FORENSIC FAIL?

AS RESEARCH CONTINUES TO UNDERSCORE THE FALLIBILITY OF FORENSIC SCIENCE, THE JUDGE'S ROLE AS GATEKEEPER IS MORE IMPORTANT THAN EVER.

VOLUME 102 NUMBER 1 SPRING 2018

JUDICATURE

Published by the Bolch Judicial Institute at Duke Law. judicialstudies.duke.edu/judicature

INTRODUCTION

BY BRANDON L. GARRETT

THIS YEAR MARKS THE 25TH ANNIVERSARY OF THE U.S. SUPREME COURT'S DECISION IN *DAUBERT V. MERRELL DOW PHARMACEUTICALS, INC.*, which fundamentally reshaped how judges evaluate scientific and expert evidence.[1] This volume of *Judicature*, with three wonderful contributions by Jay Koehler, Pate Skene, and an expert team led by William Thompson, comes at an ideal time to reconsider how successful the modern judicial approach to expert evidence has been. That approach is now reflected in Federal Rule of Evidence 702, revised in 2000 to comport with the *Daubert* ruling, and in state judicial rulings and state rules of evidence, which have followed suit in most states.[2]

The Supreme Court's *Daubert* ruling coincided with a surge in scientific research relevant to criminal cases, including the development of modern DNA testing that both exonerated hundreds of individuals and provided more accurate evidence of guilt. Since then, leading scientific commissions have pointed out real shortcomings in the use of forensic evidence in the courtroom. They also have noted that judges have largely abdicated their responsibility as gatekeepers.[3] Moreover, we have learned that those same DNA exonerations are not a one-sided triumph of modern forensic science, because over half of those innocent people were originally convicted by flawed, overstated, and unreliable forensics.[4] A flood of scandals have led to audits of thousands of state and federal cases, lab closures, and review commissions. In response to such concerns, the Judicial Conference Advisory Committee on Evidence Rules has solicited comments on potential revisions to Rule 702 addressing forensic expert testimony.[5]

In this volume, William Thompson and his coauthors describe how we are undergoing a sea change in forensics, particularly in the pattern disciplines. This is a time of crisis but also a time of great promise in forensic science. The terminology used to express conclusions, error-rate statistics, and the fundamental conception of what experts are doing are all in flux. Traditional disciplines like latent fingerprinting, which has been done the same way for over a hundred years, are on the cusp of a transformation. In that field, analysts would typically state without qualification that a print was a "match" and came from the defendant, and that there was a zero probability of an error. Now it is well understood that no human expertise is immune from error, any subjective comparison is inherently probabilistic, and expertise depends on the proficiency of the person doing the analysis.

Today, as Thompson and his colleagues describe, forensic expert conclusions are becoming more appropriately cautious in many disciplines. The 2016 White House Presidential Council of Advisers on Science and Technology (PCAST) report emphasized the need to validate forensics, including by studying error rates and informing jurors of those error rates.[6] The 2017 American Association for the Advancement of Science (AAAS report), which Thompson co-authored, recommended such changes as well and made more detailed recommendations concerning the language to be used in latent fingerprinting.[7] Indeed, forensic conclusions may soon be quantitative; Thompson describes the move to incorporate statistics in forensics. Researchers are hard at work on methods to use algorithms to supplement or even supplant the subjective judgment of individual forensic analysts.

Next, Jay Koehler focuses on reliability: What are the error rates for forensics? As Koehler notes, many people assume that forensics are nearly infallible. If jurors think that forensic experts are infallible but judges know they are not,

then what is the obligation of a judge to ensure that the jury is informed about the limitations of the science?

Rule 702 states that an expert may testify if using "reliable principles and methods," which are "reliably applied" to the facts.[8] Or as the Advisory Committee states, judges shall "exclude unreliable expert testimony."[9] Koehler is right that now is the time to ask whether the "reliability rule" adopted in *Daubert* and in Rule 702 is being appropriately used by the judiciary.

Koehler also highlights the (sometimes quite aggressive) responses to the PCAST report. Some members of government agencies and professional organizations called the report unfounded and biased for suggesting that a range of forensic disciplines lack empirically validated reliability. They suggest there is no problem with continuing to rely on an expert's experience and subjective professional judgment.

In response, Koehler emphasizes judges should not admit evidence just because an expert claims to have experience. Judges should not admit evidence just because other judges have done so for a long time. Judges should not admit evidence just because experts take (extremely unrealistic and easy) proficiency tests. Experts should have to show that their work is reliable and that they are truly proficient. That is, after all, what Rule 702 demands. The problem, Koehler concludes, is not with the text of the rule, but its laissez-faire application by judges.

Finally, Pate Skene further explores what the proper role of judges is at a time when empirical evidence to support so many forensics can be mixed or lacking. Skene describes how judges have themselves raised real questions about the reliability of commonly used forensic techniques. Skene focuses on the problem that for many forensic techniques, well-designed empirical studies have not yet been conducted to validate the reliability of the techniques. The PCAST report emphasized as much.

Turning back to jurors, Skene highlights how important it is for judges not to just exercise their role as gatekeeper, but also to ensure that when forensic evidence is admitted, jurors hear about its limitations. Jurors are highly receptive to information about error rates in forensic techniques and information about the proficiency of particular forensic analysts, as Greg Mitchell and I have shown in several studies.[10] Skene suggests that such information may be conveyed by jury instructions or by additional experts who can explain error rates or reliability concerns to the jury. Skene also suggests that the need for judicial intervention to educate jurors will be greatest when there is less known about the reliability of a forensic technique.

Twenty-five years after *Daubert,* the reliability revolution is still nascent. In an era of plea bargaining and the vanishing criminal trial, it is all the more important that judges safeguard reliability, since it will be the rare occasion when a fact-finder can scrutinize reliability in the courtroom. It is equally important that crime labs themselves incorporate blind proficiency and error management as part of routine quality control. In response to quite complex problems, these thought-provoking contributions from Koehler, Thompson, and Skene set out a clear agenda to bring reliability back into our criminal courtrooms.

BRANDON L. GARRETT is the White Burkett Miller Professor of Law and Public Affairs and Justice Thurgood Marshall Distinguished Professor of Law at the University of Virginia. He joins the Duke Law faculty in July. He recently convened a symposium on "The 25th Anniversary of *Daubert* and the Path Forward for Forensic Science in the Courts" with the *Virginia Journal of Criminal Law* and the Center for Statistics and Applications in Forensics.

1 509 U.S. 579 (1993).

2 Fed. R. Evid. 702; *see* Brandon L. Garrett & M. Chris Fabricant, *The Myth of the Reliability Test,* 86 Fordham L. Rev. 1559 (2018) (summarizing state expert evidence rules, listing states that adopted the revised federal Rule 702, and analyzing state court rulings).

3 *See* Nat'l Research Council, Strengthening Forensic Science in the United States: A Path Forward 95–97 (2009).

4 *See generally* Brandon L. Garrett & Peter J. Neufeld, *Invalid Forensic Science Testimony and Wrongful Convictions*, 95 Va. L. Rev. 1 (2009).

5 Daniel J. Capra, *Foreword: Reed Symposium on Forensic Expert Testimony,* Daubert*, and Rule 702*, 86 Fordham L. Rev. 1459 (2018).

6 President's Council of Advisors on Sci. & Tech., Exec. Office of the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods (2016).

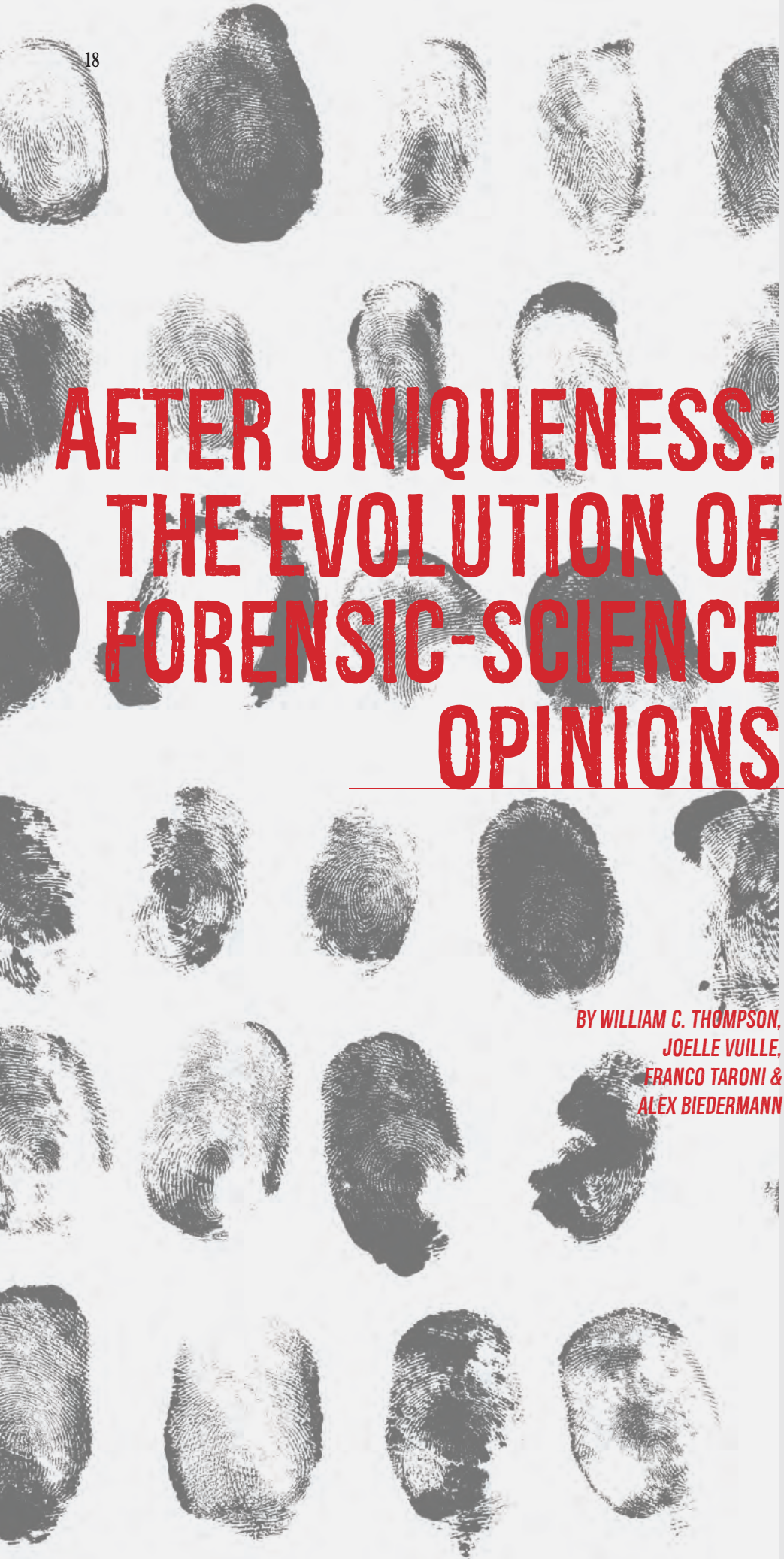
7 Am. Ass'n for the Advancement of Sci., Forensic Science Assessments: A Quality and Gap Analysis: Latent Fingerprint Examination (2017).

8 Fed. R. Evid. 702(c)–(d).

9 Fed. R. Evid. 702 advisory committee's notes to 2000 amendment.

10 Brandon L. Garrett & Gregory Mitchell, *The Proficiency of Experts*, 166 U. Pa. L. Rev. (forthcoming 2018); Gregory Mitchell & Brandon L. Garrett, *The Impact of Proficiency Testing Information on the Weight Given to Fingerprint Evidence* (draft on file with author); Brandon Garrett & Gregory Mitchell, *How Jurors Evaluate Fingerprint Evidence: The Relative Importance of Match Language, Method Information and Error Acknowledgement*, 10 J. Empirical Leg. Stud., 484 (2013).

VOL. 102 NO. 1



AFTER UNIQUENESS: THE EVOLUTION OF FORENSIC-SCIENCE OPINIONS

BY WILLIAM C. THOMPSON, JOELLE VUILLE, FRANCO TARONI & ALEX BIEDERMANN

BIG CHANGES ARE OCCURRING IN FORENSIC SCIENCE, PARTICULARLY AMONG EXPERTS WHO COMPARE THE PATTERNS FOUND IN FINGERPRINTS, FOOTWEAR IMPRESSIONS, TOOLMARKS, HANDWRITING, AND THE LIKE. FORENSIC EXAMINERS ARE REACHING CONCLUSIONS IN NEW WAYS AND CHANGING THE LANGUAGE THEY USE IN REPORTS AND TESTIMONY. THIS ARTICLE EXPLAINS THESE CHANGES AND THE CHALLENGES THEY POSE FOR LAWYERS AND JUDGES.

Although testimony about forensic comparisons has been offered in court for over a century, it has recently become controversial. Questions have emerged about the scientific foundation of the pattern-matching disciplines and about the logic underlying forensic scientists' conclusions. The traditional assumption that items like fingerprints and toolmarks have unique patterns that allow experts to accurately determine their source has been challenged and is being replaced by a new logic of forensic reporting. The new logic requires experts to evaluate and weigh probabilities rather than claim certainty. Forensic experts must now moderate the claims they make about their own accuracy and, increasingly, use numbers to describe the strength of their conclusions. Because these changes have important implications for the probative value of the conclusions that forensic experts offer in court, it is important that judges understand them.

THE DEMISE OF THE THEORY OF DISCERNIBLE UNIQUENESS

As recently as a decade ago, forensic scientists in the pattern-matching disciplines told a common story when asked to explain how they reached conclusions. Their analytic process began with the assumption that the items they examined had unique patterns: For example, every finger was said to have a unique set of friction ridges, and thus every print left by a given finger (if sufficient in size and clarity) was expected to be different from the print made by any other finger. Similarly, every gun barrel was thought to be unique; hence the pattern of marks found on bullets fired through a given barrel (if sufficient in size and clarity) was expected to differ from the pattern found on bullets fired through any other gun barrel. The soles of shoes and human dentition also were presumed to be unique, and thus the impressions left by a given shoe, or a given set of teeth (if sufficiently clear and detailed) were assumed to differ from the impressions left by any other shoe or set of teeth. Applying the same analysis, everyone's handwriting was presumed to be unique, and hence a sample of handwriting from a given individual (if sufficiently extensive) was presumed to be distinguishable from the handwriting of any other individual. These presumptions have been called *the theory of discernible uniqueness*.[1]

According to this traditional account, the job of the forensic examiner was first to assess whether the patterns seen in impressions contained sufficient detail to allow a determination of source and, second, to compare the impression patterns. If sufficient detail was available, then a "match" between the patterns meant the source of the impressions must necessarily be the same, and a mismatch (failure to match) meant that the source of the impressions must necessarily be different. If insufficient detail was available to make a definitive determination, then the examination was inconclusive.

Examiners in a number of forensic disciplines have testified that this analysis allows them to make source determinations with complete certainty. A prominent fingerprint examiner explained the matter as follows:

> *Fingerprint examiners routinely claim to have "identified" or "individualized" an unknown mark to a single known print. This identification is often characterized as being "to the exclusion of all others" on earth to a 100 {percent} certainty, and the comparison method used is claimed to have a zero percent error rate. These claims are based on the premises that friction ridge skin is unique and permanent.*[2]

Unfortunately, these claims have not withstood scientific scrutiny. Indeed, commentary on the issue in the broader scientific and academic communities (beyond the community of forensic science practitioners) has been nearly unanimous in dismissing such claims as unwarranted.[3]

Consider the claim that the ridge patterns on every finger are unique. Like similar claims about snowflakes, it is impossible to demonstrate empirically that this claim is true because one cannot conduct a systematic comparison of every finger against every other. Furthermore, there is a difference between the claim that the ridge pattern on each finger is unique and the claim that a fingerprint examiner can accurately determine whether two fingerprints were made by the same finger. The validity of the latter also depends on the quality of the prints and the level of analysis employed during the comparison. Even if the ridge detail of every finger were unique, it does not follow that every impression made by every finger will always be distinguishable from every impression made by any other finger, particularly when the impressions are of poor quality (e.g., limited detail, smudged, distorted, or overlaid on another impression). By analogy, it may be that every human face is unique, but we can still mistake one person for another, particularly when comparing poor-quality photos.[4]

This is a limitation that most fingerprint examiners now acknowledge:

> *When fingerprint comparisons are being made, they are not being made from friction ridge skin to friction ridge skin. They are being made from one imperfect, incomplete recording to another. . . . {Hence} correctly associating a degraded mark to its true source is by no means a certainty, even were one to presume absolute uniqueness of all friction ridge skin.*[5]

Consequently, the key scientific question is not whether the ridge pattern of ▸

each finger is unique, but how well an examiner can distinguish the impressions of different fingers at the level of analysis applied in a forensic examination. That question cannot be answered by assertions about the uniqueness of ridge patterns; it can only be answered by empirical research.

This critique also applies to other forensic pattern-matching disciplines, such as toolmark analysis, footwear analysis, handwriting analysis, and bitemark analysis. Although some practitioners in these fields persist in making the injudicious claim that their conclusions *must* be accurate because they are comparing patterns that are unique, the broader scientific community has called for empirical studies to put such claims to the test.

