FORENSICS OR FAUXRENSICS? Ascertaining Accuracy in the Forensic Sciences

Jonathan J. Koehler*

ABSTRACT

Forensic science—which includes such techniques as DNA analysis, fingerprint examination, and firearms comparison—plays a crucial role in our criminal justice system by helping to convict the guilty and acquit the innocent. However, our confidence in forensic science conclusions must be tempered by the odds that those conclusions are wrong. What are those odds? Nobody knows the answer because no disinterested researchers have conducted the appropriate studies in any of the forensic science disciplines. This is a serious problem because, without this information, legal decision makers cannot properly assess the validity or probative value of forensic evidence. In this paper, I examine the institutional forces and misunderstandings that are responsible for our ignorance about the accuracy of forensic science conclusions. I then recommend a new type of proficiency testing regimen (Type II proficiency testing) that is designed to measure error rates under appropriate test conditions in the various forensic subfields. Unless and until such studies are undertaken, legal decision makers will continue to fly blind when it comes to assessing the reliability of a reported forensic match.

INTRODUCTION.....1370

I. NO RELIABLE DATA: WHO'S TO BLAME AND WHY? 1378
A. The Scientific Community: Inconsistent Error Rate Guidance ...1378
1. 1992 NAS Report: DNA Evidence (NRC I).....1379
2. 1996 NAS Report: DNA Evidence (NRC II).....1381
3. 2009 NAS Report: Non-DNA Forensic Sciences 1383
4. 2010 NAS Report: Biometrics 1384
B. Legal Rules: Admissibility Standards That Don't Require Proof of Accuracy 1386

* Beatrice Kuhn Professor of Law, Northwestern Pritzker School of Law. Ph.D., University of Chicago; M.A., University of Chicago, B.A., Pomona College. This research was supported, in part, by the Northwestern Pritzker School of Law Faculty Research Program. The author thanks Ed German, David Kaye, and Deborah Tuerkheimer for comments on an earlier draft.

C. The Judges: "Utterly Ineffective" Gatekeepers 1389
1. Fingerprints 1392
2. DNA 1393
3. Firearms and Toolmarks 1394
D. The Forensic Science Culture: Insufficiently Scientific 1395

II. PROFICIENCY TESTS TO ESTIMATE ERROR RATES 1398
A. Two Types of Proficiency Tests 1399
1. Type I Proficiency Tests 1400
2. Type II Proficiency Tests 1400
3. A Non-Forensic Illustration of Type I and Type II Proficiency Tests 1401
4. Responding to Opposition to Type II Proficiency Tests 1403
B. Getting to Type II Proficiency Tests 1406
C. The Best Existing Studies Fall Short 1409

CONCLUSION 1414

INTRODUCTION

> *When a scientist doesn't know the answer to a problem, he is ignorant. When he has a hunch as to what the result is, he is uncertain. And when he is pretty darn sure of what the result is going to be, he is still in some doubt. We have found it of paramount importance that in order to progress we must recognize the ignorance and leave room for doubt. Scientific knowledge is a body of statements of varying degrees of certainty—some most unsure, some nearly sure, but none absolutely certain.*[1]
>
> *– Richard P. Feynman*

As the quotation above from Nobel Prize-winning physicist Richard Feynman makes clear, uncertainty is an inescapable part of every scientific pronouncement. Yet for one hundred years this humble sentiment has been conspicuously absent from forensic science testimony in legal cases.[2]

---

1. RICHARD P. FEYNMAN, WHAT DO YOU CARE WHAT OTHER PEOPLE THINK? 245 (1988).

2. Beginning with the first case in which fingerprints were admitted as evidence in a criminal case in the U.S., *People v. Jennings*, 96 N.E. 1077, 1082 (Ill. 1911), fingerprint examiners reporting a match between a questioned and known print have generally testified without a hint of uncertainty that the source of the known print is also the source of the questioned print. This has been the case for most other forensic sciences as well. DNA examiners have

Forensic scientists have long testified that they are 100% certain about who or what is the source of an evidentiary print or marking.[3] Firearms experts testify that this bullet came from that gun to the exclusion of all other guns in the world.[4] Fingerprint experts exclude every person in the world as a potential source of a recovered latent print other than the defendant.[5] The risk of error is said to be 0%.[6] Numerous forensic authorities and respected textbook authors encourage such hyperbole.[7]

---

arguably provided more scientifically cautious testimony than their non-DNA forensic science counterparts. DNA examiners typically admit uncertainty about who is the source of a questioned DNA sample by identifying the chance of a coincidental match between the questioned and known samples. However, even DNA examiners are prone to exaggeration. Jonathan J. Koehler, *Error and Exaggeration in the Presentation of DNA Evidence*, 34 JURIMETRICS 21, 23 (1993) [hereinafter *Error and Exaggeration*] (discussing early DNA cases in which experts testify, for example, that they are 100% certain that a particular trace came from a particular person); *see also* McDaniel v. Brown, 558 U.S. 120, 132 (2010) (considering, and then rejecting, the argument that a DNA examiner's exaggerated statistical testimony was reversible error); Jonathan J. Koehler, *Linguistic Confusion in Court: Evidence from the Forensic Sciences*, 21 J.L. & POL'Y 515, 522–29 (2013) [hereinafter Koehler, *Linguistic Confusion in Court*] (reviewing statistical probability errors in *McDaniel*).

3. When pressed for levels of certainty for their source identifications (or individualizations) at trial, fingerprint examiners, until recently, were compelled by the guidelines of their primary professional organization to testify with 100% certainty. *See, e.g.*, Maryland v. Bryan Rose, No. K06-0545, slip op. at 24–25 (Md. Cir. Ct. Oct. 19, 2007) ("Mr. Meagher [a top FBI latent print examiner] has stated that the FBI testifies to 'a 100 percent certainty that we have an identification.' . . . Mr. Meagher claimed that there is no error rate for ACE-V [the fingerprint method]."); OFFICE OF THE INSPECTOR GEN., U.S. DEP'T OF JUSTICE, A REVIEW OF THE FBI'S HANDLING OF THE BRANDON MAYFIELD CASE 8 (2006) ("Latent fingerprint identifications are subject to a standard of 100 percent certainty."); *see also Resolution VII*, IDENTIFICATION NEWS, Aug. 1979, at 1 ("[F]riction ridge identifications are positive, and [the International Associate for Identification] officially oppose[s] any testimony or reporting of possible, probable or likely friction ridge identification." The resolution went on to indicate that examiners who did indicate that a fingerprint identification was merely "possible, probable or likely . . . shall be deemed to be engaged in conduct unbecoming such member, officer or certified latent print examiner . . . and charges may be brought . . . ."). This resolution was rescinded in 2010. INT'L ASS'N FOR IDENTIFICATION, IAI RESOLUTION 2010-18, at 2 (2010) [hereinafter IAI RESOLUTION], https://www.theiai.org/member/resolutions/2010/Resolution_2010-18.pdf.

4. Morgan v. Bradt, No. 6:13-CV-6643(MAT), 2016 WL 1188438, at *3 (W.D.N.Y. Mar. 28, 2016) (noting that at trial, the expert testified that "all 12 of the fire cartridge cases were fired in the Taurus pistol to the exclusion of all other firearms").

5. Ohio v. Cruz, No. CA2012-03-059, 2013 WL 311333, at *2 (Ohio Ct. App. Jan. 28, 2013) ("[O]ne fingerprint was consistent with appellant's fingerprints, to the exclusion of all others.").

6. *60 Minutes: Fingerprints* (CBS television broadcast Jan. 5, 2003) (responding to a question by 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.").

7. WILLIAM J. BODZIAK, FOOTWEAR IMPRESSION EVIDENCE 347 (2d ed. 1999) ("An identification means the shoe positively made the questioned impression and no other shoe in the world could have made that particular impression."); MARIA JOSEFI, HANDBOOK OF FORENSIC

Exaggerated expert testimony of this sort is problematic not only because it is unscientific and lacks empirical support, but also because it forecloses inquiry by the legal decision maker into matters related to the reliability and accuracy of a forensic scientist's conclusions. Indeed, such testimony is the very antithesis of Dr. Feynman's message above because the testifying forensic scientists themselves seem to be telling jurors that their scientific conclusions and opinions cannot be wrong.

Of course, many legal decision makers know that a scientist should not be 100% certain or that an error rate cannot be 0%. And an increasing number of forensic experts and professional organizations are beginning to concede as much.[8] But legal decision makers and forensic experts who look for reliable data pertaining to forensic science error rates will be out of luck. *In most forensic science disciplines, there simply are no data pertaining to the rates at which forensic science procedures and forensic scientists err*.[9]

Nobody knows how accurate the opinions and conclusions offered by DNA analysts, firearms examiners, odontologists, document examiners, blood spatter specialists, or any other forensic scientists are. In most areas of forensic science, we can't even begin to estimate accuracy rates (or error rates) because none of the requisite studies have been conducted.[10]

There is plenty of blame to go around for this shameful state of affairs. The scientific community deserves blame because it has failed to articulate and emphasize the importance of testing forensic claims. Indeed, an important National Academy of Sciences (NAS) panel that advised the nation

---

SCIENCE 60 (1994) ("Toolmark identification is a microscopic side-by-side comparison that attempts to link a particular tool with a particular mark to the exclusion of any other tool produced. Such singular identification can be accomplished . . . ."); RICHARD SAFERSTEIN, CRIMINALISTICS 73 (9th ed. 2007) ("Balthazard has mathematically determined that the probability of two individuals having the same fingerprints is one out of 1×10^{60} . . . . This probability is so small as to exclude the possibility of any two individuals having the same fingerprints.").

8. IAI RESOLUTION, *supra* note 3, at 2 (rejecting the previous requirement that examiners disavow probabilistic identifications; examiners may now rely on "mathematically based models to assess the associative value of the evidence"); NAT'L COMM'N ON FORENSIC SCI., U.S. DEP'T OF JUSTICE, PRESENTATION OF EXPERT TESTIMONY POLICY RECOMMENDATIONS 2 (2017), https://www.justice.gov/sites/default/files/pages/attachments/2014/10/20/draft_on_expert_testimony.pdf ("Experts should not use misleading terms that suggest that the methodology or the expert is infallible when testifying." (footnotes omitted)).

9. Within the last few years, several error rate studies were conducted in the area of fingerprints. Two such studies, which are discussed in the last section of Part II, show good intentions but fall short of what is needed.

10. In one forensic area—fingerprint analysis—a few well-intentioned studies have appeared in recent years. But these studies fall short of what is needed to estimate casework error rates. *See infra* note 199.

on DNA evidence actually went so far as to suggest that there was no need to conduct scientific tests that measure error rates for DNA evidence.[11]

Those responsible for developing legal standards and rules deserve blame for serving up weak and malleable admissibility standards over the years that did not prevent untested and inaccurate junk forensic science from becoming the centerpiece of a criminal case. Things might have changed when the Supreme Court decided *Daubert v. Merrell Dow Pharmaceuticals, Inc.* in 1993. In *Daubert*, the Court expressly identified the rate of error as a factor that might be helpful to trial judges seeking to determine if a scientific method is sufficiently reliable to be admissible.[12] However, *Daubert* offered up the error rate factor in vague language ("the known or potential rate of error") with no explanation, no details, and with the caveat that neither the error rate factor nor any other reliability factors identified in *Daubert* necessarily applies in a given case. Not surprisingly, then, *Daubert*'s error rate factor has rarely been invoked to block exaggerated or unproven forensic science evidence.[13]

Trial judges deserve blame for repeatedly crediting the unsupported testimony of forensic scientists and historical precedent on matters of reliability. If trial judges had demanded proof of accuracy from the forensic science community—as indicated by low error rates in appropriately designed studies—as a condition of admissibility, the requisite studies probably would have been conducted.

Finally, the forensic science leadership deserves blame for failing to create and promote a scientific culture within the profession in which the study, measurement, and reporting of error is an integral part of the work performed. Instead, many of the forensic sciences have evolved, in the words of Professors Michael Saks and David Faigman, into "nonsciences" whose "primary claims for validity rest on anecdotal experience and proclamations of success over time."[14] Until recently, the nonscience forensic sciences that

---

11. COMM. ON DNA FORENSIC SCI., NAT'L ACAD. OF SCIS., THE EVALUATION OF FORENSIC DNA EVIDENCE 185 (1996) [hereinafter NRC II] (failing to endorse a recommendation from an earlier National Academy of Sciences report on DNA evidence that expressly called for laboratory error rates to be measured and disclosed to juries, stating "we attempt no such policy judgment"). NRC II also suggested that disclosing error rates was unnecessary because "[t]he risk of error is properly considered case by case." *Id.* at 87.

12. Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 594, 597 (1993).

13. For a recent and rare exception, see Almeciga v. Ctr. for Investigative Reporting, Inc., 185 F. Supp. 3d 401, 421–22 (S.D.N.Y. 2016), reviewing available error rate studies on handwriting analysis and concluding that for the task at hand, "the available error rates for handwriting experts are unacceptably high."

14. Michael J. Saks & David L. Faigman, *Failed Forensics: How Forensic Science Lost Its Way and How It Might Yet Find It*, 4 ANN. REV. L. & SOC. SCI. 149, 150 (2008).

Saks and Faigman reference (fingerprints, handwriting, bitemarks, voiceprints, firearms, tire tracks, shoe prints, etc.) not only failed to examine their own propensity to err under various conditions, but they did not even entertain the possibility of a non-zero error rate.

Perhaps the oddest part about the lack of error rate data in the forensic sciences is that, until very recently, hardly anyone seemed concerned. Why, for example, wasn't the broader scientific community shouting about this problem from the rooftops? Maybe these scientists simply assumed that the data were there. Likewise, the general public didn't seem to notice that there was a problem. If anything, the public has steadily gained confidence in forensic science evidence thanks, perhaps, to the wildly popular CSI television crime series shows which, for fifteen years,[15] "taught" the world that that a good-looking detective could link any print or marking to its one and only source. Whatever the reasons, it appears that, until recently, few people gave much thought to how we know that forensic opinions and conclusions are accurate.

Interest in examining the accuracy of forensic science claims grew following the appearance of the 2009 NAS Report on the state of the non-DNA forensic sciences in the U.S.[16] This report concluded that the non-DNA forensic sciences suffer from a lack of basic research, untested assumptions, and a tendency to offer exaggerated claims. The report pointedly called for research to help identify the accuracy of forensic science opinions and conclusions.[17]

Although some forensic science organizations modified their standards and practices following this report, disturbing revelations about the forensic sciences have continued to appear: massive crime lab scandals in Massachusetts,[18] an acknowledgment by the Justice Department and FBI that

---

15. Todd Leopold, *"CSI" Being Laid to Rest After 15 Years*, CNN (Sept. 25, 2015, 2:11 PM), http://www.cnn.com/2015/09/25/entertainment/csi-finale-immortality-feat/.

16. COMM. ON IDENTIFYING THE NEEDS OF THE FORENSIC SCI. CMTY., NAT'L ACAD. OF SCIS., STRENGTHENING FORENSIC SCIENCE IN THE U.S., at xix (2009) [hereinafter 2009 NAS Report].

17. *Id.* at 122 ("The assessment of the accuracy of conclusions from forensic analyses and the estimation of relevant error rates are key components of the mission of forensic science."); see also Jonathan J. Koehler & John B. Meixner, *An Empirical Research Agenda for the Forensic Sciences*, 106 J. CRIM. L. & CRIMINOLOGY 1, 8–9 (2016) for identification of a series of forensic science studies that should be conducted to increase our understanding of what forensic examiners are doing, how well they are doing it, and how cognitive bias may affect the accuracy of their conclusions.

