JENNIFER L. MNOOKIN

OF BLACK BOXES, INSTRUMENTS, AND EXPERTS: TESTING THE VALIDITY OF FORENSIC SCIENCE

ABSTRACT

This paper argues that judges assessing the scientific validity and the legal admissibility of forensic science techniques ought to privilege testing over explanation. Their evaluation of reliability should be more concerned with whether the technique has been adequately validated by appropriate empirical testing than with whether the expert can offer an adequate description of the methods she uses, or satisfactorily explain her methodology or the theory from which her claims derive. This paper explores these issues within two specific contexts: latent fingerprint examination and the use of breath tests for the detection of alcohol. Especially in the forensic science arena, I suggest courts have often been seduced by superficially plausible explanations and descriptions of a technique or method, and permitted these to serve as a substitute for empirical testing. Thinking through these two examples illustrates both why evaluating the extent of testing should be the most important method by which courts assess reliability, and why, when other forms of explanatory evidence are readily available, we may nonetheless elect to make use of them. This paper suggests that these descriptions and explanations may at times usefully supplement evidence of testing, but should not generally be substituted for it. Finally, this paper embraces a kind of evidentiary pragmatism, in which the quantum of evidence required to establish legal reliability is determined not in the abstract, but in relation to the evidence that is, or ought to be, available as a result of reasonable research and investigation.

In the now-seminal 1993 case of *Daubert v. Merrell Dow*, the Supreme Court delineated judges' responsibilities vis-à-vis scientific evidence introduced in federal courts. The Court said that federal trial court judges must serve as gatekeepers, assuring that parties' proffered evidence is sufficiently reliable. "Evidentiary reliability will be based on scientific validity," the Court explained. (*Daubert* 1993, 590, n.9)

How, then, ought judges to assess reliability? *Daubert* provides a set of flexible and non-exclusive criteria for judges to consider. These include whether the evidence in question has been tested; whether it is generally accepted in the relevant

DOI: 10.3366/E1742360008000440

https://doi.org/10.3366/E1742360008000440 Published online by Cambridge University Press

expert community; whether it has standards governing its operation; whether it has a known error rate; and whether it has been peer reviewed. As others have noted, these various criteria are something of a theoretical hodgepodge – and indeed, the opinion simultaneously cites with approval thinkers as various as Karl Popper and Sheila Jasanoff (593). But particulars aside, what *Daubert* and its progeny fundamentally stand for is the idea that courts should be in the business of screening expert evidence for validity, while delegating to a large extent questions regarding the particulars of how to do this screening to the sound discretion of the trial judge (*Daubert* 1993; *General Electric Co. v. Joiner* 1997; *Kumho Tire Co. v. Carmichael* 1999).

In what follows I want to focus on a tension which has received little direct attention in the voluminous *Daubert* literature. I wish to examine the relative importance which judges assessing validity ought to give to an *explanation or description* of the method or technique at use, on the one hand, and to the adequate *testing* of its results, on the other. To what extent must a judge aiming to fulfill her responsibility under *Daubert* understand the underlying mechanism, method, or technique that the expert uses to reach her results? For that matter, to what extent need the expert herself – or, for that matter, anyone – understand it in order for its use as legal evidence to be acceptable? Can sufficient testing of the technique be a substitute for this understanding? Is an explanation, an understanding of the method, essential? If not essential, is it sufficient?

My basic argument is that a formal delineation or explanation of method is neither essential nor sufficient. Rather, legal evaluators of the reliability of scientific techniques of evidence, particularly techniques of forensic science, ought to be more concerned with testing than with the adequacy of the explanation of the method. Their inquiry should focus first and foremost on validation – or, more precisely, on the extent to which appropriate empirical testing supports the claims made by the expert – rather than on whether the expert (or, more broadly, the community of experts) can offer a plausible account of the underlying mechanism at work. More specifically, I want to suggest that sufficient testing alone may constitute an adequate method for establishing reliability under *Daubert*, even in the absence of a persuasive explanation of the underlying mechanism or method by which the expert or the technique reaches conclusions. I also want to argue that judges should be wary of being lulled by explanations of how a method works or descriptions of a method's operation that sound intuitively plausible when they are not accompanied by evidence that the method, in actual practice, achieves what it sets out to do. My focus and my examples come from areas of the forensic sciences, though the arguments may also apply beyond this realm as well.[1]

To put the point differently, let us conceive of a technique or method of forensic science, from fingerprint identification to DNA fingerprinting to handwriting identification as if it were a mysterious "black box" that provided us with information relevant to legal cases using methods that remained unknown to us, hidden inside the box and out of our sight. My argument is that it is significantly

https://doi.org/10.3366/E1742360008000440 Published online by Cambridge University Press

less important for judges to pry open this black box than it is for them to establish whether its operation has been tested under conditions similar to those at issue in court. Peering inside the black box to see *how* it works is less critical for an assessment of validity than assessing whether input/output testing shows *that* it works.

