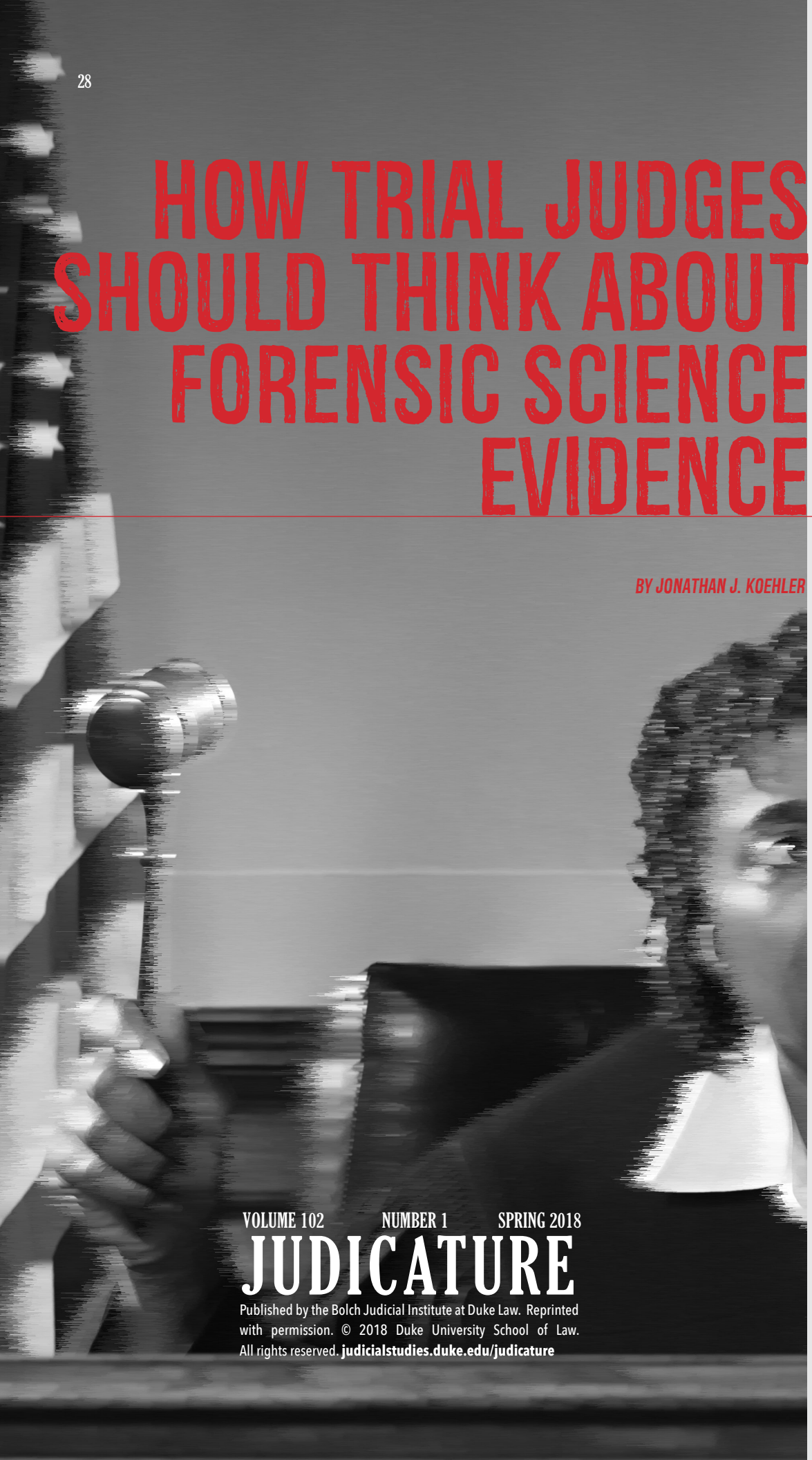
HOW TRIAL JUDGES SHOULD THINK ABOUT FORENSIC SCIENCE EVIDENCE

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.

VOLUME 102 NUMBER 1 SPRING 2018
JUDICATURE
Published by the Bolch Judicial Institute at Duke Law. Reprinted with permission. **judicialstudies.duke.edu/judicature**

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).