18. For a review of the Massachusetts scandals, see Dahlia Lithwick, *Crime Lab Scandals Just Keep Getting Worse*, SLATE (Oct. 29, 2015, 5:21 AM), http://www.slate.com/articles/news_and_politics/crime/2015/10/massachusetts_crime_lab_scandal_worsens_dookhan_and_farak.html.

microscopic hair testimony was exaggerated in more than 95% of cases,[19] statistical errors in the FBI's DNA database,[20] and a moratorium on bite mark evidence in Texas.[21]

Based on this spate of bad news, some may be tempted to infer that the conclusions reached by forensic scientists are unlikely to be accurate. But this inference assumes too much, and it is not my claim here. *The problem is not inaccuracy per se as much as it is uncertainty about what the risk of inaccuracy is*.[22]

Shortly after the present manuscript was accepted for publication, the President's Council of Advisors on Science and Technology (PCAST) issued a report on the forensic sciences that rely on subjective feature-matching techniques (e.g., fingerprints, bitemarks, hair, firearms, shoeprints, and various DNA analyses). PCAST "is an advisory group of the Nation's leading scientists and engineers, appointed by the President to augment the science and technology advice available to him from inside the White House and from cabinet departments and other Federal agencies."[23] The PCAST Report

---

19. Spencer S. Hsu, *FBI Admits Flaws in Hair Analysis Over Decades*, WASH. POST (Apr. 19, 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=.7860438dc58b ("The Justice Department and FBI have formally acknowledged that nearly every examiner in an elite 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. Of 28 examiners with the FBI Laboratory's microscopic hair comparison unit, 26 overstated forensic matches in ways that favored prosecutors in more than 95 percent of the 268 trials reviewed so far . . . .").

20. Spencer S. Hsu, *FBI Notifies Crime Labs of Errors Used in DNA Match Calculations Since 1999*, WASH. POST (May 29, 2015), https://www.washingtonpost.com/local/crime/fbi-notifies-crime-labs-of-errors-used-in-dna-match-calculations-since-1999/2015/05/29/f04234fc-0591-11e5-8bda-c7b4e9a8f7ac_story.html?utm_term=.9ddb2bf85c12 ("The FBI has notified crime labs across the country that is has discovered errors in data used by forensic scientists in thousands of cases to calculate the chances that DNA found at a crime scene matches a particular person . . . .").

21. Brandi Grissom, *Texas Science Commission is First in the U.S. to Recommend Moratorium on Bite Mark Evidence*, DALL. MORNING NEWS (Feb. 12, 2016), https://www.dallasnews.com/news/politics/2016/02/12/texas-science-commission-is-first-in-the-u-s-to-recommend-moratorium-on-bite-mark-evidence.

22. In a similar vein, see Nathan J. Robinson, *Should We Trust Forensic Science?*, BOS. REV. (Feb. 18, 2016), http://bostonreview.net/books-ideas/vernon-nirenberg-respond-robinson-forensic-pseudoscience ("[T]he problem with forensic science is not that it is wrong, but that it is hard to know when it is right.").

23. 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, at iv (2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.

picks up where the 2009 NAS Report left off. The PCAST Report emphasized that rigorous empirical tests of forensic methods that rely on subjective human judgment are "required"[24] to "provide valid estimates of the method's accuracy (that is, how often the method reaches an incorrect conclusion)."[25] PCAST spells out what those tests should look like and cautions that, "[n]othing—not training, personal experience nor professional practices—can substitute for adequate empirical demonstration of accuracy."[26]

So what should be done? Clearly the answer is to begin the process of testing the accuracy of forensic science conclusions using rigorous scientific techniques. Simple, right? Wrong. The central problem is that the forensic science community has long claimed that such "proficiency tests"[27] *have already* been conducted, are currently being conducted, and the results show accuracy rates at (or nearly at) 100%. But here's the catch: the "proficiency tests" that the forensic science community refers to—and that courts have pointed to as proof of low error rates—were designed with purposes other than measuring accuracy in mind. As a result, those tests are wholly inadequate for estimating the accuracy of forensic conclusions.

A different type of test is needed to measure the risk of forensic science error. Without the information that can be gleaned from such tests, consumers of forensic science evidence have no way to evaluate the probative value of forensic match evidence. This paper identifies the parameters of such a test, which I identify for the first time as a "Type II proficiency test." A Type II proficiency test is quite different from existing proficiency tests—which I refer to as "Type I proficiency tests"—in both purpose and design. Regarding purpose, a Type I proficiency test serves purposes that are *internal* to forensic science. How well are the examiners able to follow laboratory protocols? Are the examiners adequately trained? How closely do the results achieved by examiners in one laboratory mirror those of examiners in other laboratories? These questions are important, but they do not speak to the needs of

---

24. *Id.* at 46.
25. *Id.* at 5.
26. *Id.* at 46.
27. In the world of forensic science, a proficiency test is "the evaluation of practitioner performance against pre-established criteria." FED. SCI. WORKING GRP. FOR FORENSIC ANTHROPOLOGY, U.S. DEP'T OF JUSTICE, PROFICIENCY TESTING 1 (2012), https://web.archive.org/web/20160803111859/http://swganth.startlogic.com/Proficiency%20Testing%20Rev0.pdf. The proficiency testing process involves providing known samples to examiners for analysis. The performance of those examiners is then assessed in some way. As discussed *infra* in Part II, the construction and procedure associated with a particular proficiency test will vary depending on the goal of the test.

consumers of forensic science information (e.g., judges, jurors, members of the general public) who are primarily concerned with knowing how trustworthy are the results and conclusions provided by forensic scientists. A Type II proficiency test directly addresses this concern by employing test procedures that are specifically designed to identify error rates for the various forensic sciences under various real-world conditions. Type I proficiency tests, which account for nearly all of the proficiency testing in the forensic sciences, cannot achieve this goal because Type I tests use samples that are "more artifact than real world,"[28] and are otherwise designed in ways that present a distorted picture of the frequency with which errors occur.[29]

The remainder of this paper focuses on (a) the institutional forces and misunderstandings that are responsible for our ignorance about the accuracy of forensic science conclusions, and (b) identifying a scientific approach to estimating and measuring the error rates associated with our forensic science conclusions. In Part I, I identify four culprits who share responsibility for our current ignorance about error rates. The first culprit is the scientific community, which has failed to provide clear, consistent guidance on the matter. The second culprit is the relevant evidentiary rules and legal standards, which do not allow us to distinguish between worthy and unworthy forensic science evidence at trial. The third culprit is the courts, which have not required the forensic sciences to produce scientific data in support of their claims. The fourth culprit is the leadership within the forensic science community, which has failed to create a scientific culture in the various forensic disciplines in which self-scrutiny, empirical study, and conservatism are valued and practiced. In Part II, I call for the development and implementation of a new type of proficiency testing regimen (Type II proficiency testing) for all forensic sciences. The purpose of these tests is not to improve the forensic sciences *per se*. Instead, the purpose is to provide consumers of forensic science evidence with an empirical basis for estimating error rates in the various forensic subfields under various conditions. Such data are critical because those estimated error rates will tell us nearly everything that we need to know about a reported forensic science match for assessing the trust we can place in that match report. In the Conclusion, I note that some reform efforts that are under way in the forensic sciences. However,

---

28. COLLABORATIVE TESTING SERVS., INC., CTS STATEMENT ON THE USE OF PROFICIENCY TESTING DATA FOR ERROR RATE DETERMINATIONS 2 (2010), https://www.ctsforensics.com/assets/news/CTSErrorRateStatement.pdf. Collaborative Testing Services provides testing materials to laboratories across the forensic sciences.

29. *Id.* at 3 ("The design of an error rate study would differ considerably from the design of a proficiency test.").

error rate studies are urgently needed as part of these efforts because they are the best way to identify the probative value of forensic science evidence.

I. NO RELIABLE DATA: WHO'S TO BLAME AND WHY?

As noted above, Part I examines why we know virtually nothing about the accuracy of forensic science conclusions (including conclusions based on an examination of DNA evidence). Below, I assign responsibility for this state of affairs to the scientific community, the evidentiary rules and legal standards, the courts, and the forensic science leadership.

A. The Scientific Community: Inconsistent Error Rate Guidance

DNA typing is widely regarded to have the strongest scientific foundation of any forensic science.[30] However, the scrutiny DNA typing received has *not* included a close examination of the overall risk of a false positive error (i.e., the risk that an examiner reports a match between samples that, in truth, came from different people). Instead, the scientific community focused more narrowly on one element of the false positive error risk, namely, the risk of coincidental matches. There are undoubtedly many reasons for this focus, one of which is that estimating this particular risk was directly in the wheelhouse of the population geneticists who were at the forefront of the DNA typing movement. As a result, a great deal of attention was given to matters such as whether the best estimate of a given DNA profile is, say, 1 in 10,000,000 or 1 in 10,000,000,000. The participants in this theoretical debate, which played out in the pages of the prestigious journal *Science*,[31] were not concerned with identifying laboratory error rates more generally or with documenting the mundane ways in which human error could produce a false positive result.

The best chance for receiving guidance from the general scientific community on forensic error rates came from reports issued by the National Academy of Science (NAS) panels that addressed forensic science evidence. The NAS was established in 1863 as a non-profit society of distinguished

30. 2009 NAS REPORT, *supra* note 16, at 41 ("DNA analysis—originally developed in research laboratories in the context of life sciences research—has received heightened scrutiny and funding support. That, combined with its well-defined precision and accuracy, has set the bar higher for other forensic science methodologies, because it has provided a tool with a higher degree of reliability and relevance than any other forensic technique."); Michael J. Saks & Jonathan J. Koehler, *The Coming Paradigm Shift in Forensic Identification Science*, 309 SCIENCE 892, 893 (2005) ("DNA typing technology was an application of knowledge derived from core scientific disciplines.").

31. For a review, see William C. Thompson, *Evaluating the Admissibility of New Genetic Identification Tests: Lessons from the "DNA War,"* 84 J. CRIM. L. & CRIMINOLOGY 22 (1993).

scholars who provide nonpartisan advice to the nation on matters of science, technology, and health policy. From time to time, NAS convenes blue ribbon panels of esteemed specialists to assist with their advisory work. The reports issued by these panels, which are regarded as an indicator of the views of the relevant scientific community, often have an immediate and significant impact on society, including the courts. The sections below briefly review the views of four NAS panels that weighed in to various extents on the question of error rate in forensic science.

1. 1992 NAS Report: DNA Evidence (NRC I)

NRC I was published in 1992 just a few years after DNA evidence first appeared in U.S. courtrooms.[32] NRC I concluded that DNA technology had enormous value as an investigatory tool, and should continue to be introduced in litigation as powerful evidence of identity.[33] However, NRC I also noted that DNA typing may be "vulnerable to error"[34] and that "[l]aboratory error rates should be measured with appropriate proficiency tests and should play a role in the interpretation of results of forensic DNA typing."[35]

In recommending "appropriate proficiency tests" to measure laboratory error rates, NRC I was not merely paying lip service to the notion that all scientific and human processes have error rates. Instead, NRC I was signaling that the identification of error rates in DNA typing is critically important precisely because the theoretical power of this technology was so great: "Especially for a technology with high discriminatory power, such as DNA typing, laboratory error rates must be continually estimated in blind proficiency testing and must be disclosed to juries."[36]

Unfortunately, NRC I did not clearly explain *why* it is important for a forensic technology that has high discriminatory power to measure and disclose its error rate. This task was left to a small group of statisticians,[37]

---

32. COMM. ON DNA TECH. IN FORENSIC SCI., NAT'L ACAD. OF SCIS., DNA TECHNOLOGY IN FORENSIC SCIENCE (1992) [hereinafter NRC I].

33. *Id.* at 98 ("[I]t is now clear that DNA analytic methods are a most powerful adjunct to forensic science for personal identification . . . .").

34. *Id.* at viii.

35. *Id.* at 15, 94–95.

36. *Id.* at 14, 89.

37. *See* David J. Balding & Peter Donnelly, *Inference in Forensic Identification*, 158 J. ROYAL STAT. SOC'Y 21, 21 (1995).

psychologists,[38] legal scholars,[39] and multi-disciplinary scholars[40] who echoed NRC I's calls for proficiency tests to identify DNA error rates. These scholars pointed out that the mathematics are such that the impact of the false positive error rate on the probative value of a highly discriminating technology such as DNA typing is larger than it is on less discriminating technologies. In other words, the more discriminating the technology, the more likely it is that the false positive error rate will restrict the ability of that technique to achieve its theoretical potential.[41] Thus, even if a DNA analysis indicates that a DNA profile is so rare there is probably only one person on the planet who could be its source, the reliability of that reported match will ultimately turn on the overall risk of a false positive error (which includes sample switches, misrecordings, and related human errors), and not merely on the small risk that the match is coincidental.

Though this theoretical point was not controversial, the forensic science community was not particularly enthusiastic about participating in tests that could be used to measure error rates that would subsequently be disclosed at

---

38. *See* Jonathan J. Koehler, *DNA Matches and Statistics: Important Questions, Surprising Answers*, 76 JUDICATURE 222, 228–29 (1993); *Error and Exaggeration*, *supra* note 2, at 24–25.

39. *See* David H. Kaye, *DNA Evidence: Probability, Population Genetics and the Courts*, 7 HARV. J.L. & TECH. 101 (1993).

40. *See* Richard Lempert, *Some Caveats Concerning DNA as Criminal Identification Evidence: With Thanks to the Reverend Bayes*, 13 CARDOZO L. REV. 303, 310 (1991); Thompson, *supra* note 31, at 104 (concluding that NRC I's recommendation that DNA laboratories participate in externally administered proficiency tests has "great merit").

41. COLIN AITKEN & FRANCO TARONI, STATISTICS AND THE EVALUATION OF EVIDENCE FOR FORENSIC SCIENCES 425 (2004) ("If the probability of an error . . . is much greater than the probability of matching profiles . . . then the latter probability is effectively irrelevant to the weight of the evidence."); DAVID J. BALDING, WEIGHT-OF-EVIDENCE FOR FORENSIC DNA PROFILES 35 (2005) ("If the false-match probability (ii) is judged to be much larger than the chance-match probability (i), then the latter probability is effectively irrelevant to evidential weight . . . . [I]t is not the absolute but the *relative* magnitude of the false-match to the chance-match probabilities that determines whether or not the former can be safely neglected."); Jonathan J. Koehler, Audrey Chia & J. Samuel Lindsey, *The Random Match Probability (RMP) in DNA Evidence: Irrelevant and Prejudicial?*, 35 JURIMETRICS 201, 201 (1995) ("RMPs contribute little to an assessment of the diagnostic significance of a reported DNA match beyond that given by the false positive laboratory error rate when RMPs are several orders of magnitude smaller than this error rate."); Richard Lempert, *After the DNA Wars: Skirmishing with NRC II*, 37 JURIMETRICS 439, 447 (1997) ("[T]he probative value of a DNA match is always limited by the chance of false positive error."); William C. Thompson, Franco Taroni & Colin G.G. Aitken, *How the Probability of a False Positive Affects the Value of DNA Evidence*, 48 J. FORENSIC SCI. 1, 1 (2003) ("[H]aving accurate estimates [of] the false positive probabilities can be crucial for assessing the value of DNA evidence."). Laboratory error includes all types of error that might result in a reported match on a person who is not, in fact, the source of the evidentiary item.

trial.[42] Without a clear explanation from NRC I about why error rate proficiency tests are so critical, few blind error rate tests were conducted.