Too often, judges may find an explanation of a technique's premises and methods seductive – and a superficial and superficially plausible explanation of a technique may blind judges to the lack of actual data supporting the expert's claims. Certainly in the forensic science arena, this dynamic is both familiar and all-too frequent. Judges assessing forensic science techniques, for example, would do better to focus on evidence about the conditions under which a technique or method succeeds, and how much testing supports the expert's claims. I do not mean to claim that there is no role whatsoever for explanation in the assessment of scientific validity in court. Evidence that unseals the black box and permits the judge to peek inside may indeed be a valuable *supplement* to input/output testing. But it ought not to be a *substitute*.

Simultaneously, it is appropriate for judges to approach their assessments of reliability with a certain dose of pragmatism: specifically, they ought to couple a primary focus on testing with an inquiry into what evidence of reliability is and ought to be available to the court with respect to the specific technique and claims at issue.

I of course recognize that the terms with which I am concerned here – validity, explanation, testing – are at root philosophical and that they are terms which have been analyzed and parsed with tremendous sophistication and detail in the philosophical literature. However, because I am a law professor trained in history and sociology rather than a philosopher, I approach these questions from the bottom up rather than in the abstract. In what follows, I examine two areas within forensic science in which disputes over the relative importance of testing versus method has generated controversy: first, the assessment of the reliability of latent fingerprint identification, and second, disputes over the importance of software source code for assessing the reliability of breath test detection devices.

(1) LATENT FINGERPRINT IDENTIFICATION AND THE ACE-V "METHOD"

In recent years, latent fingerprint identification has been subject to increasing scrutiny. Indeed, in the last decade, there have been numerous legal challenges to the reliability of fingerprint evidence. It is fair to say that these challenges have nearly all failed, in that very few of them have persuaded judges to exclude or even to limit the technique. Nonetheless, these admissibility contests, combined with, first, several high profile cases in which fingerprint identifications turned out to be erroneous, and second, a number of academic critiques of latent fingerprint

https://doi.org/10.3366/E1742360008000440 Published online by Cambridge University Press

examination by academics in a variety of fields, have left the reputation of what was once thought of as forensic science's gold standard somewhat tarnished.

And rightly so. For the method used by latent fingerprint examiners offers less than meets the eye, and the extent to which the method (or the examiners' ability to reach accurate results) has been empirically tested and validated is shockingly limited.

If a latent fingerprint examiner is asked about the method she employs in her work, she will typically assert that she uses the ACE-V methodology. ACE-V stands for analysis, comparison, evaluation, and verification. Of what does this method consist? Let us imagine a latent fingerprint examiner wants to decide whether a print found at the crime scene matches that of a suspect. First, the examiner looks at each of the two prints closely (analysis). Then she examines the prints in relation to each other, noting both similarities and any potential differences (comparison). Next, she evaluates these similarities and possible differences to reach a conclusion about whether or not the two prints come from the same source (evaluation). Finally, a second examiner takes the prints and independently re-analyzes them (though, in many jurisdictions, this second examiner engages in this analysis in full knowledge of the original print examiner's conclusions) (verification).

Latent fingerprint examiners routinely insist that ACE-V is a version of the scientific method and offers a reliable methodology that establishes that fingerprint evidence is indeed a valid science. And many courts have agreed that ACE-V passes muster under *Daubert*. In fact, ACE-V's relationship to the scientific method is tenuous at best; as a methodology it amounts, more or less, to having two people look carefully at fingerprints. While the careful observation and recording of one's observations may be a necessary part of many scientific practices, in and of itself, careful observation does not constitute a method. The simple act of labeling this process of careful looking as a 'methodology' does not make it so. Nor does labeling it scientific tells us anything about its likely validity or error rate.

Surely ACE-V amounts to more than I am suggesting? Yes and no. Latent print examiners do have norms about what kinds of print ridge detail they ought to be looking at, and examiners are trained to search both for relevant details and to assess their contextual relationship with and position on a fingerprint. But latent fingerprint examination lacks any formalized specifications about what is required in order to declare a match. There is no required minimum number of points of resemblance, nor any required quantum of any specific kind of ridge detail. Instead, examiners decide for themselves how much similarity is sufficient to declare a match based on their expertise and experience.

Additionally, latent fingerprint examiners do not employ any statistical information or models in their process. They do not have statistically-based information about the frequency of different friction ridge characteristics. They do not use empirical studies showing how common or how rare different fingerprint details might be. Nor do they have any statistically-validated standard to justify how

https://doi.org/10.3366/E1742360008000440 Published online by Cambridge University Press

many identifying characteristics must be the same on two prints in order to warrant a finding that they 'match.' Many technical obstacles exist to being able to develop a satisfactory statistical measure of the frequency of various ridge characteristics. As of now, there simply is no specified statistical model that is available for latent fingerprint examiners to employ.

