PNAS PERSPECTIVE OPEN ACCESS

The scientific reinvention of forensic science

Jonathan J. Koehler[a,1], Jennifer L. Mnookin[b], and Michael J. Saks[c]

Edited by Thomas Albright, Salk Institute for Biological Studies, San Diego, CA; received April 14, 2023; accepted May 26, 2023

Forensic science is undergoing an evolution in which a long-standing "trust the examiner" focus is being replaced by a "trust the scientific method" focus. This shift, which is in progress and still partial, is critical to ensure that the legal system uses forensic information in an accurate and valid way. In this Perspective, we discuss the ways in which the move to a more empirically grounded scientific culture for the forensic sciences impacts testing, error rate analyses, procedural safeguards, and the reporting of forensic results. However, we caution that the ultimate success of this scientific reinvention likely depends on whether the courts begin to engage with forensic science claims in a more rigorous way.

forensic science | paradigm shift | error rate | testing | courts

1. The Transformation of Forensic Science

It would be hard to overstate the importance of the transformation that is underway throughout most of the forensic sciences. For much of the 20th century, evidence from a variety of forensic sciences was routinely admitted in state and federal courts with very little scrutiny of whether it had either substantial validity or a genuine scientific foundation. Experts, usually associated with law enforcement and often without any formal scientific training, testified in court to the validity and outsized accuracy of the techniques and their conclusions. Courts admitted their testimony, generally without limitation or careful scrutiny, based on assurances from the forensic science community that the techniques were accurate, effective, and broadly accepted as valid. Assertions unsupported by empirical validation sufficed. The scientific authority of forensic science testimony rarely faced significant challenge from the opposing party, and the occasional challenges that were offered were nearly always unsuccessful.

The story began to change when DNA evidence emerged in the late 1980s and early 1990s. After initial breathless enthusiasm by courts about this transformative new identification technique, highly credentialed scientists identified meaningful concerns regarding how to "translate" laboratory DNA assessments for courtroom use. Several judges excluded DNA evidence to ensure adequate vetting by the scientific community. In the 1990s, scientists from various core disciplines including genetics, statistics, and psychology engaged in lively and sometimes contentious debates in peer-reviewed, scientific journals about the forensic use of DNA profiling, including such matters as population genetics, error rates, standards for defining a DNA match, and communicating the evidentiary meaning of a match. Those debates, and two DNA reports issued by the National Academy of Sciences (NAS), impacted the way DNA evidence was treated in court, creating a greater focus on scientific validity than existed for prior forensic techniques. Also in the 1990s, the Supreme Court decided a trio of critical cases on the use of scientific and other expert evidence in the courts. These cases emphasized that the Federal Rules of Evidence gave judges the responsibility to engage in judicial "gatekeeping" to determine whether that scientific and expert evidence was sufficiently reliable and valid to be admitted in court (1–3).

By the early part of the 21st century, a shift to a more scientific paradigm for the forensic sciences was observable, though still in its infancy (4). This shift represented a move from a framework of "trusting the examiner" to "trusting the method." Rather than relying on untested foundational assumptions, and assurances from witnesses that their training and experience makes their confident conclusions accurate and trustworthy, legal scholars, scientists, and some forensic practitioners began endorsing a more scientific model that prioritizes common and detailed protocols, empirical testing, and more moderate, data-driven knowledge claims. Some have hinted that a scientific paradigm shift has already occurred (5, 6); others see little evidence of a shift (7). Most likely, the transformation remains a work in progress: Notable progress has been made on some fronts, but significant concerns remain (8).

In some areas, when scientific reviews established that available empirical science did not support experts' claims, entire subfields of forensic science that had contributed to criminal convictions for decades ceased (e.g., bullet lead analysis) or ceased using discredited principles (e.g., fire and arson analysis). In other areas, scrutiny led to reduced credibility and a shift away from exaggerated claims (e.g., microscopic hair analysis). However, other fields, such as bitemark identification, continued despite adverse scientific reviews (9).

Some forensic subfields, such as single-source DNA identification, survived scientific scrutiny quite well. Latent fingerprint identification, which has been scrutinized more than most other forms of pattern identification evidence, has survived as well, although it has scaled back on its claims in recognition of the role that human factors and subjectivity play in reaching conclusions (10). Firearms evidence is gaining attention from the scientific community, and weaknesses

Author affiliations: [a]Northwestern Pritzker School of Law, Chicago, IL 60611; [b]Office of the Chancellor, University of Wisconsin-Madison, Madison, WI 53706; and [c]Sandra Day O'Connor College of Law, Arizona State University, Phoenix, AZ 85004

Author contributions: J.J.K., J.L.M., and M.J.S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: jay.koehler@northwestern.edu.

Published October 2, 2023.

in its scientific foundation and reporting traditions have been identified (11).

In what follows, we discuss how the move to a more empirically grounded scientific culture in the forensic sciences impacts testing, error rate analyses, procedural safeguards, and the reporting of results. Whereas there can be no debate that forensic science claims must be grounded in both relevant testing and data, legitimate open questions remain about how best to make the forensic sciences "scientific." How should errors and mistakes by forensic practitioners be defined and counted? How should conclusions be reported? These questions are currently being discussed and debated by the scientific community. Responsibility for implementing recommendations from the scientific community ultimately rests with the courts. Unfortunately, few courts have undertaken serious gatekeeping of forensic science evidence. We discuss this problem and conclude by examining how to build on institutional and structural opportunities to assure that this vital reinvention of forensic science proceeds.

2. Testing

The shift to a truly scientific framework in the forensic sciences requires attention to empirical testing of the techniques and methods employed under realistic conditions. As PCAST (12) notes, "Scientific validity and reliability require that a method has been subjected to empirical testing, under conditions appropriate to its intended use, that provides valid estimates of how often the method reaches an incorrect conclusion" (p. 27 and p. 118). Empirical testing is a sine qua non for moving from a "trust the examiner" to a "trust the methods" ethos.

Although scientifically-minded people understand the importance of empirical testing in any scientific endeavor, calls to test the accuracy of forensic science claims are relatively recent. For most of the 20th century, few asked forensic scientists to provide empirical proof that they could do what they claimed. The training, knowledge, and experience of the examiner, coupled with assurances that the method used was generally accepted in the forensic community, were deemed sufficient to admit nearly every forensic science that was proffered in court in the 20th century. Once admitted, forensic scientists commonly offered conclusions with 100% confidence and claimed, with little evidence, a 0% error rate (13). Although some optional forms of certification existed, little attention was paid to whether, or how, forensic examiners should be required to pass proficiency tests or what those tests should include. Nor did judges require any form of testing or certification as a prerequisite to allowing forensic testimony.

2.1. History. Most forensic sciences were raised, if not always born, in the world of law enforcement for the purpose of helping police identify criminals. The granddaddy of forensic identification, anthropometry was invented by Alphonse Bertillon in the Paris Prefecture of Police in the 1880s. This technique involved making systematic measurements of bodies of prisoners to assist with their identification at a later date if they were using aliases (14). Fingerprints soon proved to be a more useful means of identifying criminals, and courts eagerly admitted this evidence without serious inquiry into the scientific underpinnings of the claim that experts could accurately identify the source of partial prints recovered from crime scenes. At no point did the fingerprinting method face the rough-and-tumble questioning of a scientific discipline where everything is questioned and tested, progress is incremental, and cautious, tentative claims are the norm. Over time, other forensic science techniques were invented and introduced on the basis of assurances from practitioners rather than persuasive evidence from rigorous scientific tests.

