any case, it is unambiguously clear that mainstream academic scientists uniformly question the foundational

[69] UPTON SINCLAIR I, CANDIDATE FOR GOVERNOR: AND HOW I GOT LICKED (1994).

validity of firearms identification; those accepting it are limited to those self-interested in accepting it.

118. An alternative measure of general acceptance advanced by some is acceptance among courts. In one respect, this is a reasonable standard to adopt. If scientific findings have come to be accepted by courts generally, relitigating claims of lack-of-validity in case after case is an obvious waste of time. But this criterion assumes two basic premises are true. First, that the science was accepted by the appropriate group of scientists in the first place, thus providing an adequate foundation on which courts came to accept it. And, second, that a new understanding has not come to replace the one that courts had previously relied on for reaching their conclusion of acceptance. These are integrally related in the case of firearms and I consider them in turn.

119. Although, as noted above, a small group of scholars had challenged the adequacy of the research literature prior to 2009, the first in-depth questioning of that research basis came in the NRC's Report that year. Hence, courts had long relied on practitioners in the field, most of whom are not trained in statistics or research methods, for the conclusion that firearms identification was generally accepted. As noted above, there is an obvious fallacy in relying on practitioners whose livelihoods depend on the acceptance of their practice for determining the acceptance of their practice.

120. In any case, even if there was a general view prior to 2009 that was adequate for courts to rely on, that is no longer true. One of the hallmarks of science is that knowledge progresses as researchers study and evaluate the research basis for existing beliefs. In the case of firearms, the two Reports described above were written by accomplished academic scientists. They studied in depth the research basis for firearms identification and concluded that it was not adequate to support claims of practitioners that they could match a particular bullet or cartridge

case to a specific gun. By any measure of what is meant by "general acceptance," the fact that a Committee of the National Academies of Science and the President's Council of Advisors on Science and Technology—both well representing the mainstream scientific community in the United States—concur that the basis for firearms identification does not support the testimony offered, should give courts pause.

V. CONCLUSIONS

121. In light of the substantial criticism of the firearms research literature as it pertains to the courtroom task of identification, the question remains regarding what a firearms expert should be allowed to say in his or her testimony. At the extremes, there appears to be little question. On the one hand, the research does not support an examiner's ability to determine a "match" to the practical exclusion of all guns in the world with a zero error rate. On the other hand, the information contained in the marks left by a tool are data that have relevance to a fact "of consequence to the determination of the action."[70]

122. The issue of identification, of course, is endemic to the courtroom. Close consideration of the nature of any identification reveals two basic levels of complexity. The first involves possible errors in perception. The witness who believes that the perpetrator drove a mint green 1964 Buick Skylark convertible could have been mistaken if the perpetrator was actually driving a mint green 1963 Pontiac Tempest.[71] The second complexity involves what scientists refer to as the base-rate. This concerns the frequency or size of the phenomenon of interest. If the suspect car was the 1964 Buick, how many such cars were manufactured or might have been in the vicinity of the crime? The base-rate, for example, of a red Ford Mustang is much greater than

[70] MARYLAND RULE 5-401.

[71] *See generally My Cousin Vinny* (20th Century Fox, 1992).

a red Ferrari, and this base-rate obviously affects the probative value of the information about the color and make of the suspect's car.

123. The assessment of toolmarks left by a firearm is fundamentally similar to problems such as identifying a suspect's car, except that the evidence is offered by an expert. When the identification is made by an expert, the two levels of complexity—perception and base-rate—are multiplied.

124. In regard to possible perceptual errors, while it is true that fact-finders will have the toolmarks before them, they nonetheless are likely to be greatly influenced by the perception of the expert. If the toolmarks ultimately relied on by the firearms examiner are misperceived, fact-finders are unlikely to be able to correct that error. Indeed, given the inherently subjective nature of the AFTE's guidelines on selecting the marks of relevance, this perception problem is largely not amenable to cure even by the ordinary processes of cross-examination. An expert's misperception is not a product of deceit or deception—the ordinary targets of cross-examination; it is instead a sincerely held mistake of fact. Misperceptions offered as a product of scientific proof are precisely the sort of error that the rules of evidence were meant to minimize.

125. The second complexity involving base-rates is hugely problematic. As the 2008 NRC Report first pointed out, we cannot assume uniqueness of the underlying phenomenon, either as a theoretical or practical matter. And unlike 1964 Buick Skylarks, we have no grounds on which to estimate the base-rate of the marks, either individually or collectively. Undocumented and unverified anecdotal "experience" is simply not adequate to estimate base-rates. Hence, statements of comparison or identification in firearms analysis are open to significant doubt and potentially present the danger of substantial unfair prejudice.

126. Although the comparison of toolmarks to a particular tool such as a gun are certainly relevant, their probative value is largely unknown and potentially subject to baseless speculation. The most that can be said about such marks is that they could have been left by the category of gun in question.

127. Therefore, on the basis of the existing literature, firearms examiners should not be permitted to offer an opinion that a particular bullet or cartridge case came from a particular firearm. A firearms examiner should be limited to testifying only that a particular bullet or cartridge case came from a general type or class of firearms.

128. Researchers have largely not studied forensic firearms identification rigorously or in ways relevant to how it is practiced. This is the lesson of the reports published by the NRC and PCAST. Still, these reports give some hope that this situation might yet change, as academic researchers begin to examine the challenges of pattern identification in this and other contexts. But courts have largely not demanded better research from the forensic community.[72] As PCAST made plain, research could be done that would test the foundational validity of current practices; and such research might yet discover improved practices.

129. Until courts demand that such work be done, however, better research will not be forthcoming.

I solemnly affirm under the penalties of perjury that the contents of this document are true to the best of my knowledge, information, and belief. Signed: May 18, 2021:

David L. Faigman

[72] *See* Stephanie L. Damon-Moore, *Trial Judges and the Forensic Science Problem*, 92 N.Y.U. L. REV. 1532 (2017)

APPENDIX A

May 2021

DAVID L. FAIGMAN

University of California
Hastings College of the Law
200 McAllister Street
San Francisco, CA 94102
(415) 565-4739
faigmand@uchastings.edu

EMPLOYMENT

Regular

2017-present	Chancellor & Dean, University of California, Hastings College of the Law
2016-2017	Acting Chancellor & Dean, University of California, Hastings College of the Law
2007-present	John F. Digardi Distinguished Professor of Law, University of California, Hastings College of the Law
2009-present	Professor, Department of Psychiatry, School of Medicine, University of California, San Francisco
2014-2020	Co-Founder, JuriLytics, LLC

****

2009-2015	Founding Director, UCSF/UC Hastings Consortium on Law, Science & Health Policy
2006-2007	Distinguished Professor of Law, University of California, Hastings College of the Law
1993-2006	Professor of Law, University of California, Hastings College of the Law
1997-1998	Harry H. and Lillian H. Hastings Research Chair, University of California, Hastings College of the Law (One-year appointment)
1990-1993	Associate Professor of Law, University of California, Hastings College of the Law
1987-1990	Assistant Professor of Law, University of California, Hastings College of the Law

1986-1987 Judicial Clerkship with The Honorable Thomas M. Reavley of the United States Court of Appeals for the Fifth Circuit, Austin, Texas

Visiting

2000 (May) Universita Piemonte Orientale Amedeo Avogadro, Allassandria, Italy

1998 (May) Universita Degli Studi di Trento, Trento, Italy

1995 (April-June) Universite D'Aix-Marseille, Aix-en-Provence, France

1995 (January-March) Universita Degli Studi di Trento, Trento, Italy

EDUCATION

1986 J.D., University of Virginia, School of Law.
Member of the Editorial Board, VIRGINIA LAW REVIEW.
Member, Order of the Coif.

