ARTICLE

# THE PROFICIENCY OF EXPERTS

BRANDON L. GARRETT† & GREGORY MITCHELL††

*Expert evidence plays a crucial role in civil and criminal litigation. Changes in the rules concerning expert admissibility, following the Supreme Court's* Daubert *ruling, strengthened judicial review of the reliability and the validity of an expert's methods. Judges and scholars, however, have neglected the threshold question for expert evidence: whether a person should be qualified as an expert in the first place. Judges traditionally focus on credentials or experience when qualifying experts without regard to whether those criteria are good proxies for true expertise. We argue that credentials and experience are often poor proxies for proficiency. Qualification of an expert presumes that the witness can perform in a particular domain with a proficiency that non-experts cannot achieve, yet many experts cannot provide empirical evidence that they do in fact perform at high levels of proficiency. To demonstrate the importance of proficiency data, we collect and analyze two decades of proficiency testing of latent fingerprint examiners. In this important domain, we found surprisingly high rates of false positive identifications for the period 1995 to 2016. These data would qualify the claims of many fingerprint examiners regarding their near infallibility, but unfortunately, judges do not seek out such information. We survey the federal and state case law and show how judges typically accept expert*

† White Burkett Miller Professor of Law and Public Affairs, Justice Thurgood Marshall Distinguished Professor of Law, University of Virginia School of Law.

†† Joseph Weintraub–Bank of America Distinguished Professor of Law, University of Virginia School of Law. Many thanks to Barbara Armacost, Simon Cole, Aliza Cover, Brian Feinstein, Kim Ferzan, Kim Forde-Mazrui, Aziz Huq, Sharon Kelley, Jay Koehler, Genevieve Lakier, Richard McAdams, Daniel McConkie, John Monahan, Dan Murrie, John Rappaport, Diego Zambrano, and participants at faculty workshop at the University of Virginia School of Law, a public law workshop at the University of Chicago School of Law, and an ABA CJS Academic Roundtable for their invaluable comments on earlier drafts, to the Center for Statistics and Applications in Forensic Evidence and the National Institute for Standards and Technology for their research support, and to Andrew Bae, Stephanie Boutsicaris, Elizabeth Hoffman, and Tess Sewell for invaluable research assistance.

*credentials as a proxy for proficiency in lieu of direct proof of proficiency. Indeed, judges often reject parties' attempts to obtain and introduce at trial empirical data on an expert's actual proficiency. We argue that any expert who purports to give falsifiable opinions can be subjected to proficiency testing and that proficiency testing is the only objective means of assessing the accuracy and reliability of experts who rely on subjective judgments to formulate their opinions (so-called "black-box experts"). Judges should use proficiency data to make expert qualification decisions when the data is available, should demand proof of proficiency before qualifying black-box experts, and should admit at trial proficiency data for any qualified expert. We seek to revitalize the standard for qualifying experts: expertise should equal proficiency.*

## INTRODUCTION

Expert witnesses appear in a vast number of cases every year.[1] In civil cases, experts often address questions central to liability and damages, and in criminal cases, they address questions touching on both guilt and punishment. Following the Supreme Court's ruling in *Daubert v. Merrell Dow Pharmaceuticals*,[2] and subsequent revisions to Federal Rule of Evidence 702 dealing with expert evidence,[3] judges now have a much greater authority and responsibility to inquire into the reliability and validity of expert's methods. The threshold question, however, is whether a person is "qualified" to be an expert based on "knowledge, skill, experience, training, or education."[4] To answer this question, courts routinely accept a witness's own self-serving statements of expertise buoyed by educational credentials, professional training, or experience, rarely spending much time on this threshold question before moving on to examine the methods used and conclusions reached by the putative expert.[5] In this Article, we seek to revitalize the expert qualification inquiry and encourage greater reflection on what should be required of expert witnesses.

What does it mean to label someone an "expert"? From a social scientific perspective, the label "expert" means something different than simply having specialized education or experience. It is not a matter of credentials or work history but rather a question of *performance*: "Expertise is defined as a sequence of mastered challenges with increasing levels of difficulty in specific areas of functioning."[6] Experts are those who are particularly proficient on a task or who

---

1 Samuel R. Gross, *Expert Evidence*, 1991 WIS. L. REV. 1113, 1119 (finding experts in 86% of civil trials sampled); Andrew W. Jurs, *Expert Prevalence, Persuasion, and Price: What Trial Participants Really Think About Experts*, 91 IND. L.J. 353, 355 (2016) ("[T]he data reveals that expert witnesses appear in 86% of the cases in the study, which is an identical percentage as in two prior research studies.").

2 509 U.S. 579, 592-95 (1993) (holding that the trial judge must make "a preliminary assessment of whether the reasoning or methodology underlying the testimony is scientifically valid").

3 FED. R. EVID. 702 (requiring expert testimony to be based on "sufficient facts or data" and the product of "reliable principles and methods").

4 *Id.*

5 *See infra* Section II.A; *see also* DAVID FAIGMAN ET AL., 5 MOD. SCI. EVID. § 43:2 (2016–2017 ed.) (summarizing that "courts typically are generous in finding that a proposed expert's training or experience satisfies Federal Rule of Evidence 702 or its state equivalent").