2. 1996 NAS Report: DNA Evidence (NRC II)

In 1993, at the behest of the FBI Director, the NAS convened a second group of scientists and scholars to take yet another look at forensic DNA evidence. NRC II, which was published in 1996, focused largely on statistical and population genetics matters.[43] However, it also reviewed and rejected NRC I's call for proficiency tests to identify laboratory error rates. NRC II concluded that error rate estimates from proficiency tests are "almost certain to yield wrong values. When errors are discovered, they are investigated thoroughly so that corrections can be made."[44] NRC II also concluded that proficiency tests are not designed to measure error rates,[45] that reliance on the results of such tests "is almost certain to yield wrong values"[46] and unfair to the better laboratories.[47]

Though NRC II's position on proficiency tests and error rates was quickly and forcefully criticized,[48] few were listening to the critics. In the two decades since NRC II was published, *there have been virtually no blind proficiency tests designed to estimate case-relevant DNA match error rates*.[49] Further, attempts to introduce error rate evidence at trial or to discuss the absence of good forensic error rate data through expert testimony generally fail.[50]

NRC II's skeptical position on error rates and its unwillingness to endorse NRC I's common sense rule that "[l]aboratory error rates should be measured

---

42. Thompson, *supra* note 31, at 98 ("[E]fforts to obtain discovery of [DNA proficiency test data] have met strong resistance . . . .").

43. NRC II, *supra* note 11, at vi.

44. *Id.* at 86.

45. *Id.* at 80.

46. *Id.* at 86.

47. *Id.* ("The pooling of proficiency-test results across laboratories . . . could penalize the better laboratories.").

48. *See generally* Symposium, *The Evaluation of Forensic DNA Evidence*, 37 JURIMETRICS 395 (1997).

49. In 2003, Joseph Peterson and his colleagues conducted a study to assess the feasibility of establishing a program of blind proficiency tests for DNA laboratories throughout the U.S. Although this study was unique in that participants were unaware that they were being tested, the study was not designed to provide an estimate of error rate in actual DNA cases. *See* Joseph L. Peterson et al., *The Feasibility of External Blind DNA Proficiency Testing. II. Experience with Actual Blind Tests*, 48 J. FORENSIC SCI. 32, 32 (2003).

50. *See, e.g.*, United States v. Shea, 957 F. Supp. 331, 340 (D.N.H. 1997) ("A laboratory's error rate is a measure of its past proficiency that is of limited value in determining whether a test has methodological flaws.").

with appropriate proficiency tests and should play a role in the interpretation of results of forensic DNA typing"[51] reverberated across the forensic land. After all, if error rates need not be computed in cases involving DNA analyses, then surely they were unnecessary for cases involving fingerprint analyses and all other forensic methods. In *United States v. Mitchell*, a federal public defender challenged the admissibility of fingerprint evidence, in part, on grounds that such evidence failed the *Daubert* error rate factor.[52] Prosecutors sought to derail this challenge by questioning the value of error rates. In response to a request from the prosecutor to discuss "error, and error rate," a prominent forensic science expert testified as follows:

> We have to understand that error rate is a difficult thing to calculate. I mean, people are trying to do this, it shouldn't be done, it can't be done . . . . An error rate is a wispy thing like smoke, it changes over time because the real issue is, did you make a mistake, did you make a mistake in this case? If you made a mistake in the past, certainly that's valid information that someone can cross-examine or define or describe whatever that was, but to say there's an error rate that's definable would be a misrepresentation.[53]

Arguments of this sort are illustrative of a well-known logical fallacy known as the base rate fallacy.[54] Contrary to the expert's assertion in *Mitchell*, the general rate of error is not only relevant to an assessment of the chance that an error occurred in a specific case, but the risk of error in the instant case cannot be estimated accurately without knowing the error base rate.

In sum, NRC II not only paved the way for such illogical arguments at trial, but it effectively eliminated any hope that the forensic sciences would be forced to provide empirical data that would help legal decision makers assess the accuracy of their conclusions. At least until 2009.

---

51. NRC I, *supra* note 32, at 15, 86.

52. 365 F.3d 215, 220 (3d Cir. 2004). As discussed *infra* Section I.B, *Daubert v. Merrill Dow Pharm., Inc.*, 509 U.S. 579, 594 (1993), identified a new reliability standard for the admissibility of scientific evidence. One factor that was expressly identified by the *Daubert* court as potentially relevant to a trial judge's reliability inquiry is the "known or potential rate of error" for the technique under consideration. *Id.*

53. Transcript of Record at 122–23, United States v. Mitchell, 199 F. Supp. 2d 262 (E.D. Pa. 2002) (No. CR.A. 96-407-1).

54. Jonathan J. Koehler, *Fingerprint Error Rates and Proficiency Tests: What They Are and Why They Matter*, 59 HASTINGS L.J. 1077, 1089 n.34 (2008); Saks & Koehler, *supra* note 30, at 894–95.

3. 2009 NAS Report: Non-DNA Forensic Sciences

In 2009, the National Academy of Sciences published a blistering report on the scientific status of the non-DNA forensic sciences in the U.S.[55] Among other things, the 2009 NAS Report concluded that the non-DNA forensic sciences suffered from a lack of basic research, unproven assumptions, and a tendency to offer exaggerated claims.

Regarding the lack of basic research, the report concluded that, "[l]ittle rigorous systematic research has been done to validate the basic premises and techniques in a number of forensic science disciplines."[56] In particular, the report noted that "no studies have been conducted of large populations to establish the uniqueness of marks or features" in most forensic science disciplines.[57] Regarding exaggerated testimony, the report suggested a more scientifically defensible approach to claims about having "individualized" the source of a particular print or marking: "The concept of individualization is that an object found at a crime scene can be uniquely associated with one particular source. By acknowledging that there can be uncertainties in this process, the concept of 'uniquely associated with' must be replaced with a probabilistic association . . . ."[58]

To remedy the relative absence of science in modern day forensic science, the report called for research that focuses on assessing the validity of various forensic claims. According to the 2009 NAS Report, error rates must play a central role in this effort: "The assessment of the accuracy of conclusions from forensic analyses and the estimation of relevant error rates are key components of the mission of forensic science."[59] The report details the types of errors and error rates that could be computed and emphasized their importance.[60]

The position taken by the 2009 NAS Report on error rates was similar to that provided in NRC I, and quite different from that provided in NRC II. The 2009 NAS Report treated research on and understanding of error rates as crucial; it made it clear that the accuracy of forensic science conclusions requires careful and controlled study. It also noted that a reliable error rate estimate requires that *all* sources of potential error be measured and accounted for, including the (a) risk that two samples match by chance alone,

---

55. 2009 NAS REPORT, *supra* note 16, at 43.
56. *Id.* at 189.
57. *Id.* at 188–89.
58. *Id.* at 184.
59. *Id.* at 122.
60. *Id.* at 117–22.

and (b) the risk that the samples do not actually match.[61] This is an important point because it signals that the coincidental match probabilities of the sort that typically accompany DNA match reports would not provide a sufficient indicator of the risk that the match report is in error. However, like NRC I, the 2009 NAS Report did not go so far as to say that the absence of rigorous testing to identify forensic science error rates is intolerable. It did not advise the courts to be skeptical of proffered forensic science evidence when such evidence is not accompanied by evidence of error rate. Instead, as indicated above, it simply stated that error rate estimation is a "key" component of the forensic science mission.[62] By failing to offer a stronger statement about the *necessity* of error rate data, the 2009 NAS Report offered plenty of wiggle room for judges and forensic science proponents to continue introducing forensic science evidence of unknown accuracy into trials.

In sum, what the 2009 NAS Report said about error rates was very good, but not good enough or explicit enough to effect meaningful change in this area.

4. 2010 NAS Report: Biometrics

In 2010, another NAS Report was issued on biometric systems.[63] Biometrics refers to technologies that measure and analyze human body characteristics for authentication purposes. Some biometric technologies, such as DNA and fingerprints, overlap with forensic technology and forensic science. The preface of the 2010 NAS Report points out that "biometric recognition is an inherently probabilistic endeavor . . . . Consequently, even when the technology and the system it is embedded in are behaving as designed, there is inevitable uncertainty and risk of error."[64]

The significance of this point for our purposes here is that another NAS panel has directly acknowledged that even the best forensic technologies—including DNA and fingerprints—yield erroneous conclusions from time to time. Like NRC I and the 2009 NAS Report, the 2010 NAS Report thoroughly embraced the notion of error rate[65] and indicated that it should

---

61. *Id.* at 121 ("Both sources of error need to be explored and quantified in order to arrive at reliable error rate estimates for DNA analysis." (footnote omitted)). The point holds for non-DNA forensic analyses as well.

62. *Id.* at 122.

63. WHITHER BIOMETRICS COMM., NAT'L RESEARCH COUNCIL, BIOMETRIC RECOGNITION: CHALLENGES AND OPPORTUNITIES, at viii (Joseph N. Pato & Lynette I. Millett eds., 2010), https://www.nap.edu/read/12720/chapter/1 [hereinafter 2010 NAS REPORT].

64. *Id.* at viii–ix.

65. The phrase "error rate" appears forty-three times in the 2010 NAS Report. *Id. passim*.

continue to be measured. The 2010 report also warns that "[f]ield error rates are likely to be higher than laboratory testing suggests,"[66] and that "the largest components [of error rates] are the human interaction and environmental components."[67]

In sum, it seems fair to conclude that the scientific community as a whole, as represented through distinguished NAS reports, has repeatedly indicated that the measurement and disclosure of accuracy and error rates is an essential part of the work of forensic sciences.[68] However, it is not difficult to see why the NAS reports that support this view have had little impact. None of the reports clearly explains *why* it is so important to identify error rates in the forensic sciences. None of the reports explains what the parameters of the desperately needed error rate studies should look like. None of the reports provides guidance to trial judges on how to interpret existing or future error rate studies. And none of the reports explains how we should treat forensic science evidence if such tests are not performed.[69] To make matters worse, the lone NAS report that adopts a different stance on testing for error rates, the 1996 NAS Report (NRC II), actually does detail a series of objections to error rate computations and disclosure. Although NRC II's arguments have been widely rebutted,[70] for many years this report eliminated any pressure the forensic science community may have felt to measure and disclose error rates. Hence my conclusion that the scientific community has failed on the matter of error rates and accuracy in the forensic sciences.

---

66. *Id.* at 9, 83.

67. *Id.* at 66.

68. A fifth National Academy of Sciences Report, this one on ballistic imaging, is less on point as it focused on the use of computerized ballistic imaging technologies as a search tool for firearms examiners as opposed to, say, how that technology is or should be used by examiners to draw conclusions in legal cases. Nevertheless, the position offered in this report on testing and error rates appears to be consistent with those of three of the four NAS reports reviewed above (i.e., the reports from 1992, 2009, and 2010). *See* COMM. ON LAW AND JUSTICE, NAT'L RESEARCH COUNCIL, BALLISTIC IMAGING 82, 85 (Daniel L. Cork et al. eds., 2008), https://www.nap.edu/read/12162/chapter/1 ("[S]tatements on toolmark matches (including legal testimony) should be supported by the work that was done in the laboratory, by the notes and documentation made by examiners, and by proficiency testing or established error rates for individual examiners in the field and in that particular laboratory, but should not overreach to make extreme probability statements.").

69. The 1992 NAS Report comes closest to offering a view about what should happen to untested forensic science evidence. NRC I, *supra* note 32, at 55 ("No laboratory should let its results with a new DNA typing method be used in court, unless it has undergone such proficiency testing via blind trials."). But this statement and others in the report fall short of indicating a recommendation that untested forensic science evidence be inadmissible or be admitted with the caveat that jurors be told that the results are untested.

70. *See, e.g.*, Koehler, *supra* note 54, at 1089 n.34; Thompson, Taroni & Aitken, *supra* note 41, at 1; Symposium, *supra* note 48.

B. Legal Rules: Admissibility Standards That Don't Require Proof of Accuracy

Turning our attention to the legal standards that American courts use when deciding whether to admit or exclude forensic science evidence, I argue below that these standards have generally not been up to the task of blocking untested or exaggerated claims of accuracy.

When fingerprint evidence was first approved for admission by the Supreme Court of Illinois in *People v. Jennings* (1911),[71] the court articulated a standard for admitting expert evidence (including forensic science evidence) that resonates in the modern day Federal Rules of Evidence. The *Jennings* court wrote, "[e]xpert evidence is admissible when the witnesses offered as experts have peculiar knowledge or experience not common to the world, which renders their opinions, founded on such knowledge or experience, an aid to the court or jury in determining the questions at issue."[72] In other words, expert testimony is admissible when a qualified witness has something to say that helps a factfinder in the instant case. But what standard should a trial judge use when deciding whether an expert's statements will help the court or jury? What role does the risk of expert error play in this process? *Jennings* is silent on these questions.

A dozen years later, the U.S. Court of Appeals for the District of Columbia provided a standard for at least a subset of expert statements, namely those pertaining to novel scientific evidence. In *Frye v. United States* (1923), the Court held that, to be admissible, proffered scientific evidence must generally be accepted in the relevant scientific community.[73] *Frye* did not concern itself with *how* scientific matters gain general acceptance among scientists in the relevant field, or what role error rates play in this process. The idea was simply that the law should rely upon the judgments of the knowledgeable scientific community when deciding whether to admit novel scientific evidence. For the next fifty years or so nearly all U.S. courts admitted evidence from a broad range of forensic sciences using this relatively easy-to-execute standard.[74]

When the Federal Rules of Evidence (FRE) were adopted in 1975, none of the rules addressed the admissibility of novel scientific evidence per se. But FRE 702 did address the admissibility of expert evidence more broadly

---

71. 96 N.E. 1077, 1081–83 (Ill. 1911).

72. *Id.* at 1083.

73. 293 F. 1013, 1014 (D.C. Cir. 1923).

74. But for a list of forensic techniques that were screened out by the *Frye* standard, see DAVID H. KAYE, DAVID E. BERNSTEIN & JENNIFER L. MNOOKIN, THE NEW WIGMORE: A TREATISE ON EVIDENCE—EXPERT EVIDENCE, SECOND EDITION § 6.3.2.

and, as indicated above, this rule was consistent with the standard identified in *Jennings*. FRE 702 stated that expert testimony, offered by a qualified expert, is in principle admissible when it "will assist the trier of fact to understand the evidence or to determine a fact in issue."[75] Like *Jennings*, FRE 702 is silent on what role, if any, the judge should assign to the risk of error.