1983 M.A., University of Virginia (Psychology) (Degree awarded May 1984).

1979 B.A., State University of New York, College at Oswego.
Majors: Psychology and History.

PUBLICATIONS

General Non-Fiction Books

2008 CONSTITUTIONAL FICTIONS: A UNIFIED THEORY OF CONSTITUTIONAL FACTS (Oxford Univ. Press 2008).

2004 LABORATORY OF JUSTICE: THE SUPREME COURT'S 200-YEAR STRUGGLE TO INTEGRATE SCIENCE AND THE LAW (Henry Holt & Co. (Times Books), 2004 (Paperback, 2005 (Owl Books)).

1999 LEGAL ALCHEMY: THE USE AND MISUSE OF SCIENCE IN THE LAW (W.H. Freeman & Co., 1999 (Paperback, 2000)).

Treatises, Course Books & Manuals

2021 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with Edward Cheng, Jennifer Mnookin, Erin Murphy, Joseph Sanders and Christopher Slobogin) (West/Thomson Publishing Co., 2020-2021 Edition) (Volumes 1-5) (forthcoming).

2020 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with Edward Cheng, Jennifer Mnookin, Erin Murphy, Joseph Sanders and Christopher Slobogin) (West/Thomson Publishing Co., 2019-2020 Edition) (Volumes 1-5).

2019 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with Edward Cheng, Jennifer Mnookin, Erin Murphy, Joseph Sanders and Christopher Slobogin) (West/Thomson Publishing Co., 2018-2019 Edition) (Volumes 1-5).

2018 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with Edward Cheng, Jennifer Mnookin, Erin Murphy, Joseph Sanders and Christopher Slobogin) (West/Thomson Publishing Co., 2018-2019 Edition) (Volumes 1-5).

2017 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with Edward Cheng, Jennifer Mnookin, Erin Murphy, Joseph Sanders and Christopher Slobogin) (West/Thomson Publishing Co., 2017-2018 Edition) (Volumes 1-5).

2016 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with Edward Cheng, Jennifer Mnookin, Erin Murphy, Joseph Sanders and Christopher Slobogin) (West/Thomson Publishing Co., 2016-2017 Edition) (Volumes 1-5).

2015 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with Edward Cheng, Jennifer Mnookin, Erin Murphy, Joseph Sanders and Christopher Slobogin) (West/Thomson Publishing Co., 2015-2016 Edition) (Volumes 1-5).

2014 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with Jeremy Blumenthal, Edward Cheng, Jennifer Mnookin, Erin Murphy and Joseph Sanders) (West/Thomson Publishing Co., 2014-2015 Edition) (Volumes 1-5).

2013 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with Jeremy Blumenthal, Edward Cheng, Jennifer Mnookin, Erin Murphy and Joseph Sanders) (West/Thomson Publishing Co., 2013-2014 Edition) (Volumes 1-5).

2012 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with Jeremy Blumenthal, Edward Cheng, Jennifer Mnookin, Erin Murphy and Joseph Sanders) (West/Thomson Publishing Co., 2012-2013 Edition) (Volumes 1-5).

2011 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with Jeremy Blumenthal, Edward Cheng, Jennifer Mnookin, Erin Murphy and Joseph Sanders) (West/Thomson Publishing Co., 2011-2012 Edition) (Volumes 1-5).

2010 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with Michael J. Saks, Joseph Sanders and Edward K. Cheng) (West/Thomson Publishing Co., 2010-2011 Edition) (Volumes 1-5).

2009 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with Michael J. Saks, Joseph Sanders and Edward K. Cheng) (West/Thomson Publishing Co., 2009-2010 Edition) (Volumes 1-5).

2008 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with Michael J. Saks, Joseph Sanders and Edward K. Cheng) (West/Thomson Publishing Co., 2008-2009 Edition) (Volumes 1-5).

2007 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with David H. Kaye, Michael J. Saks, Joseph Sanders and Edward K. Cheng) (West/Thomson Publishing Co., 2007-2008 Edition) (Volumes 1-4).

2006 MODERN SCIENTIFIC EVIDENCE: STANDARDS, STATISTICS AND RESEARCH ISSUES (Student Edition); MODERN SCIENTIFIC EVIDENCE: FORENSICS (with David H. Kaye, Michael J. Saks, Joseph Sanders, and Edward Cheng) (West/Thomson Publishing Co. 2006).

2006 ANNOTATED REFERENCE MANUAL ON SCIENTIFIC EVIDENCE (with Michael J. Saks, David H. Kaye and Joseph Sanders) (West/Thomson Publishing Co., 2006).

2005	MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with David H. Kaye, Michael J. Saks, and Joseph Sanders) (West/Thomson Publishing Co., 2005-2006 Edition) (Volumes 1-4).
2004	ANNOTATED REFERENCE MANUAL ON SCIENTIFIC EVIDENCE (with Michael J. Saks, David H. Kaye and Joseph Sanders) (West Publishing Co., 2004).
2003	MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with David H. Kaye, Michael J. Saks, Joseph Sanders, eds., Supplement 2003, West Publishing Co.).
2002	MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with David H. Kaye, Michael J. Saks and Joseph Sanders, eds., 2d ed., 2002, West Publishing Co.) (Volumes 1-4).
2002	SCIENCE IN THE LAW: STANDARDS, STATISTICS AND RESEARCH ISSUES; SCIENCE IN THE LAW: SOCIAL AND BEHAVIORAL SCIENCE ISSUES; SCIENCE IN THE LAW: FORENSIC SCIENCE ISSUES (with David H. Kaye, Michael J. Saks and Joseph Sanders) (West Publishing Co.) (reprinted chapters from M.S.E. for student edition).
2000	MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with David H. Kaye, Michael J. Saks, Joseph Sanders, eds., Supplement 2000, West Publishing Co.).
1999	MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with David H. Kaye, Michael J. Saks, Joseph Sanders, eds., West Publishing Co.) (Volume III).
1999	MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with David H. Kaye, Michael J. Saks, Joseph Sanders, eds., Supplement 1999, West Publishing Co.).
1997	MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (with David H. Kaye, Michael J. Saks, Joseph Sanders, eds., 1997, West Publishing Co.) (Volumes I & II).

In Preparation

NINO & CLARENCE'S EXCELLENT ADVENTURE: THE TALE OF TWO SUPREME COURT JUSTICES AND THEIR JOURNEY THROUGH TIME TO DISCOVER THE ORIGINAL MEANING OF ORIGINAL INTENT

Using Scientific Data to Make Individual Decisions in the Law (G2i) (with multiple coauthors).

Articles, Essays, Comments & Book Reviews

2021 *Using Burdens of Proof to Allocate the Risk of Error in Juvenile Sentencing*, ___ WM. MARY L. REV. ___ (forthcoming 2021) (with Kelsey Geiser) (Symposium Issue).

2021 *Differential Etiology: Inferring Specific Causation in the Law from Group Data in Science*, ___ ARIZONA L. REV. ___ (forthcoming 2021) (with Joseph Sanders, Peter B. Imrey & Philip Dawid).

2021 *Fact-Finding in Constitutional Cases*, *in* A DIALOGUE BETWEEN LAW AND HISTORY (Baosheng Zhang, Thomas Man, Jing Lin, eds.) (Springer 2021).

2019 *Psychological Assessments in Legal Contexts: Are Courts Keeping "Junk Science" Out of the Courtroom?* 20(3) PSYCHOLOGICAL SCIENCE IN THE PUBLIC INTEREST 135 (2019) (with Tess M.S. Neal, Christopher Slobogin, Michael J. Saks & Kurt F. Geisinger).

2019 *A Man for All Seasons: A Remembrance of Geoffrey C. Hazard*, 70 HASTINGS L. J. 949 (2019).

2018 *Intellectual Disability, the Death Penalty, and Jurors*, 58 JURIMETRICS J. 437 (2018) (with Emily V. Shaw & Nicholas Scurich).

