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

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

Christopher Chyba
Professor, Astrophysical Sciences and International Affairs
Princeton University

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

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

Christine Cassel
Planning Dean
Kaiser Permanente School of Medicine

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

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

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

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

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

Eric Schmidt
Executive Chairman
Alphabet, Inc.

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

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

Craig Mundie
President
Mundie Associates

Staff

Ashley Predith
Executive Director

Diana E. Pankevich
AAAS Science & Technology Policy Fellow

Jennifer L. Michael
Program Support Specialist

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

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

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

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

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

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
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™, 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™, 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™, 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