The list of difficulties goes on. It is a tenet of latent fingerprint examination that if two prints have even a single dissimilarity, then they must necessarily come from different sources. This, however, begs the question of how one decides what constitutes a dissimilarity. Latent fingerprint examiners lack any formalized criteria for determining when a 'difference' between two prints is merely an artifact of the processes for rendering latent prints visible, and hence can appropriately be ignored, and when an apparent difference actually counts as such. Most of the time, latent print examiners are also not shielded from potentially biasing information about the broader legal case and its context when undertaking their examination. If an examiner knows about other apparently strong evidence linking the defendant to the crime, might this knowledge risk informing the examination of the prints and whether they match? The concerns over a large range of scientific fields that have led to procedures to 'blind' researchers to information that might inadvertently bias them has generally not yet reached forensic science. Both the psychological literature on 'observer effects' and small amounts of empirical testing of the importance of context in the evaluation of fingerprint evidence suggests, however, that we ignore these issues at our peril (Risinger et al. 2002; Dror 2006). Additionally, until recently, the 'verification' phase of ACE-V was also typically not blind – that is, an examiner engaging in a verification knew she was verifying prints declared by an earlier examiner to match, turning verification into at best a kind of 'double-checking,' rather than a second, independent analysis.

Furthermore, although fingerprint examiners do, at times, undergo proficiency tests, these have for the most part been extremely easy, far easier than the challenges that can be faced by examiners in actual casework. In a significant *Daubert* challenge of a few years ago, a fingerprint examiner from the U.K. characterized the FBI's proficiency evaluations as so easy as to be "a joke," (*Llera Plaza* 2002(A), 74) and though the trial court judge did elect to admit the fingerprint identification evidence at issue, he too concurred that the FBI's proficiency tests were "less demanding than they should be" (*Llera Plaza* 2002(B), 566).

Furthermore, in some laboratories, examiners are permitted to take their proficiency tests in groups rather than individually – and these tests are also often not conducted blind. When examiners are aware that they are undertaking a proficiency test, they may be motivated to use a different degree of care in their analysis as a result of this knowledge, making it difficult to extrapolate from these results to their casework performance. Thus, in their current form, proficiency tests may possibly be a check on gross incompetence, but they do not pose challenges equal to those faced by examiners in their actual practices. As a result, we really have very little idea how often fingerprint examiners might make errors within their

https://doi.org/10.3366/E1742360008000440 Published online by Cambridge University Press

casework. Several recent high-profile cases (the most notorious of which involved Brandon Mayfield, held as a material witness in the Madrid bombing on the basis of the FBI's erroneous analysis of a fingerprint) make clear that mistakes can and do happen. How often? The honest answer to that question is that we really do not know (Cole 2005).

In addition, in the typical case, there are not any obvious feedback mechanisms through which an examiner would learn if she had made an error, or even by which she can confirm the correctness of her identification. To be sure, a defendant may, when confronted with the evidence against her, decide to plead guilty or to confess. These might be understood to offer a positive feedback loop, as these actions by the defendant would at first glance seem to suggest the correctness of the fingerprint identification that constituted a portion of the evidence that spurred the defendant's actions. But this is a highly problematic assumption. Fingerprint identification is still perceived by most people – and hence by most jurors – as extremely powerful evidence. That, combined with the large, indeed often enormous, disparities between the sentences offered via plea bargain and those a defendant may face upon conviction, makes a defendant's decision to plea virtually uninterpretable. Perhaps the defendant took the plea because she was guilty. But perhaps she was innocent but risk averse: after assessing the likelihood of conviction based on the apparent strength of the inculpatory evidence and the expected sentence if convicted, perhaps she decided that she would rather plead guilty than risk trial. Nor, with the burgeoning research on false confessions, can we simply presume that the defendant who confesses is necessarily guilty (Drizin & Leo 2004). To take either a confession or a guilty plea as straightforward evidence of the correctness of the fingerprint match is therefore too simple. Moreover, even if we do assume that in the great majority of cases in which a defendant confesses or pleads guilty, the fingerprint match is accurate, that tells us nothing about whether we can be confident of the correctness of the fingerprint match in those cases in which a defendant does neither.

One might expect that this array of limitations and concerns would lead fingerprint examiners to wrap their identifications in a cloak of interpretive modesty. Quite the contrary. Fingerprint examiners claim not merely to be able to describe the similarities between two prints, but that they have the skills to individualize. When they declare a match, they assert that two prints come from a common source to the exclusion of *all other possible sources in the world.* Unlike DNA evidence, which is typically introduced through a statistical measure of the random match probability, fingerprint examiners entirely shun the language of probability. Fingerprint examiners instead claim absolute certainty. They assert that when two prints 'match' they simply must have come from the same person.