2018 *The Curious Case of* Wendell v. GlaxoSmithKline, Inc., 48 SETON HALL L. REV. 607 (2018) (with Jennifer Mnookin).

2016 *Scientific Gatekeeping: Using the Structure of Scientific Research to Distinguish Between Admissibility and Weight in Expert Testimony*, 110 NW. U. L. REV. 859 (2016) (with Christopher Slobogin and John Monahan).

2015 *Where Law and Science (and Religion) Meet*, 93 U. TEX. L. REV. 1659 (2015).

2015 *Organized Common Sense: Judge Jack Weinstein's Uncommonly Sensible Approach to Expert Evidence*, 64 DEPAUL L. REV. 421 (2015) (with Claire

Lesikar) (invited paper for the Clifford Symposium in honor of Judge Jack Weinstein).

2015 *Toward a Jurisprudence of Psychiatric Evidence: Examining the Challenges of Reasoning from Group Data in Psychiatry to Individual Decisions in the Law*, 69 U. Miami L. Rev. 685 (2015) (with Carl Erik Fisher & Paul Appelbaum).

2014 *Group to Individual (G2i) Inference in Scientific Expert Testimony*, 81 U. Chi. L. Rev. 417 (2014) (with John Monahan & Christopher Slobogin).

2014 *Promises, Promises for Neuroscience and Law*, 24 Current Biology 861 (2014) (with Joshua W. Buckholtz).

2014 *Fitting Science into Legal Contexts: Assessing Effects of Causes or Causes of Effects?* 43 Sociological Methods & Research 359 (2014) (with Philip Dawid & Stephen Fienberg). With additional commentary in:

Authors' Response to Comments on Fitting Science into Legal Contexts: Assessing Effects of Causes or Causes of Effects?, 43 Sociological Methods & Research 416 (with Philip Dawid & Stephen Fienberg);

On the Causes of Effects: Response to Pearl, 44 Sociological Methods & Research 165 (2015) (with Philip Dawid & Stephen Fienberg).

2014 *"My Older Clients Fall Through Every Crack": Geriatrics Knowledge Among Legal Professionals*, 62 Journal of the American Geriatrics Society 734 (2014) (with Tacara Soones, Cyrus Ahalt, Sarah Garrigues & Brie A. Williams).

2013 *Neuroscientists in Court*, 14 Nature Reviews Neuroscience 730 (2013) (with Owen Jones, Anthony Wagner & Marcus Raichle).

2013 *Wading Into the* Daubert *Tide*: Sargon Enterprises, Inc. v. University of Southern California, 64 Hastings L. Rev. 1665 (2013) (with Edward Imwinkelried).

2013 *The* Daubert *Revolution and the Birth of Modernity*, 46 UC Davis L. Rev. 101 (2013).

2012 *Implicit Bias in the Courtroom*, 59 UCLA L. Rev.1124 (2012) (with Jerry Kang, et. al).

2010 *Evidentiary Incommensurability: A Preliminary Exploration of the Problem of Reasoning from General Scientific Data to Individualized Legal Decision Making*. 75 Brooklyn L. Rev. 1115 (2010) (Festschrift for Professor Margaret Berger).

2009 *Defining Empirical Frames of Reference in Constitutional Cases: Unraveling the As-Applied versus Facial Distinction in Constitutional Law*, 36 Hastings Const. L. Q. 631 (2009).

2009 *Standards of Legal Admissibility and Their Implications for Psychological Science*, *in* Psychological Science in the Courtroom: Controversies and Consensus (Jennifer Skeem, Kevin Douglas & Scott Lillienfeld, eds. 2009) (with John Monahan).

2009 *Evidence Code Section 802: The Neglected Key to Rationalizing the California Law of Expert Testimony*, 42 Loyola L. Rev. 427 (2009) (with Edward Imwinkelried).

2008 *A Matter of Fit: The Law of Discrimination and the Science of Implicit Bias*, 59 Hastings L.J. 1380 (2008) (with Nilanjana Dasgupta and Cecilia L. Ridgeway).

2008 *Failed Forensics: How Forensic Science Lost Its Way and How It Might Yet Find It*, 4 Annual Review of Law and Social Science 149 (2008) (with Michael J. Saks).

2008 *Scientific Realism in Constitutional Fact-Finding*, 73 Brooklyn L. Rev. 1067 (Symposium) (2008).

2008 *Anecdotal Forensics, Phrenology, and Other Abject Lessons from the History of Science*, 59 Hastings L.J. 979 (2008).

2008 Three entries in The Encyclopedia of Psychology and Law: (1) Expert Testimony, (2) Expert Testimony, Qualifications of, (3) Expert Testimony, Forms of (Brian Cutler, ed. 2008 (Sage Publications)).

2008 *Admissibility Regimes: The "Opinion Rule" and Other Oddities and Exceptions to Scientific Evidence, the Scientific Revolution, and Common Sense*, 36 Southwestern L. Rev. 699 (Symposium) (2008).

2007 *Fact-Finding in Constitutional Cases*, *in* How Law Knows (Stanford University Press, ed., Austin Sarat et al. 2007) (Amherst College Speaker Series).

2007 *The Limits of Science in the Courtroom*, *in* Beyond Common Sense: Psychological Science in the Courtroom (Eugene Borgida & Susan T. Fiske, eds. 2007 (Blackwell Publishers)).

2006 *Judges as "Amateur Scientists,"* 86 Boston Univ. L. Rev. 1207 (2006).

2006 *Amicus Brief of Constitutional Law Professors David L. Faigman and Ashutosh A. Bhagwat, et al. in the Case of* Gonzales v. Carhart, 34 Hastings Const. L. Q. 69 (2006) (with Ashutosh A. Bhagwat & Kathryn M. Davis).

2005 *Expert Evidence After* Daubert, 1 Annual Rev. Law & Soc. Sci.105 (2005) (with Michael Saks).

2004 *Psychological Evidence at the Dawn of the Law's Scientific Age*, 56 Annual Rev. Psychol. 631 (2004) (with John Monahan).

2003 *Expert Evidence: The Rules and Rationality the Law Applies (or Should Apply) to Social Science Expertise*, in Handbook of Psychology in Legal Contexts (John Wiley & Sons Inc., David Carson & Ray Bull, eds., 2003).

2003 *The Limits of the Polygraph*, 20 Issues in Science & Technology 40 (Fall 2003) (with Stephen E. Fienberg & Paul C. Stern).

2003 *Making Moral Judgments Through Behavioral Science: The "Substantial Lack of Volitional Control" Requirement in Civil Commitments*, 2 Law, Probability And Risk 309 (2003).

2003 *Expert Evidence in Flatland: The Geometry of a World Without Scientific Culture*, 34 Seton Hall L. Rev. 255 (2003).

2002 *Is Science Different for Lawyers?,* 297 Science 339 (2002).

2001 *The Tipping Point in the Law's Use of Science: The Epidemic of Scientific Sophistication that Began with DNA Profiling and Toxic Torts*, 67 Brooklyn L. Rev. 111 (2001).

2001 *Embracing the Darkness:* Logerquist v. McVey *and the Doctrine of Ignorance of Science is an Excuse,* 33 ARIZONA ST. L. REV. 87 (2001).

2000 *The Law's Scientific Revolution: Reflections and Ruminations on the Law's Use of Experts in Year Seven of the Revolution*, 57 WASH. & LEE UNIV. L. R. 661 (2000).

2000 *How Good is Good Enough? —Expert Evidence Under* Daubert *and* Kumho Tire, 50 CASE W.L. REV. 645 (2000) (with David H. Kaye, Michael J. Saks, and Joseph Sanders).

1998 *Looking for Policy in All the Wrong Places: A Comment on the Strategies of the Race and Gender Crowd Toward Evidence Law*, 28 SOUTHWESTERN L. REVIEW 289 (1998).

