VOLUME 102 NUMBER 1 SPRING 2018
JUDICATURE

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.

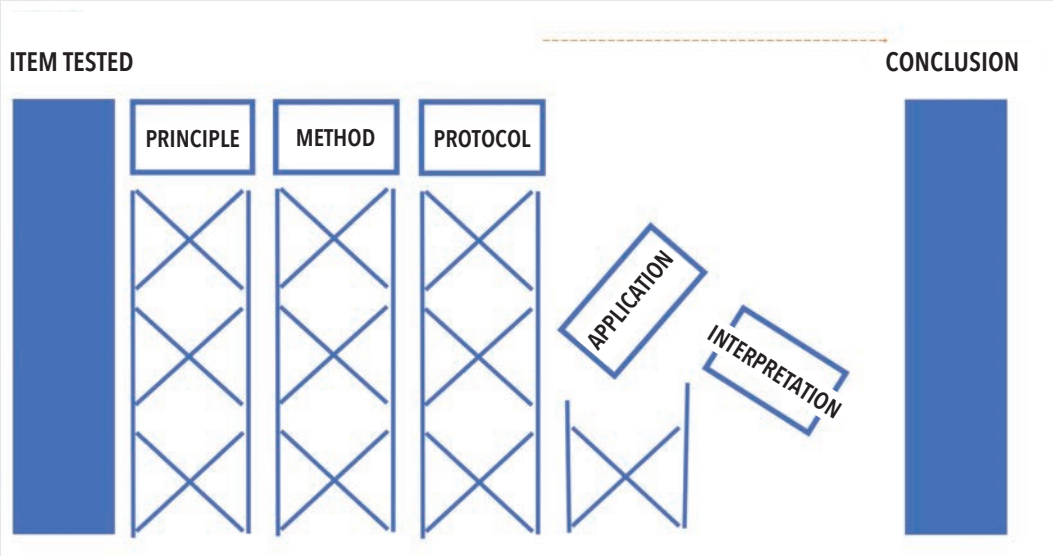
LIMITING TESTIMONY

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.

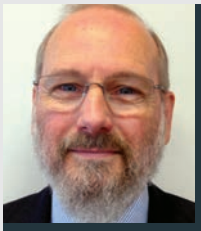
The need for these tools will be lowest where empirical studies provide the most extensive and granular information about sources of variation, limitations, and error rates for a forensic method in a form that jurors can understand and apply directly to the testimony in a particular case. That would 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.