The Disregarded Necessity: Validity Testing of Forensic Feature-Comparison Techniques

*Michael J. Saks**

I. INTRODUCTION

In growing knowledge about the material world, is anything as important as insightful, clever, and brilliant ideas that aim to penetrate to the underlying logic of the way things work? Yes: the testing of those ideas. In normal science, generating ideas and testing their validity go together like a horse and carriage, a hand in a glove, and a chicken and an egg. Even the most beautiful of ideas must be discarded if they cannot survive fair and robust attempts to validate them[1]—or, more to the point: rigorous tests designed to disconfirm them if they are not valid.[2] Without the latter, the former is illusory.

Normal science recognizes empirical testing as central and indispensable to the scientific enterprise. Whenever normal scientists take a good look at forensic science, they are astonished at the paucity of such

* Professor, Sandra Day O'Connor College of Law at Arizona State University.

[1] For discussion of forensic science efforts to get by with poor validity testing, see D. Michael Risinger & Michael J. Saks, *A House With No Foundation: Litigation-Directed Research in the Criminal Justice System*, 20 ISSUES SCI. & TECH. 35 (2003).

[2] D. Michael Risinger et al., *Exorcism of Ignorance as a Proxy for Rational Knowledge: The Lessons of Handwriting Identification "Expertise,"* 137 U. PA. L. REV. 731 (1989). *See generally* DAVID L. FAIGMAN ET AL., MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY (2017).

testing.[3] In normal science, the centrality of empirical testing has led to a massive infrastructure that exists to carry out such testing—from the education of scientists in research skills to the eventual publication of their research findings and all the steps in between—with immense resources devoted to enabling that constellation of research activities.

For manifold reasons, one of the most striking things about the forensic sciences is the absence of any remotely comparable infrastructure, or even an appreciation of the need for testing. As one wag expressed the problem: the custom of forensic science is to testify first and validate later, if ever. This problem formed the title of a prominent article devoted to the problem: "The Need for a Research Culture in the Forensic Sciences."[4]

In their way, the courts have long shared science's recognition of the importance of validation. For as long as Anglo-American courts have been entertaining expert testimony, they have tried to evaluate what was being proffered by developing various criteria in an effort to admit the valid and exclude the non-valid.[5] The clearest modern expression of that necessity is found in *Daubert*: "In a case involving scientific evidence, evidentiary reliability will be based upon scientific validity."[6]

For a century, though, the forensic sciences did not attempt to empirically validate their techniques or the ideas underlying them, and by all indications, the courts were unable to recognize those deficiencies.[7] In the wake of the reports by the National Research Council (NRC)[8] and the President's Council of Advisors on Science and Technology (PCAST),[9] and the advent of the National Commission on Forensic Science (NCFS) and the Organization of Scientific Area Committees (OSACs), things might be

[3] NAT'L RES. COUNCIL OF THE NAT'L ACAD. OF SCIS., STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES: A PATH FORWARD (2009), https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf [hereinafter NAS REPORT]; Sylvester James Gates, *New FSSB Member Dr. Sylvester James "Jim" Gates, Jr. on the Importance of Science in Standards*, OSAC NEWSLETTER (Mar. 2017) ("I bring to the conversation an outsider's opinions and perspective on how to best align the practices of forensic science with the mores, standards, and experimental designs common to other fields that use the designation of 'science.'").

[4] Jennifer L. Mnookin et al., *The Need for a Research Culture in the Forensic Sciences*, 58 UCLA L. REV. 725 (2011).

[5] FAIGMAN ET AL., *supra* note 2, § 1.

[6] Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579, 590 n.9 (1993).

[7] *See* FAIGMAN ET AL., *supra* note 2, §§ 29–44; Michael J. Saks, *Merlin and Solomon: Lessons from the Law's Formative Encounters with Forensic Identification Science*, 49 HASTINGS L.J. 1069 (1988).