This troubling state of affairs has led an array of critics – of whom I am one – to believe that latent fingerprint identification ought to have serious problems under *Daubert.* (Courts, to be sure, have generally not agreed with this assessment.) But at the same time, virtually everyone, even the most adamant critics of

https://doi.org/10.3366/E1742360008000440 Published online by Cambridge University Press

latent fingerprint examination, recognizes that human fingerprints of course have substantial variability and can therefore offer useful information about identification. What, then, ought to be necessary prerequisites to the use of fingerprint evidence in court?

Some of the critics of fingerprint evidence in its current form have suggested that it needs to become a 'science.' For example, one recent article focuses on the need for a significantly greater degree of formalization of method, with criteria for identification that could be specified with precision and therefore tested (Haber & Haber 2008). Others have emphasized that fingerprint examination must be placed on a secure statistical foundation (e.g., Zabell 2005), so that the probabilistic meaning of 'match' can be adequately determined. Both of these suggestions are designed to improve our understanding of *how* fingerprint identification works – they are designed to provide better explanations of the method in use and a deeper understanding of its possible limits.

By contrast, I believe that the single most important requirement should be to require regularly-conducted, well-designed, double-blind proficiency tests to assess fingerprint examiners' actual performance. I certainly do not oppose efforts either to specify the methods employed by fingerprint examiners more formally and with greater detail, and the same goes for endeavors to develop a statistical model of ridge detail that might permit probabilistic assessments of the likelihood that two prints came from a common source. But I do not believe that either of these two developments – each of which is, at present, beyond our collective capacity – ought to be necessary to permit use of fingerprint identification in court, nor that standing alone they ought even to be sufficient.

Rather, 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. Our focus ought to be on developing the best information we can about fingerprint examiners' error rate in actual practice, and carefully designed proficiency tests could largely achieve this goal (Mnookin 2008; Koehler 2008).

In other words, it ought to be acceptable for latent print examiners themselves to operate as a kind of black box. If we know they are getting the right answers a very high proportion of the time over a range of circumstances that mimic what they encounter in actual cases, that ought to be adequate to justify the use of the information they provide as legal evidence in court. Even if we do not fully know how their method operates – indeed, even if they do not fully understand it themselves – this ought not to prevent us from making use of their conclusions if we have good information about how accurate they are and what circumstances might increase the risk of error.

By contrast, we should not settle for explanation of method standing alone. Indeed, the history of the identification sciences in court shows a repeated and dangerous pattern: when judges are provided with an intuitively appealing description of a science of 'matching,' they frequently let poorly specified

https://doi.org/10.3366/E1742360008000440 Published online by Cambridge University Press

explanations of method substitute for a more searching assessment of validity and reliability. From handwriting identification evidence to voiceprint analysis to hair identification and forensic odontology, courts have frequently found a description of how an expert goes about finding a match to be plausible enough that they do not require more (Saks 1998). To be fair, these methodologies came into court well before *Daubert*, and judges did not then conceive their role as requiring anything akin to a careful validity assessment. But these claims to be able to make a 'match' of various kinds had – and still have – a deep intuitive appeal. Moreover, though they required expert interpretation, these identification techniques were nonetheless 'semi-legible' to the layperson, the judge or jury member who could look herself at the specimens at issue and see for herself indicia of similarity or difference. Early experts in handwriting identification, for example, made much of both this intuitive plausibility and the seeming transparency of their method. By appearing to teach judges (and juries) how to engage in handwriting identification, and by offering what looked, at least superficially, like a kind of explanation of their method, they shored up their own claims to offering a legitimate kind of expertise (Mnookin 2001).

One could argue, to be sure, that what forensic sciences have thus far provided to the court is not truly explanation – that ACE-V, for example, is better understood as jargon that masks the lack of either a carefully-specified method or a persuasive explanation of why the technique succeeds. But my point is as much institutional as philosophical: courts have tended to accept at face value the claims of forensic science to be providing something like explanation. Judges are not well-positioned, either institutionally or epistemically, to see easily past claims of these kinds. They have frequently treated ACE-V as if it were a methodology that helps to establish latent fingerprint identification's validity, and as if the existence of this methodology can excuse the dearth of actual testing of the technique's accuracy. In practice, judges operating in all good faith have been seduced by these intuitively plausible claims of an ability to make identifications. Inviting courts to focus instead on the extent of testing might lead judges to frame their inquiry with a somewhat more skeptical eye. Moreover, it ought to incentivize the development of more and better proficiency tests. This would quite clearly increase our knowledge of how well the forensic sciences are actually operating, and it might also help us to understand what conditions and circumstances increase the probability of an erroneous identification.