1998 *Truth with a Small t*, 49 HASTINGS LAW JOURNAL 1185 (1998).

1998 *Should Forensic Science be "Scientific"?* in POLICE, TECHNIQUÉS MODERNES D'ENQUÊTE OU DE SURVEILLANCE ET DROIT DE LA PREUVE, Faculte de droit, Université de Sherbrooke, Quebec (Canada) (28 et 29 mai 1998).

1997 *The Battered Woman Syndrome in the Age of Science*, 39 ARIZONA L. REV. 67 (1997) (with Amy J. Wright).

1997 *Appellate Review of Scientific Evidence Under* Daubert *and* Joiner, 48 HASTINGS L.J. 969 (1997).

1996 *"And the Republic for Which it Stands": Guaranteeing a Republican Form of Government in the Individual States*, 23 HASTINGS CONST..L.Q. 1057 (1996) (with Catherine A. Rogers).

1996 *The Syndromic Lawyer Syndrome: A Psychological Theory of Evidentiary Munificence*, 67 COLORADO L. REV. 817 (1996).

1996 *Making the Law Safe for Science: A Proposed Rule for the Admission of Expert Testimony*, 35 WASHBURN L. REV. 401 (1996).

1995 *The Evidentiary Status of Social Science Under* Daubert*: Is it "Scientific," "Technical," or "Other" Knowledge?,* 1 PSYCHOLOGY, PUBLIC POLICY AND LAW 960 (1995), *revised and reprinted in* 1999 WILEY EXPERT WITNESS UPDATE: NEW DEVELOPMENTS IN PERSONAL INJURY LITIGATION (Eric Pierson, ed., Aspen, 1999).

1995 *Mapping the Labyrinth of Scientific Evidence*, 46 HASTINGS L.J. 555 (1995).

1994 *Modeling Constitutionality Transactionally*, 45 HASTINGS L.J. 753 (1994).

1994 *Check Your Crystal Ball at the Courthouse Door, Please: Exploring the Past, Understanding the Present and Worrying About the Future of Scientific Evidence*, 15 CARDOZO L. REV. 1799 (1994) (with Elise Porter & Michael J. Saks).

1994 *Madisonian Balancing: A Theory of Constitutional Adjudication*, 88 NORTHWESTERN U.L. REV. 641 (1994).

1993 *Constitutional Adventures in Wonderland: Exploring The Debate Between Rules and Standards Through the Looking Glass of the First Amendment*, 44 HASTINGS L.J. 829 (1993).

1992 *Reconciling Individual Rights and Government Interests: Madisonian Principles Versus Supreme Court Practice*, 78 VA. L. REV. 1521 (1992).

1992-2016 *A Flow Chart of the Federal Rules of Evidence*, *in* FEDERAL RULES OF EVIDENCE FOR UNITED STATES COURTS AND MAGISTRATES (Thompson/West, 1992-2006).

1992 *Like Socrates, But Hold the Hemlock: Teaching Law*, *in* FULL DISCLOSURE: DO YOU *REALLY* WANT TO BE A LAWYER (compiled by Susan J. Bell, Peterson's Guides 2d Ed. 1992).

1992 *Struggling to Stop the Flood of Unreliable Expert Testimony*, 76 MINN. L. REVIEW 877 (1992).

1991 *"Normative Constitutional Fact-Finding": Exploring the Empirical Component of Constitutional Interpretation*, 139 U. PA. L. REV. 541 (1991).

1990 *By What Authority: Reflections on the Constitutionality and Wisdom of the Flag Protection Act of 1989*, 17 HASTINGS CON.L.Q. 353 (1990); *reprinted in* THE FLAG AND THE CONSTITUTION, VOL. II, FLAG BURNING AND THE LAW (Michael Kent Curtis ed. 1993).

1989 *To Have and Have Not: Assessing the Value of Social Science to the Law as Science and Policy*, 38 EMORY L.J. 1005 (1989).

1988 *Bayes' Theorem in the Trial Process: Instructing Jurors on the Value of Statistical Evidence*, 12 LAW & HUMAN BEHAV. 1 (1988) (with A.J. Baglioni, Jr.).

1987 *Discerning Justice When Battered Women Kill* [Book Review], 39 HASTINGS L.J. 207 (1987).

1986 Note, *The Battered Woman Syndrome and Self-Defense: A Legal and Empirical Dissent*, 72 VA.L.REV. 619 (1986); *reprinted in* REPRESENTING ... BATTERED WOMEN WHO KILL (S.L. Johann & F. Osanka eds. 1989).

Government Reports/Books

2018 *Translating Scientific Concepts into Legal Constructions (and Vice Versa)*, in *Neuroforensics: Exploring the Legal Implications of Emerging Neurotechnologies: Proceedings of a Workshop*, National Academies of Science, National Academies Press; PDF available at http://nap.edu/25150

2002 *The Polygraph and Lie Detection* (Final Report of the Committee to Review the Scientific Evidence on the Polygraph, National Academies of Science/National Research Council) (National Academy Press).

PROFESSIONAL ACTIVITIES

2017-2020 Member, ABA National Commission on the Future of Legal Education.

2016 Senior Advisor, President's Council of Advisors on Science and Technology (PCAST), Office of Science and Technology Policy, Executive Office of the President, for its Report, "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods."

2011-2016 Member, MacArthur Foundation Law and Neuroscience Network

2011-2016 Principal Investigator, MacArthur Foundation G2i Committee (a Subcommittee of the MacArthur Foundation Law and Neuroscience Network)

2010-present Advisory Council (Law), Institute for Linguistic Evidence

2006-present Member, American Law Institute

2005-present Member, National Commission on Forensic Science and Public Policy, American Judicature Society

2002-present Outside Reviewer, JUDICATURE

2002-present Outside Reviewer, JURIMETRICS

2002-present Member, Editorial Advisory Board, ENCYCLOPEDIA OF FORENSIC AND LEGAL MEDICINE, Academic Press, London.

2008-2011 Member, MacArthur Foundation Law and Neuroscience Project

2001-2002 Committee Member, National Academy of Sciences/National Research Council, Committee to Review the Scientific Evidence on the Polygraph.

2001-2004 Member, Editorial Review Board, JOURNAL OF LEGAL EDUCATION.

2000-2006 Member, Editorial Review Board, PSYCHOLOGY, PUBLIC POLICY AND THE LAW.

2000-present Outside Reviewer, National Science Foundation, Wash. D.C.

2000 Guest Editor, AMERICAN PSYCHOLOGIST.

1998-99 Chair, Section on Law and Social Science, American Association of Law Schools.

1996 Contributing Editor, CRIMINAL LAW BULLETIN.

1995-present Member, Editorial Review Board, LAW AND HUMAN BEHAVIOR.

1995-96 Chair, Section on Law and Mental Disability, American Association of Law Schools.

1989-90 Member, Multistate Bar Essay Exam Committee, National Conference of Bar Examiners.

1982-83 Chair, Committee on Student Involvement, Division of Law and Psychology of the American Psychological Association: The Committee was formed in August 1982 to facilitate student participation in the Division of Law and Psychology.

MEMBERSHIPS, BOARD OF DIRECTORS

2018-2020 Member, Board of Directors, Lowboknow, LLC

2013-2018 Member, Board of Directors, Institute for Responsible Nutrition (IRN)

2013-present Chair, Board of Directors, Lawyers for America (LfA)

OTHER ASSOCIATED ACTIVITY

2014-2018 The Exploratorium, San Francisco; Host, "In the Balance: Bringing Science to Justice" – a bimonthly program on subjects at the intersection of law and science.

BAR MEMBERSHIP

1992-present Member, California Bar

MISCELLANEOUS AWARDS AND HONORS

2015 Innovator of the Year Award, awarded by *The Recorder*, for co-founding JuriLytics, LLC., a company formed to bring expert academic peer review to expert testimony (with Dr. Amit Lakhani).