[8] NAS REPORT, *supra* note 3.

[9] PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., FORENSIC SCIENCE IN CRIMINAL COURTS: ENSURING SCIENTIFIC VALIDITY OF FEATURE-COMPARISON METHODS (2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf [hereinafter PCAST].

changing.[10]

II. NECESSITY AND AVOIDANCE

A bit more might be said regarding that climate and whether it is changing. A perfect storm has prevented the forensic sciences from developing a research culture and producing a stream of validation studies on their most important assumptions. As long as the conditions that prevented validation research from being undertaken continue to exist, such work will continue to not be undertaken.

Among the ingredients of this perfect storm was that the inventors and major adopters of the techniques were not, for the most part, scientists. Police agencies were not staffed with scientists. The vast majority of forensic scientists continue to have minimal science education, and certainly are not trained as researchers. In addition to a lack of personnel with the skill to design and conduct validation research, there also was a lack of time and funding to do such work. Even if practicing forensic scientists had the skills, the time, and the support (financial and otherwise) to undertake validation research, incentives all pushed against undertaking such research.

What, they might have asked, could be gained? The courts already universally admitted their offerings, and the public could not have been more credulous. As one judge reflected on the situation in regard to one pattern-comparison field: "[F]ingerprint evidence has been afforded a near magical quality in our culture. In essence, we have adopted a cultural assumption that a government representative's assertion that a defendant's fingerprint was found at a crime scene is an infallible fact . . . ."[11] Research that confirmed their claims and the public's beliefs could not make things look any better to judges or jurors than they already did. Research that showed the claims to be overblown, as it almost certainly would, and for some of them perhaps wildly overstated, could only reduce their current status (as perceived by judges and jurors).

---

[10] Or not yet. *See* Spencer S. Hsu, *Sessions Orders Justice Dept. To End Forensic Science Commission, Suspend Review Policy*, WASH. POST (Apr. 10, 2017), 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=.2ac849625783. But maybe. *See* Sadie Gurman, *Justice Dept. Tries to Shore Up Forensic Science, Testimony*, ASSOCIATED PRESS (Aug. 7, 2017), https://www.apnews.com/ef513b48286345c9b0a210afd1a37fe1. No, not really. *See* Radley Balko, *Deputy AG Announces New Forensic Science Working Group but Still Doesn't Grasp the Extent of Problem*, WASH. POST (Aug. 7, 2017), https://www.washingtonpost.com/news/the-watch/wp/2017/08/07/deputy-ag-announces-new-forensic-science-working-group-but-still-doesnt-grasp-the-extent-of-problem/?utm_term=.3d2e570e1eec.

[11] State v. Quintana, 103 P.3d 168, 171 (Utah Ct App. 2004) (Thorne, J., concurring) (citations omitted).

The behavior of the courts did not set forensic science on its unscientific path, but it does perpetuate the problem. If the courts create no disincentive for offering junky expert testimony, then the courts create no incentive for improvement (that is, validating, testing, and bringing reports and testimony within the bounds of validated knowledge). Barry Fisher, former director of the Los Angeles County Crime Laboratory made these points candidly and not infrequently. At a conference in April 2009, discussing the NRC Report, he spoke of having expected the "free ride" that the courts had given to him and his colleagues over the decades to come to an end. He recalled that on reading the *Kumho Tire*[12] opinion the day it was issued, he gasped, "Oh my God, we're gonna get slammed in the courts." But, as he told the conference audience a decade later, "I'm still waiting."[13] In short, if the courts are content with junky expert testimony, then validity testing, when it comes, will not have been prompted by fear of unfavorable rulings by district judges, notwithstanding what the Supreme Court has said about validation being the touchstone of expert evidence admissibility.