The risk of error *does* arise as a potential judicial consideration in the landmark U.S. Supreme Court case *Daubert v. Merrell Dow Pharmaceuticals, Inc.* (1993).[76] This case introduced a new standard for the admissibility of scientific evidence. Noting that FRE 702 requires that the subject of admissible scientific expert testimony must reflect "scientific . . . knowledge,"[77] the Court asserted that the hallmark of scientific knowledge is its reliability (or validity). Accordingly, the Court reasoned that this reliability requirement should be extended to proffered scientific evidence at trial.[78] Under the *Daubert* standard, trial courts have an affirmative duty to ensure that the methods underlying scientific testimony are reliable (or "scientifically valid"), and that those methods are applicable to the focal case.[79]

To help guide a trial judge's inquiry into the reliability of proffered scientific evidence, the *Daubert* Court offered guidance in the form of five flexible "general observations" that are widely referred to as the *Daubert* factors.[80] The factors are (a) extent to which the underlying theory has been tested,[81] (b) the existence of peer-reviewed publications,[82] (c) the "known or potential rate of error" of the scientific process,[83] (d) the existence of "standards controlling the technique's operation,"[84] and (e) general acceptance within the scientific community.[85]

Although many courts that reviewed forensic science evidence under the *Daubert* standard have tried to consider how well each of the five factors are

---

75. FED. R. EVID. 702.
76. 509 U.S. 579, 597 (1993).
77. *Id.* at 589–90.
78. *Id.* at 590.
79. In 2000, FRE 702 was revised to incorporate the essential elements of the *Daubert* admissibility standard. FED. R. EVID. 702.
80. *Daubert*, 509 U.S. at 593–94.
81. *Id.* at 593.
82. *Id.* at 593–94.
83. *Id.* at 594.
84. *Id.*
85. *Id.* In Kumho Tire Co. v. Carmichael, the Supreme Court extended the *Daubert* standard to all expert testimony and subtly clarified that courts should be concerned with whether the known or potential error rate for a technique is high. 526 U.S. 137, 149 (1999) ("Whether, in respect to a particular technique, there is a high 'known or potential rate of error.'").

met in the target case, discussions of the error rate factor have largely been superficial.[86] One reason for this is that neither *Daubert* nor its progeny clarified what courts are supposed to look for when they consider the "known or potential rate of error" of a forensic method.[87] A natural interpretation would seem to be that courts should check to see if the casework error rate[88] for a challenged forensic method is sufficiently low[89] in cases where error rate is a relevant consideration.[90] But even if this interpretation were correct and adopted by courts, key questions would remain. What type of proof should courts rely on as proof of a low casework error rate? How low is low enough? Does the evidentiary opponent have an obligation to show that the error rate is insufficiently low? These questions, which have not been addressed by the Court let alone answered, are all the more crucial because *error rate is the single most important component of a reliability assessment*. Whereas the presence of testing, peer review, the existence of common standards, and acceptance within the scientific community should generally boost our confidence in a forensic technique, none of these *Daubert* factors speak with any precision to the probative value of a reported match. In contrast, the error rate—specifically, the false positive error rate—not only speaks directly to probative value,[91] it reveals nearly all of what we need to know about it. The reason for this is that the false positive error rate places an upper limit on the probative value of a match report.[92] Thus, even though

86. Simon A. Cole, *More Than Zero: Accounting for Error in Latent Fingerprint Identification*, 95 J. CRIM. L. CRIMINOLOGY 985, 1048 (2005) (discussing how trial courts make "no attempt to responsibly estimate the 'practitioner error rate'").

87. John B. Meixner & Shari Seidman Diamond, *The Hidden Daubert Factor: How Judges Use Error Rates in Assessing Scientific Evidence*, 2014 WIS. L. REV. 1063, 1071 ("[M]ost importantly, the *Daubert* Court did not specify whether the error rate factor is intended to apply only to quantitative error rates that can be identified by the expert (or the field more generally) or whether it can apply more broadly to the chance that the expert may have made a mistake in his methods that could lead to erroneous testimony being given to the trier of fact.").

88. Jennifer L. Mnookin, *Fingerprint Evidence in An Age of DNA Profiling*, 67 BROOK. L. REV. 13, 60 (2001) ("Of course, what *Daubert* must mean when it refers to an error rate is the error rate in practice.").

89. *Kumho Tire Co.*, 526 U.S. at 149 (citing petitioner's brief which inquires about whether "there is a high 'known or potential rate of error'" for a particular technique).

90. Although trial courts are not required to rely on each and every *Daubert* factor when making admissibility rulings, the error rate factor would seem to be particularly applicable in cases involving scientific evidence which, if believed, could be dispositive in many cases.

91. Meixner & Diamond, *supra* note 87, at 1075 (noting that the error rate "is the only [*Daubert*] factor that speaks directly to the probative value of the evidence itself").

92. The mathematical details of this point are outside the bounds of this article, but are spelled out in detail elsewhere. *See, e.g.*, Koehler, *supra* note 54, at 1079 ("[T]he false positive error rate limits and controls the probative value of the match report."); Koehler, *Linguistic Confusion in Court*, *supra* note 2, at 533 ("Simply put, the probative value of a DNA match is

the *Daubert* standard improves on previous admissibility standards in the abstract by focusing on evidentiary reliability rather than proof by authority, it fails to provide judges with guidance for employing the error rate factor, and does little to signal the tight relationship between error rate and reliability.

In sum, the courts have relied on various standards over the past century to guide the admissibility of forensic science evidence. Only the *Daubert* standard, which was handed down by the Supreme Court in 1993, expressly identifies "error rate" as an admissibility factor. And even in *Daubert*, the error rate factor was offered up in vague language ("the known or potential error rate"), with no explanation or details, and with the caveat that this and all other *Daubert* factors may or may not apply to a given case. Not surprisingly, then, *Daubert*'s error rate factor has only rarely been used to block exaggerated or unproven forensic science evidence.[93]

C. The Judges: "Utterly Ineffective" Gatekeepers

Whether or not *Daubert* provides sufficient guidance to trial judges for judging evidentiary reliability, it clearly confers upon them an obligation to act as "gatekeepers" when it comes to challenged forensic science evidence.[94] The judge should admit forensic science evidence that is derived from methods that are demonstrably reliable. Proponents of the evidence would presumably satisfy this burden at an admissibility hearing by supplying a sufficient number of high quality scientific studies that find error rates for the method at issue to be sufficiently low. But, as I argue below, this is not what happens. In practice, error rates have only played a small role in trial courts' admissibility decisions.[95] Indeed, in a study that included 208 district court

---

capped by the frequency with which false positive errors occur."); Thompson, Taroni & Aitken, *supra* note 41, at 1 ("[H]aving accurate estimates [of] the false positive probabilities can be crucial for assessing the value of DNA evidence.").

93. For a recent and rare exception, see Almeciga v. Ctr. for Investigative Reporting Inc., 185 F. Supp. 3d 401, 422 (S.D.N.Y. 2016) (reviewing available error rate studies on handwriting analysis and concluding that for the task at hand, "the available error rates for handwriting experts are unacceptably high").

94. Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 597, 597 (1993) (referencing "a gatekeeping role for the judge").

95. *See* Mark Haug & Emily Baird, *Finding the Error in Daubert*, 62 HASTINGS L.J. 737, 739 (2011); Meixner & Diamond, *supra* note 87, at 1063 ("One factor of the *Daubert* test, the 'known or potential rate of error' of the expert's method, has received considerably less scholarly attention than the other factors, and past empirical study indicates that judges have a difficult time understanding the factor and use it less frequently in their analyses as compared to other factors.");

cases, Meixner and Diamond (2014) found very little discussion of error rates in the subset of cases that included forensic experts.[96] Is it any wonder that the 2009 NAS Report concludes that "the courts have been *utterly ineffective*" in dealing with the lack of scientific data on the accuracy of forensic conclusions?[97]

As noted earlier, many courts have held *Daubert* admissibility hearings in which they systematically stepped through all five *Daubert* factors as they pertained to the case under consideration. When considering the error rate factor in forensic science cases prior to the 2009 NAS Report (which called for greater testing to examine forensic science claims), courts rarely searched for and reviewed error rate studies. Instead, they typically offered vague references to "very low" error rates based on unsupported conclusory testimony from forensic scientists.[98] In a detailed review, Simon Cole concluded that judges "gullibly accept the claim of the zero 'methodological error rate' . . . [and] parrot totally unsupported assertions from latent print examiners that the so-called 'practitioner error rate' is 'vanishingly small,' 'essentially zero,' 'negligible,' 'minimal,' or 'microscopic.'"[99]

A few months after the 2009 NAS Report was released, the Tenth Circuit ruled that fingerprint analyses were sufficiently reliable to be admissible.[100] In doing so, this court expressly considered the potential error rate of the ACE-V fingerprint method,[101] and found that this factor "strongly supported

---

Munia Jabbar, Note, *Overcoming Daubert's Shortcomings in Criminal Trials: Making the Error Rate the Primary Factor in Daubert's Validity Inquiry*, 85 N.Y.U. L. REV. 2034, 2037 (2010).

96. Meixner & Diamond, *supra* note 87, at 1113 ("Implicit error rate discussion of forensic experts was rare, accounting for just over 15% of all Daubert analysis [sic] and lower than any other category except for natural science.").

97. 2009 NAS REPORT, *supra* note 55, at 53 (emphasis added); *see also* Jennifer L. Mnookin et al., *The Need for a Research Culture in the Forensic Sciences*, 58 UCLA L. REV. 725, 758 (2011) ("[E]ven after *Daubert v. Merrell Dow Pharmaceuticals, Inc.* . . . most judges confronted with pattern identification evidence have continued to admit it without restriction."); Saks & Faigman, *supra* note 14, at 166 (concluding that "courts have so utterly failed to carry out their gatekeeping duties" in matters related to proffered forensic science evidence).

98. Simon A. Cole, *Grandfathering Evidence: Fingerprint Admissibility Rulings from Jennings to Llera Plaza and Back Again*, 41 AM. CRIM. L. REV. 1189, 1189 (2004); Cole, *supra* note 86, at 1043.

99. Cole, *supra* note 86, at 1048.

100. United States v. Baines, 573 F.3d 979, 992 (10th Cir. 2009).

101. ACE-V is the standard method used for fingerprint analyses. The ACE-V acronym stands for Analyze, Compare, Evaluate, and Verify. EXPERT WORKING GRP. ON HUMAN FACTORS IN LATENT PRINT ANALYSIS, U.S. DEP'T OF COMMERCE, LATENT PRINT EXAMINATION AND HUMAN FACTORS: IMPROVING THE PRACTICE THROUGH A SYSTEMS APPROACH 1 (2012) [hereinafter LATENT PRINT] (stating that ACE-V is "[t]he conventional procedure for associating impressions of friction ridge skin by a latent print examiner").

the judge's decision to admit the expert testimony."[102] The court based its conclusion on two points. First, the examiner testified that he had not made errors on his proficiency tests. But as discussed in Part II below, excellent performance on proficiency tests that are not designed to measure error rates tells us little, if anything, about error rate.[103] Second, there was testimony from the former head of the FBI's fingerprint unit that the FBI had "an error rate of one per every 11,000,000 cases."[104] The expert witness arrived at this estimate by claiming that the FBI makes about 1,000,000 identifications per year, and that the agency was known to have made just one mistake over the past eleven years (i.e., over the past 11,000,000 identifications).[105] In other words, this expert witness simply *assumed* the accuracy of millions of FBI fingerprint identifications,[106] and then relied on this assumption to bolster the identification he offered in the case. The logic employed here is flawed, of course, and the expert's error rate estimate of one per 11,000,000 was rightfully and severely criticized by others.[107] Still, the expert's estimate has been cited by other courts as "proof" of a very low rate of error for the ACE-V fingerprint method.[108]

This misguided proof of the error rate for fingerprint evidence is important because it was one of the first federal cases that examined the role of forensic error rate after the 2009 NAS Report appeared. The fact that such a poorly reasoned error rate argument could hold sway in a prestigious appellate court did not bode well for how error rate would be treated in a post-2009 NAS Report world. Since that time, federal appellate courts have mainly considered the forensic error rate issue in three domains: fingerprints, DNA, and firearms/toolmarks. Although few in number, these recent cases paint a relatively consistent—and depressing—picture of how courts are dealing with error rate challenges in the aftermath of the 2009 NAS Report.

---

102. *Baines*, 573 F.3d at 991.

103. *See* LATENT PRINT, *supra* note 101, at 33 ("[N]ormal proficiency tests are neither designed for nor particularly suitable for estimating error rates.").

104. *Baines*, 573 F.3d at 991.

105. *Id.* at 984.

106. Because ground truth in casework is absent, claims about the accuracy of casework conclusions lack any scientific basis.

107. LATENT PRINT, *supra* note 101, at 33 ("Historical inquiry is simply not a viable way to estimate how low the false positive rate has been.").

108. *See, e.g.*, United States v. Rose, 672 F. Supp. 2d 723, 725 (D. Md. 2009) (citing the error rate analysis in *Baines* in support of the conclusion "that the known error rate [for fingerprinting] is very low").

1. Fingerprints

In *United States v. Aman* (2010), a U.S. district court acknowledged the absence of fingerprint error rate data.[109] However, the court ruled that this fact was an "appropriate topic(s) for cross-examination, not grounds for exclusion."[110] The court is mistaken. Error rate shortcomings, like the other *Daubert* factors that provide a check on reliability, are indeed grounds for exclusion. Any suggestion to the contrary would effectively eviscerate the judicial gatekeeping responsibilities of *Daubert*.

Other courts have suggested that the absence of error rate data—or even the presentation of false error rate information—is not especially important to admissibility decisions. In *United States v. Watkins* (2011), the Sixth Circuit Court of Appeals held that even if a fingerprint examiner falsely claims that the ACE-V fingerprint method has a zero error rate, such testimony does not "negate(s) the scientific validity of the ACE-V method given all the other factors that the district court was required to consider."[111]

A third judicial approach to fingerprint error rates, documented by Simon Cole more than a decade ago, takes the form of simple assertions that fingerprint errors are "very rare"[112] or "extremely rare."[113] Support for these assertions in judicial opinions is weak, ranging from observations about the infrequency with which DNA exonerations have occurred in fingerprint cases[114] to string cites to other cases that offered a similar pronouncement. In *United States v. Campbell* (2012), a federal court satisfied itself that the error rate was low through reference to a 2010 Eleventh Circuit ruling that the use of a second examiner in fingerprint cases to verify the match report "reduces" the risk of error.[115]

In *United States v. Love* (2011), a federal court in California cited much of the available, but limited data, pertaining to the fingerprint error rate.[116] This court acknowledged shortcomings in the data, but it appeared to reverse the burden of production associated with this *Daubert* factor. The court wrote

---

109. 748 F. Supp. 2d 531, 542 (E.D. Va. 2010).

110. *Id.*

111. 450 F. App'x 511, 516 (6th Cir. 2011).

112. United States v. Herrera, 704 F.3d 480, 487 (7th Cir. 2013) (citing Greg Hampikian et al., *The Genetics of Innocence: Analysis of 194 U.S. DNA Exonerations*, 12 ANN. REV. GENOMICS & HUMAN GENETICS 97, 106 (2011)).

113. *Rose*, 672 F. Supp. 2d at 726.

114. *Herrera*, 704 F.3d at 487.

115. No. 1:11–cr–00460–AT–RGV, 2012 WL 2374528, at *5 (N.D. Ga. April 19, 2012) ("[Although] there is no scientifically determined error rate, the examiner's conclusions must be verified by a second examiner, which reduces, even if it does not eliminate, the potential for incorrect matches." (citing United States v. Scott, 403 F. App'x 392, 397–98 (11th Cir. 2010)).

116. No. 10cr2418–MMM, 2011 WL 2173644, at *5 (S.D. Cal. June 1, 2011).

that because there was no evidence in the record to suggest error rates higher than those identified in the various flawed studies the court reviewed, then the error rate factor favored admissibility. Taken together, these fingerprint cases do not suggest a sea change in the way error rate is viewed as an admissibility factor in the post-2009 NAS Report era.