2.1.1. *DNA evidence*. When DNA technology burst onto the legal landscape in the late 1980s—a technology that, unlike most forensic disciplines that came before it, derived from basic scientific disciplines—the broader scientific community took notice. Initially, this impressive technology was received with great enthusiasm. But questions about its courtroom use soon emerged. In People v. Castro (15), through the involvement of talented defense counsel and distinguished scientists as defense experts, substantial concerns about how laboratory DNA science was being "translated" for courtroom use gained prominence (16). In the wake of Castro and several cases that followed, the National Research Council of the National Academy of Sciences convened a blue-ribbon committee to examine DNA evidence, and a flurry of additional scientific activity ensued. Geneticists, statisticians, evolutionary biologists, psychologists, and others debated, tested, and wrote about various aspects of this new technique in prestigious scientific journals. It was not forensic science business as usual; this time there would be no deference to authority or to the say-so of a narrowly defined forensic community.

The National Research Council (NRC) ended up writing two reports, four years apart, about DNA evidence (17 [NRC I] and 18 [NRC II]). We do not focus on the reports as a whole but limit our attention to their respective treatments of testing in the forensic sciences.

Two types of proficiency tests were needed to legitimate the use of DNA profiling in court. One type of test would address issues that were internal to the forensic sciences. These tests address matters such as whether examiners can follow the protocols for a particular technique and whether different examiners and different laboratories obtain identical (or nearly identical) results on identical samples. A second type of test focused more on matters external to the day-to-day workings of forensic science analyses, such as helping triers of fact assign appropriate weight to DNA evidence. This goal is best accomplished through another type of proficiency test designed specifically to identify accuracy and error rates under various casework-like conditions (19). As NRC I noted, "Interpretation of DNA typing results depends not only on population genetics, but also on laboratory error" (17, p. 88). This report referenced the results of a DNA proficiency test conducted a few years earlier that identified a false positive error rate of 2%. Noting that some of the early proficiency tests were "less than ideal," NRC I stressed that for DNA typing, "laboratory error rates must be continually estimated in blind proficiency testing and must be disclosed to juries" (17, p. 89).

This testing recommendation was largely ignored by the forensic science community and the courts. Moreover, some

influential forensic science voices actively counseled against error rate testing on the specious grounds that error rates are irrelevant to individual cases because they change over time (testimony from a leading FBI scientist in United States v. Llera Plaza (20, p. 510). At trial, prosecutors argued that the source opinions of DNA examiners were reliable. With few exceptions, trial judges gave little weight to defense arguments that DNA evidence should be limited or excluded when error rate tests had not been performed.

NRC II offered a different perspective on tests designed to measure laboratory error rates than that taken by NRC I. NRC II offered four arguments against performing such tests: 1) error rates are unknowable because they are always in flux, 2) error rates never translate directly into an estimate of error for a given case because each "particular case depends on many variables," 3) general error rate estimates "penalize the better laboratories," and 4) an "unrealistically large number of proficiency trials" would be required to obtain reliable error rate estimates (18, p. 85–86). Although these arguments were widely rebutted (21–23), this report stifled calls for empirical testing and made it difficult for defense attorneys to argue that the reliability of any proffered forensic science method is unknowable without such data.

Fourteen years later, yet another National Research Council report was issued (24 [NAS]). This report examined a variety of non-DNA forensic science disciplines (latent prints, shoeprints, toolmarks, hair, etc.) and concluded that nearly all had failed to test their fundamental premises and claims. According to NAS, testing requires an "assessment of the accuracy of the conclusions from forensic analyses and the estimation of relevant error rates" (24, p. 122). A follow-up report by the President's Council of Advisors on Science and Technology (PCAST) argued even more forcefully for empirical error rate testing programs: "Without appropriate estimates of accuracy, an examiner's statement that two samples are similar—or even indistinguishable—is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact" (12, p. 6).

We thus see a variety of particularized approaches to proficiency testing in the forensic sciences across blue-ribbon analyses of the topics. Three of the four reports noted above emphasized the importance of proficiency testing and the development of empirically grounded error rates. Although there are challenges to developing meaningful error rates, the program of proficiency testing called for in the PCAST and various NAS reports is an indispensable part of the evolving scientific framework in the forensic sciences. Error rate proficiency tests have now been conducted with forensic examiners in various subfields including latent prints (25, 26), firearms and toolmarks (27, 28), and footwear (29). These studies are important steps forward and have prompted interest in how error rates should be computed and reported. A consensus has not yet emerged. Far from signaling a discipline in disarray, ongoing research and sophisticated debates depict a field that is undergoing a scientific transformation.

2.2. Evolving Error Rate Studies. In the late 1900s, proficiency testing in the forensic sciences focused mainly on the issue of examiner competence. Could the examiner conduct a proper analysis using simple exemplars, and did the conclusions reached by different examiners agree? To the extent error rates were computed from these proficiency tests, it was clear that those rates should be considered with a grain of salt. The study participants were usually volunteers who knew that they were being tested and who may or may not have collaborated with others or otherwise examined the test samples differently than they treat casework samples. The test providers often were not disinterested parties, and the samples used were less challenging than many that appear in actual cases. Although some of these testing problems remain, efforts have been made in recent years to employ realistic samples and to blind examiners to the fact that they are working with test samples rather than casework samples (30, 31).

2.3. Inconclusives. A focus on testing and accuracy raises important correlative questions: Precisely what counts as an error and how should error rates be computed? There is no single "correct" error rate (32, 33). False-positive error rates, false-negative error rates, and false discovery rates are all different, legitimate error rates. But even when there is agreement about which error rate is of interest, scientists might not agree about what "counts" as an attempt (or trial) and what "counts" as an error. If examiners always reached either an identification conclusion (i.e., that two patterns derive from the same source) or an exclusion (i.e., they come from different sources) for all sample pairs in a test situation, it would be a simple matter to compute, say, a false-positive error rate. It would be the number of times the examiner reached a "same source" conclusion divided by the number of sample pairs that were known to have been produced by a different source.

But forensic examiners do not always reach a firm binary source decision. Depending on the subfield, they might reach more limited judgments, such as leaning toward identification, high degree of association, association of class characteristics, limited association of class characteristics, inconclusive, indications of nonassociation, and leaning toward exclusion.* We discuss the wisdom of categorical conclusions later. For now, we simply note that error rate computations are not straightforward when an examiner reaches a conclusion other than identification or exclusion for a given paired comparison. Because all pairwise samples are, as a matter of ground truth, either produced by a common source (corresponding to a conclusion of identification) or by different sources (corresponding to a conclusion of exclusion), any conclusion other than identification or exclusion cannot be factually correct. This raises the question: Should conclusions other than identification or exclusion be classified as errors? If not, should these comparisons be included in the error rate denominator?