2014 Innovator of the Year Award, awarded by *The Recorder*, for formulating and developing the idea for Lawyers for America (with Prof. Marsha Cohen).

2009 Visiting Consortium Professor of Law, University of Minnesota Consortium on Law and Values in Health, Environment and Life Sciences.

2008 Elected Honorary Distinguished Member - American Psychology-Law Society (AP-LS).

1991 Awarded Honorable Mention (2d Place) in The Annual AALS Call for Scholarly Papers.

1986 Received the Roger and Madeline Traynor Prize, awarded to acknowledge the best written work by a graduating student, University of Virginia, School of Law.

1979 Awarded the William G. McGarvey Award presented to the Outstanding Senior in the Department of Psychology, State University of N.Y., College at Oswego.

1978 Member of the National Honor Societies in Psychology (Psi Chi) and History (Phi Alpha Theta).

PAPERS AND PRESENTATIONS

2019 Presented at, and participated in, day-long conference sponsored by the American Association for the Advancement of Science (AAAS) at the National Judicial College in Reno, NV, on integrating science into legal decision making, with special emphasis on issues arising with algorithms and addiction. November 22, 2019.

2019 Presented "Fact-Finding in Constitutional Decision-Making," at a conference held at Peking University School of Transnational Law, Shenzhen, China, November 16-17, 2019.

2018 Presentations to State and Federal judges, National Judicial College, on "The Intersection of Law and Science," "Forensic Science," and "Medical Causation," Reno, Nevada, August 20, 2018.

2018 Keynote, Developing a Framework for the Use of Scientific Evidence in the Law, European Association of Psychology & Law, Turku, Finland, June 26, 2018.

2018 Developing a Framework for Use of Scientific Evidence: The Abiding Challenge of Translating Scientific Concepts into Legal Constructions (and Vice Versa), Forensic Science Conference, University of Virginia, Charlottesville, VA, March 26, 2018.

2018 Developing a Framework for Use of Evidence From Emerging Neurotechnologies – A Way Forward, National Academies of Science Workshop: Neuroforensics: Exploring the Legal Implications of Emerging Neurotechnologies, Washington, DC, March 6, 2018.

2018 Evaluating Scientific Evidence and Considering When – If Ever – Judges Can Ethically Get It For Themselves, Program sponsored by the Texas Court of Criminal Appeals, Austin, Texas, February 2018.

2018 *Daubert* After 25 Years: A Prospective Look at the Next Great Challenges in Expert Reliability, American Association of Law Schools Annual Meeting, January, 2018.

2016 Presented lectures to state judges on various areas of scientific evidence (forensics, behavioral science, medical causation, and predictions of violence) for a program sponsored by the National Judicial College, Clearwater, Florida, September, 2016.

2016 Presented “Managing Scientific Evidence in the Age of Science,” to a group of Illinois Judges, sponsored by the Illinois Judicial Education Administration, Chicago, IL, April, 2016.

2016 Presented “Operationalizing Law/Operationalizing Science: The Abiding Challenge of Translating Scientific Concepts into Legal Constructions (and Vice Versa) to the National Academies of Science, Committee on Science, Technology and Law, Newport Beach, CA, March, 2016.

2016 Presented “Operationalizing Neuroscience,” at the “Law and Neuroscience” Symposium at Fordham University School of Law, February, 2016.

2016 Presented “Changes in Evidence Science or, Bringing Scientific Sensibilities to Scientific Evidence,” to a group of Louisiana State judges, sponsored by the Louisiana Judicial College and the Louisiana Association for Justice, New Orleans, LA, February, 2016.

2016 Invited panelist, Georgetown University School of Law, program on “Rational Basis Review,” February, 2016.

2016 Presented “Managing Scientific Evidence in the Age of Science,” to a group of Illinois Judges, sponsored by the Illinois Judicial Education Administration, Chicago, IL, February, 2016.

2015 Presented paper entitled “Disinterested Science and Interested Scientists: Some Thoughts on the Perils and Pitfalls at the Intersection of Science and Policy,” at the inaugural UC Consortium on Social Science and Law Conference, Irvine, CA., Oct. 2015.

2015 Presentation on Neuroscience and the Law at the ABA/NJC Conclave for state and federal judges on contemporary issues and findings in neuroscience and law, Chicago, IL, Oct. 2015.

2015 Taught in Master of Judicial Studies Program, 1 week, on topics in Science and Law for the University of Reno, Nevada, July, 2015.

2015 Presented “Judges as Amateur Scientists? Using Scientists to Measure the Science in Scientific Evidence,” at the 2015 Advanced Judicial Academy for Illinois State judges, University of Illinois College of Law, June, 2015.

2015 Presentation on “The Role of the Courts in Improving Forensic Science,” in a program for State and Federal Judges organized by the ABA, April, 2015, Chicago, IL.

2015 Presentation entitled “’Forensic “Science’: How to Determine When it Is Scientific and When it is Not,” at the 2015 Criminal Justice Conference, organized by the Texas Court of Criminal Appeals, February, 2015,Austin, TX.

2015 Two presentations at the CACJ & CPDA 2015 Capital Case Defense Seminar, entitled “*Daubert*, *Frye* & *Sargon*,” and “Statistics for Lawyers: Intellectual Disability as a Case-in-Point,” Monterey, CA, February, 2015.

2015 “Where Law and Science (and Religion) Meet,” Keynote Presentation at a Symposium on Science and Law at the University of Texas, Austin, TX, Jan., 2015.

2014 Two presentations, organized by the National Judicial College, to the California Personnel Board, entitled “Opinion Testimony: Experts,” and “Medical Evidence: The Challenge of Determining Causation Through Toxicology, Epidemiology and Misc. Other Scientific Methods,” Sacramento, CA, Dec. 2014.

2014 Presentation entitled “The Challenge of Scientific Expert Testimony in the 21st Century: Neuroscience as a Case-in-Point,” at a conference sponsored by the Universita Cattolica, Milan, Italy, Oct. 2014.

2014 Presented lecture on expert testimony and medical causation, to the New Mexico Judicial Conclave (all NM state court judges), Albuquerque, NM, June, 2014.

2014 Presentation on “Why Avoiding *Daubert* at the Trial Level Creates Problems at the Appellate Level” at the 2014 Criminal Justice Conference for Texas State Judges, organized by the Texas Court of Criminal Appeals, Dallas, May, 2014.

2014 Presented lectures on “Forensic Identification ‘Science,’” “Predictions of Violence in Civil Commitment and SVP Cases,” and “Medical Causation,” for a judicial education program at the National Judicial College, Reno, NV, May 2014.

2014 Presented paper “Organized Common Sense: Judge Jack Weinstein’s Uncommonly Sensible Approach to Expert Evidence,” at the Clifford Symposium, in honor of Judge Jack Weinstein, at DePaul University, Chicago, April, 2014.

2014 Presentation on “The Challenge of Reasoning from Group Data in Science to Individual Decision Making in the Law (G2i),” at Faculty Workshop, University of San Diego Law School, San Diego, April, 2014.

2013 Presented ‘work in progress’ on “The Challenge of Reasoning from Group Data in Science to Individual Decision Making in the Law (G2i),” at The Mind Research Network, University of New Mexico, Albuquerque, November, 2013.

2013 Invited lecture on “Reasoning from Group Data to Individual Decision Making,” at Harvard University School of Law, October, 2013.

2013 Presentation on “Current Trends in Admissibility of Expert Testimony,” at the 2014 Criminal Justice Conference for Texas State Judges, organized by the Texas Court of Criminal Appeals, Dallas, May, 2013.

2013 Presentation on “Reasoning from Group Data in Science to Individual Decision Making in the Law,” at an ABA Conference on Law and Neuroscience, organized by the MacArthur Foundation, Chicago, April, 2013.