6 Barry J. Zimmerman, *Development and Adaptation of Expertise: The Role of Self-Regulatory Processes and Beliefs*, *in* THE CAMBRIDGE HANDBOOK OF EXPERTISE AND EXPERT PERFORMANCE 705, 706 (2006) (citation omitted); *accord* Barbara A. Spellman, *Judges, Expertise, and Analogy*, *in* THE PSYCHOLOGY OF JUDICIAL DECISION MAKING 149, 152 (David Klein & Gregory Mitchell eds., 2010) ("Due to study, training, and practice—often in addition to talent and motivation—experts are better than nonexperts in some domain of performance."); David J. Weiss & James Shanteau, *Decloaking the Privileged Expert*, 18 J. MGMT. & ORG. 300, 307 (2012) ("Rather than thinking generically of people as experts, we prefer to say that a person has demonstrated expertise in a specific set of tasks."); *see also* EXEC. OFFICE OF THE PRESIDENT, PRESIDENT'S COUNCIL OF ADVISORS ON SCI. AND TECH., FORENSIC SCIENCE IN CRIMINAL COURTS: ENSURING SCIENTIFIC

are especially knowledgeable about a subject. The expert interpreter can translate one language into another with a high degree of accuracy across language samples that would remain inscrutable to non-experts. The expert cytologist can differentiate cancerous cells from non-cancerous cells with a high degree of accuracy and with a high degree of reliability when those samples are submitted for retesting, while most non-experts examining the samples would perform at levels no better than chance. An expert on American history can answer questions on arcane historical topics with a degree of accuracy and reliability far beyond that exhibited by the average person. Expertise may be acquired in many different ways, but anyone who claims to be an expert should be able to prove that expertise empirically through superior performance within the domain of purported expertise.[7] Ideally, assessments of expertise make use of a performance measure that indisputably separates good performance from bad. Where such a "gold standard" for good performance exists,[8] an expert can be subjected to what is commonly called proficiency testing.[9] In proficiency testing, the putative expert's response on a test can be objectively scored as correct or incorrect. Thus, a candidate to become a court interpreter can be presented with a number of foreign phrases whose English meanings are known and her interpretations can be evaluated for their accuracy. Or a cytologist can be presented with cell samples known to include cancerous and non-cancerous cells to assess how accurately the cytologist distinguishes the two types of cells. Proficiency testing also permits an assessment of an individual's reliability, or consistency, in performance: to what extent does the person give the same answers across like items and

---

VALIDITY OF FEATURE-COMPARISON METHODS 6 (Sept. 2016) [hereinafter PCAST Report] ("Demonstrating that an expert is *capable* of reliably applying the method is crucial—especially for subjective methods, in which human judgment plays a central role.").

7 The Federal Rules of Evidence take a functionalist, relativistic approach to expertise: experts are those who possess knowledge that will enable them to provide information that is not generally known by non-expert laypersons. *See* FED. R. EVID. 702 advisory committee's note ("Whether the situation is a proper one for the use of expert testimony is to be determined on the basis of assisting the trier. 'There is no more certain test for determining when experts may be used than the common sense inquiry whether the untrained layman would be qualified to determine intelligently and to the best possible degree the particular issue without enlightenment from those having a specialized understanding of the subject involved in the dispute.'" (citation omitted)).

8 David J. Weiss & James Shanteau, *Empirical Assessment of Expertise*, 45 HUM. FACTORS 104, 104 (2003) ("The ideal is to correlate action with a *gold standard*, an unequivocally valid, universally accepted outcome measure that directly reflects the behavior under scrutiny."). When a gold standard of performance does not exist, other benchmarks can be used to distinguish expert from non-expert performance, but disagreements may arise with respect to the suitability of these other benchmarks.

9 *See* Jonathan J. Koehler, *Fingerprint Error Rates and Proficiency Tests: What They Are and Why They Matter*, 59 HASTINGS L.J. 1077, 1091 (2008) ("A proficiency test is an assessment of the performance of laboratory personnel using samples whose sources are known to the proficiency test administrator but unknown to the examinee.").

different answers across different items?[10] Increasingly difficult tests can be used to identify those most expert at a task.

Unlike judges who focus on the experience or credentials of an expert, social scientists who study experts emphasize performance over self-serving statements and credentials because the latter are often unreliable guides to true expertise. "Experts have often been identified by self-proclamation or acclamation by other experts as well as by experience, titles, and degrees. However, these methods can be misleading when searching for an expert."[11] In some domains, governments require that persons and organizations claiming to be experts on some task demonstrate that expertise empirically, through performance on proficiency tests tailored to that task. Thus, persons seeking to serve as court interpreters must pass proficiency tests designed to ensure expertise in the languages to be interpreted,[12] and clinical labs that screen human samples for diagnostic testing must participate in regular proficiency testing that produces results the public can examine and compare.[13]

In one large and important domain, however, neither federal nor state governments require performance-based evidence of expertise: in order to

---

10 We focus in this Article on experts that provide conclusions based on analysis of facts in a case. Experts may also testify about more general scientific research, in order to provide a "framework" to educate factfinders, and they may explain industry or professional norms as well. *See* Laurens Walker & John Monahan, *Social Frameworks: A New Use of Social Science in Law*, 73 VA. L. REV. 559, 570 (1987) (defining the "social framework" as "the use of general conclusions from social science research in determining factual issues in a specific case" (emphasis omitted) (footnote omitted)). Such expertise can also be tested; the person's knowledge of the relevant research and standards can be assessed. However, it is particularly important that "black box" experts be tested, since such experts claim to reach conclusions using methods that may be opaque.