Some scholars have argued that under particular circumstances, uncertain conclusions (e.g., "inconclusive") should be scored as correct or incorrect and should be included in error rate computations (34). According to this argument, inconclusives should be scored as errors when the available information—as judged by qualified experts or by the set of

*Similarly, examiners are often permitted to conclude that samples are "unsuitable" or "insufficient" for reaching any conclusion.

tested examiners themselves in aggregate—suggests that one of the two conclusive decisions could in fact be reached by a competent examiner. Dror (35, pp. 1036–1037) goes so far as to say that, even when an examiner correctly concludes that two samples came from the same source, that decision should be scored as a false-positive error when a panel of experts or group of other examinees regard the comparison to be inconclusive.

Others have argued that inconclusives should not be scored as errors or counted in error rate computations on grounds that when examiners fail to offer a conclusive decision, they are neither wrong nor right because they have not made a claim about the underlying state of nature (36, 37). According to this view, neither a panel of independent experts nor a wisdom-of-the-crowd approach provides a dependable gold standard for ascertaining when a pairwise comparison should be deemed inconclusive (38). Indeed, experts are most likely to disagree with one another on hard cases which, of course, are also the cases where examiners will be tempted to offer an inconclusive decision.

Resolution of this debate is complicated by the practical reality that forensic scientists might be motivated to minimize their reported error rates. If inconclusives are not treated as errors, then examiners might be incentivized to minimize their reported error rates in known test situations by deeming all but the most obvious comparisons inconclusive, even if they might reach a definitive conclusion about many or even most of those same stimuli in real-world casework. Conversely, if inconclusives are treated as errors, examiners might be incentivized to reach conclusions on even the most difficult cases and thereby increase the risk that innocent people are convicted based on faulty forensic science. Misuse of the inconclusive category is likely to be reduced when blind testing is broadly implemented and when examiners provide weight-of-evidence testimony rather than source conclusion testimony. This very debate, and the sophistication of the engagement with this set of questions about measuring error, is a welcome development.

3. Procedural Reforms

For more than a century, the forensic science enterprise in the United States has been controlled and often staffed by law enforcement agencies. This may not be surprising given that police are responsible for investigating crimes, and forensic scientists have the ability to collect and examine evidence in a wide range of cases. But forensic science should not be the exclusive tool of law enforcement for several reasons. First, for the adversary system to work as intended, all parties—including criminal defendants—need to have equal access to forensic science resources. Second, the scientific status of the forensic sciences is compromised by its close association with one side. If crime laboratories are beholden to the needs of law enforcement, they might be discouraged from pursuing scientific investigations that are not aligned with the interests of law enforcement (24, pp. 78–79; 39, p. 775). Relatedly, if forensic scientists see themselves as working in partnership with police and prosecutors, subtle contextual and cognitive biases might creep into their work at various stages.

3.1. Adversarial Allegiance. There has long been concern that expert witnesses who are retained by one side or the other in legal cases will, intentionally or unintentionally, slant their conclusions and testimony in favor of the party retaining them (40). Psychologists theorize that experts see themselves as part of a team and often develop a so-called "myside bias" (41) or "adversarial allegiance" to their team and teammates (42). In one controlled experiment, 108 forensic psychologists evaluated the risk posed by certain sex offenders at the request of either the prosecution or the defense. After reviewing and scoring four case files using standard risk-assessment instruments, the psychologists who thought that they had been hired by the prosecution viewed the offenders as posing greater risks than did the psychologists who thought that they had been hired by the defense (43).

3.2. Bias. The tendency to favor one's own side in an adversarial setting is one of many demonstrated psychological influences (or biases) on human judgment and decision. These biases may be perceptual, cognitive, or motivational in nature. Perceptual biases commonly refer to situations in which a person's expectations, beliefs, or preferences affect their processing of visual stimuli (44). For example, a latent print examiner might "see" a point of similarity between two prints after having noted several other points of similarity between the prints, whereas another examiner—or even the same examiner—might not see the similarity absent an expectation that the two prints share a common source. Cognitive biases refer to systematic distortions in thinking that occur when people are processing information. Confirmation bias is a well-known cognitive bias in which people seek, interpret, and recall information in ways that tend to confirm their prior beliefs (45). Motivational biases, such as motivated reasoning, refer to the phenomenon in which our wishes distort our interpretations of events (46). The significance of these overlapping biases for forensic science work is that they might affect what examiners choose to look at, what they see when they look, and the conclusions that they reach about what they have seen.

Research shows that irrelevant contextual, cognitive, and motivational factors can alter the judgments and decisions of forensic scientists in many areas, including fingerprint (47), handwriting (48), firearms (49), DNA (50), pathology (51), forensic anthropology (52), digital forensics (53), bloodstain pattern (54), and forensic odontology (55). The takeaway point of these studies is not that forensic science evidence is fatally flawed. The point is that forensic scientists, like other scientists (56, 57), are subject to potentially significant biases that should be examined empirically and minimized where possible.

3.3. Reforms to Minimize Bias. Despite the ubiquity of subtle biases in human judgments (58), people do not readily recognize that their own judgments and decisions could be biased (59). Unsurprisingly, this reluctance has been observed in the forensic science community. When a small group of psychologists and forensic scientists debated the risk of bias in forensic judgment in a scientific journal in the late 1990s, some forensic scientists argued that their disciplines were objective (hence unbiased) and that potentially biasing

information therefore need not be withheld from examiners (60). Two decades later, a survey of 403 forensic scientists suggested that this view may still be common. Most of the survey respondents did not think that their own judgments were influenced by cognitive bias, and most did not agree that examiners in their domain "should be shielded from irrelevant contextual information" (61, p. 455). Regardless of whether practicing forensic scientists support efforts to guard against unwanted influences, it is incumbent on the broader scientific community to continue researching potential sources of bias and to continue proposing reforms designed to blunt the impact of bias on forensic judgments.

Perhaps the most important reform is blind testing and blind review. Training in most scientific fields includes learning how scientific judgments and choices might be tainted by subtle psychological forces. This problem is best addressed in human research by blinding investigators and participants alike to the participants' condition (e.g., placebo or treatment). Similarly, in fields that rely heavily on subjective judgments—as many pattern-matching forensic sciences do—it would seem important to prevent analysts from receiving extraneous information that could affect their judgments about the patterns they analyze. In forensic science, blind analysis requires an administrator or case manager to provide examiners with case information on a need-to-know basis. Trace samples recovered from crime scenes (i.e., unknown samples) should be examined thoroughly prior to the introduction of reference samples (i.e., known samples). Knowledge about features of known samples, like knowledge about other aspects of the case, could inadvertently cause an examiner to see features in the unknown sample that are not there or fail to see features that are there (17).

Similar precautions should be taken for verifiers, i.e., examiners who are called on to provide a second opinion. These examiners should be unaware of their role as verifier of the conclusions offered by another examiner. Such knowledge could create a confirmation bias that affects the verifier's forensic perceptions and judgments.