A key event in the evolution of forensic science opinion was a 2009 report by the United States National Academy of Sciences (NAS), which called for the development of "quantifiable measures of the reliability and accuracy of forensic analyses" that reflect "actual practice on realistic case scenarios . . . ."[6] It called for research to establish "the limits of reliability and accuracy that analytic methods can be expected to achieve as the conditions of forensic evidence vary."[7] The report concluded that "much forensic evidence — including, for example, bitemarks and firearm and tool mark 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."[8]

In response to this high-level scientific criticism, forensic scientists made some efforts to study the accuracy of their methods, although these efforts have been limited. The FBI commissioned an important series of studies on the accuracy of latent print analysis, but relatively little research has been conducted on the accuracy of other forensic science disciplines. In 2016, the President's Council of Advisors on Science and Technology (PCAST) issued a report that reviewed scientific research published to that point on the accuracy of six forensic science disciplines that rely on "feature comparison": DNA analysis, latent print analysis, firearms analysis, bitemarks analysis, footwear analysis, and microscopic hair analysis.[9] PCAST found that adequate research had been done to establish the "foundational validity" of latent print analysis and DNA analysis of single-source and simple mixture samples. "Foundational validity" means the method in question is capable of producing accurate results when properly performed. PCAST concluded, however, that too little research had been published to establish the "foundational validity" of firearms analysis, bitemarks analysis, footwear analysis, microscopic hair analysis, and DNA analysis of complex mixtures.

Moreover, even if latent print examination has "foundational validity," the studies do not show that it is infallible (as examiners have claimed). The studies reviewed by PCAST showed that latent print examiners have:

> *. . . 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 {one} error in 306 cases {based on an FBI study} and {one} error in 18 cases based on a study by another crime laboratory.*[10]

The studies reviewed by PCAST also showed substantial numbers of false exclusions.[11]

In light of these developments, forensic scientists have begun to change the way they describe their analytic process and report their conclusions. They can no longer credibly claim the ability to infallibly discern whether two compared sets of features share a unique pattern and thus have a common source. Professional associations and standards-setting bodies in various branches of forensic science have recommended that examiners avoid asserting that their conclusions are infallible and avoid claiming that they can discern whether a pattern is unique.[12] Experts are now discussing a variety of new approaches to reporting.

THE LOGIC OF FORENSIC INFERENCE

To understand and evaluate the new approaches to reporting, it is necessary to understand the logic of forensic inference — that is, the logical steps by which a forensic examiner proceeds from observations to conclusions. Let's consider, as an example, the logical steps that lead a latent print examiner from the observation that two fingerprints have similar ridge patterns to conclusions about whether they were made by the same finger. If examiners can no longer credibly claim that prints must necessarily have a common source if they appear to have "matching" ridge patterns, what conclusions can they reasonably draw?

The new approaches all recognize that forensic inference requires an inductive line of reasoning, which entails consideration of probabilities. The examiner must consider the probability of seeing the patterns observed in the impressions under two alternative hypotheses about their origin: (1) that the impressions have the same source (e.g., same finger, same tool); and (2) that the impressions have a different source.

Suppose, for example, that a latent print examiner observes that two fingerprints have similar patterns but with slight discrepancies. The examiner must consider how probable it would be to observe those particular patterns

(including both similarities and discrepancies) *if* the prints were made by the same finger. This might involve consideration of the likelihood that slipping or torsion of the finger, or some other process, could have distorted one or both of the prints enough to produce the discrepancies. The examiner must also consider how probable it would be to observe those particular patterns (including both similarities and discrepancies) *if* the prints were made by different fingers. This would involve consideration of the rarity of the shared features, hence how likely or unlikely it would be to observe so much similarity in prints made by different fingers.

In order to draw inferences and reach conclusions about whether two impressions have a common source, the expert must consider the balance between the two key probabilities: (1) the probability of the observed patterns if the impressions have the *same* source; and (2) the probability of the observed patterns if the impressions have a *different* source. The ratio between these two probabilities provides an index of the probative value of the evidence for distinguishing the two hypotheses. The evidence favors a particular hypothesis to the extent that the observed results are more probable under that hypothesis than under the alternative hypothesis. For example, if a latent print examiner thinks the observed ridge patterns (including both similarities and discrepancies) would be more probable if the prints have the same source (same finger) than if they have a different source (different fingers), then the evidence supports the hypothesis that the prints have the same source.

This logic is fundamental and inescapable. It is the basis for any conclusions that examiners choose to report.

APPROACHES TO REPORTING

There are several schools of thought about how examiners should report their conclusions regarding the balance of probability. In this section of the article, we will outline the different approaches and discuss their strengths and weaknesses.

PROFESSIONAL ASSOCIATIONS AND STANDARDS-SETTING BODIES IN VARIOUS BRANCHES OF FORENSIC SCIENCE HAVE RECOMMENDED THAT EXAMINERS AVOID ASSERTING THAT THEIR CONCLUSIONS ARE INFALLIBLE AND AVOID CLAIMING THAT THEY CAN DISCERN WHETHER A PATTERN IS UNIQUE.

Likelihood Ratios. One approach that is popular in Europe allows examiners to use numbers called likelihood ratios to describe their perception of the balance of probabilities.[13] The likelihood ratio represents the expert's view of the relative probability of the observed features under the alternative hypotheses about the source of the impressions. A likelihood ratio of 1000, for example, represents the expert's view that the observed patterns are 1000 times more probable under one hypothesis (e.g., same source) than under the alternative hypothesis. Experts typically make the favored hypothesis the numerator of the likelihood ratio so that reported values range from one to infinity. A value of one means the expert thinks the observed patterns are equally likely under the two hypotheses, and hence the evidence has no value for distinguishing the hypotheses. A value greater than one means the expert thinks the observed patterns are more likely under one hypothesis than the alternative, and thus the forensic evidence supports the favored hypothesis. The larger the likelihood ratio, the greater the expert's perception of how strongly the balance of probabilities supports the favored hypotheses. European latent print experts sometimes report very high likelihood ratio values, such as one million or even ten million.

The European Network of Forensic Science Institutes (ENFSI) and the U.K. Royal Statistical Society promote the use of likelihood ratios to describe experts' assessments of the strength of forensic evidence.[14] Many forensic scientists in Europe, New Zealand, and parts of Australia also have adopted this approach.[15]

The question most commonly asked about likelihood ratios is how the experts come up with the numbers they report. In some disciplines, experts can rely on databases and statistical modeling. This is most common in fields like forensic DNA analysis and forensic voice comparison, where extensive databases exist and methods for statistical modeling have been evaluated in the scientific literature.[16] Likelihood ratios have been presented in the United States for many years in connection with forensic DNA evidence. The expert typically says something like the following:

> *The genetic characteristics observed in the evidentiary sample are X times more likely if the defendant was a contributor than if the contributor was instead a random unknown Caucasian.*

In the past, there has been insufficient data on the rarity of the features observed by experts in most pattern-matching ▸

disciplines to allow statistical estimates, but that is starting to change. Recently the Defense Forensic Science Center (DFSC) of the Department of the Army began presenting probabilities in connection with fingerprint evidence. In March 2017, the laboratory announced that future reports would include statements like the following:

> *The latent print on Exhibit ## and the standards bearing the name XXXX have corresponding ridge detail. The probability of observing this amount of correspondence is approximately ## times greater when impressions are made by the same source rather than by different sources.*[17]

The laboratory uses a software program to score the similarity of the prints being compared based on "the spatial relationship and angles of the ridge details."[18] The program then uses a database to evaluate how much more common it is to observe a given similarity score when comparing prints from the same finger than prints from different fingers. Although this is a novel method that has not yet been adopted by other forensic laboratories, the DFSC has reportedly offered to share this software with any government forensic laboratory in the United States, and other labs are evaluating this approach. Similar software-based, quantitative methods for assessing toolmark and handwriting evidence also are under development, although it may be a few years before they are ready for the courtroom. As experts begin offering testimony based on these new methods in United States courtrooms, litigants are likely to challenge admissibility under the *Daubert* or *Frye* standards, which will require judges to scrutinize whether the new methods are reliable and generally accepted.

Likelihood ratios also can be reported in forensic science disciplines that have not developed databases and statistical models. In those fields, experts may rely on their training and experience to come up with a likelihood ratio. In some instances, a likelihood ratio can be based partly on empirical data and partly on the expert's judgment.[19] While some commentators have derided such estimates as "subjective" and questioned their validity (one commentator called them "numbers from nowhere"[20]), the practice of presenting likelihood ratios based on expert judgment (rather than a database) appears to have taken hold in many European countries.[21] Whether such testimony should be admitted in the United States is an issue judges may soon need to contemplate.

IF THE EXAMINER DOES NOT KNOW ENOUGH TO ASSESS THE RELEVANT PROBABILITIES, THEN THE EXAMINER DOES NOT KNOW ENOUGH TO EVALUATE THE STRENGTH OF THE FORENSIC EVIDENCE — AND HENCE NOTHING THE EXAMINER SAYS ABOUT THE VALUE OF THE EVIDENCE SHOULD BE TRUSTED.

Those who support the use of likelihood ratios based on expert judgment (rather than databases) point out that a forensic examiner must make subjective judgments of probability in order to draw *any* conclusions about whether two items have a common source.[22] If the examiner does not know enough to assess the relevant probabilities, then the examiner does not know enough to evaluate the strength of the forensic evidence — and hence nothing the examiner says about the value of the evidence should be trusted. It makes no sense, proponents say, to allow experts to testify about conclusions they reached based on a subjective judgment of the balance of probabilities but not allow the expert to use a likelihood ratio to say what their judgment was. When experts report their judgments of the likelihood ratio, proponents argue, the expert's judgmental process is more transparent, and hence the value of the expert's conclusions is easier to evaluate.[23]

Verbal Equivalents of Likelihood Ratios. Examiners may nevertheless be reluctant to put specific numbers on their subjective judgments, even if those judgments are well grounded. An examiner may justifiably believe that the observed results are more probable if the items being compared have the same source than a different source, for example, without being able to say with any precision how much more probable. Forcing examiners to articulate numbers may lend a false air of precision to a subjective assessment.

One way to avoid this problem is to allow examiners to express conclusions about the balance of probabilities using words rather than numbers. In a 2012 report, a group of experts assembled by the National Institute of Standards and Technology (NIST) recommended that latent print examiners report their conclusions using statements like the following:

> *It is far more probable that this degree of similarity would occur when comparing the latent print with the defendant's fingers than with someone else's fingers.*[24]

TABLE 1. PROPOSED LIKELIHOOD RATIO TERMINOLOGY *(AFSP, 2009)*

NUMERICAL EXPRESSION OF PROBATIVE STRENGTH *(likelihood ratio)*	VERBAL EXPRESSION OF PROBATIVE STRENGTH
1 – 10	Weak or limited
10 – 100	Moderate
100 – 1,000	Moderately strong
1000 – 10,000	Strong
10,000 – 1,000,000	Very strong
> 1,000,000	Extremely strong

This approach allows examiners to substitute an imprecise verbal statement ("far more probable") for a number, while still explaining the strength of the forensic evidence in terms of the balance of probabilities. Of course lawyers can (and should) ask experts testifying in this manner to explain what they mean by statements like "far more probable" and what basis they have for that conclusion.

Another approach that has been popular in Europe substitutes words for numerical likelihood ratios. The U.K.-based Association of Forensic Science Providers (AFSP) has proposed that forensic scientists use the "verbal expressions" shown in Table 1 (above) to describe how strongly their evidence supports a particular hypothesis about the evidence (e.g., the hypothesis that two items have a common source).[25] Under this approach, forensic scientists first come up with a likelihood ratio that reflects their perception of the balance of probabilities, and then use one of the verbal expressions in the table instead of (or in addition to) the number to describe their conclusions in reports and testimony.

For example, a forensic scientist who concludes (by whatever means) that the results observed in a forensic comparison are 500 times more likely if the items have a common source than if they have a different source would report that the comparison provides "moderately strong" support for the conclusion that the items have a common source. A forensic scientist who concluded that the results are 100,000 times more likely if the patterns being compared have a common source would say that the evidence provides "very strong support" for the hypothesis of a common source. Statements of this type are not common in U.S. courts, but they have been discussed extensively in the academic literature.[26] They offer one possible answer to the question of how to report source conclusions.

Match Frequencies / Random Match Probabilities. When a comparison reveals matching features in two items, forensic scientists sometimes estimate and report the frequency of the matching features in a reference population. This occurs most commonly in forensic DNA analysis, where genetic databases provide an empirical basis for assessing the proportion of a population that has a particular genetic feature. Forensic DNA analysts sometimes refer to these estimates as *match frequencies* (e.g., "The blood stain at the crime scene and the reference blood sample from the suspect have the same DNA profile. This profile is estimated to occur in one person in 10 million among Caucasian-Americans."). Alternatively, they may present these estimates as *random match probabilities (RMPs)* (e.g., "The probability that a random Caucasian-American would match this DNA profile is 0.0000001 or 1 in 10 million."). As forensic scientists develop databases that can be used to quantify the rarity of pattern features, we are likely to see similar testimony in other pattern-matching disciplines.

Even without empirical data, experts sometimes make statements about the random match probability based on training and experience. These subjective-match probabilities are typically reported with words rather than numbers. An examiner might say, for example, that the set of features shared by two items is "rare" or "unusual."

One drawback of this approach is that it addresses only one of the two questions needed to evaluate the balance of probabilities reflected in the likelihood ratio. It addresses the probability of the observed patterns under the hypothesis that they have a different source. It fails to consider the probability of the observed patterns if the impressions have the same source. Consequently, this approach may be misleading in cases in which the latter probability is low, when, for instance, the patterns have important discrepancies as well as similarities. Likelihood ratios, which consider both probabilities, arguably offer a more balanced and complete account of the strength of such evidence.

Source Probabilities. In the United States, forensic examiners often present opinions on the probability that two items have a common source. Opinions of this type can be expressed quantitatively, using probabilities or percentages. For example, a forensic scientist might say there is a 99 percent chance that two items have a common source. It is more common, however, for examiners to express such conclusions with words rather than numbers. For example, the forensic scientist might say it is "moderately probable," "highly probable," or "practically certain" that two items have a common source.

Lawyers and judges tend to like source probabilities because they are ▸

easy to understand; they address the exact question that the trier of fact needs to assess — how likely it is that the two impressions (e.g., two fingerprints) come from the same source? The problem, unfortunately, is that the information forensic scientists can glean from a comparison of impressions is not, by itself, sufficient to allow them to reach conclusions about source probability. As we will explain, examiners can logically draw conclusions about source probabilities *only* by combining conclusions drawn from a comparison of the impressions with assumptions or conclusions about the strength of other evidence that bears on the question of whether the impressions being compared have a common source.[27]

To illustrate, consider the Elvis Problem discussed in the sidebar. What is the probability that Elvis Presley was the source of the evidence left at the crime scene? As explained, this question cannot be answered based on the forensic science evidence alone. It is only by making assumptions or drawing conclusions about the likelihood of Elvis being at the crime scene — a matter having nothing to do with the forensic science evidence — that the forensic examiner can draw conclusions about the probability that Elvis was the source.

The same problem arises whenever forensic scientists express opinions on source probabilities. The opinion must, of logical necessity, depend in part on conclusions or assumptions about matters having nothing to do with forensic science, such as whether the person who is alleged to have left a trace (e.g., a fingerprint or shoeprint) at the crime scene is a likely or unlikely suspect and how many other people had access to the crime scene. Forensic examiners are not in a good position to make such judgments and have no business doing so anyway.

IN ORDER TO UNDERSTAND THE EXPERT'S CONCLUSIONS, THE TRIER-OF-FACT WILL NEED TO KNOW HOW THE EXPERT EVALUATED THE RELEVANT PROBABILITIES, AND HOW, WHERE, AND WHY THE EXPERT SET THE THRESHOLD FOR REPORTING A PARTICULAR DECISION. THE TRIER-OF-FACT ALSO WILL NEED INFORMATION ABOUT THE ACCURACY OF DECISIONS REACHED IN THIS MANNER.

Identification and Exclusion. In the United States, the most popular method of reporting results of forensic comparisons is to state a bottom-line conclusion about whether two traces have a common source. The conclusion that two traces have the same source is often described as "identification" or "individualization," while a conclusion that they have a different source is "exclusion." These conclusions can be seen as extreme examples of source probabilities, corresponding to either a 100 percent or a zero percent chance that the traces being compared have the same source.

The demise of the theory of discernible uniqueness has made these conclusions more difficult to justify. Most experts now acknowledge that these conclusions require the examiner to *make a decision* about whether the evidence is strong enough to support a definitive conclusion, but there does not appear to be a generally accepted theory regarding how experts should make that decision.

One approach requires experts to make an assessment of the source probability. They report "identification" when their assessed source probability exceeds some high threshold and "exclusion" when their assessment falls below some low threshold. As discussed in the previous section, however, the assessment of source probabilities requires the expert to make assumptions or draw conclusions about matters beyond the forensic comparison in question. Experts cannot draw conclusions about source probabilities without facing the Elvis Problem, which renders such conclusions problematic. If courts allow experts to present conclusions reached in this manner, they should also require experts to disclose the factual basis for their asserted source probabilities. To evaluate the expert's conclusion, the trier-of-fact will need to know the extent to which the expert's decision was influenced by assumptions or conclusions about matters beyond the realm of forensic science.