In addition, a focus on testing rather than explanation may lead courts to look more carefully at issues relating to the 'fit' between the claims made for a technique and the testing that validates such claims. Fingerprint examiners frequently testify at present not only that matches are absolute but that the error rate of the technique is zero. Better proficiency tests could lead us to be able to approximate an error rate with greater accuracy – and whatever this error rate turns out to be, we can say with some confidence that it will not be zero.

https://doi.org/10.3366/E1742360008000440 Published online by Cambridge University Press

Judges, therefore, would be well advised to focus on the degree of testing associated with the claims made by experts rather than emphasizing whether the expert has offered a seemingly plausible explanation of her technique and her conclusion. This is not to suggest that a focus on testing does not generate some potential problems of its own. For example, emphasizing testing raises obvious questions about just how much testing ought to be required. How much testing is enough? What tests – and what results – ought to be required as a precondition for admissibility? How much can the results of testing be extrapolated to similar but non-identical circumstances? At what point does an error rate become too high to warrant admissibility of the technique in question? While these questions do not have a straightforward answer, the point is that they are precisely the right questions for the courts to be asking.

(2) BREATH TESTS FOR THE DETECTION OF ALCOHOL AND THE PROBLEM OF SOURCE CODE

I turn now to a somewhat analogous set of issues that has arisen in a handful of recent DUI prosecutions. The issue of the sufficiency of input/output testing for assessing reliability has been taken up explicitly in a spate of recent cases about the reliability of breath test devices for detecting alcohol. These devices, a mainstay of DUI prosecutions, measure the quantity of alcohol found in a test subject's breath.

In general, the scientific techniques at work within these devices are well-understood. Moreover, DUI, as a general matter, has a more well-developed and complex statutory and regulatory regime than virtually any other area within forensic sciences. While the details of the statutes proscribing driving while intoxicated vary in significant ways across jurisdictions, all 50 states have a *per se* law prohibiting driving with a certain breath or blood alcohol concentration (BAC). Under these *per se* laws, driving or operating a motor vehicle with a BAC above the statutorily defined limit is, by definition, an offense, without regard to any particularized showing of actual impairment. Every state at present prohibits driving with a BAC greater than 0.08 percent, with lower thresholds for commercial vehicles and minors. The passage of these *per se* laws in the states was strongly incentivized by the federal government which tied the passage of such laws to the receipt of federal highway funds.

These statutory schemes put significant pressure on the question of the reliability of the instruments used to detect concentrations of alcohol. Practically speaking, the results of these tests are often wholly dispositive of the question of whether an offense was committed. While preliminary screening devices, typically hand-held breath test detection devices, are frequently used by police officers who suspect intoxication, the results from these devices are generally not admissible in court to prove the requisite *per se* level of alcohol. Only officially approved "evidentiary" devices meet this standard. These evidentiary devices and their use is

https://doi.org/10.3366/E1742360008000440 Published online by Cambridge University Press

extensively regulated by statute, more so than virtually any other kind of forensic science evidence. Indeed, a Federal agency, NHTSA (the National Highway Traffic Safety Administration), evaluates the reliability of new breath detection instruments and determines whether instruments meet designated performance standards, and many states conduct their own supplementary investigations as well.

Certainly it is fair to say that the scientific methods behind these breath test detectors are well understood. Take, for example, the Alcotest 7110, recently adopted as the evidentiary instrument of choice in the state of New Jersey. This instrument makes use of two distinct methods for measuring alcohol. It uses infrared analysis, in which a source inside the device emits an infrared light. This light is sensed by a detector prior to measuring the test subject's sample to establish a baseline. When a breath sample containing alcohol goes into the chamber, the infrared light is partially absorbed by the molecules of alcohol. Therefore not all of the infrared light reaches the detector, and by comparing the difference, the instrument can measure the concentration of alcohol in a person's breath. Simultaneously, this device uses a second method as well, making use of a fuel cell which produces an electrical current. Alcohol causes an increase in the flow of electricity in the current, and this too can be measured and quantified by the machine.

However, while the general principles underlying the device are well-understood, and while the testing of the instrument's accuracy is quite extensive, the instrument itself operates as a kind of 'black box' from the perspective of both the operator and the test-taker. Breath detection devices are, at present, astonishingly automatic. It is not an exaggeration to say that the operator does little more than enter in identification data about the subject (such as name and driver's license number) and press a button. The test subject blows into a mouthpiece (usually twice) and the instrument does the rest, spitting out a ticket with a printed BAC result upon conclusion of the test. Given this high degree of automaticity, it is fair to say that the instrument is controlled far more meaningfully by its internal software than it is by the human being who oversees it.

Over the last few years, defendants in several jurisdictions have argued that any assessment of the instrument's reliability ought to include an evaluation of this internal operating software, or source code. A number of courts have rejected these requests on procedural grounds, arguing, for example, that the source code is not discoverable because it is not in the state's possession or control, or is a protected trade secret of the manufacturer (See, for example, *People v. Robinson* 2008; *City of Fargo v. Levine* 2008).