Similarly, we do not focus in this Article on experts whose work is not a "black box," but where they perform a test that uses a scientific method, or even an automated method. There may be questions whether a test to analyze material for whether it contains controlled substances is the correct test, or whether the drug-testing machine was properly calibrated, but the method itself is not a black box, so long as its processes are disclosed in litigation. *See* PCAST Report, *supra* note 6, at 5 (distinguishing between objective methods, for which foundational validity can be studied by measuring accuracy, and subjective methods, for which "black box" evaluation must be conducted). When experts report the results of automated or machine methods, then the focus should be on the validity and reliability of the machine, not the expert. *See* Andrea Roth, *Machine Testimony*, 126 YALE L. J. 1972, 1979 (2017) (criticizing how testimony by human experts "might create a veneer of scrutiny when in fact the actual source of the information, the machine, remains largely unscrutinized").

11 Weiss & Shanteau, *supra* note 8, at 104.

12 For federal English–Spanish interpreter certification rules, which require both written and oral examinations, see *Federal Court Interpreter Certification Examination*, U.S. COURTS, http://www.uscourts.gov/services-forms/federal-court-interpreters/federal-court-interpreter-certification-examination [http://perma.cc/BE3X-LU74]; *see also, e.g.*, FLORIDA RULES FOR CERTIFICATION AND REGULATION OF SPOKEN LANGUAGE COURT INTERPRETERS 4-6, http://www.flcourts.org/core/fileparse.php/419/urlt/formatted-interpreter-rules-May-1.pdf [https://perma.cc/EJH4-5H7Q] (establishing multiple levels of expertise for court interpreters, with increasing qualifications for each level).

13 42 U.S.C. § 263a(b) (2012). Section III.A describes these regulations in greater detail.

qualify as an expert witness at trial, a person need only proclaim that she has specialized knowledge or ability, obtained through education or experience, that would enable her to supply relevant information that a non-expert witness could not supply.[14] The testimony that the expert hopes to give is supposed to be the product of reliable methods applied to sufficient data, but there is no requirement that the expert demonstrate her proficiency at giving correct answers to the kinds of questions counsel will pose to her at trial.[15]

Take the example of fingerprint experts, who have testified in criminal cases for over a hundred years.[16] The latent fingerprint examiner uses some objective criteria initially, when categorizing a print as having a "whorl" or a "loop" pattern.[17] However, the analysis that follows is "purely subjective," requiring an evaluation of details in a print and comparing it to details in another print.[18] Such a subjective method can be valid and reliable.[19] But based on what we currently know, fingerprint analysis has a "substantial" error rate, making it especially important to assess how "expert" an individual examiner is at fingerprint identifications.[20] Yet, judges do not ask fingerprint examiners to come forward with evidence that they can correctly match latent prints to known prints with a high (or any) degree of proficiency. The examiner seeking to testify need only describe training and familiarity with the methods in the field to determine whether latent fingerprints from a crime scene match the fingerprints taken from known individuals. Judges do not inquire further even if she boasts that she has a zero or near-zero error rate in fingerprint identifications. Under present practice, as we will describe, both federal and state courts regularly accept proxies for expert performance in lieu of actual performance data. Judges also deny opposing parties access to proficiency testing data when it exists, despite unsupported claims by experts regarding their supposed proficiency.[21]

---

14 Weiss and Shanteau label experts whose performance is not subjected to empirical scrutiny "privileged experts," and they include expert witnesses within this category. Weiss & Shanteau, *supra* note 8, at 300-02.

15 See Section I.A for a description of how courts do not examine proficiency when deciding whether to qualify experts.

16 PCAST Report, *supra* note 6, at 9 ("Latent fingerprint analysis was first proposed for use in criminal identification in the 1800s and has been used for more than a century.").

17 *Id.* at 89 n.253.

18 *Id.* at 9.

19 *Id.* (calling fingerprint analysis a "foundationally valid subjective methodology").

20 *Id.*

21 See Section II.A for more on the trend of judicial disregard for proficiency testing in evaluating expert witnesses. For prominent criticism of such testimony in the fingerprint area, see PCAST Report, *supra* note 6, at 87-103; and NAT'L RESEARCH COUNCIL, COMMITTEE ON IDENTIFYING THE NEEDS OF THE FORENSIC SCIENCES COMMUNITY, STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES: A PATH FORWARD 136-45 (National Academies