Scientists have recommended various blinding procedures for the forensic sciences. These include sequential unmasking (62), case manager models (63), and evidence line-ups (64). Sequential unmasking minimizes bias by blinding examiners to information about known samples until after the examiners have completed an initial review of the unknown samples. Information related to the known samples that is required for the examiner to draw additional conclusions is "unmasked" as needed. Whereas separate analyses of unknown and known samples will generally work well for DNA and fingerprint analysis, a modified version of this procedure is needed for fields such as firearms and handwriting where the known sample provides information needed for a proper examination of the unknown sample. Sequential unmasking has been implemented on occasion in the United States (65) and is employed as a working standard for fingerprint and DNA evidence at the Netherlands Forensic Institute and at the Dutch National Police for DNA (66). Recently, extensions of this technique have been proposed (67, 68).

The case manager method minimizes bias by assigning a forensic "manager" to interact with investigators and to participate in decisions related to what is tested and how a "blind" examiner conducts those tests. The manager then tells an examiner what to do without revealing other case-relevant (or potentially biasing) information. In evidence line-ups, known reference samples that are not the source of the unknown sample are provided to the examiner at the comparison stage along with a reference sample from the suspected source of the unknown. In the context of an eyewitness lineup, this "filler-control procedure" (69) purportedly reduces errors that incriminate innocent suspects by spreading the errors among a set of fillers as well as the innocent suspects (70). This technique, which could be costly to implement broadly (69), may reduce false positive errors in forensic contexts as well (71).

Growing attention to bias-reducing reforms, though implemented only to a limited degree thus far, suggests that the forensic sciences are beginning to recognize that examiners may be influenced by irrelevant contextual knowledge. Behavioral science research holds the key to identifying procedural guardrails that should be erected to reduce unintentional bias.

4. Examiners' Conclusions and Reporting

4.1. Categorical Reporting. Forensic scientists in many subfields offer one of three categorical conclusions when comparing an unknown (questioned) sample to a known (reference) sample: exclusion (the paired samples come from different sources), individualization (the paired samples come from the same source), or inconclusive (insufficient basis for excluding or individualizing). Exclusions arise when an examiner determines that there are important identifiable features in one of the samples that are not present in the other sample. That determination is left to the judgment of the individual examiner (72). When examiners feel that they lack sufficient evidence that two samples come from different sources, they must decide whether there is enough evidence to conclude that the pair come from the same source. An individualization—sometimes referred to as an identification—is a conclusion that a particular item or person is the one and only possible source of an unknown item of forensic evidence.[†] Despite the long history of reaching individualization conclusions in most forensic sciences, it is an unscientific practice that should be abandoned.

4.2. Individualizations Are Not Scientific. Individualization has long been central to the forensic science enterprise.[‡] Examiners make individualizations in most of their casework (73). Until recently, such testimony was routinely offered with "100% certainty"[§] and assurances of a 0% error rate.[¶] Although vestiges of this type of hyperbole remain, several

[†]For shoeprint evidence, "An identification means the shoe positively made the questioned impression and no other shoe in the world could have made that particular impression" (98, p. 347).

[‡]"The concept of individualization is clearly central to the consideration of physical evidence. Our belief that uniqueness is both attainable and existent is central to our work as forensic scientists" (99, p. 123).

[§]"Latent fingerprint identifications are subject to a standard of 100% certainty" (100, p. 8).

[¶]Responding to a question by 60-Minutes interviewer Leslie Stahl, Stephen Meagher, the former head of the FBI's latent print unit, said that the chance that a reported fingerprint match is in error is "zero" (101).

forensic professional associations now warn their members not to engage in these practices.

However, the individualization claims themselves are nearly as problematic from a scientific standpoint as the exaggerated ways in which those claims are sometimes made. Individualization claims exaggerate what the underlying science can reveal (7, 74–76). A scientist cannot determine that there is no chance that any object other than a particular known sample could be the source of an unknown sample simply because the known and unknown samples share many features (77). When forensic scientists offer individualization conclusions, they are merely offering personal speculation that markings on one of the samples that are not shared by the other sample are unimportant for source determination purposes and that they believe that the samples show sufficient similarity to conclude that they share a common source.

4.3. Abandon Source Opinions and Source Probabilities. The individualization problem cannot be solved by adding a caveat that an individualization is a personal opinion rather than a scientific statement or that it is made to "a reasonable degree of scientific certainty," as had become common in recent years (78). An examiner who offers such an opinion would still be engaged in an unwarranted "leap of faith" (76). Moreover, empirical research shows that such caveats have little impact on the weight that people assign to the forensic testimony (79, 80).

Furthermore, if individualization testimony is abandoned, it should not be replaced by a statement that provides an estimate of the probability that the samples in question were produced by a common source. First, most forensic disciplines do not have extensive data on the frequency with which the various markings appear in various populations or statistical models that reveal the frequency with which particular markings appear in particular combinations. Therefore, no scientific basis exists for estimating the chance that observed similarities between items were merely coincidental. Second, even in disciplines where such data have been collected (e.g., DNA) or are being collected (e.g., fingerprints), it would still be inappropriate to use those data to provide source probability estimates. According to Bayesian logic, these estimates require the examiner to take account of the prior probability that the known source is the actual source of the unknown sample before reaching a conclusion about the source probability in question. The prior probability is informed by a variety of nonforensic considerations, including the existence and strength of other evidence in the case that the forensic scientist should not and likely would not know. Even when the forensic scientist does know the nonforensic facts of a case, that knowledge and its corresponding impact on the forensic scientist's beliefs are not relevant at trial. Instead, jurors' own prior beliefs about the source of the forensic evidence, based on other evidence in the case, should inform their source probability estimates.

4.4. Provide Weight of the Evidence. How then should forensic examiners provide information to a factfinder? There is broad agreement in the scientific community that forensic scientists can and should confine their testimony to providing information pertinent to the weight of the forensic evidence (81, 82). The question to be addressed is how much support do the results of the forensic analysis provide for the proposition that the unknown and known samples share a common source? Note that this is a different question from how likely it is that the two samples share a common source. Triers of fact should make the latter judgment for themselves by updating their initial beliefs about the common source hypothesis with the additional weight provided by the results of the forensic analysis.

4.4.1. Likelihood ratios. There is also an emerging consensus in the scientific and statistical communities that likelihood ratios (LRs) are the most appropriate tool for identifying the strength of forensic evidence (10, 83–85).[#] In its most common form, the LR measures the strength of support that the forensic findings provide for the hypothesis that two samples share a common source relative to the alternative hypothesis that the two samples do not share a common source. If E denotes the evidence from the forensic analysis and CS denotes the hypothesis that the two samples share a common source, then the LR is P(E|CS)/P(E|-CS). In words, the LR is the probability of obtaining this forensic evidence if the two samples came from a common source divided by the probability of obtaining this evidence if the two samples did not come from a common source.

At an abstract level, the LR is an appealing way to report forensic science evidence. In practice, however, it raises a set of challenges. Aside from a relative dearth of data, a significant obstacle to employing LRs to assess evidentiary weight is that it often is not obvious what values to use for the LR numerator and denominator. Even when LRs are computed using reliable data, human judgment usually plays a significant role. For example, reasonable people might disagree about the size and composition of the reference population used to inform the denominator of the LR. Consequently, the size of the LR may vary, sometimes by orders of magnitude.