Courts, of course, are not the only institutions that could evaluate what particular forensic sciences do and then influence what they offer. In modern times, two forensic sciences were found to be so wanting in validity that they were withdrawn by the FBI and, in consequence, withdrawn altogether. They disappeared from the courts without the courts doing a thing. Those fields were voice spectrography[14] and comparative bullet lead analysis[15]—sent to their graves by NRC reviews and the resulting reports. These are blunt consequences. One would expect that an active program of validity testing would determine which beliefs and techniques are sound (leading to their retention) and which are unsound (leading to their abandonment). In that way, fields would evolve with better knowledge continually replacing poorer knowledge.

A third area, arson investigation, looks more like that. The field belatedly did its own testing of many of the "arson indicators" on which its members had come to rely to evaluate whether a fire was accidental or had been set intentionally. Nearly twenty of those indicators were found incapable of distinguishing arson from accident, and were removed from the field's official toolbox.[16] Presumably, in the wake of those changes, fewer

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

[13] Barry A.J. Fisher, Crime Lab. Dir., L.A. Cty. Sheriff's Dep't, Comments at the Forensic Science for the 21st Century: The National Academy of Sciences Report and Beyond Conference at Arizona State University (Apr. 3, 2009).

[14] NAT'L RES. COUNCIL, ON THE THEORY AND PRACTICE OF VOICE IDENTIFICATION (1979), https://www.nap.edu/read/19814/chapter/1.

[15] NAT'L RES. COUNCIL, FORENSIC ANALYSIS: WEIGHING BULLET LEAD EVIDENCE (2004), https://www.nap.edu/read/10924/chapter/1.

[16] John Lentini, *Fires, Arsons and Explosions*, *in* FAIGMAN ET AL., *supra* note 2, § 37.

errors of both the false positive and false negative kind are made. Additional disciplines might join those already in the forensic graveyard.[17]

These examples of now-deceased forensic sciences and misinforming techniques make clear that validation, in science and in law, can lead to far more than finicky tweaks. For decades, junk had been offered to courts under a false flag of science, and erroneous verdicts followed. Post-mortems of DNA exoneration cases found forensic science to be implicated in more false convictions than any factor other than eyewitness errors.[18] Neither the forensic science establishment nor the courts performed, respectively, their scientific or legal validity testing duties, and so the science and the nonsense were offered and received as an undifferentiated mix. Only future validity studies can discover how much more nonsense remains to be filtered out.

III. THREE APPROACHES TO VALIDITY TESTING OF FORENSIC FEATURE-COMPARISON

Speaking at the National Institute of Justice's Annual Conference on Science and the Law in 2001, I outlined three approaches to validity testing of forensic feature-comparison techniques: the black-box model, the DNA-model, and the basic-research model.[19]

A. *The Black-Box Model*

Building on the proficiency studies pioneered by Joseph Peterson and his colleagues,[20] the black-box model involves presenting forensic examiners with samples of known origin (that is, known to have been fired by the same or different weapons, bitten by the same or different teeth, written by the same or different persons, and so on). The value of such studies depends on the quality of their design (as with all research). Most obviously, if trivially easy tasks are presented, and examiners perform spectacularly well, or if absurdly difficult tasks are presented and examiners do spectacularly poorly, we don't learn much.

To learn the most from such studies, the samples to be examined would have to be representative of the population of items at issue, in all of that population's dimensions. That probably would require quite a large sample

---

[17] Perhaps the most likely candidate is bitemark identification. Michael J. Saks et al., *Forensic Bitemark Identification: Weak Foundations, Exaggerated Claims*, 3 J. L. & BIOSCIENCES 538 (2016).

[18] *See* BRANDON L. GARRETT, CONVICTING THE INNOCENT: WHERE CRIMINAL PROSECUTIONS GO WRONG (2012); Michael J. Saks & Jonathan J. Koehler, *The Coming Paradigm Shift in Forensic Identification Science*, 309 SCI. 892 (2005).

[19] NAT'L INST. OF JUST., SCIENCE AND THE LAW: 2001 AND 2002 NATIONAL CONFERENCES (2004), https://www.ncjrs.gov/pdffiles1/nij/202955.pdf.