New Jersey, however, took a different path. The state's decision to switch from the Breathalyzer to the Alcotest 7110 led to a consolidated hearing on the reliability of the new device before the New Jersey Supreme Court. In the course of this proceeding, the defendants requested access to the source code in order to evaluate the underlying procedures through which the instrument operated. Draeger, the manufacturer of the device, refused to hand over the source code. The defendants

https://doi.org/10.3366/E1742360008000440 Published online by Cambridge University Press

"argued that the manufacturer's intransigence forced the Special Master and, by extension, this Court, to rely on 'black box' testing, when only a complete and thorough analysis of the source code used to operate the device would suffice for constitutional purposes." (*State v. Chun* 2008).

Eventually, the Court ordered the manufacturer to turn over the source code information to experts designated by the defendants and by the state, simultaneously establishing procedures designed to protect Draeger's interest in maintaining confidentiality. The New Jersey Supreme Court's description reveals obvious lingering ambivalence about whether this was actually necessary:

> [T]he refusal of Draeger to intervene precluded the Special Master from permitting any testing of the manner in which the device operates, and required him to rely on tests that at best could only demonstrate that the machine reliably appeared to be able to identify correctly, or at least acceptably within the established parameters, the alcohol concentration of a known test sample. There is some logic to that method of proceeding. *If a breath testing device can, reliably and consistently over time, correctly analyze a sample of known alcohol concentration, one might argue that it matters little how the device is able to do so.* Notwithstanding the rather considerable force of that logic, we were persuaded that, in light of the constitutional dimension of the issues before us, Draeger's eventual election to intervene in this matter afforded us the opportunity to permit defendants to engage in the technical analysis of the source code that they had asserted was so necessary to the adequate protection of their rights. (*Chun* 2008, 133, italics added)

The Court recognized that what was at issue was precisely the question of the sufficiency of input/output testing for the assessment of reliability. If the device, when extensively tested using samples of known alcohol concentration, consistently reached accurate results, why did it matter by what precise means the instrument conducted its operations? What was the justification for peering into the "black box" and examining the source code directly? The Court recognizes "the rather considerable force" of the logic suggesting that input/output testing is sufficient. Its decision to permit the defendants' experts access reads as if it was the result of the Court wanting to give the defendant the benefit of every doubt rather than seeming like something the Court independently deemed necessary to a fair evaluation of reliability.

Without quite saying so, the Court was wrestling with the problem of induction. Just because the instrument has proved accurate in past tests, how can we know that it will continue to be so in the future? What if an evil or mischievous programmer had built some quirk into the operating software that would emerge only rarely, or only after it had been in operation for an extensive period? Unlikely, perhaps, but not inherently impossible. More benignly, what if incompetence or oversight meant that some rarely-occurring confluence of particular testing circumstances could lead to inaccuracy? No amount of testing alone could fully eliminate such possibilities.

I confess that I share the *Chun* Court's ambivalence. On the one hand, extensive input/output testing is far more likely to provide useful information about validity

https://doi.org/10.3366/E1742360008000440 Published online by Cambridge University Press

than explanation standing alone. A general explanation of the principles underlying the Alcotest 7110 may be plausible and even sound persuasive, but it will utterly elide the question of whether the device, in actual practice, can succeed in doing what it is supposed to do both accurately and consistently across a range of circumstances. Access to the source code would not be an acceptable substitute for testing either. Even unpacking in detail through a source code examination the steps that the system was supposed to take under a variety of circumstances would be a poor substitute for seeing the device in operation and testing whether the software actually worked up to its specifications. If we had to choose between testing and examining the source code, we will learn more from direct testing and observation of the instrument itself than from peering at the design of its internal mechanisms.

On the other hand, is it not possible that the source code could usefully supplement direct black box testing? Examination of the source code, for example, would likely put to rest any concerns about a mischievous programmer that had purposefully built the instrument to mislead after a certain number of tests or in specific circumstances. In fact, in the *Chun* litigation, the experts for each side reached strikingly different conclusions about the integrity of the source code (as is typical for warring experts in our adversarial system). The Special Master in the case largely accepted the conclusions of the State's experts, and did not find that the source code revealed any problems that fundamentally undermined the instrument's validity. The source code examination did reveal, however, according to the court, several minor problems worth correcting. It found, for example, a buffer overflow error which could lead the instrument to record an inaccurate result in those instances in which a third breath sample was required (which occurs when the first two breath sample results diverge to an extent beyond normal measurement tolerances), as well as several other details in need of correction. Minor? Perhaps. But all things considered, not a bad set of things to know. Or to put it differently, the examination of the source code and the fact that very few significant problems were revealed by it, when coupled with the extensive input/output testing of the device, could appropriately increase the Court's confidence in the instrument's validity, compared with the assessment it could have reached from the input/output testing standing alone.[2]

It matters quite a lot to this analysis that the source code in fact *existed* – that it was a tangible, human-made thing that existed in the world and could be revealed with only a modicum of effort. By contrast, no fully-specified statistical model of fingerprinting is available. It simply does not yet exist. Perhaps additional time, research and effort will eventually lead to one, but right now, it is not something available in the world but being kept from the courts. Rather, it lies outside our collective knowledge and expertise.