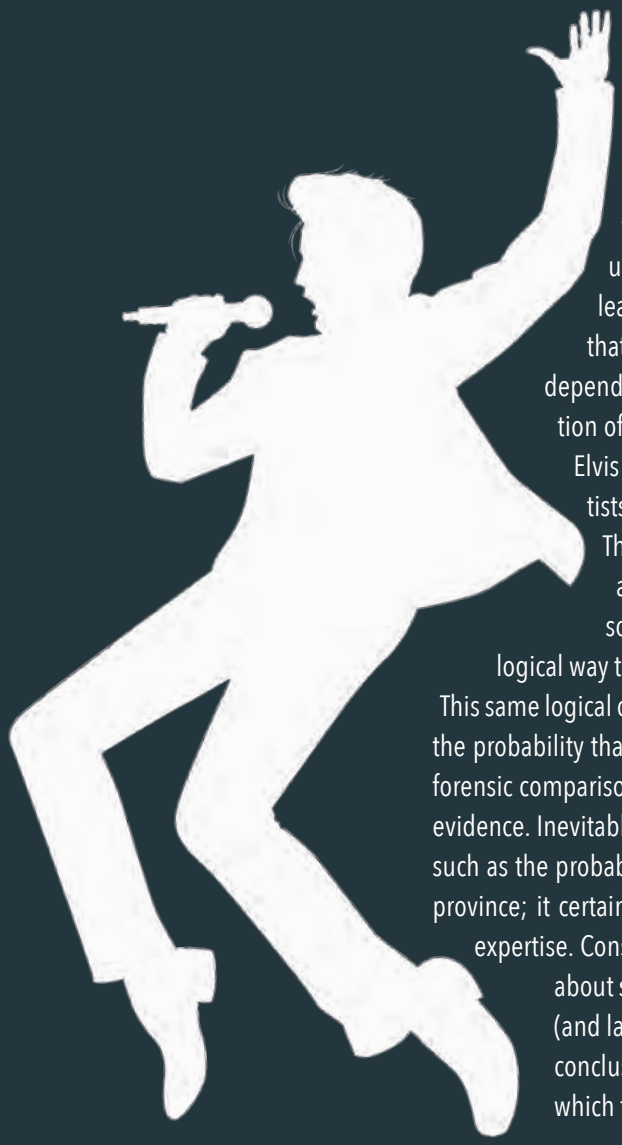
To avoid the Elvis Problem, forensic scientists might instead base their decision on their judgment of the balance of probabilities. If they believe the balance weighs strongly enough in favor of the hypothesis that the items being compared have the same source, then they might report "identification." If they believe the balance weighs strongly enough in favor of the hypothesis that the items have a different source, then they might report "exclusion." This approach avoids the need for

ELVIS'S ALIBI

IMAGINE THAT A BLOODSTAIN OF RECENT ORIGIN IS FOUND AT THE SCENE OF A CRIME. IMAGINE FURTHER THAT THE DNA PROFILE OF THE BLOODSTAIN IS SOMEHOW DETERMINED TO BE THE SAME AS THE DNA PROFILE OF ROCK-AND-ROLL LEGEND ELVIS PRESLEY. FINALLY, IMAGINE THAT THE DNA PROFILE IN QUESTION IS ONE MILLION TIMES MORE LIKELY TO BE OBSERVED IF THE SAMPLE CAME FROM ELVIS THAN IF IT CAME FROM A RANDOM PERSON. BASED ON THE DNA EVIDENCE, WHAT CAN THE EXAMINER LOGICALLY INFER ABOUT THE PROBABILITY THAT THE CRIME SCENE STAIN CAME FROM ELVIS PRESLEY?

A moment of reflection should be sufficient to realize that the examiner can draw no conclusion about the probability that the crime scene stain came from Elvis based on the DNA evidence alone; the examiner must also consider other matters, such as whether Elvis could plausibly be the source. In this case, the suspect (Elvis) has a strong alibi – he was widely reported to have died in 1977. If the forensic scientist believes this "alibi," then the probability that the bloodstain came from Elvis is necessarily zero.

An examiner who believes Elvis is dead might decide to report that there is a zero percent chance the crime scene sample came from Elvis. Notice, however, that this conclusion is not based on the strength of the DNA evidence. It depends entirely on the expert's assessment of matters beyond the realm of forensic science – in this case Elvis's alibi.



The expert might try to take a neutral position on the alibi – assuming, for example, that the question of whether Elvis could have been the source is a toss-up or 50:50 chance. When this seemingly neutral assumption about the truth of the alibi is taken as a starting point, the expert can update the initial assessment in light of the DNA evidence. That approach leads logically to the conclusion that there is more than a 99 percent chance that Elvis was the source of the blood.[32] Notice, however, that this conclusion depends only partly on the DNA evidence; it also depends critically on the assumption of a 50 percent chance *a priori* that the blood at the crime scene came from Elvis (an assumption many people will view as fanciful). Should forensic scientists be basing their conclusions on assumptions of this type?

The problem (as should now be clear) is that no assumption about the probability of an alibi's veracity can truly be considered "neutral." Yet without some assumption about the probability of the alibi's veracity, there is no logical way to assess the probability that Elvis was the source.

This same logical conundrum arises in any case in which a forensic scientist is asked to assess the probability that a particular suspect was the source of a crime scene sample based on a forensic comparison. The expert can never answer the question based solely on the forensic evidence. Inevitably the expert must make assumptions or take a position on other matters, such as the probability that the suspect's alibi is true. Doing that may well invade the jury's province; it certainly requires the expert to delve into matters beyond his or her scientific expertise. Consequently, judges should consider carefully whether to admit statements about source probabilities into evidence. If such statements are admitted, judges (and lawyers) should try to make clear to the jury the extent to which the expert's conclusions depend on comparison of the items in question, and the extent to which they depend on assumptions or conclusions about other matters.

the expert to evaluate source probabilities, but it still raises many questions. In order to understand the expert's conclusions, the trier-of-fact will need to know how the expert evaluated the relevant probabilities, and how, where, and why the expert set the threshold for reporting a particular decision. The trier-of-fact also will need information about the accuracy of decisions reached in this manner.

In the past, expert forensic science testimony about "identification" and "exclusion" often went unchallenged, with lawyers on both sides assuming such testimony was reliable and uncontroversial. As lawyers become more aware of the issues discussed in this article, we expect they will examine the logic and basis of such conclusions far more closely than they have in the past.

THE GROWING IMPORTANCE OF STATISTICAL DATA ON ERROR RATES

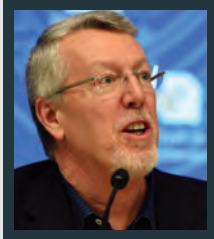
Regardless of how forensic scientists choose to present their conclusions, we also expect in the near future to see more testimony about the error rates of pattern-matching disciplines. The 2016 PCAST report argued forcefully that empirical research is the only way to assess the accuracy (and hence the probative value) of examiners' source conclusions:

> *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.*[28]

PCAST called for a continuing program of research in which examiners are tested by having them compare samples from known sources. PCAST recommended that the samples used in the research be representative of the samples encountered in casework, that examiners have no information about the correct answer, that independent groups with no stake in the outcome conduct multiple studies, and that the data be available to other scientists for review.[29] Courts will need to consider the results of such studies when deciding whether testimony about forensic comparisons is sufficiently trustworthy to be admitted — whether, in the words of Rule 702(c) of the Federal Rules of Evidence, it is "the product of reliable principles and methods."[30]

When such testimony is admitted, error-rate data will be relevant for assessing its probative value. PCAST suggested that testimony about error rates of the relevant forensic method, as research has shown on samples like those in the case at hand, should always be presented in conjunction with testimony about the results of forensic comparisons. Experts are likely to be asked about error rates during cross-examination even if the proponent of the forensic evidence elects not to present error-rate data in direct testimony. Lawyers are likely to debate the implications and significance of error-rate data for evaluating the probability that an error occurred in the case at hand.

ALEX BIEDERMANN is associate professor at the University of Lausanne (Switzerland) in the Faculty of Law, Criminal Justice and Public Administration. Previously, he worked as a scientific advisor in the Federal Department of Justice and Police in Berne in cases investigated by the Office of the Attorney General of Switzerland.



WILLIAM C. THOMPSON is professor emeritus of criminology, law and society at the University of California, Irvine. He studies human judgment and decision making with a particular focus on cognitive and contextual bias in scientific assessments, the use of scientific and statistical evidence in the courtroom, and lay perceptions of scientific and statistical evidence.

JOËLLE VUILLE is a senior researcher at the University of Lausanne School of Criminal Justice. She has written extensively on the use of scientific evidence in criminal courts and the use of probabilities in the judicial context.

FRANCO TARONI is professor of forensic statistics at the University of Lausanne, also in the Faculty of Law, Criminal Justice and Public Administration. He has authored and coauthored several books on the use of probabilities in the judicial context and is editor of the journal *Law, Probability and Risk* (Oxford University Press).

Vuille and Biedermann gratefully acknowledge the support of the Swiss National Science Foundation through grants no. PZ00P1_154955 / 1 and BSSGI0_155809. Thompson gratefully acknowledges support from the Center for Statistics and Applications in Forensic Evidence (CSAFE).

We are thus on the cusp of a new era for forensic science — an era in which statistics will inevitably play a greater role. Oliver Wendell Holmes once declared that "the man of the future is the man of statistics . . . ."[31] In the pattern-matching disciplines of forensic science, that future has arrived.

---

[1] Michael J. Saks & Jonathan Koehler, *The Coming Paradigm Shift in Forensic Identification Science*, 309 Science 892 (2005), at 892.

[2] Heidi Eldridge, *The Shifting Landscape of Latent Print Testimony: An American Perspective*, 3 J. of Forensic Sci. & Med. 72 (2017), at 72.

[3] *See*, Nat'l Acad. of Sci., Nat'l Research Council, Strengthening Forensic Science in the United States: A Path Forward (2009) [hereinafter NAS Report] at 44, 108, 162, 169, 176; President's Council of Advisors on Sci. & Tech., Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods (2016) [hereinafter: PCAST Report], at 19, 30, 54.

[4] Simon Cole, *Forensics Without Uniqueness, Conclusions Without Individualization: The New Epistemology of Forensic Identification*, 8 Law Probability & Risk 233 (2009), at 236-237.

[5] Eldridge, *supra* note 2, at 76.

[6] NAS Report, *Recommendation 3(b)*, *supra* note 3, at 23.

[7] *Id.*

[8] *Id.*, at 108.

[9] PCAST Report, *supra* note 3.

[10] *Id.*, at 9–10.

[11] *See e.g.,* Igor Pacheco, et al., Miami-Dade research study for the reliability of the ACE-V process: Accuracy & precision in latent fingerprint examinations (2014), at 53-55.

[12] *See e.g.*, Sci. Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST), Document # 4: Guideline for the Articulation of the Decision-Making Process for the Individualization in Friction Ridge Examination (Latent/Tenprint) r. 11.2.3 (2013), at 11.2.3.

[13] Colin G.G. Aitken & Franco Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists (2d ed. 2004), at 95 (providing a simple mathematical description of the likelihood ratio that lawyers and judges may encounter when reviewing forensic evidence. Let E represent the observed features of two traces that a forensic scientist is asked to compare; let H_s represent the proposition (hypothesis) that the items have the same source and H_d the proposition that they have a different source. The likelihood ratio is then $p(E|H_s)/p(E|H_d)$, which is read as "the probability of E given H_s over the probability of E given H_d").

[14] European Network of Forensic Sci. Insts., Guideline for Evaluative Reporting in Forensic Science (2015), at 2.4; *see also*, Royal Statistical Soc'y, http://www.rss.org.uk/practitioner-guides (last visited Jan. 7, 2018) (providing reports on this issue).

[15] Alex Biedermann, et al., *Development of European Standards for Evaluative Reporting in Forensic Science: The Gap Between Intentions and Perceptions*, 21 The Int'l J. of Evidence & Proof 14 (2017), at 26.

[16] *See,* John Butler, Fundamentals of Forensic DNA Typing (2009); Geoffrey S. Morrison & William C. Thompson, *Assessing the Admissibility of a New Generation of Forensic Voice Comparison Testimony,* 18 Colum. Sci. & Tech. L. Rev. 326 (2017).

[17] Def. Forensic Sci. Ctr., Dep't of the Army, Information Paper: Modification of Latent Print Technical Reports to Include Statistical Calculations (2017), at 2.

[18] *Id., at 2.*

[19] Alex Biedermann, et al., *How to Assign a Likelihood Ratio in a Footwear Mark Case: An Analysis and Discussion in the Light of R v T*, 11 Law, Probability & Risk 259 (2012), at 265-270.

[20] D. Michael Risinger, *Reservations About Likelihood Ratios (and Some Other Aspects of Forensic 'Bayesianism')*, 12 Law, Probability & Risk 63, 72 (2012).

[21] Charles E. H. Berger, et al., *Evidence Evaluation: A Response to the Court of Appeal Judgment in R v T*, 51 Sci. & Just. 43 (2011), at 43-44.

[22] Marjan Sjerps & Charles E. Berger, *How Clear is Transparent? Reporting Expert Reasoning in Legal Cases*, 11 Law, Probability & Risk 317 (2012).

[23] *Id.*; Biedermann, *supra* note 19, at 259; William C. Thompson, *Discussion Paper: Hard Cases Make Bad Law – Reactions to R v T*, 11 Law, Probability & Risk 347 (2012), at 351-353.

[24] Expert Working Grp. on Human Factors in Latent Print Analysis, Latent Print Examination and Human Factors: Improving the Practice Through a Systems Approach (2012), at 134.

[25] Ass'n of Forensic Sci. Providers, *Standards for the Formulation of Evaluative Forensic Science Expert Opinion*, 49 Sci. & Just. 161 (2009), at 163.

[26] Raymond Marquis et al., *Discussion on How to Implement a Verbal Scale in a Forensic Laboratory: Benefits, Pitfalls and Suggestions to Avoid Misunderstandings*, 56 Sci. & Just. 364 (2016).

[27] *See*, Bernard Robertson, et al., Interpreting Evidence – Evaluating Forensic Science in the Courtroom (2d ed. 2016), at 16-18; Ian W. Evett, *Towards a Uniform Framework for Reporting Opinions in Forensic Science Casework*, 38 Sci. & Just. 198 (1998), at 200-201 (explaining that after comparing two items, a forensic examiner may be able to judge the probability of the observed results under the alternative hypotheses: $p(E|H_s)$ and $p(E|H_d)$. But these probabilities are not the same as source probabilities; source probabilities are the inverse of these conditionals — i.e., $p(H_s|E)$ and $p(H_d|E)$. To infer source probabilities from the probability of the observed evidence, the examiner must take into account the prior probability that the items have the same source, $p(H_s)$, or different source, $p(H_d)$.).

[28] PCAST Report, *supra* note 3, at 46.

[29] *Id.*, at 66.

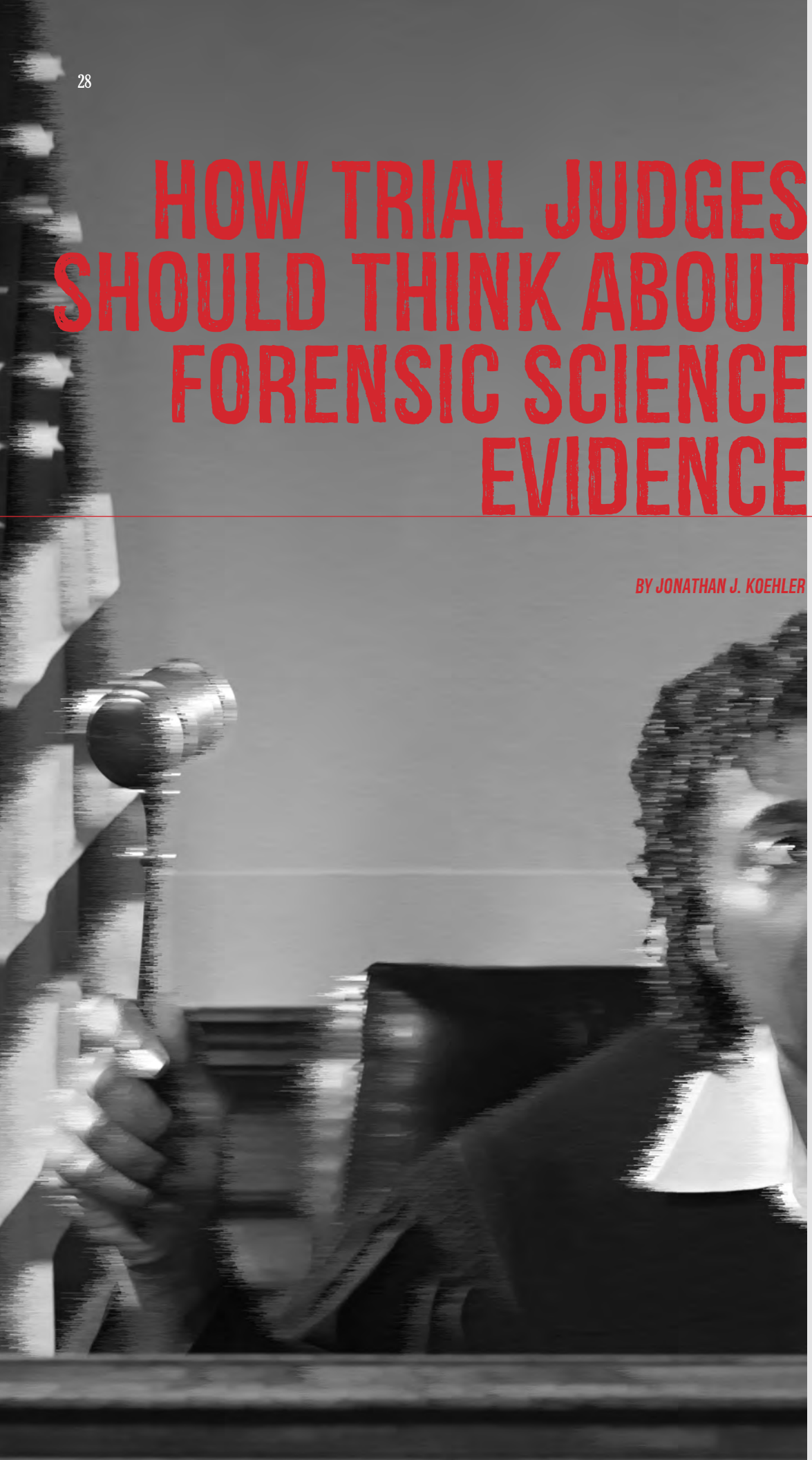
[30] Fed. R. Evid. 702(c).

[31] O.W. Holmes, *The Path of the Law, 8* Harv. L. Rev. 457, 469 (1897).

[32] *E.g.*, David J. Balding & Christopher D. Steele, Weight-of-Evidence for Forensic DNA Profile (2015).

HOW TRIAL JUDGES SHOULD THINK ABOUT FORENSIC SCIENCE EVIDENCE

BY JONATHAN J. KOEHLER



HERE IS A FORENSIC-SCIENCE TEST FOR YOU. PLEASE ANSWER EACH OF THE THREE QUESTIONS BELOW *TRUE* OR *FALSE*.