2. DNA

A persistent myth associated with DNA evidence is that the error rate has been identified and it is extremely low. In *United States v. Scott* (2010), the Eleventh Circuit wrote, "[u]nlike DNA evidence, there is no known percentage error rate for fingerprint examination."[117] In fact, there is no known percentage error rate for either fingerprint or DNA evidence. The relevant studies have not been conducted. This is probably news to the public and to most trial judges.[118] Perhaps the Eleventh Circuit and others believe that we know the error rates for DNA evidence because they think that the tiny random match probability (RMP) that usually accompanies a DNA match report *is* the error rate. But error rate and RMP are conceptually distinct. The RMP identifies the frequency of a DNA profile in a target population (e.g., Caucasians). As such, it provides an indication of the risk that a reported match is merely coincidental. It provides no indication about the risk of human error that could produce a false positive result. Nevertheless, a recent study suggests that more than half of the public confuses the DNA RMP with the error rate.[119] This is problematic not only because it falsely leads the public to believe that the error rate for DNA match reports is known, but also because it leads the public to believe that the error rate is as small as the RMP—which is commonly one in billions, trillions, or quadrillions.

In *United States v. Wrensford* (2014), a district court took a cue from *United States v. Aman* (2010) discussed above by claiming that the absence of DNA error rate data is not relevant for an admissibility determination.[120] In *United States v. Williams* (2013), a Hawaii district court followed NRC II

---

117. 403 F. App'x 392, 395 (11th Cir. 2010).

118. For all the talk about how accurate DNA and other forensic science evidence supposedly is, there are no scientific data that provide a reasonable estimate of casework accuracy or error rates in *any* of the forensic sciences.

119. Jonathan J. Koehler, *Intuitive Error Rate Estimates for the Forensic Sciences*, 57 JURIMETRICS 153, 163 (2017).

120. No. 2013–0003, 2014 WL 1224657, at *11 (D.V.I. Mar. 25, 2014) ("[T]he lack of a specific error rate does not weigh against admissibility.").

by waving off the error rate issue altogether, holding that "the risk of error is considered on a case by case basis."[121] In *United States v. Pritchard* (2014), a California district court applied the "error rate" factor to the random match probability statistics rather than the technique, and then misleadingly cites NRC II for the proposition that the rule used to generate those RMPs has "an acceptable rate of error."[122] In *United States v. McCluskey* (2013), a New Mexico district court cited the 2009 NAS Report for the proposition that DNA evidence has "a low error rate."[123] However, the 2009 NAS Report provides no data to support this claim, focusing instead on the state of the traditional, non-DNA forensic sciences. *McCluskey* also asserts that the FBI has a "low to zero error rate"[124] and cites a 2001 case in support.[125] However, the cited case merely asserts a similarly exaggerated claim without scientific foundation.[126] As in the recent fingerprint cases noted earlier, recent cases that considered the admissibility of DNA evidence relied on weak arguments and poor logic when discussing the error rate factor.

3. Firearms and Toolmarks

In the area of firearms and toolmarks, the courts appear to be more focused on data when considering the error rate *Daubert* factor in recent years than are courts that consider fingerprint and DNA evidence.[127] Unfortunately, the error rate data that the courts rely on are those generated by proficiency tests conducted by Collaborative Testing Services (CTS). This is a problem because CTS has long advertised that the data they compile "are not intended to be an overview of the quality of work performed in the profession and

---

121. No. 06–00079 JMS/KSC, 2013 WL 4518215, at *7 (D. Haw. Aug. 26, 2013).

122. 993 F. Supp. 2d 1203, 1211 (C.D. Cal. 2014).

123. 954 F. Supp. 2d 1224, 1243 (D.N.M. 2013).

124. *Id.*

125. United States v. Trala, 162 F. Supp. 2d 336, 347 (D. Del. 2001).

126. *See id.* ("The FBI methodology has been developed to result in a zero error rate within acceptable measurement error conditions . . . .").

127. United States v. Ashburn, 88 F. Supp. 3d 239, 246 (E.D.N.Y. 2015) ("Studies have shown that the error rate among trained toolmark and firearms examiners is quite low."); United States v. Johnson, No. 14–cr–00412–TEH, 2015 WL 5012949, at *4 (N.D. Cal. Aug. 24, 2015) ("The data show that the error rate in matching sample casings and bullets to particular firearms . . . is sufficiently low."); United States v. Wrensford, No. 2013–0003, 2014 WL 3715036, at *17 (D.V.I. July 28, 2014) ("low error rates"); United States v. Otero, 849 F. Supp. 2d 425, 434 (D.N.J. 2012) ("[T]he information derived from the proficiency testing is indicative of a low error rate."); United States v. Taylor, 663 F. Supp. 2d 1170, 1177 (D.N.M. 2009) ("[P]roficiency testing done by the Collaborative Testing Service (CTS) . . . suggests that the error rates are quite low.").

cannot be interpreted as such."[128] CTS test results are inappropriate to use as indicators of "low" error rates because the design and conduct of CTS "tests" are quite unlike those of tests a scientist would design to measure error rates. For example, CTS tests do not use realistic samples, do not use blind testing, and they do not control the way laboratories or examiners use their tests. Nevertheless, courts continue to rely heavily on these data in firearms and toolmark cases to justify the conclusion that error rates are sufficiently low.

Whether discussing fingerprints, DNA, or firearms and toolmarks, the common thread that runs through recent federal appellate opinions on the admissibility of forensic science evidence is that the lack of reliable error rate data is not a serious problem given the secondary ways that one might infer that those error rates are low. Unless there is some sort of jolt to the system—a system that continues to take the *Daubert* error rate factor less seriously than other *Daubert* factors[129]—progress on this front promises to be slow.[130]

D. The Forensic Science Culture: Insufficiently Scientific

"Lawyers and judges should not be counted on to fix the science problem. What we need is for the forensic science community to improve so that it better serves the needs of justice."

---

128. COLLABORATIVE TESTING SERVS., INC., TOOLMARKS EXAMINATION: TEST NO. 15-528 SUMMARY REPORT 1 (2015) [hereinafter TOOLMARKS EXAMINATION]; *see also* COLLABORATIVE TESTING SERVS., INC., *supra* note 28, at 3 ("The design of an error rate study would differ considerably from the design of a proficiency test. Therefore, the results found in CTS' Summary Reports should not be used to determine forensic science discipline error rates.").

129. Meixner & Diamond, *supra* note 87, at 1063 ("One factor of the *Daubert* test, the 'known or potential rate of error' of the expert's method, has received considerably less scholarly attention than the other factors, and past empirical study indicates that judges have a difficult time understanding the factor and use it less frequently in their analyses as compared to other factors."); *see also* Mnookin, *supra* note 88, at 60 ("[A] major argument leveled by those challenging fingerprinting is that the error rate for fingerprinting has received insufficient attention and study.").

130. Nancy Gertner, a retired federal judge who was one of few trial judges who did not give a pass to the forensic sciences when it came to the absence of error rate data, faults the judiciary for not requiring more from the forensic science community. *See* Judge Nancy Gertner, *Commentary on the Need for a Research Culture in the Forensic Sciences*, 58 UCLA L. REV. 789, 790 (2011) ("[U]ntil courts address the deficiencies in the forensic sciences—until courts do what *Daubert v. Merrell Dow Pharmaceuticals, Inc.* requires that they do—there will be no meaningful change here."); *see also* Alex Kozinski, *Criminal Law 2.0*, 44 GEO. L.J. ANN. REV. CRIM. PROC., at iii, iv–v, xxxv (2015) (arguing that the infallibility of forensic science results is a myth, and that "courts must be far more rigorous in enforcing *Daubert* before allowing experts to testify in criminal trials").

– Judge Harry T. Edwards[131]

When contemplating the dearth of information about the accuracy of the forensic sciences, the elephant in the room is the forensic sciences themselves. As the 2009 NAS Report indicated, most of the forensic sciences simply have not conducted scientific studies to identify how well they can do what they say they can do.[132] Why not?

An easy explanation focuses on time: forensic scientists are so busy getting trained and doing casework that they simply don't have the time to indulge in general scientific validation work. This explanation will not suffice. Nobody argues that bench-level forensic scientists should be designing and conducting research that addresses fundamental questions about their disciplines.[133] This work should be performed by trained *researchers*, a cohort that is generally restricted to those with graduate level training in scientific methodology. So why haven't researchers conducted rigorous studies that identify forensic science accuracy rates?[134]

As a number of scholars have pointed out, one reason may be cultural. Whereas most sciences share a research culture in which empiricism, transparency, and an ongoing critical perspective are paramount, the forensic sciences operate in more of a "quasi-adversarial" culture in which key claims are simply accepted as true without empirical support.[135] For example, examiners are taught that markings are unique and that a competent examiner can individualize a marking to its unique source.[136] Such teachings contribute

---

131. Harry T. Edwards, *Solving the Problems that Plague the Forensic Science Community*, 50 JURIMETRICS 5, 13 (2009).

132. 2009 NAS REPORT, *supra* note 16, at 22, 45 ("[Recommending research] to address issues of accuracy, reliability and validity in the forensic science disciplines . . . [and lamenting] the lack of good data on the accuracy of the analyses conducted in forensic science disciplines."); *see also* Edwards, *supra* note 131, at 6 ("[W]e were also trying to determine the extent to which there is any peer-reviewed, scientific research to support the validity and reliability of existing forensic disciplines; in particular, we were looking for scientific studies that address the level of accuracy of forensic disciplines that rely on subjective assessments of matching characteristics. We invited experts in each discipline to refer us to any such research; however, apart from the materials on nuclear and mitochondrial DNA and drug analysis, we received little in the way of compelling scientific research assessing the accuracy of forensic disciplines.").

133. Mnookin et al., *supra* note 97, at 741 ("[Forensic science] practitioners need not, and indeed often should not, be the primary producers of the research themselves.").

134. Over the past few years, a few fingerprint researchers have conducted studies that ostensibly were designed to identify error rates. These studies are discussed in the last section of Part II.

135. *See* Mnookin et al., *supra* note 97, at 731.

136. This may be changing in some subfields, but claims of unique source identification are still common.

to a perception within the profession that error is an indication of individual examiner incompetence as opposed to, say, an inevitable risk associated with all scientific processes that should be measured and reported. This perception, in turn, effectively ensures that studies to measure error and error rates will not be treated as part of the profession's basic research agenda.

In an adversarial system, the substitution of forensic dogma for scientific studies favors the side that proffers the forensic-match evidence. In criminal cases, this side is usually the prosecution. And, of course, the prosecution works closely with law enforcement which, in turn, controls most forensic science laboratories.[137] The lack of independence between forensic science laboratories and law enforcement poses enormous risks to the overall quality of forensic science work.[138] It may also create disincentives to measure and publicize forensic errors. If criminal defendants are denied access to forensic error rate data, they will have few weapons available to challenge the forensic evidence which, as noted earlier, is widely believed to be valid and extremely accurate.

In addition to cultural obstacles within the discipline that impede error rate studies, political considerations loom large.[139] Forensic science currently enjoys a reputation as highly accurate. A recent study of 210 members of the jury-eligible population indicated that the average person believes that the chance of a false positive error for various forensic sciences ranges from 1 in 100,000 for document examination to 1 in 10,000,000 for DNA.[140] Study participants who were provided with a small DNA random match probability estimated the false positive error risk for DNA to be lower still (median estimate: 1 in 1,000,000,000). Similarly, a 2013 study that included 305 members of an Orange County California jury pool, reported that the median juror estimated the chance of a false positive DNA result to be 1 in 1,000,000.[141] In face-to-face interviews with 1,000 German adults, 63% of those interviewed indicated that fingerprint analyses were "absolutely

---

137. 2009 NAS REPORT, *supra* note 55, at 183 ("The majority of forensic science laboratories are administered by law enforcement agencies, such as police departments, where the laboratory administrator reports to the head of the agency.").

138. *Id.* at 183–84.

139. The historian of science Simon Cole has documented various maneuvers employed by the forensic science community to help maintain its status as a highly accurate science. *See, e.g.*, Cole, *supra* note 98, at 1263 (noting that the fingerprint and forensic testing communities "have taken pains to ensure that proficiency tests results that they do not like cannot be construed as indicative of the accuracy of the technique in real casework").

140. Koehler, *supra* note 119, at 162.

141. William C. Thompson et al., *Do Jurors Give Appropriate Weight to Forensic Identification Evidence?*, 10 J. EMPIRICAL LEGAL STUD. 359, 382 (2013).

certain," and 78% believed the same is true for DNA analyses.[142] These data suggest that the public believes that the risk of forensic science error is extremely—and unrealistically—low. It is a near certainty that a properly designed error rate study in any of the forensic disciplines would yield error rates that are orders of magnitude larger than these median estimates. From a reputational standpoint, the forensic sciences may think they have little to gain by embracing error rate studies that are sure to yield values that are higher than those the public anticipates.

Judge Harry Edwards, co-chair of the 2009 NAS Report, has suggested that progress in forensic science will require a major cultural shift within the forensic sciences.[143] Such a shift will likely require a corresponding shift in the willingness of the broader scientific and legal communities to trust less and verify more.

II. PROFICIENCY TESTS TO ESTIMATE ERROR RATES

In most areas of forensic science, we can't even begin to estimate accuracy rates (or error rates) because none of the requisite studies have been conducted. Part I identified and examined the institutional forces and misunderstandings that are responsible for why this is the case. The culprits identified in this part include the scientific community, the legal rules and standards pertaining to evidentiary admissibility, the trial judges who are charged with gatekeeping responsibilities, and the forensic science culture. Having pointed out the problem and those responsible for it, the next big question is what to do about it. The answer is obvious: testing. We need to test all of the forensic science methods to find out just how accurate they really are. Does anyone really disagree? Surprisingly, the answer is yes.

In Part II, I explain that discussions of proficiency testing in the forensic sciences must begin with a recognition that there are different types of proficiency tests, each having different goals and different designs. Consequently, the results of proficiency tests that were designed to measure, say, the adequacy of laboratory protocols, cannot be used to answer questions about the rate at which examiners err. With this point in mind, I distinguish sharply here between proficiency tests that serve internal purposes (e.g., validating a laboratory's methods) and proficiency tests that serve external purposes (e.g., providing jurors with an estimate of the casework error rate for a method). I respond to anticipated criticisms of the latter type of

142. GERD GIGERENZER, RISK SAVVY: HOW TO MAKE GOOD DECISIONS 18–19 (2014).

143. Edwards, *supra* note 131, at 14 ("I am also convinced that the forensic science community will never change for the better unless certain cultural habits are broken.").

proficiency test (which I refer to as Type II proficiency tests), and identify key features of those tests. I also explain why proficiency tests conducted to date, including two tests in the fingerprint area that have been offered as proof that fingerprint error rates are very low, are inadequate.

A. Two Types of Proficiency Tests

Everyone claims to support testing in the forensic sciences. Moreover, everyone claims to support a kind of testing called "proficiency testing." Proficiency tests are urged by all professional forensic science organizations, by all of the relevant NAS reports cited herein,[144] and by nearly every scholar who has written about them. The problem, though, is that there is no uniform agreement about the underlying goal of a proficiency test. For some, the goal of a proficiency test is to measure the accuracy rate of examiners working with various samples under various conditions. For others, the goal is to determine whether an examiner can follow laboratory procedures. For still others the goal is to assess whether a laboratory's procedures are adequate. To complicate matters, some people distinguish proficiency tests from performance tests, achievements tests, competency tests, and accuracy tests, whereas others do not.[145] Because the proficiency testing language is such an integral part of the forensic science profession and scholarly literature, I adopt this terminology as well. However, I distinguish sharply between two broad types of proficiency tests identified here for the first time as Type I and Type II proficiency tests. The two types of proficiency tests can be distinguished by their different purposes.[146]

---

144. NRC II, *supra* note 11 at 4, 37 (explaining that "[l]aboratories should participate regularly in proficiency tests[,]" though not for the purpose of identifying error rates).

