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

The President’s Council of Advisors on Science and Technology

No
philosophers
of science.

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

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

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

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

Craig Mundie
President
Mundie Associates

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

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

Eric Schmidt
Executive Chairman
Alphabet, Inc.

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

Staff

Ashley Predith
Executive Director

Jennifer L. Michael
Program Support Specialist

Diana E. Pankevich
AAAS Science & Technology Policy Fellow

PCAST Working Group

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

Working Group

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

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

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

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

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

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

Staff

Diana E. Pankevich
AAAS Science & Technology Policy Fellow

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

Writer

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

Senior Advisors

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

Senior Advisor Co-Chairs

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

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

Senior Advisors

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

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

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

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

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

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

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

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

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

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

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

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

Executive 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.

Main conclusions:

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.

[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.

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*

repeatable,
reproducible,
and accurate

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.

Expert's judgment no substitute for "empirical evidence."

"empirical evidence is thus a sine qua non. Nothing can substitute for it."

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]

Terms of art defined.

"Repeatable" and "Reproducible" have slightly different meanings in the context of controlled trials.

[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.

> *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.

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.

Science
versus
pseudoscience

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.

What science is.

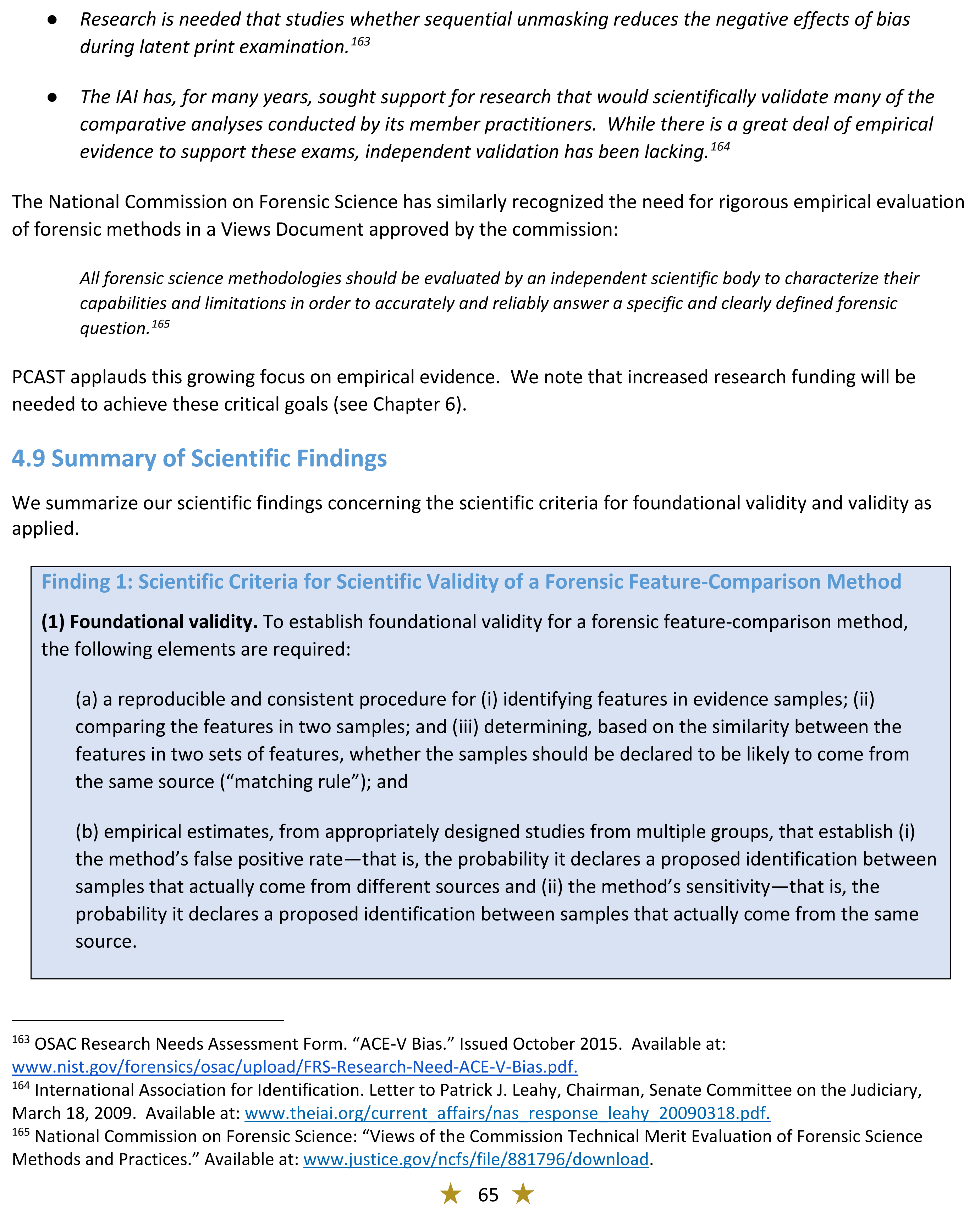
More importantly, the stated method is circular. It declares that an examiner may state that two toolmarks have a "common origin" when their features are in "sufficient agreement." It then defines "sufficient agreement" as occurring when the examiner considers it a "practical impossibility" that the 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).