2013 Presented paper on reasoning from group data in law to individual decision making in the law at the University of Oregon, School of Law, April, 2013.

2013 Presentation on Standards of Admissibility for Expert Evidence at a CLE organized by the New Mexico State Bar, Albuquerque, NM, March, 2013.

2013 Presentation on standards of admissibility and the problem of reasoning from group data in science to individual decision making in the law to a group of federal judges at a conference on Law, Neuroscience and Criminal Justice at Stanford University School of Law, organized by the Federal Judicial Center and the MacArthur Foundation, March, 2013.

2012 Invited Commentator, Empirical Legal Studies Conference, Stanford University, School of Law, November, 2012.

2012 Invited Lecturer, The Current State of Art (and Science) of Forensic Science, ASTAR conference for state judges, Albuquerque, NM, September, 2012.

2012 Keynote Speaker, "ACA's Supreme Court Decision: What's Next?" sponsored by The Greenlining Institute; at CA State Association of Counties (CSAC) conference center in Sacramento, July, 2012.

2012 Presenter, San Francisco PD Office: 2012 San Francisco Justice Summit, at the Koret Auditorium at the Main Library in San Francisco’s Civic Center, May, 2012.

2012 Presentation on “Forensic Science: The State of the Art and Future Directions," California Judges Association (Mid-year meeting), Palm Springs, CA, April, 2012.

2011 Presented lectures on Forensic Science and the Law and Science of Predicting Violence to a program on “Handling Capital Cases," organized and sponsored by the National Judicial College, held in Phoenix, AZ, September, 2011.

2011 Presentations on Scientific Evidence, at the National Judicial College, August, 2011.

2011 Participated in a meeting of legal scholars, forensic scientists, lawyers and judges to discuss ways to begin reforming the field of forensic identification science, held at the offices of the MacArthur Foundation, Chicago, IL, March, 2011.

2011 Presentation on “Implicit Bias in the Courtroom: State of the Field and Institutional Responses So Far," at the UCLA Conference on Implicit Bias, Los Angeles, CA, March, 2011.

2011 Presentation on "Science in the Supreme Court: The Constitutional Significance of ‘Real Differences' Between Men and Women," at the "Frontiers in Women's Health" Conference, February, 2011.

2011 Participated in Federalist Society debate regarding the constitutionality of the Affordable Care Act, February, 2011, at UC Hastings.

2011 Participated in a meeting of legal scholars and scientists to discuss proposals to be submitted to the MacArthur Foundation for Phase II of the Law & Neuroscience Project, Vanderbilt University, Nashville, TN, February, 2011.

2010 Participated in a two-day charrette for the San Francisco Exploratorium, with the purpose of identifying new exhibits to be developed when the museum moves to its new location on Fisherman’s Wharf, February, 2010.

2010 Participated on three person inspection team (with Sandra Johnson & George Annas) to review the University of Minnesota’s Consortium on Law and Values in Health, Environment & the Life Sciences, March, 2010.

2010 Presented Lecture on Medical Causation for the National Judicial College, with simultaneous Webinar, in San Diego, CA, April, 2010.

2010 Presentation on Forensic Identification Evidence for the Texas Criminal Defense Lawyers Association, San Antonio, TX, June, 2010.

2010 Presentation on the "Science?" of Forensic Science, at a conference of forensic examiners organized by the National Institute of Justice (US Dept. of Justice), in Clearwater Beach Florida, August, 2010.

2010 Organized, and participated in, a meeting of legal scholars, neuroscientists and statisticians to discuss the challenge of reasoning from group data in science to individual decision making in the law, sponsored by the MacArthur Foundation, September, 2010.

2010 Participated in a meeting of legal scholars and scientists to consider the dangers associated with cognitive biases in forensic identifications, sponsored by the National Science Foundation and held at Northwestern University Law School, September, 2010.

2009 Presented lectures at the National Judicial College, Reno, NV, on 1. "Systemic Similarities and Differences Between Law and Science," 2. "Predictions of Violence in Civil Commitment Hearings Involving 'Sexually Violent Predators,'" 3. "Causation in Medical Evidence," and 4. "The Current State of the 'Art' of Forensic Identification 'Science,'" July, 2009.

2009 Presented talk on "Scientific Evidence: Theory and Practice," for the California Bar's program on Forensic Science, May, 2009.

2009 Presented talk on "Current Developments in the Law of Expert Evidence: *Daubert* and Beyond," for a program sponsored by the Texas Court of Criminal Appeals and the University of Texas Law School, May, 2009.

2009 Visiting Consortium Professor of Law, University of Minnesota Consortium on Law and Values in Health, Environment and Life Sciences. Presented public lecture to the university community, entitled "Science in the Supreme Court: Hypotheses & Hypocrisy in Constitutional Decision Making," April, 2009.

2009 Commentator at Stanford University School of Law Conference on Law and Neuroscience, February, 2009.

2009 Presented the paper "Defining Empirical Frames of Reference in Constitutional Cases: Unraveling the As-Applied versus Facial Distinction in Constitutional Law" at the Hastings Constitutional Law Symposium, February, 2009.

2008 Presentation on Standards of Admissibility, MacArthur Foundation's Network on Law and Neuroscience, St. Louis, October, 2008.

2008 Morning presentation (3 hrs.) on "Admissibility of Expert Evidence in Federal Courts," for a program sponsored by the United States District Court, Puerto, Rico, August, 2008.

2008 Discussion leader in the session "Nuts and Bolts for New Law Teachers," at the Conference on Evidence, AALS, June, 2008.

2008 Panelist discussing "Soft Science and Non-Science: Controlling Expertise in the Courtroom," at the Conference on Evidence, AALS, June, 2008.

2008 Presented lecture on "Scientific Evidence," at the Louisiana Defense Lawyers Association Conference, New Orleans, April, 2008.

2008 Presented the paper "Anecdotal Forensics, Phrenology and Other Abject Lessons from the History of Science," at the Faces of Forensics Conference, Hastings Law Journal, March, 2008.

2008 Presented the paper "A Matter of Fit: The Law of Discrimination and the Science of Implicit Bias," at a conference organized by the Hastings Law Journal, February, 2008.

2007 Lectured on the subjects of the Battered Woman Syndrome and Predictions of Violence, National Judicial College, Reno, Nevada.

2007 Public Lecture, "Adjudicating Faith: The Law's Obligation to Reconcile Religion and Science Through the Prism of the First Amendment," for the 2006-2007 Collin College Distinguished Speaker Series, Dallas, Texas.

2007 Presented paper "Scientific Realism in Constitutional Law," at the Brooklyn Law Review's symposium on "Truth," Brooklyn Law School.

2007 Presented paper "Admissibility Regimes" at the Southwestern University Law Review's symposium, Rules of Evidence: FRE v. CEC, Los Angeles, CA.

2006 Presentation on "Our Empirical Constitution," at the University of Illinois School of Law (faculty lunchtime talk), Champaign, Illinois.

2006 Lectured on the subjects of the Battered Woman Syndrome and Predictions of Violence, National Judicial College, Reno, Nevada.

2006 Panelist and presenter on the Panel "Social Science and Judging" at symposium on "Judging in the Twenty-First Century," at Boston University School of Law, Boston, MA.

2006 Attended First Annual Retreat for the American Judicature Society Commission on Forensic Science and Public Policy and gave presentation on “Good Science, Bad Science, and No Science,” to Commissioners, Greensboro, NC.

2006 Lectured on “Challenges to Forensic Science,” at the National College of District Attorney’s Annual Conference, San Francisco, CA.

2006 Presentation on “Judicial Responses to the Use of fMRI as a Lie Detector,” at a day-long symposium on Neuroscience and the Law at Stanford University School of Law.

2006 Presented “A Unified Theory of Constitutional Facts,” at Stanford University School of Law.

2006 Presentation on the “Daubert Trilogy,” at the Annual Conference of the American College of Legal Medicine, Las Vegas, Nevada.