145. *See, e.g.*, FED. BUREAU OF INVESTIGATION, QUALITY ASSURANCE STANDARDS FOR FORENSIC DNA TESTING LABORATORIES (2011), https://www.fbi.gov/file-repository/quality-assurance-standards-for-forensic-dna-testing-laboratories.pdf/view; Diana Scarborough, *Proficiency Testing Versus Competency Assessment*, N.C. PUB. HEALTH MGMT. BULL., Sept. 2013, at 1; *see also* NAT'L COMM'N ON FORENSIC SCI., U.S. DEP'T OF JUSTICE, PROFICIENCY TESTING IN FORENSIC SCIENCE 2 (2016), https://www.justice.gov/ncfs/file/831811/download ("[Proficiency testing is] an evaluation of participant performance against pre-established criteria by means of interlaboratory comparisons for the determination of service provider performance.") [hereinafter PROFICIENCY TESTING: FINAL DRAFT]; KELLY M. PYREK, PIONEERS IN FORENSIC SCIENCE: INNOVATIONS AND ISSUES IN PRACTICE 99 (2017) ("[The passing of a competency test is] the demonstration that an FSP [Forensic Science Practitioner] has acquired and demonstrated specialized knowledge, skills, and abilities in the standard practices necessary to conduct examinations in a discipline and/or category of testing prior to performing independent casework.").

146. The Type I and Type II nomenclature has been adopted in other fields as a way to distinguish between two important non-overlapping categories of a phenomenon. Physicians

1. Type I Proficiency Tests

Type I proficiency tests serve purposes that are primarily internal to forensic science, forensic laboratories, and forensic examiners. These tests are designed to identify the strengths and weaknesses in procedures and personnel for the internal purpose of improving forensic science work. As such, Type I proficiency tests have value for addressing such questions as, "Are the training programs sufficient?" "Are the laboratory protocols clear?" and "Are laboratory personnel able to follow the protocols to generate a result that competent others obtain?" Because these tests serve internal purposes, the tests may be designed in ways that reflect the particular needs or concerns of the laboratories or the examiners. For example, a laboratory that is deciding which examiners to promote might ask each to participate in a Type I testing situation in which ground truth is known. The results of such a test (i.e., the accuracy of the answers provided by each examiner) would presumably help the laboratories decide which examiner to promote. However, the test results would not provide a reasonable estimate of casework accuracy or error rates.

2. Type II Proficiency Tests

Type II proficiency tests serve purposes that are primarily external to forensic science, forensic laboratories, and forensic examiners. These tests are specifically designed to provide casework *performance information* to one or more external constituencies. Unlike Type I proficiency tests, the primary goal of Type II tests is not to generate information that will improve forensic science work. Instead, *the goal of Type II proficiency tests is to provide information about the accuracy of forensic science conclusions and opinions for the benefit of those who use this information*. Trial judges may

---

speak of Type I and Type II diabetes to distinguish between a total lack of insulin from the pancreas (Type I) and the body's failure to respond properly to insulin (Type II). *See, e.g.*, Hannah Nichols, *Diabetes: The Differences Between Types 1 and 2*, MED. NEWS TODAY (June 27, 2017), https://www.medicalnewstoday.com/articles/7504.php. Statisticians refer to Type I and Type II errors in the context of hypothesis testing, where the former describes the error of rejecting a true null hypothesis, and the latter describes the error of failing to reject a false null hypothesis. *See, e.g.*, RICHARD P. RUNYON & AUDREY HABER, FUNDAMENTALS OF BEHAVIORAL STATISTICS 268–69 (6th ed. 1988). Psychiatrists refer to Type I and Type II traumas to distinguish between traumas that result from "one sudden blow," and traumas caused by "longstanding or repeated ordeals." Lenore C. Terr, *Childhood Traumas: An Outline and Overview*, 148 AM. J. PSYCHIATRY 10, 11 (1991). In a similar vein, Type I and Type II proficiency tests describe two important types of proficiency tests that have different purposes and that are therefore designed, conducted, and interpreted in different ways.

need to know the accuracy of a forensic method before ruling on the admissibility of some item of forensic evidence. Police investigators and jurors may need to know the accuracy of a forensic method in order to assign weight to a reported match. Because the focus in Type II proficiency tests is on estimating rates of casework accuracy, it is important that relevant casework conditions be simulated as closely as possible by these tests.[147]

3. A Non-Forensic Illustration of Type I and Type II Proficiency Tests

Though Type I and Type II proficiency tests share the "proficiency test" appellation, the procedures required to satisfy the external goals of Type II tests are different than those required to satisfy the internal goals of Type I tests.[148] Consider the following analogy. When a high school student in the U.S. prepares for college, he or she generally takes the Scholastic Achievement Test (SAT). As the student prepares for the SAT, the student and the student's parents have an internal goal: they want to identify areas of strength and weakness in the student's SAT-relevant skills in order to improve those skills prior to taking the exam. The student may accomplish this goal by taking various practice tests in SAT guidebooks. The student's performance on these practice tests can help direct his or her future studies and thereby lead to improvement in selected areas. The SAT practice tests are Type I proficiency tests because their purpose is to identify areas of weakness which, in turn, can point to ways to improve the student's future performance.

When the student eventually sits for the SAT exam administered by the College Board, the goals and intended constituencies are now *external*. Unlike practice tests in SAT guidebooks, the purpose of the actual SAT exam is *not* to identify areas of weakness in the student to direct remedial efforts. Instead, the purpose of the actual SAT exam is to inform one or more external constituencies (e.g., a college admissions staff) about the student's scholastic abilities and aptitude. As such, the testing conditions required by the College Board will be substantially more rigorous than those the student used when

147. The simulation of relevant real-world conditions is referred to as "ecological validity" in the social sciences. *See* Jonathan J. Koehler & John B. Meixner, *Jury Simulation Goals*, *in* THE PSYCHOLOGY OF JURIES 161, 162 (Margaret Bull Kovera ed., 2017) (defining ecological validity in jury simulation research as "how well the experimental setting mimics real world settings of interest").

148. Angi M. Christensen, Christian M. Crowder, Stephen D. Ousley & Max M. Houck, *Error and its Meaning in Forensic Science*, 59 J. FORENSIC SCI. 123, 125 (2014) ("[I]t is not acceptable to derive error rates from practitioner proficiency tests, professional exercises, or studies that were not designed to estimate method error rates.").

preparing for the exam. In short, the actual SAT exam is a Type II proficiency test because its purpose is to identify the test-taker's performance for the benefit of an external constituency.

The take away point from this example is that just as there is a world of difference between the goals of SAT practice tests and the actual SAT exam, there is a world of difference between the Type I proficiency tests that forensic labs routinely conduct and the Type II proficiency tests that forensic labs never conduct. Whether we're talking about scholastic knowledge or forensic science skills, the results of a Type I proficiency test are a poor substitute for a Type II proficiency test.[149]

Unlike Type I proficiency tests, a well-designed Type II proficiency test *can* provide an estimate of (a) the rate at which professionals make casework errors, and (b) the circumstances under which those errors most likely arise. Errors are a natural part of any scientific endeavor, and scientists must not only strive to reduce those errors (via, for example, Type I proficiency tests), but they must also be vigilant about identifying and measuring them (via Type II proficiency tests). Even a careful, well-trained, experienced, and honest scientist who employs a reliable technology or method will not obtain the right answer every time. Human error is always possible: samples may be mixed-up, mislabeled, miscoded, altered, or contaminated. Equipment used to run samples may be miscalibrated, and technical glitches and failures may occur without warning and without being noticed. Results may be misread, misinterpreted, misrecorded, mislabeled, mixed-up, misplaced, or discarded. Type I proficiency tests may help pinpoint the types of problems that arise and may help the forensic science community identify solutions. But only

149. *See* Gary Edmond, Matthew B. Thompson & Jason M. Tangen, *A Guide to Interpreting Forensic Testimony: Scientific Approaches to Fingerprint Evidence*, 13 LAW PROBABILITY & RISK 1, 21 (2014) ("[I]t is important to distinguish the proficiency tests currently used by fingerprint examiners, such as those provided by Collaborative Testing Services . . . . Commercial proficiency tests do not adequately address the general issue of expert matching accuracy and were not designed to disentangle the factors that affect matching accuracy."). Thousands of forensic science examiners participate every year in the Type I proficiency tests provided by Collaborative Testing Services (CTS) and other commercial providers. But, as noted previously, CTS expressly states that its tests are not designed to measure error rates or otherwise provide "an overview of the quality of work performed in the profession." TOOLMARKS EXAMINATION, *supra* note 128, at 1; *see also* LATENT PRINT, *supra* note 101, at 33 ("[P]roficiency tests designed and administered for certification and quality improvement purposes bear little resemblance to actual casework . . . . [These] proficiency tests are neither designed for nor particularly suitable for estimating error rates."); *see also* Mnookin et al., *supra* note 97, at 745 ("[N]o laboratory of which we are aware regularly conducts blind proficiency tests that are given in the stream of casework in a pattern or impression discipline, or, for that matter, in any other forensic discipline.").

Type II proficiency tests can provide consumers of forensic results with an estimate of the rate at which such errors occur.

4. Responding to Opposition to Type II Proficiency Tests

Within the academic forensic science community, some have expressed opposition to Type II proficiency tests (i.e., tests designed to estimate error rates). Their primary argument is that error rates generated from such tests are unhelpful because they are insufficiently specific to the case, laboratory, or examiner under consideration.[150] Some have suggested that judges and jurors would be better served by case specific evidence, such as specific evidence that the sample in question was contaminated, or that a re-examination of the evidence produced a different result.

I offer a few points in response. First, I note that there is no significant resistance to measuring error rates in other important applied areas of science. For example, there is a mountain of data pertaining to the rates at which medical errors occur. Virtually all medical tests and procedures undergo continuous testing to identify the conditions under which they are more and less likely to provide accurate results. The studies behind these tests are typically published in high quality peer-reviewed journals and accepted as a vital part of the scientific process. In 1999, the National Academy of Sciences Institute of Medicine published a report on medical errors that had an enormous impact on health policy in the U.S.[151] This report reviewed

150. *See* JANE A. LEWIS, FORENSIC DOCUMENT EXAMINATION: FUNDAMENTALS AND CURRENT TRENDS 130 (2014) ("Forensic document examination, like many other forensic science disciplines, does not have an error rate. It is not possible to calculate because each case's evidence and every examiner examining the evidence is unique."); Bruce Budowle et al., *Perspective on Errors, Bias, and Interpretation in the Forensic Sciences and Direction for Continuing Advancement*, 54 J. FORENSIC SCI. 798, 801 (2009) ("[S]uggesting that a specific error rate must be presented adds little value to the discussion on reliability. A community-wide error rate is not meaningful, because it falsely reduces the rate of error for those who might commit the most errors and wrongly increases the rate for those who are the most proficient."); Stephen G. Bunch, Erich D. Smith, Brandon N. Giroux & Douglas P. Murphy, *Is a Match Really a Match? A Primer on the Procedures and Validity of Firearm and Toolmark Identification*, 11 FORENSIC SCI. COMM. 1, 5 (2009) ("Aggregate data do not speak directly to individual error rates, which, owing to small sample size and learning/self-correction, are difficult if not impossible to determine reliably."); Ate Kloosterman, Marjan Sjerps & Astrid Quak, *Error Rates in Forensic DNA Analysis: Definition, Numbers, Impact and Communication*, 12 FORENSIC SCI. INT'L 77, 83 (2014) ("[I]t would be essentially meaningless and potentially misleading to report general error rates from proficiency testing . . . . [T]he general numbers are not representative for the specific circumstances of the case.").

151. INST. OF MED., TO ERR IS HUMAN: BUILDING A SAFER HEALTH SYSTEM 26 (Linda T. Kohn et al. eds., 2000).

hundreds of studies pertaining to medical errors, including meta-analyses that identified error rates for physicians, hospitals, and drugs. The report found that error rates for drugs prescribed for adults and children are 0.3% and 0.5% respectively.[152] It found that the risk of "significant" medication errors is 0.2%,[153] and the error rate for manufacturers who mix drugs is 0.3%.[154] The Report also contained a lengthy appendix that summarized the results of numerous medical error rate studies.[155] Yet we don't see the argument in medicine that such tests are misleading or have little value because they are insufficiently specific to individual patients.

Second, everyone would agree that case specific evidence in the forensic domain that points to an error having occurred is relevant information. The problem, though, is that when such evidence is not available (as is typically the case), legal decision makers have no way to evaluate the probative value of reported match evidence. This is a point of some confusion. Even after judges and juries are provided with a list of the testifying expert's qualifications, along with detailed information about the technique in question, they still know virtually nothing about the accuracy of evidence. This is why *Daubert's* error rate factor is so important. But as documented above, when judges attempt to satisfy *Daubert's* error rate factor in cases that challenge the admissibility of a forensic method, they rely on the results of Type I proficiency tests (such as those prepared by Collaborative Testing Services) as a proxy for casework error rates. Forensic science experts likewise have testified that an error rate could be computed from the results of Type I proficiency tests.[156] Of course, the problem with using Type I proficiency test results as a proxy for Type II test results is that there are many important differences between the two types of tests in terms of the population sampled, the sampling procedure, the testing procedure, sample difficulty level, test blindness, and a host of other factors. In short, it makes

152. *Id.* at 33–34.

153. *Id.* at 33; *id.* app. C at 241.

154. *Id.* at 193.

155. *Id.* app. C at 215–53.

156. United States v. Baines, 573 F.3d 979, 983 (10th Cir. 2009) ("Agent Meagher testified that . . . each analyst would know his or her error rate from the proficiency examination taken at the end of training and annually thereafter."); State v. Proctor, 595 S.E.2d 480, 483 (S.C. 2004) ("At trial, Lt. Jeffcoat testified that the SLED DNA lab used proficiency testing to ensure its analysts were accurate. He was permitted to testify, over respondent's objections, 'In every occasion where we have been provided proficiency tests, we've always called the correct match.'").

no sense to rely on Type I proficiency test results as an indicator of low error rates in casework.[157]

The absence of Type II proficiency test data is problematic for jurors for the same reasons. Of course, jurors will learn something about the forensic evidence from direct and cross-exam. But, realistically, jurors will almost never have any basis for questioning the examiner's opinions and conclusions in the target case. After all, what basis could there be for suggesting that, in this particular case, the examiner misrecorded, mislabeled, switched samples, or created some other type of human error that produced a false positive match? In this respect, forensic science testimony is very different from, say, eyewitness testimony where jurors' common sense and life experiences provide them with a basis for assessing whether the eyewitness could be mistaken given the case circumstances (e.g., the witness was far away, was impaired, only had a partial view, etc.). This is a major reason why it is so important for jurors to have access to Type II proficiency test data in forensic science cases.[158]

---

157. Returning to the SAT analogy, just as no one would expect college admissions committees to rely on the results of SAT practice tests, we should not expect courts to rely on proficiency tests that the test manufacturers themselves say were not designed to measure the accuracy of work performed in the profession.