- *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

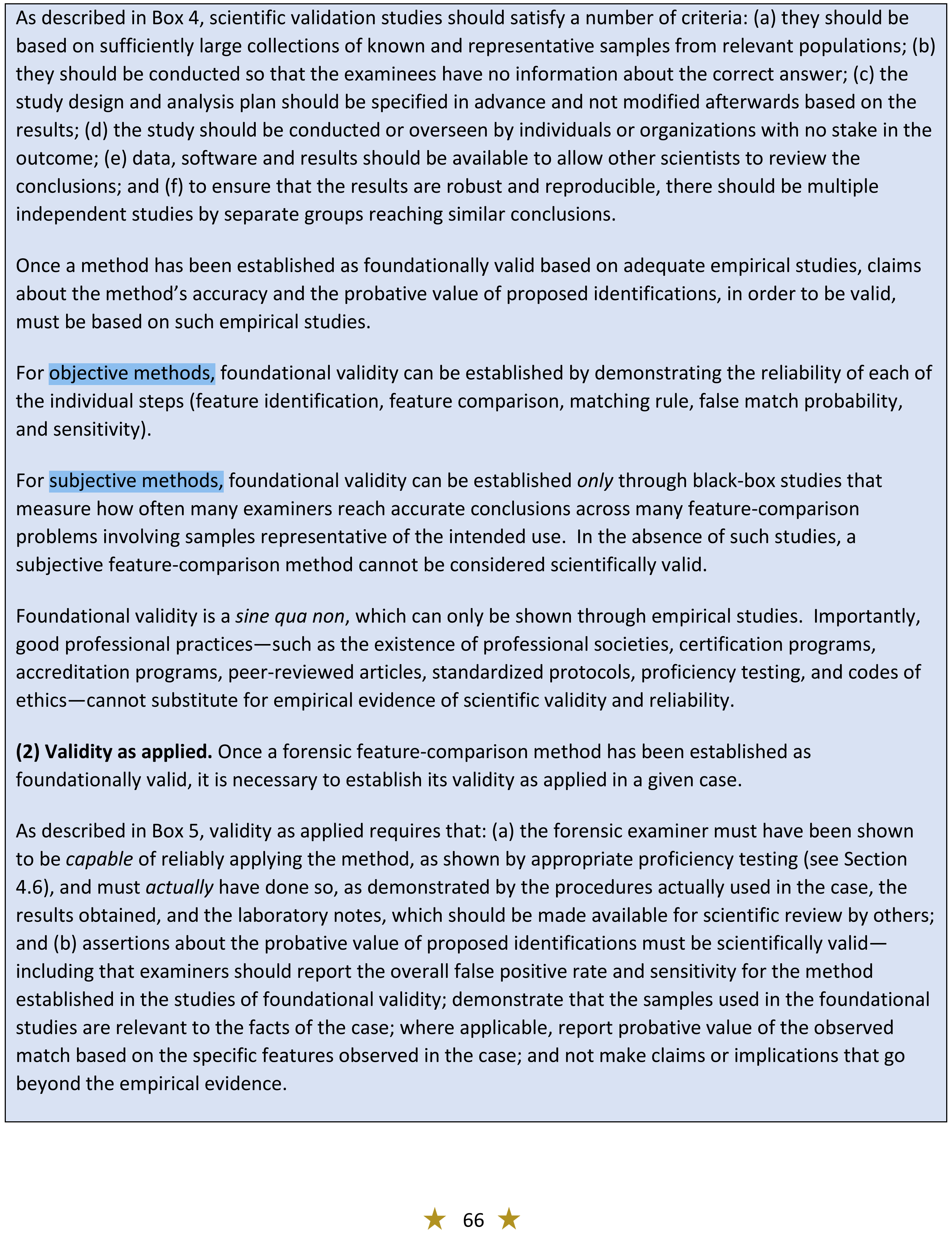
(1) Foundational validity. To establish foundational validity for a forensic feature-comparison method, the following elements are required:

(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.

The summary: too big to digest easily.



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.

Objective
versus
subjective
methhods

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.

Quantity of most interest is the posterior probability.

[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.)

Problem for Bayesian methods.

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*.")

Solution?

[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.