This odd state of affairs—in which faith is placed in self-proclamations of expertise and judges ignore the most probative evidence on expertise—cannot be attributed to the lack of gold-standard measures of performance in the areas in which experts seek to testify in court. An expert testifying in any area for which there are better and worse ways of doing a task or correct and incorrect answers to questions can be subjected to proficiency testing. These conditions apply, for instance, to medical experts in civil cases and forensic experts in criminal cases.[22] In fact, some fields of expertise already engage in proficiency testing, but only disappointingly few do so, and even fewer make their proficiency data public or willingly share it in discovery. These fields usually adopted proficiency testing when mandated to do so or as a part of an effort to regulate quality within a field to dispel credibility concerns, as with the move to proficiency testing by fingerprint examiner associations. Not only is proficiency testing commonly not done in many fields, but what little is done is often not sufficiently challenging or realistic in its design. A leading commercial provider for forensic proficiency tests candidly explained: "Easy tests are favored by the community."[23] Where an expert's primary, or perhaps only, market is the courtroom, as is the case with most forensic experts and a variety of experts in civil litigation, the expert has little to gain and much to lose from engaging voluntarily in proficiency testing. Absent a mandate from courts or regulators, widespread proficiency testing is unlikely to occur.

Only recently has this proficiency problem begun to receive real attention. Most notably, the White House Presidential Council for Advisers on Science and Technology (PCAST) issued a report in September 2016 underscoring the "essential" need for proficiency testing of forensic experts to assess "an examiner's capability and performance in making accurate judgments," in a manner that is realistic, routine, and under the supervision of a disinterested third party.[24] Unfortunately, as that report noted, in criminal cases, examiners have long testified that they were infallible without having had their proficiency rigorously assessed.[25] Several scholars have highlighted how

---

Press 2009) [hereinafter NAS Report], identifying lack of documentation and high error rate in fingerprint analysis as key problems for in-court evidentiary use.

22 Forensics involves the analysis of crime scene evidence to identify participants in a possible crime and to determine the manner in which a crime may have been committed. *See, e.g.*, NAS Report, *supra* note 21, at 35-36.

23 PCAST Report, *supra* note 6, at 57 n.133 (quoting Christopher Czyryca, President, Collaborative Testing Services, Inc.).

24 *Id.* at 102.

25 *Id.* at 3; *see also* Simon A. Cole, *More Than Zero: Accounting for Error in Latent Fingerprint Identifications*, 95 J. CRIM. L. & CRIMINOLOGY 985, 1043, 1048 (2005) (pointing to individual instances of courts accepting expert self-reporting of zero or near-zero error rates); Jonathan J. Koehler, *Forensics or Fauxrensics? Ascertaining Accuracy in the Forensic Sciences*, 49 ARIZ. ST. L.J. 1369, 1371 (2017) ("Numerous forensic authorities and respected textbook authors encourage such

existing proficiency testing suggests troubling error rates and that more rigorous proficiency is needed.[26] Few have suggested that proficiency should be taken into account by judges.[27] In this Article, we describe how proficiency should inform the legal standards for the qualification of experts and should be at the core of evidence law's understanding of expertise.

In Part I of this Article, we set out the standards for how courts currently qualify experts and then we turn to a more detailed explanation of the concept of proficiency. We propose a rethinking of the approach towards expert qualification, making the case for performance-based assessments of expertise. We proceed first with a more detailed discussion of proficiency testing and its benefits. Second, we demonstrate the value of proficiency data by consolidating all of the publicly available results of fingerprint examiner proficiency testing from the past two decades. Our analysis reveals that, contrary to common claims made by fingerprint experts in court and elsewhere, the method of fingerprint examination is far from error free. Error rates ranged from 1-2% to 10-20% per test administration with respect to false positives (i.e., erroneous linking of a latent print to an individual's known print) in proficiency tests from 1995 to 2016. We found an average 7% false positive rate over this time period and an average 7% false negative (i.e., erroneous failure to link a latent print to an individual's known print) rate. Presentation of this data at trials might well have changed the outcome in some cases by altering the jury or judge's beliefs about the near infallibility of fingerprint identifications.[28]

---

hyperbole" and citing sources). The earlier 2009 NRC Report also highlighted the need for "routine, mandatory proficiency testing that emulates a realistic, representative cross-section of casework." NAS Report, *supra* note 21, at 25.

26 *See, e.g.*, Paul C. Giannelli, *Expert Testimony and the Confrontation Clause*, 22 CAP. U. L. REV. 45, 80 (1993) ("The proficiency test results of many common laboratory examinations are disturbing."); Koehler, *supra* note 9, at 1077 ("Critics charge that fingerprint analysis lacks an empirical foundation and that examiners make exaggerated claims that are likely to mislead jurors."); Koehler, *supra* note 25, at 1369 ("Unless and until such [proficiency testing] studies are undertaken, legal decision makers will continue to fly blind when it comes to assessing the reliability of a reported forensic match."); *Seventh Circuit Upholds the Reliability of Expert Testimony Regarding the Source of a Latent Fingerprint*—United States v. Havvard, 260 F.3d 597 (7th Cir. 2001), 115 HARV. L. REV. 2349, 2356 (2002) (recommending "generation and administration" of [] proficiency tests"); *see also* Barack Obama, *The President's Role in Advancing Criminal Justice Reform*, 130 HARV. L. REV. 811, 862 (2017) (calling for ongoing work to strengthen forensic science).