1. SCIENTIFIC TESTS CONDUCTED OVER THE PAST 100 YEARS HAVE REPEATEDLY DEMONSTRATED THAT EVERY-ONE HAS A UNIQUE SET OF FINGERPRINTS.

2. RECENT SCIENTIFIC STUD-IES SHOW THAT THE CHANCE THAT DNA SAMPLES FROM TWO DIFFERENT PEOPLE WILL BE IDENTIFIED AS A "MATCH" BY A COMPETENT, WELL-TRAINED DNA EXAMINER IS LESS THAN ONE IN A MILLION.

3. DATA FROM SCIENTIFIC TESTS CONDUCTED OVER THE PAST FEW DECADES PROVIDE A RELIABLE BASIS FROM WHICH TO ESTIMATE THE ACCURACY OF MOST FORENSIC METHODS THAT HAVE BEEN ADMITTED IN U.S. COURTS.

The answer to all three of these questions is *false*. How did you do?

If you read the 2009 report from the National Academy of Sciences on forensic science[1] or the 2016 report from the President's Council of Advisors on Science and Technology (PCAST),[2] you probably got all of the questions right. These authoritative reports, investigated and written by leading scientists in the U.S., indicate that our forensic sciences are badly in need of scientific testing. They also indicate that many of the strong claims made by forensic scientists and their proponents are misleading in light of the lack of scientific data to back up those claims.

But who really reads such reports? And who really understands that there is not enough science to justify a lot of forensic science claims? Certainly not the American public. I presented those three statements to a random sample of 322 jury-eligible Americans and found that each statement was judged to be true by more than four out of five people, and nearly two out of three people (64 percent) thought that *all three* statements were true.[3] Another recent study asked people to estimate the chance that a forensic examiner will err in each of five different forensic sciences.[4] The median estimates ranged from one chance in 100,000 (for document examination) to one in 10,000,000 (for DNA). Apparently, then, most people believe that forensic science results and conclusions are extremely accurate and that reliable scientific studies back up those beliefs (see question no. 3 at left).

Whether trial judges know differently is an open empirical question. But regardless of what trial judges know (or think they know) about the forensic sciences, they should look to the broader scientific community for assistance when evaluating the reliability of any proffered forensic method, including methods that have long played an important role in our criminal justice system. If they do so, they will likely find that the (disinterested) scientific community will provide a very different perspective on the extent to which forensic science claims have stood up to empirical testing than the perspective provided by the interested examiners who provide forensic science testimony at trial.

ROADMAP TO RELIABILITY: *DAUBERT* & FRE 702

In its broadest sense, forensic science is the application of science to law. Over the past century or so, many different types of forensic science results have been admitted in U.S. courts, including evidence from fingerprints, palm prints, voice prints, DNA, microscopic hair, ballistics, toolmarks, document examination, shoe prints, tire tracks, bitemarks, soil, glass, paint chips, carpet fibers, blood spatter, and more. In the near future, prosecutors may seek to introduce biometric techniques including evidence from gaits, veins, irises, retinas, etc.

Federal Rule of Evidence (FRE) 702 (or its state equivalent) governs the admissibility of expert testimony, including testimony pertaining to forensic analyses. The first part of FRE 702 essentially requires that an expert witness be qualified and provide testimony that will assist the trier of fact. The latter part of FRE 702, adopted in a 2000 amendment, provides additional restrictions on the admissibility of expert testimony. FRE 702(b) requires that "the testimony is based on sufficient facts or data." FRE 702(c) requires that "the testimony is the product of reliable principles and methods." FRE 702(d) requires that "the expert has reliably applied the principles and methods to the facts of the case." The year 2000 amendment to FRE 702 was offered in response to *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,[5] a case that conferred a gatekeeper function to trial judges who are asked to admit expert testimony of any sort.

Like the *Daubert* opinion itself, the 2000 amendment to FRE 702 emphasizes the central role that "reliability" must play in admissibility decisions related to expert testimony. The principles that underlie an expert's testimony must be reliable, the method an expert uses must be reliable, the application of those reliable principles and methods to the instant case must be reliable, and the testimony must be based on facts and data which, presumably, must also be reliable. In short, unlike most other forms of evidence, expert testimony — including forensic science expert testimony — is inadmissible unless the evidentiary proponent can affirmatively show that it is reliable in a variety of ways.

The emphasis FRE 702 places on reliability begs the questions of what it means for expert testimony to be reliable and how a judge can determine whether or not proffered testimony is reliable. The *Daubert* decision provided some useful guidance for assessing the reliability of proffered *scientific* expert testimony in general, and this guidance is applicable to proffered *forensic science* testimony. Specifically, *Daubert* and the advisory notes that accompany FRE 702 indicate that a trial judge may consider (1) whether the expert's theory or method has been tested, (2) whether the theory or method has been subject to peer review and publication, (3) the method's error rate, (4) whether the method is a standard one with controls, and (5) whether the theory or method has been generally accepted in the scientific community.

Daubert's guidance for assessing reliability is sensible, but vague. How exactly is a trial judge to know whether a forensic theory is sound, an error rate is comfortably low, or a laboratory's controls can be trusted? Although these questions must be answered in a legal ▸

arena by a trial judge, they concern matters that are fundamentally scientific in nature. As such, *judges should look to the broader scientific community for guidance when deciding whether proffered scientific evidence is sufficiently reliable to justify its admissibility at trial*.[6] That is, trial judges should lean heavily on the broader scientific community for methodological advice about how to determine whether a scientific technique or claim is sufficiently backed up by reliable principles, methods, facts, and data. In cases involving forensic science methods and claims, such guidance is readily available from a 2009 report by the National Academy of Sciences (NAS) and a 2016 report from the President's Council on Science and Technology (PCAST).[7]

2009 NATIONAL ACADEMY OF SCIENCES (NAS) REPORT

The 2009 NAS report on the non-DNA forensic sciences sent shock waves through the criminal justice system. This report, which was written by a group of the nation's elite scientists, statisticians, judges, and other scholars, concluded that, "[l]ittle rigorous systematic research has been done to validate the basic premises and techniques" in many forensic disciplines. The report detailed how many forensic sciences — including impression evidence, toolmark and firearms analysis, microscopic hair evidence, questioned document examination, and forensic odontology — "have yet to establish either the validity of their approach or the accuracy of their conclusions"[8] and called for a "major overhaul" of the U.S. forensic science system.[9] The report repeatedly stated that there is little scientific data to indicate the reliability and accuracy of the methods used in many forensic sciences. For example, the report noted that the standard fingerprint method (ACE-V) does not guard against bias and provides insufficient guarantees that examiners will obtain the same results and draw the same conclusions.[10] The report also noted that there is no standard vocabulary to describe results,[11] which may lead to "imprecise or exaggerated expert testimony."[12] Notably, the NAS report took the courts to task for being "utterly ineffective" at pushing any of the forensic sciences to test their claims and to otherwise conduct themselves in a more scientific manner.[13]

2016 PRESIDENT'S COUNCIL ON SCIENCE AND TECHNOLOGY (PCAST) REPORT

The 2016 PCAST report picks up where the 2009 NAS report left off by providing trial judges with specific guidance for assessing the scientific reliability and validity of proffered forensic science evidence. 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."[14] In 2015, President Obama asked PCAST to provide advice and recommendations "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."[15] The focus of this report was the "forensic 'feature-comparison' methods,"[16] which include DNA, hair, fingerprints, firearms, toolmarks, bitemarks, shoe prints, tire tracks, and handwriting. The elite scientists who wrote the report — none of whom served on the 2009 NAS committee described above — indicated that their focus was on helping judges understand scientific standards for assessing scientific validity, not on dictating the legal standards pertaining to the admissibility of scientific evidence.[17] The distinction is subtle but important.

As noted earlier, in cases involving forensic science evidence (or any other form of expert evidence), FRE 702 tells judges that the principles and methods that were used to create the evidence must be reliable, and that the expert testimony related to the evidence must be backed up sufficiently by reliable facts or data. FRE 104(a) gives judges the authority to rule on the admissibility of this evidence, and judges make this determination based on a preponderance of the available evidence. The PCAST report leaves these legal standards alone, and instead weighs in on how a judge (or anyone else) can determine whether a principle, method, or purported fact is scientifically reliable and valid. In doing so, the PCAST report fills a void left by both *Daubert* and the 2009 NAS report. It offers clear guidance to courts from esteemed representatives of the scientific community. As part of this guidance, the report distinguishes "foundational validity" from "validity as applied" in practice.[18] A method is foundationally valid if and only if it has been "shown, based on empirical studies, to be repeatable, reproducible, and accurate . . . under conditions appropriate to its intended use."[19]

These words should not be construed as highfalutin', overly cautious, scientific mumbo jumbo. The foundational validity standard applies to all sciences, and it is especially important that it be understood and applied in cases involving forensic science evidence where match determinations[20] are typically subjective judgments made by individual examiners. In referring to match determinations as "subjective judgments," I do not mean to imply that there is no basis for those judgments or that the judgments are as likely to be right as wrong. I simply mean that a person, as opposed to a machine or computer program, makes one or more key determinations — such as which portion of a marking to exam-

ine, whether an element of that marking is genuine or artefactual, or whether there is enough correspondence between two markings — to conclude that they were produced by a common source.[21]

The PCAST report does not mince words when it comes to the importance of testing forensic claims and methods. Such tests, PCAST says, are "an absolute requirement"[22] for any method that purports to be scientifically reliable and valid: "[T]he *only* way to establish the scientific validity and degree of reliability of a subjective forensic feature-comparison method — that is, one involving significant human judgment — is to test it *empirically* . . . "[23] In other words, data from appropriately designed studies are *required* as part of any demonstration of foundational validity for all scientific methods, including all forensic science methods. Forensic sciences that fall short — even forensic sciences that the judiciary has long presumed to be methodologically sound — simply cannot be treated as foundationally valid.

For example, PCAST found that firearms analysis falls short on foundational validity because there is currently just one appropriately designed study that measures validity and reliability. As PCAST notes, "the scientific criteria for foundational validity requires more than one such study to demonstrate reproducibility."[24] The error rate in this study, which PCAST argues should be reported to jurors, was estimated at 1 in 66, with an upper bound of 1 in 46. Although firearms analyses are routinely admitted by courts, it is doubtful that any court has provided jurors with these error rates from this lone study. Another subjective feature-comparison method that PCAST carefully examined is bitemark analysis. PCAST found that "[f]ew empirical studies have been undertaken to study the ability of examiners to accurately identify the source of a bitemark."[25]

THE FACT THAT MANY COURTS HAVE FOUND FORENSIC TECHNIQUES AND CLAIMS TO BE SCIENTIFICALLY VALID AND ADMISSIBLE WITHOUT SUCH DATA INDICATES THAT MANY COURTS HAVE MISUNDERSTOOD SCIENTIFIC VALIDITY OR CONDUCTED INADEQUATE *DAUBERT* ANALYSES.

PCAST further notes that, "[a]mong those studies that have been undertaken, the observed false positives rates were so high that the method is clearly scientifically unreliable at present."[26] Despite the lack of science in support of bitemark evidence, trial courts routinely admit this evidence, commonly on grounds that other courts have admitted this type of evidence in the past. Perhaps this is what the 2009 NAS report had in mind when it spoke of the "utter ineffectiveness" of the judiciary to apply the appropriate admissibility criteria to proffered forensic science evidence.

The PCAST report evaluated the empirical evidence that supported various other feature-matching methods and found that there were no studies that supported the scientific validity and reliability of footwear analysis or microscopic hair comparison evidence, and just one study that supported firearms analysis.

In addition to the specific conclusions reached for various forensic feature comparison methods, the larger takeaway point from the PCAST report for judges is that investigations into the validity of forensic techniques turn *exclusively* on the availability and results of properly performed empirical studies. The fact that many courts have found forensic techniques and claims to be scientifically valid and admissible without such data indicates that many courts have misunderstood scientific validity or conducted inadequate *Daubert* analyses. Such rulings — which commonly are bolstered by reference to the fact that forensic methods have been admitted in courts for decades — should not be binding on other courts or even offered as evidence to support the admissibility of a method. As the 2009 NAS committee co-chair Judge Harry Edwards observed in a *Frontline* documentary on forensic science, "If your experience or practice has been inaccurate or wrong for many years, it doesn't become better because it's many years. It's just many years of doing it incorrectly."[27]

PCAST CRITICISMS

The sentiments expressed in the NAS and PCAST reports are not new. Academic critics of forensic science offered similar points decades earlier.[28] However, these criticisms were largely ignored by the forensic science community, perhaps because they were largely ignored by the courts. After all, as long as trial judges continued to admit forensic science evidence, and appellate courts upheld those admissions, the forensic science community had little reason to engage the critics.

However, the world has finally taken note of the serious problems afflicting the forensic sciences. High-profile errors have been made, frauds have been detected, incompetent crime labs have ▸

been exposed, forensic techniques have been abandoned, and wrongful convictions linked to forensic missteps have been reversed. In 2013, the National Commission on Forensic Science (NCFS) was established as an advisory committee for the Department of Justice to improve the reliability of the forensic sciences. An ambitious reform agenda was identified and hundreds of scientists and scholars went to work on it. Progress was slow but steady. Although the forensic science and criminal justice communities were generally pleased by the prospect of improvements and the new resources this endeavor promised, they pushed back against proposed reforms that implied or directly claimed that the foundational scientific work had yet to be done in the forensic sciences. Acknowledging such a shortcoming could risk the status of forensic evidence in court.[29]

The forensic science and criminal justice communities generally oppose the harsh conclusions that appear in the NAS and PCAST reports pertaining to the scientific validity of the various forensic sciences. Ultimately, the disagreements these communities have with the broader scientific community must be resolved by the courts. The paragraphs below outline the criticisms that various professional groups have leveled against the PCAST report in particular.[30]

NATIONAL DISTRICT ATTORNEYS ASSOCIATION

The National District Attorneys Association (NDAA) claims that the PCAST report is "scientifically irresponsible." In support, the NDAA says that PCAST "clearly and obviously . . . ignored vast bodies of research, validation studies, and scientific literature," and instead relied, "at times, on unreliable and discredited research." The NDAA also says that PCAST has "insert[ed] itself as the final arbiter of the reliability and admissibility" of forensic science evidence, and that the NDAA will defend our criminal justice system against the NAS, PCAST, and others "who would seek to undermine the role of the courts, prosecutors, defense attorneys, and juries, as we have seen in the last eight years."

THE WAY TO KNOW IF SOMETHING WORKS AS ADVERTISED IS TO SUBJECT IT TO RIGOROUS AND REPEATED EMPIRICAL TESTING UNDER CONDITIONS THAT ARE SIMILAR TO THOSE IN THE NATURAL ENVIRONMENT. THIS HAS NOT BEEN DONE FOR MOST OF THE FORENSIC SCIENCES.

In short, the NDAA suggests that the NAS and PCAST commissions chose to ignore excellent validation studies in favor of discredited research as part of a plan to remove decision-making authority from judges and juries and to otherwise undermine our criminal justice system. In a follow-up letter to President Obama, the president of the NDAA offered an additional, even more startling, argument for disregarding the PCAST report.[31] He claimed that the feature comparison methods that the PCAST report covered (e.g., toolmarks, ballistics, fingerprints, microscopic hair comparison, odontology, document examination, and tread wear) should not be held to the standards for scientific validity because the methods are not entirely scientific. The letter explains that, although the forensic sciences "incorporate aspects of science," forensic science methods also yield "'technical' and 'specialized knowledge' under Federal Rule of Evidence 702," and therefore need not be held to *Daubert's* rigorous scientific validity standard.[32]

FBI

The FBI claims that the PCAST report makes "broad, unsupported assertions regarding science and forensic science practice" and "creates its own criteria for scientific validity."[33] In support of the first claim, the FBI disagrees with PCAST's statement that proficiency tests that measure an examiner's accuracy are the only way to establish the validity of a forensic technique. Like the NDAA, the FBI claims that PCAST ignored "numerous published research studies" that establish the foundational validity of various forensic sciences, an omission that the FBI says "discredits the PCAST report as a thorough evaluation of scientific validity."

The FBI suggests that PCAST not only offered an idiosyncratic set of criteria for establishing scientific validity, but failed to consider studies that established the validity of various forensic methods using its own criteria.

THE AMERICAN CONGRESS OF FORENSIC SCIENCE LABORATORIES

The American Congress of Forensic Science Laboratories (ACFSL) characterized the PCAST report as a political document rather than a scientific one. The ACFSL criticized the PCAST membership as "imbalanced and inexperienced" and indicated that "the legitimacy of the PCAST report" is compromised by the members' motives and biases.[34] Like the NDAA and FBI, the ACFSL characterized the PCAST report as "irresponsible and inaccurate" because it "failed to objectively and completely evaluate the

overwhelming evidence of strength and reliability in forensic science."[35]

AMERICAN SOCIETY OF CRIME LABORATORY DIRECTORS

The American Society of Crime Laboratory Directors (ASCLD) challenged PCAST's definition of a scientifically rigorous "black box" validation study as "arbitrary" and "unhelpful."[36] ASCLD also argued that forensic science practitioners should have a hand in the design and conduct of the scientific studies to foster "true advancement . . . of forensic science."[37]

MIDWESTERN ASSOCIATION OF FORENSIC SCIENTISTS

The Midwestern Association of Forensic Scientists (MAFS) characterized PCAST's conclusions as "capricious."[38] Like ASCLD, MAFS suggested that the empirical testing methods that PCAST outlined "are not the only scientific way to ensure validity and reliability."[39] Also like ASCLD, MAFS indicated that the "[e]xperience and daily observation" of examiners is part of a scientific measure of reliability.[40] They wrote, "[t]o not include practitioners in the discussion would be irresponsible."[41]

OTHER FORENSIC ORGANIZATIONS

Many of the arguments raised above were echoed in responses from others in the forensic science community. The Association of Firearm and Tool Mark Examiners asserted that "[d]ecades of validation and proficiency studies have demonstrated that firearm and toolmark identification is scientifically valid."[42] The Organization of Scientific Area Committees Materials Subcommittee stated that lack of information about an error rate for microscopic hair comparison evidence "should not be interpreted to suggest that the discipline is any less scientific."[43] The International Association for Identification "finds the report lacking in basis and content, and improper in some of the statements that are made."[44] The Bureau of Alcohol, Tobacco, Firearms and Explosives expressed its "disappointment in the flawed methodology PCAST employed," saying that PCAST "did not adequately consider the numerous research studies that support the validity of firearm and tool mark forensics."[45]

CRITIQUING THE PCAST CRITICISMS

The sheer volume of professional and government organizations and representatives that have taken issue with the PCAST findings is superficially impressive. But, of course, it is the logical and scientific merit of those responses that must be critiqued, not their volume. I have identified six distinct points raised by the various PCAST critics. I comment on the merits of each of those points below.