Choices related to how to handle the risk of human error can also affect the magnitude of the LR. When the risk of such errors is ignored, LRs may become astronomically large. But when estimates of the rates at which recording errors, mislabeling errors, and sample mix-ups are incorporated into LR computations, the resultant LRs will typically be smaller (86). Whether the risk of error is expressly included in the LR computation or provided to jurors in some other way, this risk is always present, and it should place an upper limit on the weight assigned to the forensic evidence.

Misinterpretation poses another obstacle to employing LRs to describe the strength of forensic evidence (87). Studies show that people commonly transpose conditional probabilities and thereby end up treating LRs as posterior odds ratios (88). That is, rather than using LRs as a measure of the weight of evidence, people mistakenly treat LRs as if they directly answer the question, "What are the odds that these two samples come from a common source?" The error of confusing LRs with posterior odds ratios is committed by laypeople, judges, attorneys, and even the experts who present this evidence at trial.

4.4.2. Verbal scales. Some scholars have proposed using verbal scales and qualitative expressions to convey forensic conclusions. For example, a popular scale in Europe describes LRs < 10 as providing slight support/limited support for the source proposition, LRs between 10 and 100 as providing moderate support, LRs between

[#]Log-LRs provide equally rigorous measures of probative value.

100 and 1,000 as providing moderately strong support, etc. (83, p. 64). This well-intentioned idea should not be implemented absent empirical evidence that people give appropriate weight to the evidence that is described using those qualitative terms. For example, if studies show that people treat, say, a 10,000:1 LR as if it were a 100:1 LR when the term "more likely" is used, then a different qualitative phrase is needed. It is not appropriate to simply assign verbal labels to LRs without knowing how people interpret those labels. Preliminary research suggests that some verbal scale expressions are treated roughly in accordance with their corresponding LRs, but some are not (89).

Even as the forensic sciences continue to evolve, it will likely take years before conclusory individualizations are replaced by more scientifically justifiable weight-of-evidence measures such as LRs, verbal scales, or some other probabilistic indicator. A recent survey of 301 fingerprint examiners found that 98% of respondents report categorically rather than probabilistically and that a large majority regard probabilistic reporting to be inappropriate (90). To the extent that examiners in other forensic fields hold similar beliefs—and that prosecutors persuade judges that categorical reporting serves the interests of justice—change may be slow in coming. Further research on how factfinders hear and receive evidence must continue to be a priority.

5. Law

What role have the courts played in improving the scientific quality of forensic science? How can the courts do better? For centuries, courts have appreciated both the value and risk of inviting expert witnesses to help factfinders find their way to the truth of disputed facts. Where specialized knowledge can cast useful light, it would be foolish to disregard it. On the other hand, parties in our adversarial legal system are motivated to present experts only when their testimony will advance the advocate's case, regardless of whether their words illuminate underlying truths.

Courts and other rulemaking bodies have developed various legal tests calculated to facilitate the screening of expert evidence. One hundred years ago, in Frye v. United States (91), a court turned to the intellectual market for guidance. Only those propositions and techniques that had "gained general acceptance in the particular field in which it belongs" would be admissible (91, p. 1014). The Frye test, which has its merits, also exposed the courts to the substantial risk that those who stood to benefit most from the admission of certain types of expert evidence might be called upon to vouch for questionable evidence if the "particular field" was defined too narrowly. Over subsequent decades, judges variously employed the Frye test, related tests, and, often, no test at all to screen experts, including forensic science experts. As noted earlier, many different types of forensic science were admitted based simply on the say-so of the few who practiced the technique at issue.

In 1993, the US Supreme Court held that the Federal Rules of Evidence (promulgated in 1975) did not incorporate Frye's general acceptance test. Instead, judges must determine whether the methods used by proffered experts were reliable and valid, although the Court held that "general acceptance" could be one element of that inquiry. According to the Court, the "overarching subject" of "[t]he inquiry envisioned by Rule 702 … is the scientific validity and thus the evidentiary relevance and reliability—of the principles that underlie a proposed submission" (1 pp. 594–595). Daubert's focus on scientific validity is consistent with efforts to increase a scientific approach within the forensic sciences. However, judges may not have the scientific training necessary to know whether "the principles that underlie a proposed submission" have been adequately tested and validated.

Whether or not this point can serve as explanation or excuse, the fact is that when called on to evaluate the proffers of forensic science, courts have not done well. As NAS observed, "Forensic science professionals have yet to establish either the validity of their approach or the accuracy of their conclusions, and the courts have been utterly ineffective in addressing this problem" (24, p. 53). Rather than engage with the underlying science, most trial judges simply opted to follow past practice and allow proffered forensic science evidence to reach the jury. In the wake of this NAS report, numerous courts made modest gestures toward a more engaged assessment of forensic pattern evidence, limiting it around the edges (i.e., prohibiting claims of zero error rate or 100% certainty) or noting the lack of empirical support with surprise. But nearly all forensic science pattern evidence continued to be admitted.

PCAST sought to help the courts fix this problem by providing specific guidance to the courts for assessing the validity of feature-matching forensic science evidence (e.g., DNA, hair, fingerprints, firearms, toolmarks, and tire tracks). Not surprisingly, the guidance focused on rigorous empirical testing and the estimation of accuracy and error rates for the different methods.

Earlier we noted that several fields of forensic science—including bullet lead comparison, microscopic hair identification, and arson indicators—have been transformed or abolished following serious scientific reviews. Notably, the judicial system did not initiate, and barely even contributed to, these transformations. The courts have not led. Indeed, the courts have often not even followed, as some of these unvalidated techniques continue to be admitted.

Whether the courts will ultimately choose to a) follow the mandates of Daubert and the guidance provided by PCAST, or b) remain "utterly ineffective" at holding the forensic sciences scientifically accountable for their claims, is not yet clear. Although it has been business as usual in most post-PCAST cases, there are some signs of more full-throated, robust engagement, and even occasional exclusions [see, e.g., People of Illinois v. Winfield (92), excluding firearms evidence].

Thanks to Daubert, Federal Rule of Evidence 702, the 2009 NAS report and the 2016 PCAST report, judges indisputably have both the authority and the tools to insist that forensic evidence has an adequate scientific foundation. But they have only rarely availed themselves of this power. As the primary consumers of forensic science evidence, the courts can hold the forensic science community's feet to the fire by requiring that expert testimony is backed by "sufficient facts or data" (93), accompanied by relevant error rates from methodologically sound studies, and presented without exaggeration (94).

6. Successes and Challenges

The scientific reinvention of forensic science is not an all or nothing concept. Rather, it is a process of gradual and continuing change. The most important element of change currently under way in forensic science is a recognition that a framework of trusting the examiner must give way to one that trusts the empirical science. Although the training, knowledge, and experience of the examiner are important, they will not be enough to sustain the forensic enterprise going forward. Forensic science is becoming an actual science: "The debate and rigor of academic science is now influencing much of forensic science and that is the most significant change from the past" (95).

Empirical testing has proceeded rapidly in some disciplines, and efforts are under way to measure sample difficulty and to identify statistical models that capture the probative value of forensic evidence. Extreme and unsupportable claims (e.g., 0% error rate and 100% certainty), once widespread, have been rejected by numerous scientific authorities and forensic science associations. Techniques that relied on false assumptions have exited the stage, and others whose validity appears doubtful seem to be headed toward the graveyard of unsupported science as well.