[20] JOSEPH L. PETERSON ET AL., CRIME LABORATORY PROFICIENCY TESTING RESEARCH PROGRAM (1978).

for each forensic specialty. But the data collection need not be done all at once. More important than any sort of global error rate or calculation of sensitivity or specificity per field of specialization would be a map of the different kinds of patterns a specialty confronts and the field's performance in relation to each subtype.[21]

Some of the stimulus markings likely will be so obvious and easy that anyone would reach the correct conclusion and forensic experts would add nothing of substance to the fact finding. Others likely will be so challenging, and examiners consequently do so poorly, that their opinions would do nothing but mislead the factfinder into thinking something is known when it is not. (I recall one proficiency test where all of the examiners reached the same answer (100% reliability) and all of them were wrong (0% validity)).[22] Most feature-identification problems, however, would likely fall somewhere between those extremes, along a continuum of less and greater accuracy. With the help of an empirically-derived performance map, in their reports and testimony, examiners would be able to inform attorneys and factfinders of the level of accuracy/error typically attained when opining on that (apparent) sub-type of evidence.

PCAST has outlined the elements of what it would regard as well-done black-box studies of forensic feature-matching.[23] These would consist of large numbers of samples of materials (questioned plus one or more knowns), large numbers of examiners, independent researchers conducting the study and the analyses, and the calculation of an overall false-positive error rate and sensitivity for the method. Ideally, such testing would be done in the normal stream of work, and therefore not obviously be seen as a test, but merely one more case among many.

In some circumstances, it is not necessary to actually know the origins of the markings in order to test the accuracy of examiners other criteria might suffice. For example, in one study by the FBI, completed microscopic hair comparison examinations were later tested against the findings of mitochondrial DNA (mtDNA) tests. Although the mtDNA cannot definitively include a suspect as being the contributor of the hair, it can rule someone out by finding she or he could not have been a contributor.[24]

---

[21] A more focused version of this concept (focused on more difficult tasks, the point at which failure occurs) has been adopted as a recommendation by the NCFS. NAT'L COMM'N ON FORENSIC SCI., VIEWS OF THE COMMISSION: FACILITATING RESEARCH ON LABORATORY PERFORMANCE (Sept. 13, 2016), https://www.justice.gov/archives/ncfs/page/file/909311/download.

[22] D. Michael Risinger, *Handwriting Identification*, *in* FAIGMAN ET AL., *supra* note 2, § 33 (summarizing the findings of the 1984 Forensic Sciences Foundation proficiency test).

[23] PCAST, *supra* note 9.

[24] Max M. Houck & Bruce Budowle, *Correlation of Microscopic and Mitochondrial DNA Hair Comparisons*, 47 J. FORENSIC SCI. 964 (2002) (finding, *inter alia*, that of 26

The whole effort would benefit from making a small change in (or in addition to) the way examiners record their opinions in their reports. That change would be to obtain more of a continuous (and less of a categorical) measure of the examiner's confidence that the markings examined appear indistinguishably similar. That would facilitate signal-detection analysis of the data, and thereby add to the insights that can be derived from the data collected.[25]

Such an approach could be harnessed to studies of different protocols for performing examinations (allowing discovery of which produce more and which less accurate results), different training methods (same), and so on. That would contribute to continuing improvement in the discipline's techniques, training, and performance.

The chief blessing of the black-box model is that we do not need to know anything about what any given type of forensic expert does or what goes on inside the black box. We need only input well-sampled and well-characterized stimuli to examinations and collect and analyze the output. Other blessings are the possibility of mapping that performance in relation to particular kinds of stimuli and the possibility of continuous improvements.