158. An issue that is beyond the scope of this paper but worth mentioning briefly at this juncture is how the results of Type II proficiency testing might be communicated to jurors to maximize the chance that they will understand these data and use them correctly. One approach that I have championed in cases that involve a forensic match involves providing jurors with a single statistic that approximately captures the underlying probative value of the match report. In cases where the RMP is extremely low (e.g., one in millions, billions, or trillions), this statistic is approximated by the false positive error rate. A rigorous Type II proficiency test could provide a ballpark estimate of this error rate. Where data from such a test is available, an expert could testify along these lines:

> I am reporting a match between the suspect's DNA (or shoeprint, fingerprint, etc.) and DNA recovered from the crime scene. This match report doesn't necessarily mean that the crime scene DNA belongs to the suspect. There are a number of ways in which I might have found a match even if the crime scene sample belonged to someone other than the suspect. For example, there is a chance that the suspect matches by coincidence. And there is a chance that I inadvertently made an error of some sort and there really isn't a match at all. There are various ways in which such an error could occur and we know from experience that errors of this sort have occurred in the past. Taking into account the possibility of coincidence based on genetic principles and the risk of various types of errors based on proficiency tests, the approximate chance that I would report a match between the suspect and the crime scene sample if, in fact, the suspect is not the source is about 1 in ___.

B. Getting to Type II Proficiency Tests

Implementation of a broad-based Type II proficiency testing program faces an uphill political battle. Various criminal justice constituencies have much to lose if the results of those tests fail to confirm the hopes and expectations of the forensic science, legal, and public communities. But science should be indifferent to such things. Scientists are supposed to observe, measure, test, analyze, and replicate before drawing cautious inferences. Type II proficiency testing is a classic scientific endeavor. But it will not come to fruition unless and until there is consensus on at least four matters.

First, there must be consensus that judges and jurors need to know more than they currently know about the accuracy of forensic science evidence. If policy makers and judges continue to accept the weak arguments that forensic reliability is amply demonstrated by (a) a long history of admissibility in court, (b) a relatively low proportion of conclusively demonstrated casework errors, and (c) anecdotal reports that errors are "rare" and error rates are "low," then the push for Type II proficiency tests to measure error rates will fail.

Second, there must be consensus that it is important for judges and jurors to be exposed to data that speak to forensic science casework error rates, *even though such data could never identify the precise risk of error in a given individual case*. Scientists have long appreciated the value of such "base rate" data in a wide array of decision making tasks.[159] But it is also well-documented that many people fall victim to the "base rate fallacy" in which they mistakenly believe that probabilistic data collected at a group level—such as industry-wide error rate data for a given forensic technique—have little relevance for individual level probabilistic predictions.[160] The importance of identifying base rates for error in the forensic sciences has been discussed elsewhere.[161] Still, some influential forensic science scholars

---

Again, the 1 in ___ frequency estimate would most likely be similar to (or even identical to) the false positive error rate generated from a Type II proficiency test. *See* Koehler, Chia & Lindsey, *supra* note 41, at 201; Thompson, Taroni & Aitken, *supra* note 41, at 1.

159. Paul E. Meehl & Albert Rosen, *Antecedent Probability and the Efficiency of Psychometric Signs, Patterns, or Cutting Scores*, 52 PSYCHOL. BULL. 194, 195–200 (1955).

160. Daniel Kahneman & Amos Tversky, *On the Psychology of Prediction*, 80 PSYCHOL. REV. 237, 238–41 (1973); *see also* Jonathan J. Koehler, *The Base Rate Fallacy Reconsidered: Descriptive, Normative, and Methodological Challenges*, 19 BEHAV. & BRAIN SCI. 1, 1–2 (1996).

161. *See* Jonathan J. Koehler, *Proficiency Tests to Estimate Error Rates in the Forensic Sciences*, 12 LAW PROBABILITY & RISK 89, 92–93 (2013); Michael J. Saks & Jonathan J. Koehler, *Questions About Forensic Science: Response*, 311 SCIENCE 607, 609 (2006); Koehler, *supra* note 54, at 1089 n.34; Saks & Koehler, *supra* note 30, at 895.

oppose measuring and publicizing general error rates because those rates will overstate the error risk for examiners who are either better than average or working on an easy case.[162] In response, I note (again) that identification of a general error rate is not intended to be the last word on the risk of error in an individual case. It is merely a necessary, statistically appropriate, starting point. Those who believe that this rate over or understates the risk of error in a particular case would and should provide evidence of the specific factors in the target case that justify adjusting the general error rate upwards and downwards.[163]

Third, there must be consensus that Type II proficiency tests should be conducted by *disinterested parties*.[164] For example, the FBI should not have responsibility for funding, designing, or conducting Type II proficiency tests that the FBI could then use to tout the accuracy of their own forensic laboratories or scientists. Some minimal administrative role of interested parties may be required to carry out the tests (e.g., an FBI administrative insider may need to distribute and return blind proficiency test materials). But care should be taken to reduce the risk that insiders could affect the test outcomes.

The use of disinterested researchers is hardly a radical or novel idea. The importance of avoiding conflicts of interest in research is well-understood in the medical domain. Several years ago, the NAS's Institute of Medicine (IOM)[165] identified a series of recommendations to deal with potential

---

162. Christophe Champod, *Research Focused Mainly on Bias Will Paralyze Forensic Science*, 54 SCI. & JUST. 107, 107–08 (2014). *But see* Edmond, Thompson & Tangen, *supra* note 149, at 14–15 ("[T]he expression of an indicative, or general, error rate recognizes that comparison processes are fallible in circumstances where we are not entirely sure what a match actually means.").

163. For example, if an examiner was highly experienced and reliable Type II proficiency test data indicated that highly experienced examiners made 25% fewer false positive errors than less experienced examiners, those data might be helpful. Likewise, credible data that point in the opposite direction might be helpful as well.

164. Koehler, *supra* note 161, at 91. The disinterested researchers requirement is not based on a presumption that interested researchers are anything other than honest, well-intentioned scientists. But a large body of research indicates that the goals, hopes, and desires of scientists, like non-scientists, may subtly affect the way they conduct their studies, interpret their data, and describe their conclusions. For a stark analysis of the implications of this problem in the social sciences, see John P. A. Ioannidis, *Why Most Published Research Findings Are False*, 2 PLOS MED. 696, 696–97 (2005); *see also* Saul M. Kassin, Itiel E. Dror & Jeff Kukucka, *The Forensic Confirmation Bias: Problems, Perspectives, and Proposed Solutions*, 2 J. APPLIED RES. MEMORY & COGNITION 42, 42–44 (2013) (showing Itiel Dror's work on cognitive bias in the forensic sciences).

165. The Institute of Medicine was established as an independent, nonprofit organization within the National Academy of Sciences to advise the government on matters of health policy. INST. OF MED., *supra* note 151, at 3–4. In March 2016, the IOM was renamed the Health and

conflicts of interest in medical research.[166] IOM's working definition of a conflict of interest was "a set of circumstances that creates a risk that professional judgment or actions regarding a primary interest will be unduly influenced by a secondary interest."[167] The risk of such conflicts arises in forensic science research and IOM's recommendations are applicable here as well.

Fourth, there must be consensus that examiners should be unaware of when they are in a test situation. There is reason to believe that some forensic scientists, like other professionals, respond to test samples in known test situations differently from the way they respond to ordinary casework sample.[168] But if the error rates obtained from tests are intended to provide a reasonable first pass estimate for the rate of error in casework, *examiners must be blind to whether they are examining a real case or a test case*.[169] Test blindness is sometimes opposed on grounds that it is either difficult or impossible to achieve in practice.[170] However, blind proficiency tests have, on occasion, been used for DNA analyses. Joe Peterson and colleagues conducted a detailed pilot investigation in the USA, which showed that "blind tests can be constructed and successfully submitted to forensic DNA laboratories."[171] Smaller scale blind proficiency tests for DNA analyses were conducted in the early DNA evidence days as well.[172] Rand et al. (2002) also

---

Medicine Division. *About Us*, NAT'L ACAD. SCI. ENGINEERING MED., http://www.nationalacademies.org/hmd/About-HMD.aspx (last visited Dec. 9, 2017).

166. COMM. ON CONFLICT OF INTEREST IN MED. RESEARCH, EDUC. & PRACTICE, INST. OF MED., CONFLICT OF INTEREST IN MEDICAL RESEARCH, EDUCATION, AND PRACTICE 16–17 (Bernard Lo & Marilyn J. Field eds., 2009) [hereinafter CONFLICT OF INTEREST].

167. *Id.* at 46 (emphasis omitted).

168. Glenn Langenburg, *A Performance Study of the ACE-V Process: A Pilot Study to Measure the Accuracy, Precision, Reproducibility, Repeatability, and Biasability of Conclusions Resulting from the ACE-V Process*, 59 J. FORENSIC IDENTIFICATION 219, 220 (2009).

169. The virtues of blinding scientists in all academic areas from potentially biasing influences is now widely discussed. *See, e.g.*, Robert J. MacCoun & Saul Perlmutter, *Blind Analysis as a Correction for Confirmatory Bias in Physics and Psychology*, *in* PSYCHOLOGICAL SCIENCE UNDER SCRUTINY 297, 304 (Scott O. Lilienfeld & Irwin D. Waldman eds., 2017).

170. JOHN M. BUTLER, ADVANCED TOPICS IN FORENSIC DNA TYPING: METHODOLOGY 174 (2012) ("[A] number of challenges and costs are associated with blind proficiency tests." (citation omitted)); NRC II, *supra* note 11, at 24 ("[T]he logistics of constructing fully blind proficiency tests are formidable.").

171. JOSEPH L. PETERSON & R.E. GAENSSLEN, NAT'L INST. OF JUSTICE, DEVELOPING CRITERIA FOR MODEL EXTERNAL DNA PROFICIENCY TESTING, FINAL REPORT, 104 (2001); *see also* Joseph L. Peterson, G. Lin, M. Ho & R.E. Gaensslen, *The Feasibility of External Blind DNA Proficiency Testing. I. Background and Findings*, 48 J. FORENSIC SCI. 21, 22 (2003).

172. *See* MARGARET KUO, ORANGE CTY. SHERIFF CORONERS CRIME LAB., CAL. ASS'N OF CRIME LAB. DIRS., DNA COMMITTEE REPORT #6 (1988); MARGARET KUO, ORANGE CTY. SHERIFF CORONERS CRIME LAB., CAL. ASS'N OF CRIME LAB. DIRS., DNA COMMITTEE—RESULTS OF

reported the results of DNA blind trials across 129 laboratories in twenty-eight European countries.[173] Though it is not always easy to keep examiners in the dark about whether the work they are doing is for an actual case or for a proficiency test, such blindness will be easier to achieve as we continue to move toward a system where forensic scientists receive information on an as-needed basis using, for example, a sequential unmasking process.[174]

C. The Best Existing Studies Fall Short

Although most forensic sciences do not test for rates of error in any serious way, there have been a number of fingerprint error rate studies over the past twenty years. Most of these studies suffer from obvious shortcomings or design flaws that harm their credibility as indicators of casework error rates.[175] However, two recent studies deserve special attention. One study,

---

BLIND TRIAL #2 (1990); Masamitsu Honma, Tomio Yoshii, Ikuo Ishiyama, Kohnosuke Mitani, Ryo Kominami & Masami Muramatsu, *Individual Identification from Semen by the Deoxyribonucleic Acid (DNA) Fingerprint Technique*, 34 J. FORENSIC SCI. 222, 222 (1989); P. Sean Walsh, Nicola Fildes, Alan S. Louie & Russell Higuchi, *Report of the Blind Trial of the Cetus AmpliType HLA DQ-alpha Forensic Deoxyribonucleic Acid (DNA) Amplification and Typing Kit*, 36 J. FORENSIC SCI. 1551, 1551 (1991).

173. Steven Rand, Marianne Schürenkamp & Bernd Brinkmann, *The GEDNAP (German DNA Profiling Group) Blind Trial Concept*, 116 INT'L J. LEGAL MED. 199, 201 (2002); *see also* S. Rand, M. Schürenkamp, C. Hohoff & B. Brinkmann, *The GEDNAP Blind Trial Concept Part II. Trends and Developments*, 118 INT'L J. LEGAL MED. 83, 83 (2004).

174. The basic idea behind sequential unmasking is that forensic examiners perform as much of their work as possible while "blind" to case information that is not required for them to perform their analyses. Information required for the examiner to draw conclusions is "unmasked" as needed. William C. Thompson, *Interpretation: Observer Effects*, *in* WILEY ENCYCLOPEDIA OF FORENSIC SCIENCE 1575, 1577–78 (Allan Jamieson & Andre Moenssens eds., 2009); Dan E. Krane et al., *Sequential Unmasking: A Means of Minimizing Observer Effects in Forensic DNA Interpretation*, 53 J. FORENSIC SCI. 1006, 1006 (2008).

175. *See* Ralph Norma Haber & Lyn Haber, *Experimental Results of Fingerprint Comparison Validity and Reliability: A Review and Critical Analysis*, 54 SCI. & JUST. 375, 388 (2014) (reviewing thirteen published fingerprint experiments from 1996 to 2012 that purport to provide information on fingerprint accuracy and reliability, and concluding that, "[n]ot one of these 13 experiments can justify an estimate of the erroneous identification in fingerprint comparison casework, and certainly not the low rates reported in their results"). In a sharp rebuttal, three forensic science scholars countered that this review was "one-sided" and "a result of some partisan agenda." Glenn Langenburg, Cedric Neumann & Christophe Champod, *A Comment on Experimental Results of Fingerprint Comparison Validity and Reliability: A Review and Critical Analysis*, 54 SCI. & JUST. 393, 393, 395 (2014).

published in the Proceedings of the National Academy of Sciences in 2011,[176] merits attention because it was designed to respond to the 2009 NAS Report's call for attention to error rate. The other study[177] merits attention because it was funded by the National Institute of Justice, which has since offered up this study as proof that error rates in fingerprint analysis are minuscule.

In Ulery et al. (2011), 169 latent print examiners were presented with roughly 100 fingerprint pairs,[178] about 30% of which were non-mated pairs (i.e., from different fingers).[179] The authors reported that the non-mated fingerprints were selected to present a challenge to the examiners and to be similar to those that might be encountered in casework. Self-reports from the participants indicated that a majority agreed that the test was challenging. The authors reported that five of the latent print examiners committed a total of six false positive errors out of 3,628 attempts, for a false positive error rate (per sample examined) of 0.17%.[180] Ulery et al. (2011) also shows that 85% of examiners committed at least one false negative error, and that the false negative error rate (per sample examined) was 10.9%.[181] Although Ulery et al. (2011) has been cited by courts as evidence that the error rate for fingerprint identification is "quite low,"[182] its value as an indicator of false positive error rate *in casework* is questionable. The participants were volunteers and may not be representative of the fingerprint examiners who testify in court. This is important because it may be that examiners who are most likely to err in casework are less likely to volunteer to participate in a study where their shortcomings may be exposed. Another significant shortcoming in the study is that the participants were aware that they were

176. BRADFORD T. ULERY, R. AUSTIN HICKLIN, JOANN BUSCAGLIA & MARIA ANTONIA ROBERTS, ACCURACY AND RELIABILITY OF FORENSIC LATENT FINGERPRINT DECISIONS 7733 (Stephen E. Feinberg ed., 2011).

177. IGOR PACHECO, BRIAN CERCHIAI & STEPHANIE STOILOFF, MIAMI-DADE RESEARCH STUDY FOR THE RELIABILITY OF THE ACE-V PROCESS 2 (2014), https://www.ncjrs.gov/pdffiles1/nij/grants/248534.pdf.

178. ULERY, HICKLIN, BUSCAGLIA & ROBERTS, *supra* note 176, at 7733.

179. *Id.* at 7734 ("There were 520 mated and 224 nonmated pairs.").