Perhaps the most important institutional step forward thus far has been the creation of national scientific bodies whose purpose is to increase the scientific rigor of the various forensic fields. The Organization of Scientific Area Committees (OSACs)—a complex of interconnected, multispecialty entities operating mainly under the auspices of the National Institute of Standards and Technology—were established in 2014 to do the heavy lifting. These committees, which are composed of more than 800 crime lab examiners, administrators, conventional scientists, and legal experts, create standards which, when fully developed, approved, and published, are available for adoption by individual crime labs. "OSAC-approved standards must have strong scientific foundations so that the methods that practitioners employ are scientifically valid, and the resulting claims are trustworthy" (96). As of March 2023, there are 97 published standards and 37 proposed for an array of different forensic disciplines. These developments count as successes. Institutions have been built and staffed, and a process is underway.

On the other hand, it is not obvious that the emerging OSAC standards go far enough in terms of ensuring that examiners' methods are valid and that their claims are trustworthy. Rather than squarely addressing major challenges such as the individualization problem discussed above, many of the standards merely nibble around the less controversial edges. Even if the OSACs do decide to take on the most important forensic challenges, it is crucial that the standards they create be supported by an empirical foundation. But many accepted that forensic techniques remain underresearched. The scientific evolution that we have described would benefit greatly from an overarching research agenda that coordinates both the needs of standards development and the research that gets funded. For example, a gap analysis would reveal the distance between what is believed (assumed) and what has been empirically validated. Research should be aimed at filling the discovered gaps. Unfortunately, as of 2015, a report on the funding of forensic science research found that "such a research agenda has not yet been developed" (97, p. 14). To be sure, such assessments and gap analyses have begun, but they are incomplete and have yet to receive much attention from practitioners or courts.

Even if the OSACs can address these issues, a practical problem remains: The OSACs lack enforcement power. Individual crime labs are free to adopt OSAC standards as they please. Even those labs that do endorse OSAC guidelines may decide to do so only nominally and then fail to incorporate them into day-to-day work.

The solution to this practical problem lies with the courts: If judges refused to admit evidence produced by laboratories that could not demonstrate how, exactly, they have incorporated OSAC guidelines and other scientific recommendations into their work, compliance would be guaranteed. More generally, if judges took seriously their duties under the Daubert line of cases (and state equivalents) and refused to admit insufficiently validated claims, the forensic sciences would adopt scientific practices more quickly and completely. Unfortunately, few courts have been so bold. The scientific advances that have been made are largely due to initiatives by the forensic fields themselves or by the wider scientific community. However, given that most forensic disciplines have ignored calls from the broader scientific community to replace individualizations with a more appropriate weight-of-evidence measure, a push from outside the fields themselves is needed.

In short, although a scientific reinvention of the forensic sciences is underway, its ultimate success is not assured. Its success depends on consistent attention to empirical validation of methods and conclusions and that in turn requires institutional structures that can help make that focus meaningful in courts of law. One such institutional structure was proposed by the NAS report. This report called for the creation of a new federal agency that focused on forensic science. Among other things, this agency, which would operate independently of law enforcement or any other potentially interested party, would be responsible for establishing and enforcing scientific practices in the forensic sciences. Ultimately, however, such an independent agency was not created.

Courts of law provide an alternative institutional structure for advancing the forensic sciences. Although the courts may not seem like an obvious force for advancing a scientific agenda, the expert evidence gate-keeping duties imposed on trial judges by Daubert and the relevant Federal Rule of Evidence, if faithfully followed, will promote a scientific focus and culture within the forensic sciences. To be sure, the courts' record on this front does not warrant much optimism. But the scientific paradigm is young and there are signs of hope and progress. The future of forensic science is ours to choose.

Data, Materials, and Software Availability. There are no data underlying this work.