1) *The PCAST committee was biased against forensic science:* It should go without saying that ad hominem attacks on a properly convened scientific committee are inappropriate and unpersuasive. The PCAST committee, like the NAS committee before it, included some of the most accomplished scientists of our era. Few of the committee members are primarily focused on forensic science issues outside of their committee work, and there is nothing in the backgrounds of the committee members as a whole that supports a charge of bias. The PCAST committee chair, Eric Lander, co-authored a frequently cited paper in the prestigious journal *Nature* two decades ago that concluded as follows: "[T]he DNA fingerprinting controversy has been resolved. There is no scientific reason to doubt the accuracy of forensic DNA typing results . . . ." Although this conclusion may have been premature, these words do not seem to be those of a committed forensic-science foe.

2) *PCAST offered an overly narrow and idiosyncratic definition of scientific validity:* This effort by critics of the PCAST Report to broaden the scope of what constitutes scientific validity must be rejected. Part of what makes the PCAST report so helpful to courts is that it provides clear, sound, and practical guidance about exactly what judges should look for when considering the scientific matter of the foundational validity of a method that involves substantial human judgment. In a nutshell, PCAST reminds the world about the wisdom of what we learned in our high school science classes: *the way to know if something works as advertised is to subject it to rigorous and repeated empirical testing under conditions that are similar to those in the natural environment.* This has not been done for most of the forensic sciences. When PCAST critics suggest that the daily "experience" of forensic examiners vouches for the scientific validity of their work, it is important to remind ourselves that this is not how science works. Personal experience is no substitute for empirical testing. This doesn't mean that experience is worthless. If consumer reviews on Amazon indicate that a weight loss pill worked wonders for some people, a potential customer has some justification for expecting the pill to help him or her lose weight. But these reviews, which spring from the personal experiences of consumers, do not constitute scientific proof that the pill actually works. The scientific validity of a claim that a pill causes weight loss — or that a forensic method yields true results — can only be proven using justified, widely agreed upon scientific methods and standards.

3) *PCAST ignored strong evidence that proves the scientific validity of various forensic sciences:* In response to this potentially devastating charge, PCAST invited the ▸

FBI and other agencies who made this claim "to identify any 'published . . . appropriately designed studies' that had not been considered by PCAST and that established the validity and reliability of any of the forensic feature-comparison methods that the PCAST report found to lack such support." No such studies were provided. Indeed, the FBI ultimately conceded that there were no such studies after all.[46] Nevertheless, forensic scientists commonly offer sworn testimony that relevant validation studies exist and that they personally believe that the method in question is reliable. Needless to say, such testimony does not suffice as proof of scientific validity under the standards imposed by *Daubert* and FRE 702.

4) *PCAST usurped the role of judges and juries by inserting its own opinions about forensic science:* As noted previously, the PCAST report was quite clear about differentiating between scientific matters pertaining to forensic science that were clearly within its charge, and legal matters that did not concern either PCAST or the broader scientific community.[47] Whereas legal policymakers, judges, and juries must decide matters such as general admissibility standards for scientific evidence and whether a proffered method has met those legal standards, scientists are best positioned to advise on the scientific standards for scientific validity.

5) *Forensic science evidence should not be held to scientific standards of validity because the evidence includes technical or specialized knowledge:* Ordinarily, forensic science supporters are keen on promoting their disciplines as scientific. It is therefore puzzling to see the NDAA argue that their evidence should be assessed using standards that are more lenient than those used for other sciences. Whether this maneuver is regarded as clever or desperate, it should fail. As the Supreme Court noted in *Kumho Tire v. Carmichael* (1999),[48] the gatekeeping function for trial courts identified in *Daubert* extends to *all* expert testimony offered under FRE 702. This means that expert testimony, whether scientific or not, must still be reliable to be admissible. Although trial judges have latitude when assessing the reliability of expert testimony, a lower reliability standard is not automatically in play once the evidentiary proponent declares a willingness to have its evidence reviewed as non-science for admissibility purposes. Regardless of whether forensic science is characterized as 100 percent science, part science and part technical knowledge, or 100 percent technical knowledge, the reliability and validity of the methods used by examiners to reach their subjective conclusions must be demonstrated affirmatively.

6) *Practitioners' personal experiences and observations should be given weight when assessing the scientific validity of forensic science:* This claim, which also lies behind the critics' claim in point 2 above, reveals how "motivated reasoning"[49] can distort the judgments of professionals. The experiences and casual observations of forensic scientists may aid future scientific study by, for example, identifying hypotheses, ideas, patterns, correlations, etc. *But experiences and unsystematic observations must not be confused with systematic scientific study.* Judges must firmly reject the notion that experience — even a great deal of it — contributes to the scientific validation of a method. People experience and observe many things that systematic study later proves to be spurious or false.[50] When assessing the scientific validity of a method involving human judgment, systematic, rigorous empirical testing — scientific testing — is not an option: It is a requirement. There are no shortcuts, and the day-to-day work experiences of examiners are not a legitimate substitute for empirical testing. To suggest otherwise blatantly distorts the shared understanding among scientists of what it means for a method to have been scientifically validated.

In sum, a critique of the criticisms leveled against the PCAST report supports the view that PCAST and the NAS have it right: An assessment of the reliability and validity of the forensic sciences requires testing, and many of those tests have yet to be performed. As a result, we know surprisingly little about how accurate forensic science testimony is.

ERROR RATES: WHAT DO WE KNOW?

If the legal standard for admitting forensic science evidence is followed, then the evidentiary proponent must show that the methods that produced the forensic result are themselves reliable. The most important indicator of the reliability of a forensic method is the rate at which trained examiners who use that method err: the lower the error rate, the greater the reliability of the method. Of course, in an actual case in which an unknown print or marking is compared to one or more knowns, ground truth is absent. In such cases, we cannot be sure whether a correct result is achieved because there is no independent way to verify the accuracy of the examiner's conclusion. But in a properly designed test in which prints or markings are produced from recorded knowns, ground truth is available, and an examiner's error rate (or a laboratory's error rate or the error rate of a method in general) may be computed.

Unfortunately, and perhaps surprisingly, such tests are virtually nonexistent in the world of forensic science. This means that there is little basis for estimating error rates for any forensic science method. As a result, courts cannot make a properly informed judgment about the reliability of a proffered forensic method.

WHAT FORENSIC SCIENTISTS TELL JUDGES

Complicating admissibility matters for the courts, proponents of forensic science commonly tell trial judges that (a) the frequent admission of forensic science evidence by courts throughout the land is proof of the validity of their methods, and (b) examiners take various tests on a regular basis, and their success on these tests confirms the reliability of their methods.

These two points have correct premises, but false conclusions. Regarding the first point, it is true that nearly every forensic science method has been admitted by most courts for many years. But the use of forensic science evidence in court — including evidence from document examination, voice prints, bitemarks, fingerprints, bullet lead analysis, toolmarks, tire tracks, shoe prints, etc. — predates the more rigorous admissibility standard identified in *Daubert* and FRE 702. The old *Frye* standard,[51] which focused on general acceptance in the relevant scientific community, was replaced by *Daubert's* science-driven standard. The fact that forensic evidence admitted under the *Frye* standard continued to be admitted by courts after the *Daubert* standard was introduced does not necessarily speak to the methodological soundness of the forensic evidence. This would only be true if the courts that admitted the forensic evidence in question properly applied the principles outlined in *Daubert*. However, as others have pointed out for years, courts have not done this.[52] Therefore, references to the prior admission of forensic methods by courts provide little or no evidence that those methods have been vetted by the *Daubert* or FRE 702 standards.

Regarding the second point, it is true that forensic examiners in many disciplines are routinely tested. And it is true that performance on these tests is often quite good in the sense that few examiners commit major errors or otherwise fail. But it is absolutely critical for trial judges to understand that *the tests that examiners take* — tests that are commonly labeled "proficiency tests" and provided to courts as proof of a method's (or an examiner's) low rate of error — *are not designed to measure either the accuracy of a method or the accuracy of an examiner who uses that method.* Instead, these tests are "designed primarily to meet laboratory accreditation demands, not to provide individual examiners with 'real world casework-like' samples."[53] In other words, *examiners' successful performance on existing proficiency tests tells us next to nothing about the rates at which forensic scientists offer erroneous conclusions in casework.* This much is readily conceded by the test manufacturers themselves. As one leading manufacturer cautions, "The design of an error rate study would differ considerably from the design of a proficiency test. Therefore, the results found in [our] Summary Reports should not be used to determine forensic science discipline error rates."[54]

UNDER FRE 706, A TRIAL JUDGE MAY APPOINT "ANY EXPERT . . . OF ITS OWN CHOOSING" TO ASSIST WITH MATTERS RELATED TO DETERMINING WHETHER A PARTICULAR METHOD IS RELIABLE AND VALID.

Unfortunately, courts have either ignored such disclaimers or been unaware of them. This is a serious problem. *The truth is that we know next to nothing about the error rates associated with our forensic scientists or our forensic science methods — including DNA methods. No one has done the requisite studies.*

But rather than taking my word for it, or the word of some interested party at trial, judges should do their own due diligence on these issues. When doing so, judges might consider enlisting disinterested scientists who have relevant methodological expertise. Under FRE 706, a trial judge may appoint "any expert . . . of its own choosing" to assist with matters related to determining whether a particular method is reliable and valid. Importantly, a forensic scientist would not qualify as a disinterested scientist with methodological expertise. Although a small proportion of forensic scientists do have the requisite methodological skills to serve in this role, forensic scientists should not be treated as representatives of the broader scientific community. Unlike members of the broader scientific community, forensic scientists have a powerful interest in persuading judges that their methods are reliable and valid. Just as a trial judge would not rely on a polygraph examiner's opinion about the reliability of his or her polygraph method, he or she should not rely on the opinions of a blood spatter expert, a bitemark expert, or even a DNA expert when assessing the reliability of the technique the expert uses. Verbal assurances by interested experts do not fulfill the reliability mandate outlined by *Daubert* and FRE 702. Likewise, a recitation of prior courts that have admitted similar testimony does not provide adequate proof of foundational validity. As stated in the PCAST report, empirical studies specifically designed to assess reliability, validity and error rate are not just a good idea, they are required.

CONCLUSION

It is undeniable that there are serious problems with the presentation of forensic science evidence in U.S. courtrooms. In 2015, a widely-publicized review of trial transcripts found that testimony provided by FBI hair examiners prior to 2000 contained significant errors and exaggerations in more than 95 percent of cases.[55] In 2004, a NAS report examined bullet lead evidence and concluded that, contrary to what forensic experts had said in 2,500 cases since the 1960s, there was no scientific basis to support a conclusion about whether a particular bullet came from a particular box of ammunition.[56] In 2016, a forensic science commission in Texas recommended suspending the use of bitemark evidence in criminal cases because, once again, there was no scientific evidence that proved forensic dentists can do what they say they can do.[57] Problems also have been identified in our most admired forensic sciences. In 2004, four of our nation's top fingerprint examiners erroneously and very publicly matched the fingerprint of an innocent U.S. citizen to a partial fingerprint recovered from the scene of a major terrorist attack in Madrid, Spain.[58] Studies since that time have shown that fingerprint examiners can be induced to reach conclusions about whether two prints match based on considerations that have nothing to do with the prints themselves.[59] Studies also have shown disagreement among DNA examiners about whether pairs of DNA samples match or not.[60]

The point is not that forensic science is all unreliable junk science. The point is that there are compelling reasons to be concerned, these reasons are not new, and the requisite scientific testing still has not been done. Consequently, no one knows how accurate any of the forensic science conclusions are.

THE POWER TO FIX FORENSIC SCIENCE EVIDENCE — TO SUBJECT THE CLAIMS TO EMPIRICAL TESTING, TO IDENTIFY THE RISK OF ERROR ASSOCIATED WITH THE VARIOUS METHODS, TO RESTRICT EXPERT TESTIMONY TO THAT WHICH IS SUFFICIENTLY SUPPORTED BY RELIABLE FACTS AND DATA — RESIDES WITH THE JUDICIARY.

Comprehensive studies by scientific bodies find that many forensic sciences have not been validated and have not provided scientific evidence that supports a claim of low rates of error. Crime laboratory scandals in which examiners commit a variety of errors — both intentional and unintentional — are everywhere, and the problems seem to be getting worse.[61] Yet neither trial courts nor appellate courts have done anything to improve the quality of forensic science evidence that appears in court.

The problem is not the legal standards pertaining to the admission of forensic science evidence as embodied in *Daubert* and FRE 702. The problem is with the failure by courts to take the mandates of *Daubert* and FRE 702 seriously. It should be obvious that evidence should not be judged reliable simply because the evidentiary proponent says so or because other courts that used lesser standards have said so. It should be obvious that there is no burden on forensic science opponents to prove that the proffered evidence is unreliable or that the underlying methods frequently fail. *Daubert* and FRE 702 create an affirmative burden on behalf of the evidentiary proponents to produce sufficient evidence of a method's reliability before the results that spring from that method may be presented to the trier of fact. The general scientific community, the 2009 NAS report, and the 2016 PCAST report, can provide helpful guideposts to trial judges for assessing scientific reliability. Where feasible, judges also should consider getting help from a neutral expert who has strong methodological and scientific skills.

The power to fix forensic science evidence — to subject the claims to empirical testing, to identify the risk of error associated with the various methods,[62] to restrict expert testimony to that which is sufficiently supported by reliable facts and data — resides with the judiciary. As Judge Nancy Gertner has concluded, "until courts address the deficiencies in the forensic sciences — until courts do what [*Daubert*] requires that they do — there will be no meaningful change here."[63]

JONATHAN J. KOEHLER is the Beatrice Kuhn Professor of Law at Northwestern Pritzker School of Law. He has a Ph.D. in Behavioral Sciences from the University of Chicago. His areas of interest include evidence, forensic science, judgment and decision making, and quantitative reasoning in the law.

[1] Comm. on Identifying the Needs of the Forensic Sci. Cmty., Nat'l Research Council, Strengthening Forensic Science in the United States: A Path Forward (2009) [hereinafter 2009 NAS Report]

[2] Exec. Office of the President, President's Council of Advisors on Sci. & Tech., Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods (September 2016) [hereinafter 2016 PCAST Report].

[3] The data were collected online via Amazon Mechanical Turk on October 7-8, 2017. Survey participants covered a broad cross-section of people in terms of age (median age range = 40-49; 22.8 percent younger than 30, 28.1 percent older than 50), educational level (10.8 percent high school graduate or less, 21.0 percent graduate degrees), ethnicity (82.6 percent Caucasian, 17.4 percent black, Hispanic, or other), and gender (47.3 percent women). Participants were paid $0.50 for their participation. The proportions of participants who indicated that they believed that each of the three statements in the text was true were 85.0 percent, 80.2 percent, and 88.0 percent respectively. These results were substantially similar regardless of whether the analyzed sample included all participants (n=322) or only those who were both jury-eligible and not flagged for possible inattention to detail (n=167).

[4] Jonathan J. Koehler, *Intuitive Error Rate Estimates for the Forensic Sciences*, 57 Jurimetrics J. 153 (2017).

[5] *Daubert v. Merrell Dow Pharm., Inc.*, 509 U.S. 579 (1993).

[6] *Id*. at 599 (Rehnquist, C. J., concurring in part and dissenting in part) (issues pertaining to "definitions of scientific knowledge, scientific method, scientific validity, and peer review . . . [are] matters far afield from the expertise of judges").

[7] One might suggest that these two reports do not necessarily represent the views of the scientific community writ large. But absent evidence that a substantial number of other leading scientists have investigated the matters discussed in these reports, and have reached different conclusions from those expressed in the reports, it is unpersuasive to contend that these reports may be dismissed as simply one view in a divided scientific community. To the contrary, the literature strongly indicates that scientists and scholars outside of forensic science who have investigated the validity matters described in the NAS and PCAST reports agree very strongly on central matters discussed therein.

[8] 2009 NAS Report, *supra* note 1, at 53.

[9] *Id.* at 285.

[10] *Id.* at 142.

[11] *Id.* at 185-186.

[12] *Id.* at 4.

[13] *Id.* at 53.

[14] 2016 PCAST Report, *supra* note 2, at iv.

[15] *Id. at* 1.

[16] *Id. at* 1.

[17] *Id. at* 4 ("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") (emphasis in original).

[18] *Id. at* 4.

[19] *Id. at* 4–5.

[20] Some forensic sciences expressly use the term "match" to describe observed correspondences between an unknown and known print or marking, but others use terms such as identification, individualization, consistent with, similar in all respects, cannot be excluded, etc.

[21] *See also* 2016 PCAST Report, *supra* note 2, at 5 n.3 ("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.").

[22] President's Council of Advisors on Sci. & Tech., An Addendum to the PCAST Report on Forensic Science in Criminal Courts 1 (2017) [hereinafter 2017 PCAST Addendum] (emphasis in original).