27 Gary Edmond, *Forensic Science Evidence and the Conditions for Rational (Jury) Evaluation*, 39 MELB. U. L. REV. 77, 85-86 (2015) ("[R]egardless of qualifications and experience, rigorous proficiency testing tells us whether the forensic analyst performs a task or set of tasks better than non-experts or chance. A significantly enhanced level of performance is precisely what it means to be an expert.").

28 In a work in progress, we find that presentation of proficiency data does impact the weight that lay jurors attach to forensic evidence. Greg Mitchell & Brandon Garrett, *The Impact of Proficiency Testing Information on the Weight Given to Fingerprint Evidence* (work in progress) (on file with authors).

The fingerprint examiner proficiency data highlight a particular point in favor of mandatory proficiency testing: fingerprint examiners, like many other experts, rely on subjective judgment to formulate their opinions.[29] An expert who uses a wholly objective method to formulate an opinion can be shown to have erred by showing that proper use of the method does not in fact yield the result the expert claims. In contrast, whenever an expert's method involves subjective judgment, even in part, then the only means for testing the accuracy and reliability of the expert's method is through proficiency testing.[30] An expert who uses a subjective method is a "black box" into which data is fed and out of which magically pops an answer. Such an expert can never be shown definitively to have erred in applying a method because that method cannot be observed and applied by others. However, if proficiency data for such "black-box experts" exists, then we can assess their basic levels of proficiency, which provides important information about their ability to provide accurate and reliable information. Experts reaching conclusions using subjective methods may be highly reliable. But walking through the courtroom door is unlikely to transform an "expert" who regularly receives low scores on proficiency tests into a highly reliable source of information in the case at hand.

Excellence on proficiency tests does not guarantee accuracy at trial, but the opinions of the less proficient witness are, in general, more likely to be wrong. Thus, many medical residents know, in theory, how to read an X-ray to find the problems discussed in textbooks and could serve as expert witnesses under prevailing law, but that knowledge alone does not ensure expert performance at reading X-rays. Experts should be qualified based on empirical evidence of their proficiency before addressing whether their methods used and conclusions reached are valid and reliable.

In Part II, we use the contrast between claimed and actual expertise to criticize courts' reluctance to require proficiency testing, permit discovery of proficiency data, or admit proficiency data at trial. After reviewing the case law on proficiency data, we side with courts that recognize the value of

---

29 *See* PCAST Report, *supra* note 6, at 58 ("Proficiency testing is especially critical for subjective methods: because the procedure is not based solely on objective criteria but relies on human judgment, it is inherently vulnerable to error and inter-examiner variability.").

30 *Id.* Where objective evidence of innocence becomes available after trial (e.g., through DNA testing in rape cases), it is possible to determine, after the fact, whether an expert using subjective judgment erred. *See, e.g.*, Brandon L. Garrett & Peter J. Neufeld, *Invalid Forensic Science Testimony and Wrongful Convictions*, 95 VA. L. REV. 1, 1 (2009) (exploring "the forensic science testimony by prosecution experts in the trials of innocent persons, all convicted of serious crimes, who were later exonerated by post-conviction DNA testing"). Of course, this method can only be used in a relatively small number of cases and provides no protection against unreliable experts before the fact. *Id.* at 7-8 (noting that the data set is "unrepresentative of typical criminal cases," and the study does not examine whether "an examiner made a mistake or engaged in misconduct in the laboratory").

proficiency data and recognize that existing rules for expert evidence and discovery can accommodate demands for proficiency data. We argue that there is no good justification for continuing to qualify experts based on poor proxies for proficiency, like credentials or experience, when proficiency testing is possible or proficiency data already exists. In areas in which proficiency testing is voluntary, one might argue allowing discovery of proficiency data could deter voluntary testing, but this concern disappears in a world in which proficiency testing is made mandatory or it is strongly preferred. We also explain how our approach would buttress reliability analyses under *Daubert* and Federal Rule of Evidence 702 and would produce important evidence for a jury's consideration. Proficiency data provides an objective basis for excluding the opinions of any expert with unacceptably low proficiency and for assessing the weight of expert evidence.

In Part III, we describe a regulatory approach towards assuring proficiency of experts using realistic blind proficiency testing, which could help ensure that disciplines have adequate proficiency data to present in court. A federal agency could regulate proficiency more broadly, as leading scientific groups have proposed, and as is currently done for clinical laboratories, and in other countries.

A proficiency-based approach to expert qualification would improve the quality of evidence at trial, and it would simplify judicial gatekeeping of expert evidence. Judges should adopt the view that expertise equals proficiency, or legislatures should impose that view on courts.

## I. TRUE EXPERTISE = PROFICIENCY

### A. *Judicial Qualification of Experts Using Education and Experience*

The Supreme Court's most prominent ruling on expert evidence, *Daubert v. Merrell Dow Pharmaceuticals,* prescribes how judges should determine whether the opinions an expert plans to offer are sufficiently reliable to provide helpful knowledge.[31] That analysis focuses on the validity and reliability of the methods that the expert uses. That inquiry is important; if the method as a whole is invalid and error-prone, then no matter how accomplished the particular expert is at the method, the evidence should not be admitted in court. Our focus here, however, is not on the question of the validity and reliability of an entire method. Our focus is on the neglected but logically prior question of how to decide who qualifies as an expert.[32] When

31 509 U.S. 579, 592-95 (1993).