1. Daubert v., *Merrell Dow Pharmaceuticals, 509 U.S.* **579** (1993).
2. General Electric Co. v. Joiner, 522 U.S. 136 (1997).
3. Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999).
4. M. J. Saks, J. J. Koehler, The coming paradigm shift in forensic identification science. *Science* **309**, 892–895 (2005).
5. Fabricant, Carrington, The shifted paradigm: Forensic science's overdue evolution from magic to law. *Va. J. Crim. L.* **4**, 1–115 (2016).
6. I. A. Pretty, D. Sweet, A paradigm shift in the analysis of bitemarks. *Forensic Sci. Int.* **201**, 38–44 (2010).
7. A. Biedermann, The strange persistence of (source) "identification" claims in forensic literature through descriptivism, diagnosticism and machinism. *Forensic Sci. Int.: Synergy* **4**, 100222. (2022).
8. J. L. Mnookin, The uncertain future of Forensic Science. *Daedalus* **147**, 99–118 (2018).
9. K. Sauerwein, J. M. Butler, K. K. Reczek, C. Reed, Bitemark analysis: A NIST scientific foundation review (2023) https://nvlpubs.nist.gov/nistpubs/ir/2023/NIST.IR.8352.pdf.
10. Expert Working Group on Human Factors in Latent Print Analysis, NIST, Latent Print Examination and Human Factors: Improving the Practice Through a Systems Approach, 69–70 (David H. Kaye ed., 2012).
11. D. H. Kaye *et al.*, Toolmark-Comparison Testimony: A Report to the Texas Forensic Science Commission (October 9, 2022), https:/doi.org/10.2139/ssrn.4108012.
12. President's Council of Advisors on Sci. & Tech., Exec. Office of the President, Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods (2016).
13. S. A. Cole, Grandfathering evidence: Fingerprint admissibility rulings from Jennings to Llera Plaza and back again. *Am. Crim. Law Rev.* **41**, 1189–1276 (2004).
14. M. J. Saks, Merlin and Solomon: Lessons from the law's formative encounters with forensic identification science. *Hastings Law. J.* **49**, 1069–1141 (1998).
15. *People v. Castro*, 545 N.Y.S.2d 985 (N.Y. Sup. Ct., 1989).
16. J. L. Mnookin, "People v. Castro: Challenging the Forensic Use of DNA Evidence" in *Evidence Stories*, R. O. Lempert (New York: Foundation, 2006).
17. National Research Council, Committee on DNA Technology in Forensic Science, DNA Technology in Forensic Science (Washington D.C.: National Academies of Science, 1992).
18. National Research Council, Committee on DNA Forensic Science: An Update, The Evaluation of Forensic DNA Evidence (Washington D.C.: National Academies of Science, 1996).
19. J. J. Koehler, Forensics or fauxrensics? Ascertaining accuracy in the forensic sciences *Ariz. St. Law J.* **49**, 1369–1416 (2017).
20. *United States v. Llera Plaza*, 179 F. Supp. 492, 510 (E.D. Pa, 2002).
21. J. J. Koehler, Why DNA likelihood ratios should account for error (even when a National Research Council report says they should not). *Jurimetrics* **37**, 425–437 (1997).
22. R. Lempert, After the DNA wars: Skirmishing with NRC II. *Jurimetrics* **37**, 439–468 (1997).
23. W. C. Thompson, Accepting lower standards: The National Research Council's second report on forensic DNA evidence. *Jurimetrics* **37**, 405–424 (1997).
24. National Research Council, *Strengthening Forensic Science in the United States, a Path Forward* (Washington D.C.: National Academies of Science, 2009).
25. G. Langenburg, C. Champod, T. Genessay, Informing the judgments of fingerprint analyses using quality metric and statistical assessment tools. *Forensic Sci. Int'l.* **219**, 183–198 (2012).
26. B. T. Ulery, R. A. Hicklin, J. Buscaglia, M. A. Roberts, Accuracy and reliability of forensic latent fingerprint decisions. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 7733–7738 (2011).
27. S. J. Bajic, L. S. Chumbly, M. Morris, D. Zamzow, Ames Laboratory-USDOE Report: Validation Study of the Accuracy, Repeatability, and Reproducibility of Firearms Comparisons, Technical Report #ISTR-52205220, prepared for the Federal Bureau of Investigation (Oct. 7, 2020).
28. L. Scott Chumbley *et al.*, Accuracy, Repeatability, and Reproducibility of Firearm Comparisons Part 1: Accuracy. arXiv [stat.AP] [Preprint] (2021). https://doi.org/10.48550/arXiv.2108.04030.
29. R. A. Hicklin *et al.*, Accuracy, reproducibility, and repeatability of forensic footwear examiner decisions. *Forensic Sci. Int.* **339**, 111418 (2022).
30. R. Mejia, M. Cuellar, J. Salyards, Implementing blind proficiency testing in forensic laboratories: Motivation, obstacles, and recommendations. *Forensic Sci. Int. Synergy* **2**, 293–298 (2020).
31. M. Neuman *et al.*, Blind testing in firearms: Preliminary results from a blind quality control program. *J. Forensic Sci.* **67**, 964–974 (2022).
32. J. J. Koehler, Fingerprint error rates and proficiency tests: what they are and why they matter. *Hastings Law J.* **59**, 1077–1098 (2008).
33. M. J. Saks, J. J. Koehler, Questions about forensic science: Response. *Science* **311**, 607–610 (2006).
34. I. E. Dror, N. Scurich, Mis)use of scientific measurements in forensic science. *Forensic Sci. Int. Synergy* **2**, 333–338 (2020).
35. I. E. Dror, The error in "error rate": Why error rates are so needed, yet so elusive. *J. Forensic Sci.* **65**, 1034–1039 (2020).
36. H. R. Arkes, J. J. Koehler, Inconclusives are not errors: A rejoinder to Dror. *Law Probability Risk* **21**, 89–90 (2022).
37. A. Biedermann, K. N. Kotsoglou, Forensic science and the principle of excluded middle: "Inconclusive" decisions and the structure of error rate studies. *Forensic Sci. Int. Synergy* **3**, 100147 (2021).
38. H. R. Arkes, J. J. Koehler, Inconclusives and error rates in forensic science: A signal detection theory approach. *Law, Probability Risk* **20**, 153–168 (2022).
39. J. L. Mnookin *et al.*, The need for a research culture in the forensic sciences. *UCLA Law Rev.* **58**, 725–779 (2011).
40. D. Mossman, "Hired guns", "whores", and "prostitutes": Case law references to clinicians of ill repute. *J. Amer. Acad. Psychiatry Law Online* **27**, 414–425 (1999).
41. D. Simon, M. Ahn, D. M. Stenstrom, S. J. Read, The adversarial mindset. *Psychol. Public Policy Law* **26**, 353–377 (2020).
42. D. C. Murrie, M. T. Boccaccini, Adversarial allegiance among expert witnesses. *Ann. Rev. Law and Social Sci.* **11**, 37–55 (2015). 10.1146/annurev-lawsocsci-120814-121714 (February 1, 2023).
43. D. C. Murrie, M. T. Boccaccini, L. A. Guarnera, K. A. Rufino, Are forensic experts biased by the side that retained them? *Psychol. Sci.* **24**, 1889–1897 (2013).
44. E. Balcetis, D. Dunning, See what you want to see: Motivational influences on visual perception. *J. Pers. Soc. Psychol.* **91**, 612–625 (2006).
45. R. S. Nickerson, Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* **2**, 175–220 (1998).
46. Z. Kunda, The case for motivated reasoning. *Psychol. Bull.* **108**, 480–498 (1990).
47. I. E. Dror, D. Charlton, A. E. Péron, Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Sci. Int.* **156**, 74–78 (2006).
48. I. E. Dror, K. C. Scherr, L. A. Mohammed, C. L. MacLean, L. Cunningham, Biasability and reliability of expert forensic document examiners. *Forensic Sci. Int.* **318**, 110610 (2021).
49. E. J. A. T. Mattijssen, C. L. M. Witteman, C. E. H. Berger, R. D. Stoel, Cognitive biases in the peer review of bullet and cartridge case comparison casework: A field study. *Sci. Justice* **60**, 337–346 (2020).
50. I. E. Dror, G. Hampikian, Subjectivity and bias in forensic DNA mixture interpretation. *Sci. Justice* **51**, 204–208 (2011).
51. I. Dror *et al.*, Cognitive bias in forensic pathology decisions. *J. Forensic Sci.* **66**, 1751–1757 (2021).
52. S. Nakhaeizadeh, I. E. Dror, R. M. Morgan, Cognitive bias in forensic anthropology: Visual assessment of skeletal remains is susceptible to confirmation bias. *Sci. Justice* **54**, 208–214 (2014).
53. N. Sunde, I. E. Dror, A hierarchy of expert performance (HEP) applied to digital forensics: Reliability and biasability in digital forensics decision making. *Forensic Sci. Int. Digital Invest.* **37**, 301175 (2021).
54. M. C. Taylor, T. L. Laber, P. E. Kish, G. Owens, N. K. P. Osborne, The reliability of pattern classification in bloodstain pattern analysis, part 1: Bloodstain patterns on rigid non-absorbent surfaces. *J. Forensic Sci.* **61**, 922–927 (2016).
55. S.-L. Chiam, I. E. Dror, C. D. Huber, D. Higgins, The biasing impact of irrelevant contextual information on forensic odontology radiograph matching decisions. *Forensic Sci. Int.* **327**, 110997 (2021).
56. J. J. Koehler, The influence of prior beliefs on scientific judgments of evidence quality. *Organ. Behav. Hum. Decis. Process.* **56**, 28–55 (1993).
57. M. J. Mahoney, Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognit. Ther. Res.* **1**, 161–175 (1977).
58. D. Kahneman, *Thinking, Fast and Slow* (Macmillan, 2011).
59. E. Pronin, D. Y. Lin, L. Ross, The bias blind spot: Perceptions of bias in self versus others. *Pers. Soc. Psychol. Bull.* **28**, 369–381 (2002).
60. Scientific Testimony: An Online Journal, The need for blind procedures in forensic science (June 1998), http://scientific.org/open-forum/articles/blind.html.
61. J. Kukucka, S. M. Kassin, P. A. Zapf, I. E. Dror, Cognitive bias and blindness: A global survey of forensic science examiners. *J. Appl. Res. Mem. Cogn.* **6**, 452–459 (2017).
62. D. E. Krane *et al.*, Sequential unmasking: A means of minimizing observer effects in forensic DNA interpretation. *J. Forensic Sci.* **53**, 1006–1007 (2008).
63. D. M. Risinger, M. J. Saks, W. C. Thompson, R. Rosenthal, The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. *Calif. Law Rev.* **90**, 1–56 (2002).
64. M. J. Saks, D. M. Risinger, R. Rosenthal, W. C. Thompson, Context effects in forensic science: A review and application of the science of science to crime laboratory practice in the United States. *Sci. Justice* **43**, 77–90 (2003).
65. M. S. Archer, J. F. Wallman, Context effects in forensic entomology and use of sequential unmasking in casework. *J. Forensic Sci.* **61**, 1270–1277 (2016).
66. R. D. Stoel *et al.*, "Minimizing contextual bias in forensic casework" in *Forensic Sci. & Admin. Justice: Critical Issues & Direction* **vol. 67**, pp. 67–86.
67. I. E. Dror, J. Kukucka, Linear Sequential Unmasking-Expanded (LSU-E): A general approach for improving decision making as well as minimizing noise and bias. *Forensic Sci. Int. Synergy* **3**, 100161 (2021).
68. A. Quigley-McBride, I. E. Dror, T. Roy, B. L. Garrett, J. Kukucka, A practical tool for information management in forensic decisions: Using Linear Sequential Unmasking-Expanded (LSU-E) in casework. *Forensic Sci. Int. Synergy* **4**, 100216 (2022).
69. A. Quigley-McBride, Practical solutions to forensic contextual bias. *Zeitschrift für Psychologie* **228**, 162–174 (2020).
70. G. L. Wells, M. M. Wilford, L. Smalarz, Forensic science testing: The forensic filler-control method for controlling contextual bias, estimating error rates, and calibrating analysts' reports. *J. App. Rsch Memory Cog.* **2**, 53–55 (2013).
71. A. Quigley-McBride, G. L. Wells, Fillers can help control for contextual bias in forensic comparison tasks. *Law Hum. Behav.* **42**, 295–305 (2018).
72. Friction Ridge Subcommittee OSAC. Standard for friction ridge examination conclusions [DRAFT DOCUMENT]. Draft standard. Organization of Scientific Area Committees for Forensic Science (2018). https://www.nist.gov/system/files/documents/2018/07/17/standard_for_friction_ridge_examination_conclusions.pdf. Accessed March 14, 2023.
73. N. Scurich, B. L. Garrett, R. M. Thompson, Surveying practicing firearm examiners. *Forensic Sci. Int. Synergy* **4**, 100228 (2022).
74. S. A. Cole, Individualization is dead, long live individualization! Reforms of reporting practices for fingerprint analysis in the United States *Law Probab. Risk* **13**, 117–150 (2014).
75. C. E. Champod, I. W. Evett, A probabilistic approach to fingerprint evidence. *J. Forensic Identif. Alameda* **51**, 101–122 (2001).
76. D. A. Stoney, What made us ever think we could individualize using statistics? *J. Forensic Sci. Soc.* **31**, 197–199 (1991).
77. M. J. Saks, J. J. Koehler, The individualization fallacy in forensic science evidence. *Vanderbilt Law Rev.* **61**, 199–219 (2008).
78. National Institute of Standards and Technology, National Commission on Forensic Science, Testimony using the term "Reasonable Scientific Certainty," Views Document by Subcommittee on Reporting and Testimony (2016). https://www.justice.gov/archives/ncfs/page/file/641331/download.
79. B. L. Garrett, N. Scurich, W. E. Crozier, Mock jurors' evaluation of firearm examiner testimony. *Law Hum. Behav.* **44**, 412–423 (2020).
80. J. B. Kadane, J. J. Koehler, Certainty & uncertainty in reporting fingerprint evidence. *Daedalus* **147**, 119–134 (2018).