Emphasizing this distinction – between what evidence is reasonably available and what is still beyond our reach – invites courts to embrace a certain degree of

https://doi.org/10.3366/E1742360008000440 Published online by Cambridge University Press

evidentiary pragmatism in their assessments of reliability. Assessing the evidence required for determining reliability ought not to take place in a vacuum, nor is validity simply an on/off switch insensitive to a particular given context. Judges ought to assess the validity of an expert's claims in relation not only to that evidence readily available, but also with an eye toward what evidence could have been available if the parties had taken all reasonable measures to produce it. In short, *Daubert* ought to be understood as a relational standard rather than as a call for a fixed quantum of proof. To put the point differently, courts should be encouraged to take a "best evidence" approach to questions of reliability: the evidence we deem necessary may legitimately depend on the extent of evidence reasonably available with regard to reliability (Nance 2003; Nance 1998).

Recognizing this relational, gradational aspect to validity assessment helps to explain why we may, on the one hand, permit fingerprint identification evidence in the absence of any statistical model that undergirds it, and simultaneously believe that the court might in some circumstances be warranted in insisting upon an examination of the source code that undergirds the operation of a device for measuring breath alcohol. Under this approach, there is a meaningful difference between evidence we have but which a party wishes not to reveal and evidence which one party calls for but which no one on earth could provide even if they exercised all good faith and all reasonable efforts to acquire it. Availability matters.

This pragmatism should, however, be coupled with an emphasis on the primacy of testing as the key criterion for assessing validity under *Daubert*. It is true that there may, on occasion, be events so rare that subjecting them to testing is itself impractical. These cases may raise difficult policy questions about whether other methods for assessing validity ought to suffice. But except in those quite atypical instances, courts would do well to focus their inquiries regarding experts on the question of whether the experts' claims have been adequately tested. If, for example, proficiency testing establishes that the expert generally reaches accurate conclusions in similar circumstances, the lack of a fully-specified or defined methodology ought not to prompt courts to exclude the evidence.

In other words, if fingerprint identification, or questioned document identification, or toolmark identification, or ballistics analysis can be shown through appropriate – and appropriately difficult – proficiency tests to reach accurate conclusions a sufficiently high proportion of the time, then the conclusions of these experts (appropriately calibrated to the degree of accuracy suggested by these proficiency tests) ought to be deemed admissible in court – whether or not the experts can explain their methodology with a high degree of specificity or in a persuasive way. But concomitantly, precisely because there is no reason why such proficiency tests cannot be conducted, the failure to conduct them ought indeed to lead us to be skeptical of these sorts of evidence at present.

Even if an expert altogether lacks a formally specified methodology, I would suggest that she could legitimately testify in court if the outcomes she reaches

https://doi.org/10.3366/E1742360008000440 Published online by Cambridge University Press

have been validated by testing. I recognize that this analysis invites being taken to its logical extreme: to wit, ought we to permit the evidence of a clairvoyant, if she had been shown to be accurate in the preceding 99 instances in which she had attempted to see the future? My provisional answer is 'maybe' – or, more precisely, I recognize that there will inevitably, and perhaps quite appropriately, be interplay between our prior conceptions of the likelihood that the expertise at issue can produce reliable knowledge and the degree of testing that we require. We might therefore be more skeptical of the claims that the clairvoyant had passed proficiency tests than the equivalent claims made by a document examiner.

But the somewhat unsettling example of the clairvoyant does not change my two fundamental points: first, that we ought not let our prior belief about the probable existence of the expertise in question *substitute* for testing (which is, I think, a plausible way to understand what the courts are in fact doing with a variety of forensic sciences at present); and second, that the court's focus should be on whether the extent of testing that has been conducted warrants the particular kinds of claims made by the expert in question. If an arson investigator were to insist that she 'just somehow knows' whether a fire has a natural origin, but could not explain the details of her method in a convincing way, I would still permit her testimony in court if she could show that she reached accurate results a very high proportion of the time in cases that were meaningfully similar to this one.[3] True, the problem of induction would inevitably be present. Perhaps the proficiency testing that appeared to establish the accuracy of her conclusions was simply the result of very good luck, and perhaps this instance is the moment when the expert's luck would cease to hold. This is, to be sure, a risk. But under *Daubert*, trial court judges in the Federal system are supposed to do nothing more than make preliminary determinations of reliability. The question for the court is thus not whether the expert's conclusions are correct, but rather, whether they are sufficiently validated for the jury to hear them as evidence (*Daubert* 1993). Testing, rather than explanation, is the better lodestar for this inquiry.