2005 Panelist (3 Panels) at the National Academies of Science Sackler Colloquium on Forensic Science, Washington, DC.

2005 Presentation on Psychology and the Law, Predictions of Violence and the Scientific Method, to the state-wide California Judicial Conference, San Diego, California.

2005 Lectured on the subjects of the Battered Woman Syndrome and Predictions of Violence, National Judicial College, Reno, Nevada.

2005 Week-long course on Science and the Law, Master of Judicial Studies Program, University of Nevada, Reno.

2005 Presented article “A Unified Theory of Constitutional Facts,” Constitutional Law Speaker Series, University of Texas, School of Law.

2005 Presentation to faculty, “A Unified Theory of Constitutional Facts,” Washington & Lee University, School of Law.

2005 Presented work-in-progress, “Constitutional Facts: The Essential Function of Fact-Finding in Establishing Constitutional Standards,” at the Center for the Study of Law and Society, University of California, Boalt Hall School of Law.

2004 Presented lecture on “The Supreme Court’s Struggle to Integrate Science and the Law,” Oregon State University, Corvallis, Oregon.

2004 Presentation to faculty, “A Unified Theory of Constitutional Facts,” Arizona State University, College of Law, Tempe, Arizona.

2004 Presented Symposium paper, “Constitutional Facts: The Essential Function of Fact-Finding in Setting Constitutional Norms,” Amherst College, Amherst, Massachusetts.

2004 Presenter on “Evaluating National Security Surveillance Tools” at the International Biometric Society Conference, Pittsburgh, PA.

2004 Presenter on Science Policy and the Judiciary, for the Board on Behavioral, Cognitive, and Sensory Sciences, National Academies of Science, Washington, D.C.

2004 Presented “Annual Review of Federal and State Cases on Scientific Evidence,” for conference organized by the American Law Institute/American Bar Association, New Orleans, LA.

2004 Panelist on Science Program for The Ninth Circuit Judicial Conference, Monterey, CA.

2004 Lectured on “Selected Topics in Science and the Law,” National Judicial College, Reno, NV.

2003 Commentator for papers presented at Conference on Forensic Science, Univ. of CA, Irvine.

2003 Presented lectures on psychological syndromes and predictions of violence to judges at the National Judicial College, Reno, Nevada.

2003 Presented paper “Making Moral Judgments Through Behavioral Science: The ‘Substantial Lack of Volitional Control’ Requirement in Civil Commitments” at a conference at Cardozo Law School.

2003 Presented paper “Legal Perspectives on Sexually Violent Predators,” at the California Psychological Association, San Jose, CA.

2003 Presented “Annual Review of Federal and State Cases on Scientific Evidence,” for conference organized by the American Law Institute/American Bar Association, San Francisco, CA.

2003 Presented paper “Expert Evidence in Flatland: The Geometry of a World Without Scientific Culture,” at a conference at Seton Hall Law School, Newark, N.J.

2002 Presented keynote address to a group of Canadian judges, organized by Canada's National Judicial Institute, at a conference in Nova Scotia.

2002 Presented the "Annual Report on Science and the Law" for the National Conference on Science and the Law, sponsored by the National Institute of Justice, Miami, Fl.

2002 Presented "Annual Review of Federal and State Cases on Scientific Evidence," for conference organized by the American Law Institute/American Bar Association, Bermuda.

2002 Presented lecture on "Daubert and Its Progeny" to a group of judges at a conference organized by the Private Adjudication Center at Duke University. The conference was held in Miami.

2002 Presented "Current Trends in Expert Testimony," for California Judges in Pasadena.

2002 Presented "Medical Expert Testimony," for California Appellate Judges in Pasadena.

2002 Lectured on "Selected Topics in Science and the Law," National Judicial College, Reno.

2002 Week-long course on Scientific Evidence, in the Masters of Judicial Studies Program, Univ. of Nevada-Reno.

2001 Presented lecture on "Appellate Review of Expert Testimony: The Appellate Judge as Gatekeeper," for the ABA Appellate Judges Conference, Boston, MA.

2001 Presented the "Annual Report on Science and the Law" for the National Conference on Science and the Law, sponsored by the National Institute of Justice, Miami, FL.

2001 Presented a series of lectures on admissibility standards, psychological syndromes and statistics to Wisconsin state court judges, Madison, WI.

2001 Presented lectures on psychological syndromes and predictions of violence to judges at the National Judicial College, Reno, NV.

2001 Presented the paper "The Tipping Point in the Law's Use of Science: The Epidemic of Scientific Sophistication that Began with DNA Profiling and Toxic Torts" for a symposium at the Brooklyn Law School.

2001 Presented overview of recent case law for ABA section meeting on Tort and Insurance law in San Francisco.

2001 Keynote address for Mealey's Publications, conference on scientific evidence, Boston, MA.

2001 Presented lecture on "Current Trends in Forensic Science," for the Northern District of CA's annual judges conference, at Chaminade.

2001 Presented lecture on "Expert Testimony Trends in Federal Practice" for a conference organized and sponsored by ALI-ABA in San Francisco.

2001 Lecture on "Current Trends in Scientific Evidence" for a group of state and federal judges at Duke University.

2001 Presented two lectures on issues concerning the integration of science into the law for the Illinois Judicial Conference, at the University of Illinois.

2001 Presented two lectures on syndrome research and general issues concerning science in the law at a Florida judges' conference in Naples, Florida.

2000 Congressional Briefing, Standards of Admissibility for Scientific Evidence, Washington, D.C., sponsored by the American Chemical Society.

2000 Presented paper on "Current Trends in Expert Testimony" at Seton Hall University Law School, Newark, New Jersey.

2000 Presented lecture on "Integrating Scientific Research into Legal Decision Making" for the Annual Education Meeting of the Florida Conference of District Court of Appeal Judges, St. Augustine, Florida.

2000 Invited Speaker for Congressional Informational Meeting sponsored by the American Chemical Society, Washington, D.C.

2000 Panelist on a panel on science in the law at the Ninth Circuit Judicial Conference, Sun Valley, Idaho.

2000 Co-organized and presenter, program on understanding critical scientific thinking, at the American Bar Association's annual conference, held in New York.

2000 Taught at a week-long program on "Scientific Evidence and Expert Testimony," at the National Judicial College.

2000 Participated in a panel session on "Criminal Law and the Gene Revolution" at the Law and Biology Conference, organized by the Gruter Institute, held at Squaw Valley, CA.

2000 Lecture on "The *Daubert* Trilogy" to a group of state and federal judges at Duke University.

2000 Presentation to Florida State Judges on scientific evidence and the "*Frye* Test after *Daubert*" in Orlando, Fla.

2000 Moderator of panel "When the Rubber Meets the Road: Kumho Tire Past Daubert — Will the Gatekeeper Accept the World Wide Web?" Section of litigation, American Bar Association meeting, Seattle, Washington.

2000 Presented paper at conference on Kumho Tire and Expert Testimony at Washington and Lee University, School of Law, Lexington, Virginia.

2000 Invited Presidential Address, American Psychology-Law Society, New Orleans, Louisiana.

2000 Guest, Beyond Computers, National Public Radio, hosted by Maureen Taylor.

2000 Guest, Forum, KQED, hosted by Michael Krasny.

1999 Guest, Science Fridays, Talk of the Nation, National Public Radio, hosted by Ira Flatow.

1999 Guest, Technation, National Public Radio, hosted by Moira Gunn.

1999 Presented talk based on "Legal Alchemy: The Use and Misuse of Science in the Law" at the Yale Law School.

1999 Presented faculty talk based on "Legal Alchemy" at the University of Connecticut, School of Law.

1999 University talk at Arizona State University, based on "Legal Alchemy."

1999 Presented faculty talk based on "Legal Alchemy" at Arizona State University.

1999 Lecture on science in the law in the Sociology Dept., Wellesley, College.

1999 Presented lecture on Standards of Admissibility of Scientific Evidence for Florida Appellate Judges.

1999 Moderator of panel on Expert Testimony at APA-ABA joint conference, Washington, D.C.

1999 Participated in invited conference on "The State of the Field" of law and psychology at Simon Frazier University, Vancouver, B.C.