[23] *Id* at 2 (emphasis in original).

[24] 2016 PCAST Report, *supra* note 2, at 112.

[25] *Id.* at 87.

[26] *Id. at* 87.

[27] *Frontline:, The Real CSI* (PBS Apr. 17, 2012).

[28] Jennifer L. Mnookin, *Of Black Boxes, Instruments, and Experts: Testing the Validity of Forensic Science*, 344 Episteme 343, 349 (2008) ("What we ought to require as a precondition to admissibility is that the 'outputs' of fingerprint examiners – their ability to accurately identify whether fingerprints come from a common source – be tested for accuracy"); D. Michael Risinger, Mark P. Denbeaux & Michael J. Saks, *Exorcism of Ignorance as a Proxy for Rational Knowledge: The Lessons of Handwriting Identification "Expertise,"* 137 U. Pa. L. Rev. 731 (1989); Michael J. Saks & Jonathan J. Koehler, *What DNA "Fingerprinting" Can Teach the Law About the Rest of Forensic Science*, 13 Cardozo L. Rev. 361 (1991).

[29] In the spring of 2017, Attorney General Jeffrey Sessions put an end to the NCFS by declining to renew its charter. *See* Spencer S. Hsu, *Sessions Orders Justice Dept. to End Forensic Science Commission, Suspend Review Policy,* Wash. Post (April 10, 2017). It is not yet clear what, if anything, will replace NCFS.

[30] When the 2009 NAS Report appeared, this report was similarly criticized by forensic-science organizations. *See* William C. Thompson, *The National Research Council's Plan to Strengthen Forensic Science: Does the Path Forward Run Through the Courts?*, 50 Jurimetrics 35, 49–50 (2009).

[31] Michael A. Ramos, President, National District Attorneys Association, *Letter to President Obama* (Nov. 16, 2016).

[32] *Id. at* 2.

[33] Fed. Bureau of Investigation, Comments on President's Council of Advisors on Science and Technology Report to the President: Forensic Science in Federal Criminal Courts: Ensuring Scientific Validity of Pattern Comparison Methods 1 (Sept. 20, 2016).

[34] The Am. Cong. of Forensic Sci. Labs., Position Statement: The 2016 PCAST Report 1–2 (2016).

[35] *Id.* at 2.

[36] Am. Soc'y of Crime Lab. Dirs., Inc., Statement on Sept. 20, 2016 PCAST Report on Forensic Science at 1 (2016).

[37] *Id.* at 2.

[38] Midwestern Ass'n of Forensic Scientists, Response to PCAST Report 1 (2016).

[39] *Id.* at 1.

[40] *Id.* at 2.

[41] *Id.* at 2.

[42] Ass'n of Firearm and Tool Mark Exam'rs, Response to PCAST Report on Forensic Science 1 (2016).

[43] Org. of Sci. Area Comms. Materials Subcomm., Response to PCAST Call for Additional References from OSAC Materials Subcommittee 2 (n.d.).

[44] Int'l Ass'n for Identification, IAI Response to the Report to the President 'Forensic Science in Criminal Courts Ensuring Scientific Validity of Feature-Comparison Methods' Issued by the President's Council of Advisors on Science and Technology (PCAST) in September 2016 1 (n.d.).

[45] Bureau of Alcohol, Tobacco, Firearms & Explosives, ATF Response to the President's Council of Advisors on Science and Technology Report 1 (2016).

[46] 2017 PCAST Addendum, *supra* note 19, at 5.

[47] *See supra* n. 17. *See also* 2016 PCAST Report, *supra* note 2, at 21 n.11 ("In this report, PCAST addresses solely the *scientific* standards for scientific validity and reliability. We do not offer opinions concerning *legal* standards.")

[48] *Kumho Tire Co. v. Carmichael*, 526 U.S. 137 (1999).

[49] Ziva Kunda, *Motivated Inference: Self-serving Generation and Evaluation of Causal Theories*, 53 J. Pers. & Soc. Psych. 636 (1987).

[50] Examples include the Ptolemaic model of the solar system (which had the earth at the center), the flat earth theory, and even the recently discredited theory that ulcers are caused by stress (they are caused by bacteria).

[51] *Frye v. U.S.*, 293 F. 1013 (D.C. Cir. 1923).

[52] William Thompson, John Black, Anil Jain, & Joseph Kadane, Forensic Science Assessments: A Quality and Gap Analysis: Latent Fingerprint Examination 1 (2017) ("Serious questions have been raised, however, about how well judges have performed this [gatekeeping] role"); Simon A. Cole, *Toward Evidence-Based Evidence: Supporting Forensic Knowledge Claims in the Post-Daubert Era*, 43 Tulsa L. Rev. 263, 277; Peter J. Neufeld, *The (Near) Irrelevance of Daubert to Criminal Justice: And Some Suggestions for Reform*, 95 (Supp. 1) Amer. J. Pub. Health S107, S110 (2005).

[53] Collaborative Testing Services, *CTS Statement on the Use of Proficiency Testing Data for Error Rate Determinations* at 2 (Mar. 30, 2010). https://www.ctsforensics.com/assets/news/CTSErrorRateStatement.pdf. Collaborative Testing Services provides testing materials to laboratories across the forensic sciences. Occasionally, a study *is* designed to identify the rate at which examiners err. For example, a study by Ulery and his colleagues was designed to assess the rate at which latent fingerprint examiners commit different types of errors. *See* Bradford T. Ulery, R. Austin Hicklin, JoAnn Buscaglia, & Maria Antonia Roberts, *Accuracy and Reliability of Forensic Latent Fingerprint Decisions*, 108(19) Proc. Natl. Acad. Sci. 7733 (2011). However, given that the participants in the study were volunteers who knew they were being tested, that the study was paid for by an interested party (the FBI), and that some of the authors work for the FBI, the results of the study should be viewed with caution. *See* Jonathan J. Koehler, *Forensics or Fauxrensics? Ascertaining Accuracy in the Forensic Sciences* 49 Ariz St. L. J 36-38 (2018).

[54] Collaborative Testing Services, *supra* note 53. at 3.

[55] Spencer S. Hsu, *FBI Admits Flaws in Hair Analysis Over Decades*, Wash. Post, Apr. 18, 2015.

[56] Nat'l Acad. of Scis., Forensic Analysis: Weighing Bullet Lead Evidence (2004).

[57] Joe Palazzolo, *Texas Commission Recommends Ban on Bite-Mark Evidence*, Wall St. J., Feb. 12 2016.

[58] Office of the Inspector Gen., Oversight & Review Div., U.S. Department of Justice, A Review of the FBI's Handling of the Brandon Mayfield Case 1–4 (2006), https://oig.justice.gov/special/s0601/final.pdf.

[59] Itiel E. Dror, David Charlton, & Ailsa E. Peron, *Contextual Information Renders Experts Vulnerable to Making Erroneous Identifications*, 156 Forensic Sci. Int'l 74 (2006).

[60] Itiel E. Dror & Greg Hampikian, *Subjectivity and Bias in Forensic DNA Mixture Interpretation*, 51 Sci. & Justice 204 (2011); Jan W. de Keijser, Marijke Malsch, Egge T. Luining, et al., *Differential Reporting of Mixed DNA Profiles and its Impact on Jurists' Evaluation of Evidence*, 23 Forensic Sci. Int'l: Genetics 71 (2016); Sarah V. Stevenage & Alice Bennett, *A Biased Opinion: Demonstration of Cognitive Bias on a Fingerprint Matching Task Through Knowledge of DNA Test Results*, 276 Forensic Sci. Int'l 93 (2017).

[61] Dahlia Lithwick, *Crime Lab Scandals Just Keep Getting Worse*, SLATE (October 20, 2015).

[62] The issue of how and how well fact-finders will use error rate data if and when it is provided to them may be complicated. *See* Nicholas Scurich, *The Differential Effect of Numeracy and Anecdotes on the Perceived Fallibility of Forensic Science*, 22 Psychiatry, Psychology & Law 616 (2015), for a review of the empirical literature, and a finding that proficiency with numerical information (i.e., numeracy) impacts fact-finders' use of error rates.

[63] Nancy Gertner, *Commentary on the Need for a Research Culture in the Forensic Sciences*, 58 UCLA L. Rev. 789, 790 (2011).

UP TO THE COURTS: MANAGING FORENSIC TESTIMONY WITH LIMITED SCIENTIFIC VALIDITY

BY J. H. PATE SKENE

U.S. DISTRICT COURT JUDGE JED RAKOFF OF THE SOUTHERN DISTRICT OF NEW YORK TELLS THE STORY OF A FIREARMS AND TOOLMARK EXAMINER WHO APPEARED BEFORE HIM IN 2008, PROPOSING TO TESTIFY THAT THE MARKINGS ON SHELL CASINGS FOUND AT THE SCENE OF A CRIME MATCHED SHELL CASINGS FROM A GUN FOUND UNDER THE DEFENDANT'S BED "TO A REASONABLE SCIENTIFIC CERTAINTY." AS INSTRUCTED BY *DAUBERT V. MERRELL DOW PHARMACEUTICALS, INC.*, JUDGE RAKOFF INQUIRED ABOUT THE SCIENTIFIC BASIS FOR THE EXAMINER'S CLAIM:

> *I HELD A* DAUBERT *HEARING AND I ASKED HIM, FOR EXAMPLE, "WHAT'S YOUR ERROR RATE? AND WHAT'S THE ERROR RATE OF THIS METHODOLOGY THAT YOU'RE USING?"*
>
> *AND HE SAID, "ZERO."*

I said, "Zero?"
And he said, "Yes."
And I said, "How can it be zero?"
And he said, "Well, in every case I've testified, the guy's been convicted." [1]

Twenty-five years after *Daubert* made trial judges the gatekeepers of scientific evidence, leading scientists, scientific organizations, and the courts remain, in many cases, at loggerheads over standards for establishing the reliability of scientific evidence. Nowhere has this tension been more apparent than in the continuing debates over the scientific validity of long-accepted forms of forensic evidence in criminal law. From the landmark 2009 report by the National Research Council's National Academy of Sciences (NAS report)[2] to a 2016 report by the President's Council of Advisors on Science and Technology (PCAST report)[3] and a 2017 study from the American Association for the Advancement of Science (AAAS report),[4] multiple studies by leading scientists and scientific organizations continue to find that many of the most widely used forensic disciplines do not meet the standards of scientific validity that are routinely applied in scientific research. However, the field of applied forensic sciences often relies on practitioners' practical training, experience, and professional judgment. Many in the forensics community argue that the rigorous standards demanded by scientific research are neither realistic nor appropriate indicia of reliability for applied forensic sciences.[5]

Courts, for their part, have been highly reluctant to exclude forensic methods that have become integral to modern criminal investigations and prosecutions based solely on criticism by scientists outside the forensic community.[6] Different courts have cited a variety of reasons for admitting challenged forensic methods[7] consistent with the "broad flexibility" of trial courts in deciding how to assess the reliability of scientific evidence and the Supreme Court's recognition that different criteria may be appropriate for evaluating the reliability of different types of expertise.[8] Courts have been more willing to instead limit certain kinds of testimony by experts from fields with limited empirical evidence of validity; some judges, for example, have allowed latent fingerprint examiners or firearms experts to testify about the similarities between two sets of prints or shell casings but have excluded testimony about the likelihood of such similar samples arising from different sources.[9] Critics argue that this approach is ineffective and can mislead jurors. They point out, for example, that ordinary jurors generally lack the specialized experience to identify limitations in a scientific method or common sources of error in laboratory procedures. Further, in the absence of concrete information about uncertainty and the potential for error in an expert's methods, jurors tend to give excessive weight to "expert" conclusions.[10]

Further complicating admissibility decisions for trial courts, scientific validity is not a binary determination but an incremental process. Over time, many independent studies progressively define the validity of underlying principles and methods, as well as their limitations, error rates, and other variables. Empirical studies early in this process may provide meaningful evidence of validity but leave important issues unresolved.[11] Scientific reviews like the NAS and PCAST reports provide only a snapshot of the scientific validity of a particular methodology at a particular time.

In this evolving landscape, judges need a coherent framework for deciding at any given time whether the empirical evidence, as it currently stands, provides a sufficient basis for the testimony in a case. Is some minimum threshold of empirical testing and validation necessary for admitting forensic testimony? Should testimony from disciplines that just meet the threshold for admissibility be treated differently than disciplines with more rigorous scientific testing and validation? The answers lie in the discretion of judges who confront the evidence in a particular case. But the options need not be reduced to a choice between wholesale exclusion of evidence that falls just short of the most rigorous standards of scientific validity or total acceptance of methods that remain scientifically shaky. The Federal Rules of Evidence offer judges a range of tools for managing expert testimony beyond wholesale admission or exclusion. Judicious use of these tools can accommodate both the incremental nature of empirical studies of scientific validity and the need for courts to "resolve disputes finally and quickly."[12]

IT IS UP TO THE COURTS

Despite the significant response to the 2009 NAS report from many in the forensic sciences community, the 2016 PCAST report and ensuing discussions have shown that efforts to adopt more rigorous scientific standards for validation and practice have been slow and uneven, and that substantial disagreement remains over what level of empirical testing and scientific validation is appropriate for forensic evidence. Assessing the state of a subset of forensic disciplines (feature comparison methods, including DNA identifications, latent fingerprint analysis, and firearms toolmark analyses, among others) seven years after the landmark NAS report, PCAST acknowledged significant progress in some areas, including creation of the National Commission on Forensic Sciences (NCFS) and notable empirical

[T]RIAL-COURT DISCRETION IN CHOOSING THE MANNER OF TESTING EXPERT RELIABILITY . . . IS NOT DISCRETION TO ABANDON THE GATEKEEPING FUNCTION . . . [NOR] TO PERFORM THE FUNCTION INADEQUATELY. RATHER, IT IS DISCRETION TO CHOOSE AMONG REASONABLE MEANS OF EXCLUDING EXPERTISE THAT IS FAUSSE AND SCIENCE THAT IS JUNKY.

JUSTICE ANTONIN SCALIA, CONCURRING, *KUMHO TIRE CO. V. CARMICHAEL* (1999)

studies describing the reliability and accuracy of latent fingerprint and firearms toolmark analyses.[13] Nonetheless, the report concluded that most of the methods it evaluated still lack sufficient empirical evidence to demonstrate scientific validity. Although the report offered specific assessments of seven forensic disciplines, it emphasized that these would likely change over time as methods and practices evolve and new empirical studies emerge. Instead, the PCAST recommendations primarily focus on the overall criteria for evaluating scientific validity. The report's most fundamental conclusion is that empirical evidence is the *only* basis for establishing scientific validity, and thus evidentiary reliability, of forensic science methods. "Well-designed" empirical studies, according to the report, are especially important for demonstrating reliability of methods that rely primarily on subjective judgments by the examiners.

As in the larger conversation, responses to the PCAST report encompassed a wide range of viewpoints, but many of the responses from the forensics and law enforcement communities were harsh.[14] In particular, forensic scientists often pointed out that PCAST did not include active forensic scientists and argued that academic scientists with no training and experience in forensic methods cannot adequately assess the reliability of those methods. Substantively, some critics objected to PCAST's insistence on empirical studies as the only reliable basis for establishing scientific validity of empirical claims. Those critics argue that other factors, most notably training and professional experience, can be sufficient to demonstrate reliability; indeed, they argue, empirical evidence is often unnecessary and inappropriate, especially for methods that rely primarily on professional judgment that can only be acquired through extensive training and experience. Others agree that empirical evidence is important for establishing the reliability of forensic methods, but object that the criteria PCAST proposed for identifying "well-designed" empirical studies sufficient to establish scientific validity are both arbitrary and too rigid.[15]

Since the PCAST report and supplement were published, debates over the reliability of forensic sciences and testimony by forensics experts have remained in flux. Sharp rebukes of the PCAST report by critics in the forensics community continued in 2017.[16] Attorney General Jeff Sessions allowed the National Commission on Forensic Sciences, established by the Obama administration after the 2009 NAS report was published, to expire.[17] At its final meeting, the commission rejected proposals by two of its subcommittees supporting more rigorous standards for written reports and testimony by forensic practitioners.[18] The Attorney General has since appointed a special advisor on forensic sciences and established a working group within DOJ to develop guidelines for testimony by forensics experts.[19]

In response to these changes at the Department of Justice, the AAAS and other scientific societies have called on the Attorney General to establish an independent advisory group to continue to identify gaps and limitations in the scientific validity of forensic methods and to outline a research agenda to address those gaps.[20] The AAAS's 2017 report on the scientific validity of latent fingerprint analysis considered a broader range of empirical studies than did the PCAST report but concurred with PCAST that empirical studies support the foundational validity of fingerprint analysis, albeit with a greater potential for errors than previously recognized.[21] The AAAS report also emphasized that error rates may be even higher for the method as applied in many crime laboratories. Standard procedures in many laboratories allow examiners access to other information about a crime, posing a risk of "contextual bias." Both AAAS and NCFS have called for crime labs to adopt "context blind" procedures and to incorporate "blind testing" to determine the validity and error rates for various forensic methods as applied.[22] A 2017 symposium convened at the National Institute of Standards and Technology (NIST) reported promising results from such blind testing in a few crime laboratories, but also described logistical barriers to widespread implementation of similar programs.[23] In many laboratories, for example, procedures for submitting and processing samples reveal information about the crime and the submitting law enforcement agency; such processes also allow analysts to communicate with investigators involved in the case before completing ▸

the forensic analysis. For these reasons, it can be difficult to routinely introduce test samples into an examiner's workflow without detection.

As these continuing conversations illustrate, there is no clear consensus in the forensic science community about the type and extent of empirical testing necessary to establish the validity of forensic methods. The implementation of more rigorous practices and procedures remains gradual and uneven between disciplines and individual forensic laboratories. For the foreseeable future, it is likely that courts will face proffers of forensic testimony based on methods and practices that reflect a broad spectrum of empirical testing and scientific validation. As a result, it is clearly up to the courts to determine the levels of scrutiny and scientific validity required in order to admit testimony by traditional forensic science experts. Are scientists right that rigorous empirical studies are the only reliable basis for assessing scientific validity? Or are forensic scientists right that those scientific standards are too rigid, and in some cases inappropriate, in some areas of applied forensic sciences? Does it depend, as some courts have suggested, on the nature of the testimony?