81. C. Aitken, F. Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists*, chapter 3 (John Wiley & Sons, 2004), pp. 69–188.
82. American Statistical Association, American Statistical Association position on statistical statements for forensic evidence. (January 2, 2019), https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf.
83. C. C. G. Aitken *et al.*, ENFSI guideline for evaluative reporting in forensic science (2015). http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf.
84. Human Forensic Biology. Subcommittee, Biology Scientific Area Committee (OSAC for Forensic Science, Best practice recommendations for evaluative forensic DNA testimony) (January, 2022).
85. G. S. Morrison, Advancing a paradigm shift in evaluation of forensic evidence: The rise of forensic data science. *Forensic Sci. Int. Synergy* **5**, 100270 (2022).
86. J. J. Koehler, A. Chia, S. J. Lindsey, The random match probability in DNA evidence: Irrelevant and prejudicial. *Jurimetrics* **35**, 201–219 (1995).
87. W. C. Thompson, E. J. Newman, Lay understanding of forensic statistics: Evaluation of random match probabilities, likelihood ratios, and verbal equivalents. *Law Hum. Behav.* **39**, 332–349 (2015).
88. J. J. Koehler, On conveying the probative value of DNA evidence: Frequencies, likelihood ratios, and error rates. *Univ. Colo. Law Rev.* **67**, 859–886 (1996).
89. W. C. Thompson, R. H. Grady, E. Lai, H. S. Stern, Perceived strength of forensic scientists' reporting statements about source conclusions. *Law Probab. Risk* **17**, 133–155 (2018).
90. H. Swofford, S. Cole, V. King, Mt. Everest–we are going to lose many: A survey of fingerprint examiners' attitudes towards probabilistic reporting. *Law Probab. Risk* **19**, 255–291 (2021).
91. Frye v, U.S. 93 F. 1013 (1923).
92. People of Illinois v Winfield, Revised order and memorandum ruling. Circuit Court of Cook County, 14 CR 14066 01 (February 8, 2023).
93. Federal Rule of Evidence 702.
94. J. J. Koehler, Forensics source conclusions: Twenty threats to validity. *Zeitschrift für Psychologie* **228**, 149–161 (2020).
95. National Institute of Justice, The Slow But Steady March Towards a More Reliable Forensic Science. (2022). https://nij.ojp.gov/topics/articles/slow-steady-march-towards-more-reliable-forensic-science.
96. National Institute of Standards and Technology, National Commission on Forensic Science, Organization for Scientific Area Committees for Forensic Science, Instructions for Scientific and Technical review panels, Views Document by Subcommittee on Reporting and Testimony (2020). https://www.nist.gov/system/files/documents/2020/10/30/STRPs_%20Instructions.pdf.
97. National Academies of Sciences, Engineering, and Medicine, Support for Forensic Science Research: Improving the Scientific Role of the National Institute of Justice (Washington D.C.: The National Academies Press, 2015).
98. W. J. Bodziak, *Footwear Impression Evidence: Detection, Recovery and Examination* (CRC Press, ed. 2, 1999).
99. K. Inman, N. Rudin, *Principles and Practice of Criminalistics: The Profession of Forensic Science* (CRC Press, 2000).
100. Office of the Inspector General, U.S. Department of Justice, A review of the FBI's handling of the Brandon Mayfield case 8 (2006).
101. 60 Minutes: Fingerprints, CBS television broadcast, January 5, 2003.