1999 Presented paper on Genetics and Criminal Law at the Gruter Institute's Annual Conference, Squaw Valley, California.

1999 Organized and presented lectures for a week-long program on law and science for judges at the National Judicial College, Reno, Nevada.

1999 Participated in "Fred Friendly Seminar" for the California Judges Education and Research Program in San Diego, with Charles Nesson moderating.

1999 Organized and participated on a panel for the Law and Social Sciences division of AALS in New Orleans on Using Science in the Legislative Process and Before Administrative Agencies.

1998 Presented lecture, "Social Science in the Courtroom," at the National Judicial College's week-long program, Science in the Courts, Reno, Nevada, Nov. 19, 1998.

1998 Presented the paper, "Should Forensic Science be "Scientific"? at the Conference, Police, Techniques Modernes D'enquête ou de Surveillance et Droit de la Preuve, Université de Sherbrooke, Quebec.

1998 Commentator on the panel, "Race, Gender and Evidentiary Policy," at the AALS Annual Conference in San Francisco.

1997 Presented the paper, "Are the Social Sciences and Forensic Sciences 'Science'?" at the ABA Annual Conference in San Francisco.

1997 Presented the paper, "Expert Testimony Under the Federal Rules of Evidence," at a conference of state and federal judges, Duke University (sponsored by the Private Adjudication Center).

1997 Presented a lecture on standards of admissibility for scientific expert testimony at a "Bench and Bar Conference," Seattle, Washington.

1997 Presented Colloquium, "What Every Psychologist Should Know About the Law of Expert Testimony," at the University of California, San Diego (Psychology Department).

1996 Presented Colloquium, "Scientists, Sorcerers & Charlatans," at Washington & Lee University, School of Law.

1996 Presented talk, "Scientists, Sorcerers & Charlatans," as part of the Villanova Law School's Annual Lecture Series.

1996 Presented the paper, "The Admissibility of Expert Testimony in the United States," at the European Conference of Psychology and Law, in Siena, Italy, August 1996.

1996 Presented a Seminar on “The Standards of Appellate Review of Expert Testimony,” at the ABA's Appellate Judges Conference, in Portland, Oregon, August 1996.

1996 Presented the paper, “Expert Testimony Under the Federal Rules of Evidence,” at a conference of state and federal judges, Duke University (sponsored by the Private Adjudication Center).

1996 Presented the paper, “The Admissibility of Scientific Expert Testimony,” for the Texas Judicial Center, Dallas, Texas (for a conference of 350 Texas state judges).

1996 Presented the paper, “Making the Law Safe for Science...,” at The Washburn Law Journal Symposium, “Tort Reform?” Impact of 'loser pays' and 'honesty in evidence,'” March, 1996.

1996 Presented the paper, “The Syndromic Lawyer Syndrome...,” at The University of Colorado Law Review Symposium, “O.J. Simpson and the Criminal Justice System on Trial,” February, 1996.

1995 Participant in round table discussion of the role of court-appointed experts in the American trial process at Duke University (sponsored by the Private Adjudication Center).

1995 Presented the paper, “The Primary System in the United States,” Universita Degli Studi di Trento, Trento, Italy.

1995 Participant in round table discussion of “Religious Freedom in France,” at Universite D'Aix-Marseille, Aix-en-Provence, France.

1995 Presented a paper on “United States Immigration Policy: California's Proposition 187,” at Goethe University, Frankfurt Am-Main, Germany.

1995 Presented the paper, “The Admissibility of Scientific Evidence after *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,” at a conference of state and federal judges, Duke University (sponsored by the Private Adjudication Center).

1994 Participated on a panel discussing evidentiary standards for admitting expert testimony at a conference of federal judges in New Orleans (sponsored by the Federal Judicial Center).

1994 Presented the paper, “The Admissibility of Scientific Evidence after *Daubert v. Merrell Dow Pharmaceuticals*,” at a conference of state and federal judges, Duke University (Sponsored by the Private Adjudication Center).

1994 Presented the paper, “Modeling Constitutionality Transactionally,” at the Hastings Law Journal Symposium "When is a Line as Long as a Rock is Heavy: Reconciling Public Values and Individual Rights in Constitutional Adjudication," San Francisco, CA.

1994 Presenter on the panel, “Social Science and Mental Health Evidence After *Daubert v. Merrell Dow*,” at the Annual Meeting of the American Association of Law Schools, Orlando, Florida.

1994 Commentator on the panel, “Expert Testimony in the Wake of *Daubert*,” at the Annual Meeting of the American Association of Law Schools, Orlando, Florida.

1993 Presented the paper, “Check Your Crystal Ball at the Courthouse Door,” at a Symposium on Scientific Evidence at Cardoza Law School, New York, N.Y.

1993 Presented the paper, “Life After *Frye*: The Admissibility of Scientific Evidence Under *Daubert v. Merrell Dow Pharmaceuticals*,” at the annual convention of the American Psychology Association, Toronto, Canada.

1993 Moderator for panel on “The Changing Face of Constitutional Interpretation: The Supreme Court and the New First Amendment,” for a symposium sponsored by the Hastings Law Journal, San Francisco, California.

1992 Drafted the Brief for a Group of American Law Professors as *Amicus Curiae* in Support of Neither Party submitted in *Daubert v. Merrell Dow Pharmaceuticals*, No. 92-102, to the Supreme Court of the United States.

1991 Presented the paper, “Struggling to Stop the Flood of Expert Testimony: A Response to Professor Carlson,” at the Conference on Hearsay Reform at the University of Minnesota.

1991 Presented the paper, “Normative Constitutional Fact-Finding,” at the AALS Annual Convention, Washington D.C.

1991 Moderator for panel on “The Right to Die” for a symposium sponsored by the Hastings Law Journal, San Francisco, California.

1983 Presented the paper, “Setting the Odds on Justice: Statistics and Probabilities in the Trial Process,” at the annual convention of the American Psychological Association, Anaheim, California.

1983 Organized and chaired the symposium on “Psychology and Law Training Models” at the annual convention of the American Psychological Association, Anaheim, California.

APPENDIX B

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* x *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 inconclusive, 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) that 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/partly 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 examiners[332] 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-postive-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 images[339]—could attract significant interest from academic and industry.

5.6 Footwear Analysis: Identifying Characteristics

Methodology

Footwear analysis is a process that typically involves comparing a known object, such as a shoe, to a complete or partial impression found at a crime scene, to assess whether the object is likely to be the source of the impression. The process proceeds in a stepwise manner, beginning with a comparison of "class characteristics" (such as design, physical size, and general wear) and then moving to "identifying characteristics" or "randomly acquired characteristics (RACs)" (such as marks on a shoe caused by cuts, nicks, and gouges in the course of use).[340]

In this report, we do not address the question of whether examiners can reliably determine class characteristics—for example, whether a particular shoeprint was made by a size 12 shoe of a particular make. While it is important that that studies be undertaken to estimate the reliability of footwear analysis aimed at

[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/NRBTD). The database contains reflectance microscopy images and three-dimensional surface topography data acquired by NIST or submitted by users.

[340] See: SWGTREAD Range of Conclusions Standards for Footwear and Tire Impression Examinations (2013). SWGTREAD Guide for the Examination of Footwear and Tire Impression Evidence (2006) and Bodziak W. J. *Footwear Impression Evidence: Detection, Recovery, and Examination*. 2nd ed. CRC Press-Taylor & Francis, Boca Raton, Florida (2000): p 347.