CASES CITED

City of Fargo v. Levine, 747 N.W. 2d 130, 133 (N.D. 2008).
Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 573 (1993).
General Electric Co. v. Joiner, 522 U.S. 136 (1997).
Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999).
People v. Robinson, 2008 WL 2389137, 1 (N.Y.A.D., 2008).
State v. Chun, 943 A. 2d 114, 133 (N.J. 2008).
Trial Transcript, United States v. Llera Plaza (2002) (Llera Plaza (A)).
United States v. Llera Plaza (II), 188 F. Supp.2d 549 (E.D. Pa. 2002) (Llera Plaza (B)).

https://doi.org/10.3366/E1742360008000440 Published online by Cambridge University Press

REFERENCES

Cole, Simon A. 2005. "More than Zero: Accounting for Error in Latent Fingerprint Identification." *Journal of Criminal Law and Criminology* 95: 985–1078.

Drizin, Steven A. & Richard Leo. 2004. "The Problem of False Confessions in the Post-DNA World." *North Carolina Law Review* 82: 891–1007.

Dror, Itiel E. et al. 2006. "Contextual Information Renders Experts Vulnerable to Making Erroneous Identifications." *Forensic Science International* 156: 74–8.

Haber, Lyn & Ralph Norman Haber. 2008. "Scientific Validation of Fingerprint Evidence Under Daubert." *Law, Probability & Risk* 7: 87–109.

Koehler, Jonathan J. 2008. "Fingerprint Error Rates and Proficiency Tests: What They Are and Why They Matter." *Hastings Law Journal* 59: 1077–100.

Mnookin, Jennifer L. 2001. "Scripting Expertise: The History of Handwriting Identification Evidence and the Judicial Construction of Reliability." *Virginia Law Review* 87: 1723–845.

Mnookin, Jennifer L. 2008. "The Validity of Latent Fingerprint Identification: Confessions of a Fingerprinting Moderate." *Law, Probability & Risk* 7: 127–41.

Nance, Dale. 1988. "The Best Evidence Principle." *Iowa Law Review* 73: 227–97.

Nance, Dale. 2003. "Reliability and the Admissibility of Experts." *Seton Hall Law Review* 34: 191–253.

Risinger, D. Michael et al. 2002. "The Daubert/Kumho Implications of Observer Effects in Forensic Science: Hidden Problems of Expectation and Suggestion." *California Law Review* 90: 1–56.

Saks, Michael J. 1998. "Merlin and Solomon: Lessons from the Law's Formative Encounters with Forensic Identification Science." *Hastings Law Journal* 49: 1069–141.

Zabell, Sandy L. 2005. "Fingerprint Evidence." *Journal of Law and Policy* 13: 143–79.

NOTES

1 While *Daubert* invokes testing (and Carl Hempel) in an effort to define the boundaries of 'scientific knowledge,' I do not mean to suggest anything in what follows about a line of demarcation between science and non-science. My emphasis on testing does not derive from a belief that testability is what distinguishes science from other domains of knowledge; rather I want to suggest that judges should focus on validity testing because it offers the best criterion for assessing the validity of an expert's claims.

2 Whether requiring the source code to be revealed was worth it – whether the likely expected increment of additional information warranted putting the state, the manufacturer, and the courts themselves to the effort and expense associated with disclosure is a far more difficult question. How the need for evidence to improve the assessment of reliability intersects with institutional and efficiency concerns for keeping litigation costs in check is a critical question but well beyond the scope of this article.

3 I recognize that I am begging several questions here, including what error rate would be acceptable, as well as whether generalized proficiency tests results ought to be adequate, or whether we ought instead to require specific evidence of *this* particular examiner's proficiency. On the latter question, it may be that questions of the degree to which the expert community shares a method must be part of the court's inquiry. To the extent

https://doi.org/10.3366/E1742360008000440 Published online by Cambridge University Press

that a technique has a specified methodology shared by its practitioners, it may be more acceptable to use general proficiency tests, if well-designed, as a proxy for the proficiency of an individual practitioner. But when an expert lacks a formalized method, or the methods in use vary across practitioners, the fact that *some* practitioners may be extremely accurate would not seem to speak meaningfully to the question of whether *this one* is.

Jennifer L. Mnookin is Vice Dean for Faculty & Research and Professor of Law at the UCLA School of Law. She writes primarily in the area of evidence with a particular focus on both historical and contemporary analyses of expert evidence and the forensic sciences. The author thanks Scott Brewer, Mark Greenberg, Michael Pardo, and participants in the Dartmouth Conference on Evidence and Law for helpful conversations and suggestions.

https://doi.org/10.3366/E1742360008000440 Published online by Cambridge University Press