Those decisions resonate beyond the courtroom. For much of the public, crime laboratory forensics are the most visible, and often the defining, example of scientific evidence as a source of confidence and legitimacy for the criminal justice system. *Daubert* and *Kumho Tire* give trial courts wide discretion in deciding these questions, and both cases explicitly recognize that the factors appropriate for assessing the reliability of expert testimony might differ for different kinds of expertise.[24] But that leaves trial judges to resolve the competing claims from the scientists insisting that "well-designed" empirical studies are the only reliable basis for assessing scientific validity and the critics who argue that those scientific standards are too rigid, or are even inappropriate, to serve as indicia of reliability in applied forensic sciences.

EMPIRICAL STUDIES: CROSS-EXAMINING SCIENCE

The current state of empirical studies for scientific validity of the forensic sciences — what PCAST called "foundational validity" under Rule 702(c) — varies widely for different disciplines, ranging from thousands of research studies for DNA analysis of single-source samples[25] to perhaps a dozen studies for latent fingerprint analysis[26] to no empirical evidence for the validity of bitemark analysis.[27] Well-controlled empirical studies to establish error rates for those methods as applied in routine practice (Rule 702(d)) remain rare, but are beginning to be implemented in some areas.[28] As a result, at any given time, there can be a wide variation in the strength of the empirical evidence supporting the foundational validity of a forensic method and the amount of variability in the method as applied. Along this spectrum, how much is enough to admit forensic evidence? What can courts do when the empirical evidence of scientific validity for an expert's testimony is "just barely" enough — what the *Daubert* court called "shaky but admissible" evidence?[29] One way to approach that question is to ask what work empirical evidence does in science and how that relates to the goals of evidence law.

While it is common to say that empirical studies are designed to prove a scientific principle or establish the validity of a method, it is more accurate to say that the role of empirical studies in science is to probe for flaws and define the limitations of a principle or method. Embracing that point, the *Daubert* court cited the philosopher of science Karl Popper, who focused on "falsifiability" as the defining feature of science.[30] While other philosophers of science, like Thomas Kuhn and Robert Merton, differ from Popper in important ways, they embrace the essential role of empirical evidence in probing the limits of empirical claims and the unique ability of empirical studies to reveal errors or limitations in a way that cannot be ignored or rationalized away.[31]

This is especially true when the goal of a study is to test the reliability and accuracy of a widely accepted principle or method. Well-designed empirical studies probe for weaknesses and limitations, uncertainty, and the potential for error in the principle or method, just as cross-examination probes a witness's direct testimony in court. The analogy between empirical studies in science and cross-examination in law is not coincidental. In fact, 400 years ago, cross-examination in legal practice was one model for the development of empirical science. In 1620, Francis Bacon, a lawyer, former attorney general, and lord chancellor of England, articulated what would be the foundations of a new scientific method grounded in empirical observations and experiments.[32] Four centuries ahead of modern research on cognitive biases, Bacon argued that human thought is exquisitely susceptible to systematic distortions of perception, interpretation, and reasoning, which he called the "idols" or "illusions" of the human mind.[33] And, as Bacon noted and modern cognitive science confirms,[34] our strongest and most consistent cognitive biases operate primarily in one direction — systematically overweighting evidence consistent with prior beliefs and systematically ignoring or discounting evidence that conflicts with those beliefs.

Only well-designed empirical tests,[35] Bacon argued, provide a sufficient mech-

anism for revealing errors or limitations of scientific principles or methods in a way that cannot be rationalized or dismissed on the basis of subjective judgments. Bacon famously imagined his new approach to science as a kind of trial of scientific ideas.[36] And in that trial, the function of empirical studies is to test the reliability of a scientific claim, probing for weaknesses, errors, inconsistencies, limitations, or alternative explanations as a lawyer probes an ordinary witness in court. "[T]o use the language of civil procedure," he declared, "we intend, in this *Great Suit* or *Trial* . . . to *cross-examine* nature herself."[37] (Emphasis added.)

The analogy to cross-examination offers a useful framework for thinking about the role of empirical studies in deciding admissibility of expert testimony in law. The *Daubert* court itself, contemplating the possibility of admitting expert evidence that falls short of the most rigorous scientific standards, emphasized that "vigorous cross-examination, presentation of contrary evidence, and careful instruction on the burden of proof are the traditional and appropriate means of attacking shaky but admissible evidence."[38] At the same time, the Court cautioned that these tools may be less effective for experts than other witnesses: "Expert evidence can be both powerful and quite misleading because of the difficulty in evaluating it. Because of this risk, the judge in weighing possible prejudice against probative force under Rule 403 of the present rules exercises more control over experts than lay witnesses."[39]

As the *Daubert* court recognized, most jurors will lack the specialized knowledge and experience required to evaluate the reliability of an expert's principles and methods or the significance of issues raised on cross-examination, especially when dealing with scientific or technical

BY DEFINITION, JURORS WITHOUT SPECIALIZED TRAINING AND EXPERIENCE IN SCIENTIFIC ANALYSIS LACK THE FOUNDATION THEY WOULD NEED TO IDENTIFY LIMITATIONS OR WEAKNESSES IN AN EXPERT'S METHODS ON THEIR OWN.

experts. In addition to specific knowledge of their field, scientists routinely rely on procedures and modes of inference that are not typically encountered in daily life. Jurors will rarely have any basis in their own experience for recognizing the limitations that might be obvious to other scientists, or the statistical training to interpret and apply error rates correctly. Forensic methods in which the essential steps in analysis rely on the subjective judgment of an examiner magnify those concerns. Neither judges nor jurors can see inside the examiner's brain to assess consistency and accuracy, or the possible influence of cognitive biases or simple errors in an examiner's analysis. To decide what weight to give the expert's testimony, jurors must look to their own experience and intuitions, including their own preconceptions about the reliability and accuracy of forensic methods[40] and the confidence of the testifying expert. By definition, however, jurors without specialized training and experience in scientific analysis lack the foundation they would need to identify limitations or weaknesses in an expert's methods on their own. Further, because the examiner's subjective experience both is inaccessible to any outside observer and relies on the very expertise that separates her from jurors, cross-examination is unlikely to be effective in probing for those weaknesses and limitations.

From that perspective, one essential function of empirical studies is to return jurors to the process by defining limitations and the potential for error in an expert's methodology using terms laypeople can understand. This requires empirical studies that are sufficiently well designed, that define error rates and uncertainty clearly, and that are sufficiently applicable to the real-world work of the testifying expert to allow jurors to properly evaluate the testimony. In the absence of such empirical evidence, jurors have no meaningful basis for deciding what weight to give the testimony, and a court will need to consider whether the risk of confusing or misleading the jury, and the impediments to cross-examination, are too high to admit the expert's testimony.

A SPECTRUM OF SCIENTIFIC VALIDITY

Scientific critiques of forensic sciences uniformly insist on what the PCAST report called "a central tenet of science: An *empirical claim* cannot be considered valid until it has been empirically tested."[41] Yet *Daubert*, *Kumho Tire*, and the language of Rule 702 expressly recognize that expertise might be based on other factors, including training and experience.[42] Why are scientists so insistent on empirical studies? And how much is enough? Scientists make a clear distinction between principles and methods with empirical evidence of reliability and those that lack any empirical validation. But rigorous scientific validation builds incrementally over the course of multiple, independent, well-designed studies of accuracy, and there is no fixed point at which a scientific method crosses from dodgy to scientifically valid. Forensic ▸

methods that have not yet reached that level of validation might nonetheless qualify as "shaky but admissible." In deciding how much empirical validation is enough to admit forensic evidence, and how to manage evidence that is shaky but admissible, it helps to understand why scientists insist on multiple, well-designed empirical studies as the gold standard for scientific validity.

The fundamental reason scientists and engineers insist on empirical studies is simple. People whose work necessarily includes empirical feedback on the accuracy of their ideas and methods quickly discover how often that empirical feedback reveals anything from simple errors in carrying out a procedure to fundamental limitations of a principle or method they considered well established.

This is not limited to research scientists. Imagine, for example, an auto mechanic whose expertise is based entirely on practical training and experience diagnosing problems with automotive engines and fuel systems. The principles and methods she learns in her training are likely based on extensive empirical research and testing by engineers and designers. Moreover, regular practical experience in diagnosing and repairing engines and fuel systems provides continual empirical feedback on how reliably she applies those principles and methods: If the mechanic mistakenly declares that a car will not start because of a faulty fuel pump, replaces the fuel pump, and then attempts to start the car, she immediately discovers her error. If she has charged a customer for the new fuel pump, the error is likely to be brought to the mechanic's attention in a way she cannot easily overlook. A court might reasonably find that years of experience, informed by that kind of empirical feedback, is sufficient to show that the mechanic's principles and methods for diagnosing engine and fuel system failures are reliable. The same might be said of electricians, plumbers, engineers, airplane pilots, and a host of other experts whose work routinely includes empirical outcomes that reveal errors or less-than-optimal outcomes.

IN DECIDING HOW MUCH EMPIRICAL VALIDATION IS ENOUGH TO ADMIT FORENSIC EVIDENCE, AND HOW TO MANAGE EVIDENCE THAT IS SHAKY BUT ADMISSIBLE, IT HELPS TO UNDERSTAND WHY SCIENTISTS INSIST ON MULTIPLE, WELL-DESIGNED EMPIRICAL STUDIES AS THE GOLD STANDARD FOR SCIENTIFIC VALIDITY.

By contrast, forensic scientists in many disciplines get little or no empirical feedback on the accuracy of, and any errors that might result from, the ordinary course of their work. This is particularly true for disciplines in which the critical steps of analysis rely on subjective judgments by the examiners, including analyses of latent fingerprints, firearms, shoe and tire impressions, hair, and bitemarks. To be sure, examiners in those disciplines receive training and proficiency testing that includes analysis of known samples, but in those situations examiners know that they are being tested and may consciously or subconsciously adjust the way they perform their analysis.[43] In most proficiency tests, furthermore, the test samples do not represent the range of samples encountered in normal practice, but are instead designed with the expectation that all competent examiners will correctly identify the samples.[44] As a result, examiners receive little or no empirical feedback that can alert them to the possibility of errors.

In the absence of that kind of objective feedback, research across a variety of professional fields shows that training and experience without objective feedback increases the confidence of experts in their own knowledge and skills, but that confidence does not correlate with objective measures of skill or accuracy.[45] Psychological studies show that consistently following an established procedure is enough to increase confidence, even when the procedure itself produces inaccurate results.[46] Furthermore, individuals with the lowest ability to reflect on their own susceptibility to error (what researchers call "meta-cognition") tend to be the most overconfident in their own expertise and accuracy.[47] As a result, the amount of training or professional experience, in itself, provides very little information about the reliability of an expert's principles or methods as a basis for empirical statements. Rather, the value of training and experience as a proxy for reliability depends on the quality and amount of objective, empirical feedback to define the accuracy and limitations of the expert's methodology.

Because empirical testing of a scientific principle or method is a cumulative process, the quality and amount of empirical testing supporting an expert's methodology can span a wide spectrum,

from anecdotal empirical feedback acquired in the course of professional experience to rigorous empirical studies of accuracy, error rates, and other limitations. How much empirical testing is enough for admissibility? Scientific critics of forensic sciences emphasize the importance of *multiple, well-designed* empirical studies in order to establish the scientific validity of forensic disciplines and the accuracy of their methods. Centuries of experience with the scientific method, across disciplines from physics to biochemistry to psychology, have taught important lessons about basic elements of experimental design and how to conduct empirical studies of this kind in a way that minimizes and controls for a wide range of factors that can lead to misleading results, from unintended biases in sample selection to cognitive biases in interpretation to statistical flukes — what the PCAST report called "well-designed" studies. Following good study design greatly reduces the uncertainty of the results, but no single study can be definitive. Multiple studies increase reliability by gradually decreasing the range of uncertainties — indeed convergence of results from multiple studies can sometimes compensate or correct for design flaws or limitations of the individual studies.[48]

Moreover, multiple empirical studies can provide a wealth of information about the weaknesses and limitations of a method, variation in its application, and uncertainty in the results or their interpretation. What kinds of test conditions affect the accuracy of the method? How much do small variations in procedure alter the accuracy of results? Are some samples more difficult to analyze or likely to produce inaccurate results? Multiple, well-designed studies help to define the sources of variation, conditions that affect the reliability results, and error rates under various conditions, all of which can help jurors understand both the validity and limitations of results obtained in a particular case. More limited empirical testing may provide some important evidence regarding the foundational validity of a method, but will generally leave greater uncertainty about those conclusions and how they apply to the method as applied in a particular case. In those cases, courts must grapple with whether cross-examination can be an effective alternative for identifying any weaknesses or limitations of the testimony in a way jurors can understand.

SHAKY BUT ADMISSIBLE FORENSIC EVIDENCE – WHAT'S IN THE TOOL BOX?

With the exception of DNA analysis of single-source samples, none of the forensic methods reviewed by PCAST has yet met rigorous criteria for both foundational validity (Rule 702(c)) and validity as applied (Rule 702(d)).[49] Other methods, however, have reached important waypoints in the validation process. Both PCAST and the AAAS working group conclude, for example, that recent empirical studies support the foundational validity of latent fingerprint analysis, although they applied different criteria for identifying the relevant empirical studies.[50] Both groups still urge substantial caution in extrapolating from those studies to the overall validity and error rates for fingerprint analysis as applied in ordinary practice.[51] The PCAST report identified one study of firearms analysis that met its criteria for well-designed empirical studies,[52] just short of the two independent studies it recommends as a minimum criterion for scientific validity.[53] As empirical studies of these and other forensic methods continue, courts will certainly face challenges to the reliability of forensic methods supported by varying degrees of empirical evidence.

When an expert's testimony is based on principles and methods that lack any substantial empirical evidence of scientific validity, judges who embrace the widespread view of scientists that empirical studies are essential for scientific validity might use their discretion to exclude the testimony. On the other hand, given the "liberal thrust" of modern evidence law and the broad discretion of trial judges in deciding on the admissibility of expert evidence, a judge may be inclined to admit evidence supported by empirical data that falls short of the most rigorous criteria for scientific validity. In those cases, courts have a variety of tools for reducing the risk of prejudice, confusion, or misleading jurors and the related impediments to effective cross-examination.

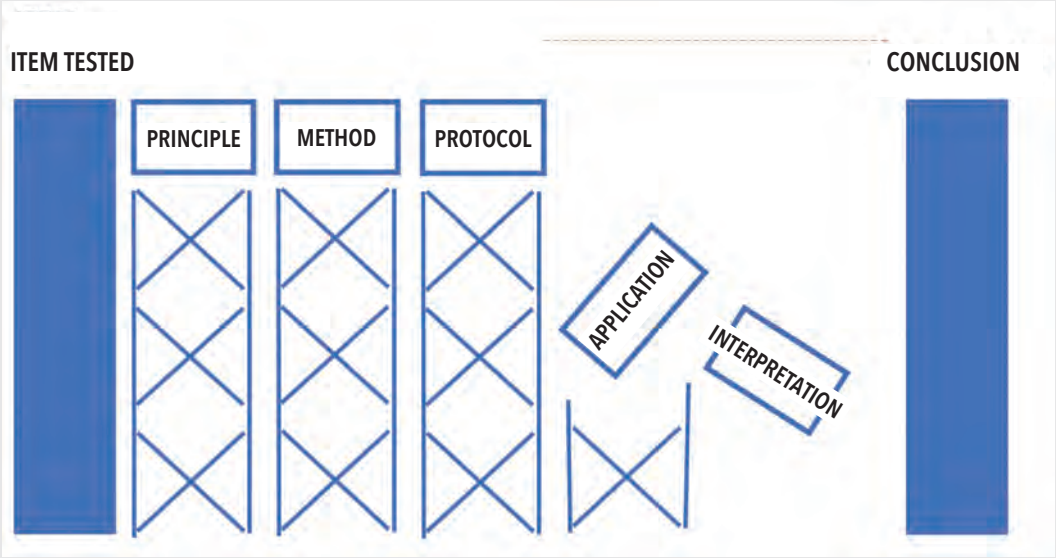
LIMITING TESTIMONY

One of the most obvious (and widely used) tools is the language of Rule 702 directed at testimony. Rather than all-or-none admission of an expert or scientific discipline, some courts have allowed forensic experts from disciplines like latent fingerprints, firearms, and handwriting analysis — whose reliability traditionally has been based on training and experience rather than empirical validation — to testify about the similarities between two sets of prints, or shell casings, or writing samples, while excluding statements about the likelihood that such a similarity might arise in samples from separate sources.[54] More recently, scientists, legal scholars, and forensic practitioners have devoted considerable attention to the importance of monitoring testimony about confidence, statistical uncertainty, error rates, and the likelihood of alternative conclusions based on the forensic results in a particular case.[55]

Is limiting the scope of testimony effective? Research on experience-based ▸

expertise does support the intuition that training and experience can improve the ability of experts to identify and categorize specific features in complex patterns and can enhance strategies for comparing those features across sample, even when those experts do not receive direct empirical feedback on the accuracy of their observations. Thus, a judge might reasonably find that the extensive training and professional experience in latent fingerprint analysis, firearms analysis, and other subjective feature comparison methods provide a reliable basis for testimony that simply points out the extent of similarities or differences between two or more samples.

FIGURE 1. CHAIN OF INFERENCES AND LOGICAL GAPS



Every step must be adequately supported by empirical evidence. If one step in the logical chain is invalid, the results are invalid. e.g., *In re Paoli R.R. Litig.*, 35 F.3d 717 (1994); *Joiner*; *In re Zoloft*, 858 F.3d 787 (2017).