32 Similarly, two decisions of the Court that built on the *Daubert* decision failed to address this question. *See* Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999); Gen. Elec. Co. v. Joiner, 522 U.S. 136 (1997).

a party proposes that a witness be allowed to testify as an expert, the trial judge must first determine whether that witness is in fact an expert. Federal Rule of Evidence 702 defines an expert as one who is qualified "by knowledge, skill, experience, training, or education" to give testimony that "will help the trier of fact to understand the evidence or to determine a fact in issue."[33] The Supreme Court has not interpreted this language, and therefore, the primary guidance for making the qualification decision comes from the text of Rule 702, advisory notes to the rule, and lower court interpretations of the rule.

The text of Rule 702 does not specify what comprises adequate "knowledge, skill, experience, training, or education," but the advisory notes emphasize that "experience alone" may be sufficient as a foundation for expert testimony and that experience is "the predominant, if not sole, basis for a great deal of reliable expert testimony" in some fields.[34] Lower courts, when they do occasionally comment on the qualification question, likewise emphasize that expertise can come from education or experience and that no special credentials are required.[35] In practice, courts often require little more than a statement from the expert that she has developed specialized knowledge in the domain in which she seeks to testify. Accordingly, expert qualification typically involves a recitation of the formal education, training, experience, and achievements that the putative expert has in a field, along with a statement of familiarity with the body of knowledge needed to formulate an opinion that may be helpful in the case.[36] Most of the work in ensuring the helpfulness of expert evidence consists in a review of the

33 FED. R. EVID. 702. State rules of evidence typically contain similar qualification language. *E.g.*, TENN. R. EVID. 702 ("If scientific, technical, or other specialized knowledge will substantially assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise."); W. VA. R. EVID. 702(a) ("If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education may testify thereto in the form of an opinion or otherwise.").

34 FED. R. EVID. 702 advisory committee's note to 2000 amendment.

35 *See, e.g.*, Tuf Racing Products, Inc. v. American Suzuki Motor Corp., 223 F.3d 585, 591, 54 Fed. R. Evid. Serv. 1492 (7th Cir. 2000) ("The notion that [*Daubert*] requires particular credentials for an expert witness is radically unsound." (citation omitted)).

36 *See, e.g.*, Warren Eginton, *A View from the Bench—The Expert in the Courtroom*, 3 PROD. LIAB. L.J. 114, 117 (1992) (stating "the curriculum vitae of the expert will be most important" to impress jurors). Some jury studies suggest that jurors place at least some weight on credentials, although it may vary based on the type of case and the presentation of the testimony. *See* Sanja Kutnjak Ivković & Valerie P. Hans, *Jurors' Evaluations Of Expert Testimony: Judging The Messenger And The Message*, 28 LAW & SOC. INQUIRY 441, 458-64 (2003) (discussing how jurors consider an expert's credentials).

sufficiency of the data considered by the expert and the reliability of the methods used and conclusions reached.[37]

Yet as we will see, for a large class of experts—those who rely in part or in whole on subjective or intuitive judgments to form their opinions—there exist no objective standards for evaluating the sufficiency of the data they considered or the reliability of their "black-box" method. For these experts, the qualification decision effectively becomes the reliability review. But credentials such as number of times previously testifying, number or type of degrees, years of experience, or membership in professional organizations—the very kind of information that judges look to in making the qualification decision—are poor indicators of whether these persons truly do have the ability to do what they claim to be expert at doing.[38]

One response to the problem of expert qualification is to suggest that Rule 702 should be revised, to focus in its text and in its Advisory Notes on empirical assessment of proficiency. We would support doing so, and in fact, the Advisory Committee is soliciting input on Rule 702 prompted by concerns raised about the reliability of expert evidence often admitted, particularly in criminal cases.[39]

However, another way to view the problem of expert qualification is that courts have not interpreted Rule 702 correctly. The text of Rule 702 does not specify what comprises adequate "knowledge, skill, experience, training, or education," and, while the Advisory Notes emphasize that "experience alone" may be sufficient,[40] the rule clearly contemplates that the witness will have uncommon knowledge and an ability to use that knowledge expertly. In other words, before a court reaches the question of whether an expert's opinions in the courtroom are reliable, the rule contemplates that the proffered witness is capable of expert performance outside the courtroom on the kinds of questions the expert will be asked to answer in the courtroom. This expertise inquiry could be much simpler and more accurate if courts did not settle for credentials and self-proclamations of expertise.

---

37 This is at least true in those jurisdictions that follow *Daubert*. *See, e.g.*, United States v. Frazier, 387 F.3d 1244, 1261 (11th Cir. 2004) ("[T]he unremarkable observation that an expert may be qualified by experience does not mean that experience, standing alone, is a sufficient foundation rendering reliable *any* conceivable opinion the expert may express.").

38 For an article making a similar observation about the standards for expert admissibility in Australia, see Edmond, *supra* note 27, at 98 ("Conventional admissibility criteria and heuristics—such as formal qualifications, a 'field', 'training, study or experience', prior legal recognition and admission—do not provide direct insight into validity, error rates and limitations, or proficiency.").