The PCAST Report relies entirely on the black-box model. But the black-box approach is cursed as well as blessed. With the black-box approach, there is no once-and-done test because people and their performance can change faster when what they are doing is entirely subjective. Human-measuring instruments generally are more variable, more volatile, and more influence-able by extraneous information and motivation. They are harder to monitor and manage.[26] Because we do not know what the examiner is actually doing (inside the black box), this kind of validity testing must become a permanent activity of forensic science. If as a group examiners changes what they do over time, performance might improve or degrade compared to previous levels. This would need to be tracked continuously, would require updating "maps," and would require updating testimony going forward to reflect current findings.

mtDNA non-matches, 9 (35%) had been erroneously identified as matches by the hair examiners).

[25] Victoria L. Phillips et al., *The Application of Signal Detection Theory to Decision-Making in Forensic Science*, 46 J. FORENSIC SCI. 294 (2001).

[26] Consider the challenges here of meeting *Daubert*'s maintenance-of-standards factor. Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579, 594 (1993).

B. *The DNA Model*

Although the PCAST Report does not go beyond the black-box model as the solution to the problem of validation in forensic feature-comparison, the black-box model fails to capture an important component of the central goal (indeed, the central traditional claim) of forensic identification. In addition to looking at two patterned images and describing their similarities and differences, forensic identification aims to reach an inference as to whether the two images were created by one and the same source—that its features locate an image in a class of n=1. If two images were created by two different objects, which happen to be indistinguishably similar, concluding that they shared a common source would constitute a false positive error.

The locus of such errors is not in the perceptual skills of examiners, but in the origins of the patterns they are examining, and the distribution of those patterns within the population of such patterns. If the examiners are wrong about something important, it is their assumption that no two patterns can be misleadingly similar, and their speculation that examiners' perceptual and cognitive skills will always enable them to distinguish patterns that were produced by different sources. That assumption—the uniqueness of forensically relevant patterns—is in the process of being discarded as unproved and probably unprovable (if it has not already been discarded) and replaced with something akin to population genetics.[27] In the instance of DNA typing, it literally is population genetics.

If the model of DNA typing is applied to forensic feature-comparison fields, it would look something like the following. Samples would be drawn from the universe of objects with which a field is concerned—fired bullets, tool marks, handwriting, footprints, tire marks, etc. The patterns observed in the sampled objects would be systematically measured. Developing workable measurement systems for those patterns, it often is said, presents a far greater challenge than measuring alleles for DNA typing ever did. For one thing, they can be highly complex images, untethered to anything as straightforward as ACGT. Unlike DNA, some of those other objects present problems of dynamics, producing changes on the fly. For example, the barrels of guns change somewhat with each firing and so the patterns produced over time change. Or consider handwriting, where a single source produces constantly fluctuating patterns within some varying range.

[27] Mark Page et al., *Uniqueness in the Forensic Identification Sciences—Fact or Fiction?*, 206 FORENSIC SCI. INT'L 12 (2011); Michael J. Saks & Jonathan J. Koehler, *The Individualization Fallacy in Forensic Science Evidence*, 61 VAND. L. REV. 199 (2008); William C. Thompson & Simon A. Cole, *Psychological Aspects of Forensic Identification Evidence*, *in* EXPERT PSYCHOLOGICAL TESTIMONY FOR THE COURTS (Mark Constanza et al. eds., 2006).

If and when the ability to systematically measure the images that constitute those other types of forensic patterns is achieved, it will undoubtedly be with the assistance of scientists from other fields, such as topology, mathematics, statistics, and computer science. Those measurements will facilitate estimation of the population frequency of each feature, or combinations of features, found in bullet striations, handwriting, fingerprints, bitemarks, and so on. With that information, it would become possible to calculate the random match probability that a conclusion regarding a questioned and known pattern were similar enough that they might have shared a common source.

With black-box methods, the random match probability (which reflects the rarity or commonness of a pattern in the relevant population of objects) gets confounded with the measurement error associated with the examiner-as-measuring-instrument. In the case of DNA typing, human error accounts for a larger share of the total error in the overall process than the variations in DNA do. With some or most of the other techniques, the sources of error might be the other way around. And that fact will never emerge from a strictly black-box approach, nor will we be able to take steps to ameliorate what is susceptible of amelioration.