That, however, does not end the inquiry. The relevance of such testimony depends on a chain of inferences leading from the expert's observations to a conclusion that makes some fact in the case more or less probable (see Figure 1 above). In order for the conclusion to be scientifically valid, every step in the chain of inferences supporting it must be valid. Courts have agreed that, if any step in the logical chain is invalid, the results are invalid.[56] In the example of feature comparison experts, the critical link in the logical chain is an empirical statement about how likely it is that similar features pointed out by the witness could arise from two different sources. Any statement by the witness on that issue would need to be based on scientifically valid empirical data. On the other hand, simply omitting any testimony on this step in the logical chain will sharply increase the risk of confusing or misleading jurors. Research on cognitive heuristics and biases show that people tend to fill in gaps in a logical chain using common heuristic devices, like the "availability heuristic"; this means that without explicit information pointing out a gap in the logical chain from observation to conclusion, jurors are more likely to link the expert's limited testimony to the implied conclusion that the similarities the expert has pointed out are very unlikely to be produced by different sources.[57]

Gaps can occur at any step in the logical chain, of course. One step that is likely to be particularly important in the near term is the link between empirical studies that address the foundational validity of a forensic method and the accuracy of that method as applied by a specific examiner using the samples in a particular case, especially for disciplines like latent fingerprint and firearms analysis, which already have significant empirical evidence of foundational validity.[58] Courts in those cases will need to ask whether the available empirical studies encompass a sufficient range of samples, test conditions, and examiner qualifications to provide a reasonable estimate of the error rate for the method as applied in the current case, or to provide a basis for effective cross-examination on that issue.

JUDICIAL INSTRUCTIONS AND BACKGROUND EXPERTS

In addition to limiting the scope of expert testimony, trial judges have broad discretion to manage the traditional tools for probing weaknesses and alternative interpretations of any testimony, including cross-examination, the presentation of other experts, and judicial instructions.[59] In the case of expert witnesses, it is always important to consider how to apply those tools so that jurors have the information they need to decide what weight to give the expert's testimony. Those considerations can be especially important in the case of forensic experts whose methods have undergone limited empirical validation and where jurors

may have particular difficulty evaluating the reliability of the testifying expert's methods and conclusions.

In the ideal case, scientific methods will have undergone rigorous empirical testing that encompasses multiple well-designed studies by independent researchers exploring a wide range of samples and test conditions, including the method as applied in normal practice. Results from these empirical studies would provide jurors with a direct and well-defined error rate for the method as applied to the same type of samples and under the same conditions as in the case at hand. Unfortunately, the current empirical testing for most forensic methods is not that extensive and is unlikely to reach that level in the near term. Where available empirical studies are more limited, jurors will have more difficulty understanding how the error rates or other measures from the available studies do or do not apply to the results and conclusions presented by the expert in the present case. A limited amount of empirical testing, for example, might be sufficient to show that a principle or method is scientifically valid in principle, but not enough to define error rates, uncertainty, or other limitations of the method as applied in the case at hand.

Cross-examination in that situation is also unlikely to be effective on its own. Jurors are unlikely to have the training or personal experience needed to evaluate the significance of limitations in the design or scope of empirical studies of a forensic method, or of any deviations from best practices in laboratory procedures or the expert's methods. Testifying experts whose training and experience is in those forensic disciplines that have not traditionally incorporated extensive empirical testing and procedural controls may not have the expertise to address questions about those limitations effectively on cross-examination to questions. This could raise potential Confrontation Clause concerns.[60]

IN ADDITION TO LIMITING THE SCOPE OF EXPERT TESTIMONY, TRIAL JUDGES HAVE BROAD DISCRETION TO MANAGE THE TRADITIONAL TOOLS FOR PROBING WEAKNESSES AND ALTERNATIVE INTERPRETATIONS OF ANY TESTIMONY, INCLUDING CROSS-EXAMINATION, THE PRESENTATION OF OTHER EXPERTS, AND JUDICIAL INSTRUCTIONS.

To provide jurors with the background information they need to evaluate the expert's testimony in those cases, and to enable effective cross-examination, courts may need to apply other available tools with particular vigor. Trial judges clearly have the option to allow testimony by experts (including neutral experts under Rule 706) to provide information about design and controls in laboratory procedures, for example, or considerations in applying error rates from the foundational studies to the methodology as applied by the testifying expert in the present case. Some courts have allowed this testimony with regard to the reliability of eyewitness identifications, where years of scientific research had found that factors affecting the formation and recall of memories by eyewitnesses differ in important ways from common preconceptions.[61] Alternatively, in the case of eyewitness identification, a number of scholars have suggested that judicial instructions might be a more concise and effective way to inform jurors about key findings from the relevant research.[62] Judges could choose to offer such instructions regarding testimony by forensics experts when jurors are likely to harbor preconceptions about the scientific validity or infallibility of forensic methods that are inconsistent with the current state of empirical studies.

The need for these tools will be lowest where empirical studies provide the most extensive and granular information about sources of variation, limitations, and error rates for a forensic method in a form that jurors can understand and apply directly to the testimony in a particular case. That would include well-designed tests of the method as applied in regular practice. Conversely, the need for expert witnesses or judicial instructions to augment jurors' understanding of the issues increases when the available empirical studies are limited or have not directly tested error rates for the method as applied in regular practice by the testifying expert in the present case. In effect, expert witnesses or judicial instructions are needed to help jurors understand what information is missing from the available empirical studies.

This is the situation described in the recent AAAS assessment of latent fingerprint analysis[63] and the PCAST review of firearms analysis.[64] Both reviews found that a limited number of empirical studies provided reasonably strong evidence of foundational validity for both meth-

ods, including specific error rates for experts under the test conditions. For both fingerprint and firearms analysis, however, PCAST and AAAS pointed out that experts in the empirical studies were aware that they were being tested, which can alter the way the examiners analyze the test samples. That does not mean that the empirical studies are not well-designed and useful, but the PCAST and AAAS reviewers emphasize that the study designs make it difficult to extrapolate directly from error rates measured in the empirical studies to the potential for error in actual practice. Based on the evidence of foundational validity for these methods, judges that have long admitted both latent fingerprint and firearms analysis are unlikely to exclude that testimony in the wake of the recent studies. But they could opt to allow expert testimony or offer judicial instructions to help jurors understand both the strength of the recent empirical studies in validating these methods and the need for caution in applying error rates from those studies to the expert testimony in a specific case.

SUMMARY

As empirical testing in forensics moves forward, courts will continue to face challenges to forensic evidence with varying degrees of empirical validation, which may include substantial empirical evidence of validity that nonetheless falls short of the most rigorous criteria for scientific validation, either foundationally or as applied. While courts have wide discretion to decide a minimum threshold of scientific validity for admitting forensic evidence, their options are not limited to wholesale exclusion or unlimited admission. However, "shaky but admissible" testimony increases the risks of prejudice or confusion resulting from juror preconceptions and cognitive biases about forensic evidence, invites jurors to draw inferences from limited testimony, and introduces the need for specialized knowledge to evaluate issues raised on cross-examination. Courts may need to take particular precautions, including the use of expert witnesses or judicial instructions, to ensure that jurors have the background information and guidance they need to appropriately evaluate a forensic expert's testimony and interpret issues raised on cross-examination.

PATE SKENE is associate research professor of neurobiology at Duke University and a member of the Duke Institute for Brain Sciences. He spent much of his career studying genes involved in brain development and repair before attending law school at Duke. His research now focuses on decision making and scientific evidence in law. He was the 2016-17 AAAS fellow at the Federal Judicial Center.

[1] Video: American Association for the Advancement of Science, *Science-based Forensic Evidence: Getting the Science Right in Criminal Investigations, at 14:50-17:14,* https://www.aaas.org/page/science-based-forensic-evidence-getting-science-right-criminal-investigations.

[2] Nat'l Acad. of Sci., Nat'l Research Council, Strengthening Forensic Science in the United States: A Path Forward (2009) [hereinafter NAS Report]

[3] President's Council of Advisors on Sci. & Tech., Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods (2016) [hereinafter: PCAST Report].

[4] American Association for the Advancement of Science (AAAS), Forensic Science Assessments: A Quality and Gap Analysis – Latent Fingerprint Examination (2017) [hereinafter AAAS report].

[5] *See* President's Council of Advisors on Sci. and Tech., Exec. Office of the President, Report to the President, An Addendum to the PCAST Report on Forensic Science in Criminal Courts (2017) [hereinafter PCAST Addendum]; *Published Statements in Response to the PCAST Report on Forensic Science in Criminal Courts,* https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_2016_public_comments.pdf [hereinafter PCAST Responses]; I.W. Evett et al., *Finding the Way Forward for Forensic Science in the US – A Commentary on the PCAST Report*, 278 Forensic Science International 16 (2017).

[6] Sarah Lucy Cooper, *The collision of law and science: American court responses to developments in forensic science*, 33 Pace Law Review 234–301 (2013); Simon A. Cole & Gary Edmond, *Science without Precedent: The Impact of the National Research Council Report on the Admissibility and Use of Forensic Science Evidence in the United States*, 4 British Journal of American Legal Studies 585–617 (2015); Jules Epstein, *Preferring the "wise man" to science: the failure of courts and non-litigation mechanisms to demand validity in forensic matching testimony*, 20 Widener Law Review 81–113 (2014); *The general assumptions and rationale of forensic identification*, *in* Modern Scientific Evidence: the Law and Science of Expert Testimony, Vol. 4 1-61 (Edward K. Cheng et al, eds., 2016-17 ed.).

[7] *Id.*

[8] *Daubert v. Merrell Dow Pharmaceuticals, Inc.* 509 U.S. 579, 594 (1993); *Kumho Tire Co. v. Carmichael*, 526 U.S. 137, 141 (1999).

[9] *United States v. Glynn,* 578 F. Supp. 2d 567 (SDNY, 2008); 4 Modern Scientific Evidence: the Law and Science of Expert Testimony. Vol. 4 23-25, Edward K. Cheng et al, eds., 2016-17 ed.

[10] Saul M. Kassin, et al, 2 Journal of Applied Research in Memory and Cognition 42, 2013; William Thompson et al., 10 Journal of Empirical Legal Studies 359, 2013; *see also Cheng et al., supra* note 9.

[11] PCAST addendum, *supra* note 5; AAAS Report, *supra* note 4.

[12] *Daubert, supra* note 8 at 595.

[13] PCAST Report *supra* note 3 at 35-37, 94-97, 109-111;

[14] PCAST Responses, *supra* note 6; see also Evett et al., *supra* note 5.

[15] *Id.*

[16] Evett at al, *supra* note 5.

[17] Spencer S. Hsu, *Sessions Orders Justice Dept. to*

End Forensic Science Commission, Suspend Review Policy, Washington Post (Apr. 10, 2017), available at https://www.washingtonpost.com/local/public-safety/sessions-orders-justice-dept-to-end-forensic-science-commission-suspend-review-policy/2017/04/10/2dada0ca-1c96-11e7-9887-1a5314b56a08_story.html?utm_term=.93ddf5b51326

[18] U.S. Dep't of Justice Archives, National Commission on Forensic Science, Final Draft Views on Documentations, Case Records and Report Contents, https://www.justice.gov/archives/ncfs/reporting-and-testimony; U.S. Dep't of Justice Archives, National Commission on Forensic Science, Final Draft Views on Statistical Statements in Forensic Testimony, https://www.justice.gov/archives/ncfs/reporting-and-testimony

[19] Press Release, U.S. Dep't of Justice, *Justice Department Announces Plans to Advance Forensic Sciences* (Aug. 7, 2017), https://www.justice.gov/opa/pr/justice-department-announces-plans-advance-forensic-science

[20] Letter from American Association for the Advancement of Science, American Chemical Society, Federation of Associations in Behavioral and Brain Sciences, and Human Factors and Ergonomics Society to the Honorable Jeff Sessions, Attorney General of the United States (June 9, 2017) https://mcmprodaaas.s3.amazonaws.com/s3fs-public/Scientific%20Society%20Comment%20on%20DOJ-LA-2017-0006-0001%20-%209%20June%202017.pdf

[21] *See* AAAS report, *supra* note 4.

[22] *Id.* at 35-42.

[23] National Institute of Standards and Technology (NIST), 2017 IFSEMS Presentations, Forensic Science Error Management International Forensics Symposium July 24-27, 2017, https://www.nist.gov/topics/forensic-science/2017-ifsems-presentations

[24] *See Daubert, Kumho Tire, supra* note 8.

[25] *See* PCAST Report, *supra* note 3.

[26] *See* AAAS Report, *supra* note 4.

[27] PCAST Report, *supra* note 3; Cheng et al., *supra* note 9.

[28] NIST, *supra* note 23.

[29] *Daubert, supra* note 8 at 596.

[30] Popper argued that it is not possible to prove a scientific principle or theory simply by making observations or producing experimental results consistent with the theory. Instead, he said, the essential test of scientific validity is to design experiments in which the principle or theory under consideration predicts one of several possible outcomes that can be confirmed by neutral observers. Falsifiability in this context means setting up tests in which it is possible for any outside observer to see when the theory or principle is wrong.

[31] David Goodstein, *How Science Works, in* Reference Manual on Scientific Evidence, Third Edition, 2011; Simon Cole, *Forensic Culture as Epistemic Culture: The Sociology of Forensic Science, 44* Studies in the History and Philosophy of Biological and Biomedical Sciences 36, 2013.

[32] Francis Bacon, The New Organon (Lisa Jardine and Michael Silverthorne eds., 2000)

[33] *Id.* at 28, 18 fn 13.

[34] Gary Edmond et al., *Thinking Forensics: Cognitive Science for Forensic Practitioners, 57* Science and Justice 144, 2017; Saul M. Kassin et al., *supra note 10; D. Michael Risinger et al., The Daubert/Kumho Implications of Observer Effects in Forensic Science: Hidden Problems of Expectation and Suggestion, 90* Cal. L. Rev. 1, 2002.

[35] Bacon, *supra* note 30 at 159.

[36] Barbara J. Shapiro, *"Fact" and the Proof of Fact in Anglo-American Law (c. 1500-1850),* How Law Knows (Austin Sarat, Lawrence Douglas, and Martha M. Umprey, eds.), Stanford University Press 2007, pp. 25-69; Harvey Wheeler, *The Invention of Modern Empricism: Juridical Foundations of Francis Bacon's Philosophy of Science, 76* Law Libr. J. 78, 1983.

[37] Bacon, *supra* note 30 at 232 (emphasis added).

[38] *Daubert, supra* note 8 at 596.

[39] *Id.* at 595, quoting Weinstein, 138 F.R.D. 631, 632. (Jack B. Weinstein, *Rule 702 of the Federal Rules of Evidence is Sound; It Should Not Be Amended,* 138 F. R. D. 631 (1991).)

[40] Kassin et al, *supra* note 10; Thompson et al., *supra* note 10; *Brandon Garrett and Gregory Mitchell, 10* Journal of Empirical Legal Studies 484, 2013.

[41] PCAST addendum, *supra* note 5 at 1 (emphasis added).

[42] *Daubert, Kumho Tire, supra* note 8.

[43] PCAST report, *supra* note 3 at 58 fn 136.

[44] *Id.*; AAAS report, *supra* note 8.

[45] Edmund et al., *supra note 32 at 149.*

[46] Elanor F. Williams, David Dunning, and Justin Kruger, 104 Journal of Personality and Social Psychology 976, 2013.

[47] *Id.*; Justin Kruger and David Dunning, 77 Journal of Personality and Social Psychology 1121, 1999; Daniel Levin, *Change Blindness Blindness: The Metacognitive Error of Overestimating Change-detection Ability,* 7 Visual Cognition 397, 2000; John T. Breidert and Jeffrey E. Fite, *Self Assessment: Review and Implications for Training, U.S. Army Research Institute for Behavioral and Social Sciences,* Research Report No. 1900 (2009).

[48] AAAS report, *supra* note 4.

[49] PCAST report, *supra* note 3; PCAST addendum, *supra* note 5.

[50] AAAS report, *supra* note 4.

[51] *Id.*

[52] AAAS report, *supra* note 4 at 109-112.

[53] PCAST addendum, *supra* note 5 at 6-8.

[54] *United States v. Glynn,* 578 F. Supp. 2d 567 (SDNY, 2008).

[55] PCAST report, *supra* note 3; AAAS Report, *supra* note 4; National Commission on Forensic Sciences, *supra* note 18.

[56] *In re Paoli R.R. Yard PCB Litig.,* 35 F.3d 717 (3d Cir. 1994); *General Electric Co. v. Joiner,* 522 U.S. 136 (1997); *In re Zoloft,* 858 F.3d 787 (2017).

[57] *See* Cheng et al., *supra* note 9 at 23-25.

[58] PCAST report, *supra* note 49; PCAST addendum, *supra* note 5.

[59] *Daubert, supra* note 8 at 596.

[60] *Crawford vs. Washington,* 541 U.S. 36 (2004); *Melendez-Diaz vs. Massachusetts,* 557 U.S. 305 (2009); *Bullcoming v. New Mexico,* 564 U.S. 647 (2011); *Williams v. Illinois, 132* S. Ct. 2221 (2012); for discussion see Cheng et al., *supra* note 9 at 56-61.

[61] National Research Council, *Identifying the Culprit: Assessing Eyewitness Identification (2014); Nancy K. Steblay, Scientific Advances in Eyewitness Identification Evidence*, 41 Mitchell L. Rev. 1090, 2015; Svein Magnussen, Annika Melinder, Ulf Stridbeck, and Abid Q. Raja, *Beliefs About Factors Affecting the Reliability of Eyewitness Testimony: A Comparison of Judges, Jurors and the General Public*, 24 Applied Cognitive Psychology 122, 2009.

[62] National Research Council, *supra* note 60 at 40-44 and 110-112.

[63] AAAS report, *supra* note 4.

[64] PCAST addendum, *supra* note 5.