39 Advisory Committee on Rules of Evidence, Spring 2017 Meeting 2 (Apr. 21, 2017), http://www.uscourts.gov/sites/default/files/advisory_committee_on_rules_of_evidence_-_spring_2017_meeting_materials.pdf [https://perma.cc/463P-UZBX] (describing a planned conference for October 2017 to discuss whether changes are necessary to Rule 702 due to concerns with its use, particularly in criminal cases).

40 *See supra* note 34 and accompanying text.

So, rather than interpret Rule 702 to endorse a liberal test of expert qualification that leaves critical scrutiny to the reliability review, an alternative interpretation that is consistent with the text and advisory note, and that has more merit from a functional perspective, is that Rule 702 cares more about expertise *per se* than about any particular route to expertise. Hence, the rule's refusal to endorse any particular expert requirements does not mean that having specialized education or experience is sufficient. Rather, what matters is being able to use that specialized knowledge in a way that could be helpful to the case. Under this performance-based interpretation, what matters is whether one has *demonstrable* expertise. To make this determination, information about credentials and experience can be helpful, but this information will always be only indirect evidence that an individual has true expertise that may be useful to the case. For any type of expert, direct evidence of expertise can be obtained.

Some experts serve only to "educate the factfinder about general principles," without reaching conclusions that attempt to "apply these principles to the specific facts of the case."[41] Proficiency can be tested even for experts who only testify about general information that may have some bearing on the case. For the expert describing background research, direct evidence of expertise would consist of proof that the individual is in fact familiar with the relevant literature that he or she seeks to summarize or translate for the jury. Courts are usually content to rely on indirect evidence of expertise to qualify such an expert, and leave direct attacks on expertise to cross-examination. That approach is defensible under Rule 702, at least where indirect evidence of expertise exists and we have reason to believe it is a good proxy for expertise.[42] But direct evidence of expertise can be obtained for the expert who just serves as a lecturer to the jury on some topic through the use of proficiency tests that assess the expert's familiarity with the relevant literature.

We should therefore see Rule 702 as imposing three basic requirements on putative expert witnesses: (a) proper tools, (b) proper data, and (c) true expertise. An ax is an excellent tool for chopping wood, but it is a terrible tool for cutting paper, and many people are not expert at using an ax for any purpose. Under Rule 702, it is not enough to know what tool should be used to analyze data to generate an answer for the case; the expert should also be required to show that she is proficient at using the tool to analyze data and

41 FED. R. EVID. 702 advisory committee's note to 2000 amendments. The Advisory Committee Notes explain that for such "generalized testimony," the court should assure that the expert is qualified, expert testimony would assist the jury, the testimony is "reliable," and the testimony fits the facts of the case. *Id.* We submit that the qualification of such an expert should be objectively assessed.

42 Though presumably the expert's lack of familiarity with the relevant literature might in some cases be so poor that the judge would treat this as a disqualification rather than a matter going to weight of the testimony.

produce correct answers. Education and experience will enable many people to claim expertise, but true expertise implies proficiency in performance.

A particular advantage of this alternative interpretation of Rule 702's qualification requirement is that it complements Rule 702's reliability requirement. For many non-scientific experts for whom the *Daubert* factors can be difficult to adapt (because there is often no research into the reliability of the methods used by experts in these domains and no journals are devoted to developing reliable knowledge within these domains), and for black-box experts whose methods cannot be directly observed for reliability, a proficiency approach to expert qualification ensures a minimal level of reliability in an expert's opinions. An expert who cannot perform significantly better than chance, or who can barely outperform non-experts on a proficiency test, is an unreliable source of information for a case. With a proficiency-based approach to qualification, courts effectively conduct an individualized reliability analysis for each expert. Just as a poor method should fail reliability scrutiny under Rule 702, an inexpert "expert" should fail qualification scrutiny under Rule 702.

### B. *Identifying Experts Through Performance*

In contrast to the traditional view of judges that having certain education, training, or experience qualifies one as an expert, among those who study experts, expertise is synonymous with performance at consistently high levels. "Experts stand out because of their superior performance and unique capabilities."[43] Much debate exists among psychologists about the origins of expertise,[44] but little debate exists regarding the need to use performance-based measures to identify experts. Alternative measures, such as reputation or years of experience, pose a risk of misidentification: "[T]here is a clear lack of association between length of experience and performance, and between *perceived expertise* and performance."[45] Therefore, where objective measures of good performance are available,

43 David Z. Hambrick & Robert R. Hoffman, *Expertise: A Second Look*, IEEE INTELLIGENT SYSTEMS, July–Aug. 2016, at 50, 54.

44 Broadly, the debate comes down to the relative contributions of deliberate practice versus innate talent. *See generally* David Z. Hambrick, Brooke N. Macnamara, Guillermo Campitelli, Fredrik Ullén & Miriam A. Mosing, *Beyond Born Versus Made: A New Look at Expertise*, 64 PSYCHOL. LEARNING & MOTIVATION 1 (2016). Our argument does not depend on our taking a side in this origins-of-expertise debate.