C. *The Basic-Research Model and Gap Analysis*

The essential idea of the basic-research model is that all of the beliefs that a forensic technique depends upon for its validity (many of which are assumptions and speculations, some well-grounded hunches, others so obviously true that they are not susceptible to reasonable challenge) would be explicated and the most important of them would be subjected to research designed to determine the extent to which the belief is valid. This could be a massive effort, undertaken by a wide range of scientists and forensic scientists.

One of the most helpful first steps would be to complete "gap analyses" on every forensic field of interest. The essential idea has been to form working groups for each forensic science subfield, composed of forensic scientists, relevant conventional scientists, and statisticians. The group would scrutinize the major beliefs of the subfield (those beliefs on which its techniques and claims of ability rest), compare those beliefs to existing empirical studies testing the beliefs, and issue a report describing the propositions that have been validated and those that have not yet been validated (or have been invalidated). The gap between what needs to be tested and what has been tested will form a research agenda for anyone and everyone concerned with the validation of each particular forensic science field.

Although these gap analyses would seem to be an essential first step,

getting them going has been a slow road for something so important to the overall project. The federal agencies working on advancing the forensic sciences have not undertaken or funded such work. So far as I am aware, it is not on their agenda. Private organizations came to the rescue in the form of funding by the Laura and John Arnold Foundation (LJAF) supporting gap analyses to be overseen by the American Association for the Advancement of Science (AAAS).[28] But before much work had been completed under that grant, LJAF withdrew its financial support. So gap analyses remain a major unmet need.

CONCLUSION

I have discussed three (I believe these are *the* three) major approaches empirical research aimed at validating forensic feature-comparison expert evidence could take, along with some discussion of the benefits and difficulties associated with each.

It should be obvious that the body of research findings (when body of research findings come into being) would provide a basis for a court to make an informed and refined admissibility decision, and juries to learn how dependable (or not) the testimony they are hearing can be expected to be. Until then, judges and juries have little alternative but to guess about validity and take the witness's opinion on faith.

It also should be obvious that validity testing cannot be the end of the challenge of providing factfinders with sound and trustworthy guidance in regard to the type of evidence under discussion. No validity testing method can prevent examiners from presenting conclusions they did not actually reach or exaggerating conclusions they did reach.[29] Entirely different tools are required to ensure that witnesses offer factfinders the results of well-conducted case-specific analyses accurately, clearly, and honestly presented.

---

[28] Initially, David Faigman, Joseph Peterson, and I presented the idea of conducting gap analyses to the John D. and Catherine T. MacArthur Foundation. MacArthur went so far as to fund a private conference where a group of forensic scientists, other scientists, scholars, judges, and lawyers met for a day to discuss with MacArthur staff the potential value of such an effort. In the end, MacArthur declined to undertake the project. Sometime later, Peterson was invited to the Laura and John Arnold Foundation (LJAF) headquarters to explain the gap analysis idea to them. LJAF then offered the project to the American Association for the Advancement of Science (AAAS), and, as noted in the text, later withdrew the funding.

[29] *See* GARRETT, *supra* note 18; Saks & Koehler, *supra* note 18. *See also* Spencer S. Hsu, *FBI Admits Flaws in Hair Analysis Over Decades*, WASH. POST (Apr. 18, 2015), https://www.washingtonpost.com/local/crime/fbi-overstated-forensic-hair-matches-in-nearly-all-criminal-trials-for-decades/2015/04/18/39c8d8c6-e515-11e4-b510-962fcfabc310_story.html?utm_term=.4868b576ecac ("The Justice Department and FBI have formally acknowledged that nearly every examiner in an elite FBI forensic unit gave flawed testimony in almost all trials in which they offered evidence against criminal defendants over more than a two-decade period before 2000.")