180. *Id.* at 7735. The authors report the number of attempts as 4,083 (and along with this a false positive error rate of 0.1%). However, that figure includes 455 instances in which examiners reported "inconclusive" and therefore did not have a "chance" to err. Therefore, the appropriate figure in the denominator is 3,628 rather than 4,083; *see also* LATENT PRINT, *supra* note 101, at 37.

181. ULERY, HICKLIN, BUSCAGLIA & ROBERTS, *supra* note 176, at 7736. Once again, the authors arrive at a lower error rate (7.5%) due to their inclusion of 1,856 "inconclusive" conclusions in the denominator. When the inconclusive conclusions (which represented nearly 1/3 of the mated samples) are set aside, the false negative error rate is 450 / 4,113 = 10.94%.

182. United States v. Love, No. 10cr2418–MMM, 2011 WL 2173644, at *5 (S.D. Cal. June 1, 2011) ("[A] false positive rate of 0.1% [is] quite low.").

being tested. We know that fingerprint examiners respond differently when they know that they are in a testing situation.[183] The representativeness of the samples used is also questionable.[184] Finally, this study fails the disinterested researcher requirement for Type II proficiency tests because it was paid for by the FBI, and two of the four authors work for the FBI.[185] The authors

183. Langenburg, *supra* note 168, at 242 (referring to a "bias loop" that arises when examiners know their work will be checked by verifiers who also know that they are merely verifying another examiner's decision). Others take issue with the suggestion raised in Ralph Norman Haber & Lyn Haber's article that examiners who know they are being tested will "perform better than when the tests are not announced and cannot be differentiated from routine work." Haber & Haber, *supra* note 175, at 386. R. Austin Hicklin et al. respond as follows:

> While participants in tests may indeed have different performance than in routine work, it is not reasonable to conclude that the results are necessarily better in the tests: a few examiners who are not taking the test seriously could have notably affected the results of a study, especially with respect to rare events. For example, we do not know if the examiner who made two erroneous individualizations was acting as s/he would have in routine work, or was just tired and apathetic, given it was just a test. It seems likely that at least some of the participants took the test less seriously than casework, given the serious implications of actual casework, and the absence of any negative implications on an anonymous test.

R. Austin Hicklin, Bradford T. Ulery, JoAnn Buscaglia & Maria Antonia Roberts, *In Response to Haber and Haber, "Experimental Results of Fingerprint Comparison Validity and Reliability: A Review and Critical Analysis,"* 54 SCI. & JUST. 390, 391 (2014).

184. BRADFORD T. ULERY, R. AUSTIN HICKLIN, JOANN BUSCAGLIA & MARIA ANTONIA ROBERTS, A STUDY OF THE ACCURACY AND RELIABILITY OF FORENSIC LATENT FINGERPRINT DECISIONS APPENDIX: SUPPORTING INFORMATION 3 (2011), http://www.pnas.org/content/suppl/2011/04/19/1018707108.DCSupplemental/Appendix.pdf (cautioning that, "the overall distribution of the fingerprint data cannot as a whole be considered as statistically representative of operational data," though they suggest that the prints used included a large proportion of poor quality prints).

185. ULERY, HICKLIN, BUSCAGLIA & ROBERTS, *supra* note 176, at 7738 ("This is publication number 10-19 of the FBI Laboratory Division. This work was funded in part under a contract award to Noblis, Inc. from the FBI Biometric Center of Excellence and in part by the FBI Laboratory Division.").

disclosed these features. But disclosure of potential conflicts of interest does not eliminate the threat that the conflicts pose.[186]

In Pacheco, Cerchiai & Stoiloff (2014), 109 fingerprint examiner volunteers were presented with various pairings from eighty latent prints that were produced from ten known sources.[187] Each examiner offered conclusions on multiple pairs of prints. Twenty-eight of the examiners committed at least one false positive error.[188] False positive errors were committed on twenty-one of the eighty latent prints.[189] Thus, the false positive errors do not appear to have been confined to a few incompetent examiners or to a few particularly difficult prints. There were forty-two erroneous identifications out of 995 chances (excluding inconclusives) for a false positive error rate of 4.2%.[190] There were 235 erroneous exclusions out of 2,692 chances (excluding inconclusives) for a false negative error rate of 8.7%.[191]

But not all false positive errors are equal, and most of those reported in this study really shouldn't "count" as false positive errors if we are concerned with *who* is the source of the fingerprint as opposed to *which finger* is the source of the fingerprint. Pacheco, Cerchiai & Stoiloff (2014) report that thirty-five of the forty-two false positive errors seemed to be "clerical errors" in which the correct person was selected but the wrong finger was

---

186. *See supra* note 164. Requiring a disinterested proficiency test administrator does not equate to a claim that non-disinterested administrators are prone to dishonesty.

> A conflict of interest is not an actual occurrence of bias or a corrupt decision but, rather, a set of circumstances that past experience and other evidence have shown poses a risk that primary interests may be compromised by secondary interests. The existence of a conflict of interest does not imply that any individual is improperly motivated.

CONFLICT OF INTEREST, *supra* note 166, at 61. As such, the point is that good scientific practice should motivate use of disinterested parties to design and conduct Type II proficiency tests.

187. PACHECO, CERCHIAI & STOILOFF, *supra* note 177, at 2.

188. *Id.* at 64 ("[Twenty-eight] of 109 participants committed an identification error.").

189. *Id.*

190. *Id.* at 53 tbl.4. Examiners concluded that 42 pairs from different sources were identifications, and that 953 pairs from different sources were exclusions. The sum of 42 and 953 is 995. The inconclusive examinations that appear in Table 4 are ignored.

191. *Id.* Examiners concluded that 235 pairs from the same source were exclusions, and that 2,457 pairs from the same source were identifications. The sum of 235 and 2,457 is 2,692. The inconclusive examinations that appear in Table 4 are ignored.

identified.[192] If we move those thirty-five minor false positives into the correct calls category, we are left with seven major false positive errors (i.e., a person who was not the source was falsely identified as the source). This translates to a 0.7% false positive error rate (i.e., about one false positive error per 142 trials).

This study also provides evidence about the value of verification for catching false positive errors. The forty-two false positives were divided up and assigned to one of three verification conditions: a group of different examiners, a group of examiners who were led to believe that they were the second verifiers, and the original examiners themselves (months later). Most, but not all, of the false positive errors were not repeated by verifiers. It is not clear from the report whether any of the seven major false positive identifications were falsely verified or not.

Unfortunately, there has been some misleading hype about this study. A Department of Justice (DOJ) press release claimed that the study showed that, "examiners had a 0% false positive . . . rate."[193] DOJ apparently arrived at this conclusion by focusing only on the fact that one of three groups of verifiers did not repeat any of the fifteen false positive errors that were committed by the 109 examiners in the first stage of the study. The authors indicate that some of the other false positive errors committed in the first stage *were* repeated by a second and third group of verifiers though, again, it is not clear whether any of the seven major false positive errors were repeated or not.

Overall, the results of the Pacheco et al. (2014) Miami-Dade study are encouraging. But, as noted above, this study has some of the same shortcomings as the Ulery et al. (2011) study that limits its value as a reasonable first-pass estimate of the false positive error rate in fingerprint

---

192. *Id.* at 64–65.

> [I]n 35 of the 42 erroneous identifications the participants appear to have made a clerical error, but the authors could not determine this with certainty. A clerical error was defined as a circumstance in which the participants chose the correct standard from the three standards presented, however, the opposite finger, . . . opposite palm, . . . or incorrect finger was reported . . . . The remaining seven errors appear to be true erroneous identifications, in which the incorrect standard was reported, or where the source was not present for that particular trial.

Id.

193. Press Release, Office of Justice Programs, Dep't of Justice, Fingerprint Examiners Found to Have Very Low Error Rates (Feb. 2, 2015), http://ojp.gov/newsroom/pressreleases/2015/ojppr02022015.pdf.

casework.[194] Specifically, the participants were volunteers,[195] the study was not blind (examiners knew they were being tested[196]), and the authors may not have been disinterested parties (all were employed by the Miami-Dade Police Department[197]). If we are serious about estimating *casework* error rates these features are not acceptable.[198]

CONCLUSION

Nobody knows how accurate the opinions and conclusions offered by DNA analysts, firearms examiners, odontologists, document examiners, blood spatter specialists, or any other forensic scientists are. As noted at the outset, we can't even begin to estimate accuracy rates (or error rates) in most areas of forensic science because none of the requisite studies have been conducted.[199]

There is plenty of blame to go around for this shameful state of affairs. The 1996 National Academy of Sciences panel that evaluated DNA evidence deserves blame both for failing to emphasize how important accuracy testing is and for going so far as to suggest that testing for error rates is a bad idea.[200] The people responsible for developing legal standards and rules deserve blame for not putting more teeth in their otherwise sensible requirement that proffered expert testimony (including forensic science testimony) must be reliable to be admissible in court. Judges deserve blame for repeatedly crediting the unsupported testimony of forensic scientists and historical precedent on matters of reliability rather than demanding proof of reliability

194. *See supra* text accompanying notes 176–186.

195. PACHECO, CERCHIAI & STOILOFF, *supra* note 177, at 23.

196. *Id.* at 6.

197. *Id.* at 34.

198. *See* Koehler, *supra* note 161, at 91–92 (discussing features that should be incorporated into error rate proficiency tests).

199. *See* PACHECO, CERCHIAI & STOILOFF, *supra* note 177; ULERY, HICKLIN, BUSCAGLIA & ROBERTS, *supra* note 176. The PCAST report is more generous than I am in terms of its evaluation of the utility of the fingerprint studies. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 23, at 101 ("[L]atent fingerprint analysis is a foundationally valid subjective methodology—albeit with a false positive rate that is substantial and is likely to be higher than expected by many jurors based on longstanding claims about the infallibility of fingerprint analysis.").

200. NRC II, *supra* note 11, at 80 ("The objective of both proficiency-testing and auditing is to improve laboratory performance by identifying problems that need to be corrected. Neither is designed to measure error rates."); *see also id.* at 185 (acknowledging that NRC I recommended that the results of high quality proficiency tests designed to measure error rates should be conducted and disclosed to juries, but distancing itself from this recommendation, saying, "we attempt no such policy judgment").

from scientific studies. The forensic science leadership deserves blame for failing to create and promote a scientific culture within the discipline that emphasizes empiricism, independence, transparency, conservatism, and an ongoing critical perspective.

The good news is that, despite all of these failings, some reform efforts are under way. Many forensic scientists and their respective professional organizations have eschewed certainty claims and 0% error claims in favor of scientifically defensible language.[201] In the version of this paper that was accepted for publication shortly before the 2016 presidential election, I included the establishment of the National Commission on Forensic Science (NCFS) in 2013 as another positive step in the march toward improving the practice and reliability of the forensic sciences as another sign of progress and enlightenment in the forensic world. The NCFS was a federal advisory committee for the U.S. Department of Justice. It included dozens of scientific area committees and subcommittees (including one on accreditation and proficiency testing),[202] hundreds of experts, and dozens of work products and recommendations that were adopted by the broader commission.[203] However, in April 2017, Attorney General Jeff Sessions put an end to NCFS by

---

201. On occasion, a court will suggest such claims by forensic scientists should be unacceptable. *See* Williams v. United States, 130 A.3d 343, 345 (D.C. 2016). In *Williams*, the prosecution's firearms expert testified with certainty that a particular bullet was fired from a particular gun. *Id.* Although the appellate court upheld the conviction because defense counsel failed to object at trial, *id.* at 351, it admonished the expert's certainty statements:

> [A] certainty statement regarding toolmark pattern matching has the same probative value as the vision of a psychic: it reflects nothing more than the individual's foundationless faith in what he believes to be true. This is not evidence on which we can in good conscience rely, particularly in criminal cases, where we demand proof—*real* proof—beyond a reasonable doubt, precisely because the stakes are so high. To uphold the public's trust, the District of Columbia courts must bar the admission of these certainty statements.

Id. at 355 (Easterly, J., concurring).

202. *See Archive of National Commission on Forensic Science Material*, U.S. DEP'T OF JUSTICE, https://www.justice.gov/archives/ncfs (last visited Dec. 9, 2017) (reviewing the NCFS's purpose, structure and goals); *see also* M. Chris Fabricant & Tucker Carrington, *The Shifted Paradigm: Forensic Science's Overdue Evolution from Magic to Law*, 4 VA. J. CRIM. L. 1, 28 (2016) (discussing NCFS efforts to set standards for forensic science and improve its reliability).

203. *Archive of Work Products Adopted by the National Commission on Forensic Science*, U.S. DEP'T OF JUSTICE, https://www.justice.gov/ncfs/work-products-adopted-commission (last visited Dec. 9, 2017).

declining to renew its charter.[204] It is not yet clear what, if anything, will replace NCFS.

A silver lining in the dissolution of NCFS is that any newly formed entity will be in a position to take a fresh look at the issue of proficiency testing in forensic science. The NCFS subcommittee tasked with this responsibility fell short. In a document entitled "Proficiency Testing in Forensic Science," the NCFS subcommittee failed to acknowledge that proficiency testing has anything to do with measuring end-result accuracy.[205] Instead, this document is replete with the definitions of proficiency tests that have nothing to do with measuring accuracy, error rates, or anything else that people who must judge the validity of forensic science evidence need to know.[206]

Whether or not the entity that replaces NCFS, a duly formed legal body, or a forward-thinking forensic science organization have the will and clout to challenge the status quo, the time has surely come for the broader criminal justice system to face the fact that consumers of forensic science evidence (judges, jurors, the public) do not have the information they need to assess the probative value of forensic science opinions and conclusions. Implementation of a broad, mandatory Type II proficiency testing program that focuses on identifying rates of error under various casework conditions—and for samples that vary in difficulty—would be an enormous step in the right direction.

---

204. 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=.287fcd771119; *see* U.S. DEP'T OF JUSTICE, NOTICE OF PUBLIC COMMENT PERIOD ON ADVANCING FORENSIC SCIENCE 3 (2017), https://assets.documentcloud.org/documents/3549200/Justice-Department-to-seek-public-comment-on.pdf (noting federal advisory committees operate on two-year, renewable terms, and the latest term was set to expire on April 23, 2017).

205. PROFICIENCY TESTING: FINAL DRAFT, *supra* note 145, at 1–2. In response to this criticism, the subcommittee wrote (in part), "The intention of the document was to educate and explain the various definitions and ways proficiency testing is currently used in forensic science." NAT'L COMM'N ON FORENSIC SCI., U.S. DEP'T OF JUSTICE, PROFICIENCY TESTING IN FORENSIC SCIENCE 3 (2016) [hereinafter PROFICIENCY TESTING: Adjudication of Final Draft Comments], https://www.justice.gov/archives/ncfs/page/file/831811/download. But focusing on "definitions and ways proficiency testing *is currently used*," only serves to reinforce the status quo which, as discussed throughout this paper, is woefully inadequate. *Id.* (emphasis added).

206. *See* PROFICIENCY TESTING: Adjudication of Final Draft Comments, *supra* note 205, at 2 ("For example, the document says that proficiency testing is . . . a mechanism for checking to see if an organization can 'adhere to the organization's procedures' . . . a tool that 'can be utilized prior to achieving accreditation' . . . [and] 'an evaluation of performance against pre-established criteria by means of interlaboratory comparisons.'") (referring to PROFICIENCY TESTING: FINAL DRAFT, *supra* note 145).