45 K. Anders Ericsson, *Expertise and Individual Differences: The Search for the Structure and Acquisition of Experts' Superior Performance*, WILEY INTERDISC. REVS. COGN. SCI., Jan.–Apr. 2017, at 1, 2; *cf.* James Shanteau, David J. Weiss, Rickey P. Thomas & Julia Pounds, *How Can You Tell If Someone Is an Expert? Performance-Based Assessment of Expertise*, *in* EMERGING PERSPECTIVES ON JUDGMENT AND DECISION RESEARCH 620, 622-24 (Sandra L. Schneider & James Shanteau, eds., 2003) (discussing the limitations of using experience, accreditation, and peer identification as the basis for identifying experts).

researchers prefer those measures to status symbols that signal specialized education or experience but can be poor proxies for expert performance.

Where clear benchmarks of superior performance exist, tests can be developed to identify high and low performers for purposes of training and quality control.[46] Such tests, which are now commonly called proficiency tests, can be used to monitor not only the performance of human workers but also the performance of machines and materials being used in production. In fact, proficiency testing had its origin in materials testing:

> [P]roficiency testing had its probable beginnings in the paleolithic [sic] age, for it can be logically assumed that Neanderthal man tested his lethal stone axes for strength, weight, and serviceability before using them for the onslaught of his enemies. In more modern times, and especially during the formative period of our country, the value of proficiency testing and quality assurance became essential for the development of our railroads and industrial corporations.[47]

The use of proficiency testing to monitor human performance became prominent after World War II, when a consortium of medical laboratories began circulating specimen samples to laboratories to determine their accuracy in identifying the specimens.[48] The impetus for the survey was the discovery of a surprising number of errors by laboratories in an experiment conducted in 1945 to assess the level of agreement across medical laboratories within Pennsylvania.[49] After viewing the results, the College of American Pathologists "recognized that, to maintain high standards, the accuracy of measurements must be under constant professional surveillance . . . . Accordingly, continuous professional assessment, or proficiency testing, eventually became accepted as the reasonable foundation upon which high standards of laboratory work

---

46 "[Proficiency] tests can serve many purposes including but not limited to training and testing personnel, improving laboratory practices and procedures, and identifying future needs for a laboratory. Properly designed, proficiency tests may also provide a reasonable estimate of the rate at which false discoveries, false positive errors, and false negative errors occur." Koehler, *supra* note 9, at 1091 (footnote omitted). International scientific organizations define proficiency testing as "an evaluation of participant performance against pre-established criteria by means of interlaboratory comparisons." NAT'L COMM'N ON FORENSIC SCI., VIEWS OF THE COMMISSION REGARDING PROFICIENCY TESTING IN FORENSIC SCIENCE 7, (Mar. 22, 2016).

47 F. William Sunderman, Sr., *The History of Proficiency Testing/Quality Control*, 38 CLINICAL CHEMISTRY 1205, 1205 (1992). This program eventually became formalized as the Sunderman Proficiency Test Service, which the American Society of Clinical Pathology took over in 1985 and continues to run. *See id.* at 1207; AM. SOC'Y FOR CLINICAL PATHOLOGY, U.S. (ONLY) PROCEDURES FOR EXAMINATION & CERTIFICATION (2017), https://www.ascp.org/content/docs/default-source/boc-pdfs/exam-content-outlines/ascp-boc-us-procedures-book-web.pdf?sfvrsn=2 [https://perma.cc/ZKE2-LXE2].

48 Sunderman, Sr., *supra* note 47, at 1206.

49 *Id.*

might be maintained."[50] Eventually, the federal Clinical Laboratory Improvement Act (CLIA), passed in 1967, mandated proficiency testing for various medical laboratories.[51]

Recognizing the value of proficiency testing as a means of quality assurance, many industries and organizations have instituted proficiency testing programs without being compelled to do so by law. The particular uses of proficiency testing results vary, but the primary uses include monitoring the performance of laboratories over time and in comparison to one another, and monitoring the performance of individuals as part of their training and as a means of ongoing quality assurance.[52]

In the area of forensic science, all accredited crime laboratories must conduct proficiency testing annually across the different disciplines that they employ, whether it is DNA testing, fingerprint testing, ballistics, toolmark identification, or some other discipline.[53] In 1974, a federal grant to the Forensic Sciences Foundation (FSF) funded administration of twenty-one proficiency tests at crime laboratories around the country.[54] The results were not heartening; FSF uncovered "serious problems" in several disciplines, leading to recommendations for improved quality assurance.[55] In 1979, forensic practitioners voted against a proposal to create a system of peer review through training, certification and proficiency testing for all types of forensics.[56] FSF continued to conduct proficiency tests through the early

---

50 *Id.* at 1206-07.

51 42 U.S.C. § 263a (2012) (applying to all clinical laboratories, defined as "a facility for the biological, microbiological, serological, chemical, immuno-hematological, hematological, biophysical, cytological, pathological, or other examination of materials derived from the human body for the purpose of providing information for the diagnosis, prevention, or treatment of any disease or impairment of, or the assessment of the health of, human beings"). An amendment in 1988 extended the reach of CLIA. *See, e.g.*, Proficiency Testing, Clinical Laboratory Improvement Amendments of 1988 (CLIA), 42 C.F.R. § 493.2 (2014).

52 Edward J. Imwinkelried, *Coming to Grips with Scientific Research in* Daubert*'s "Brave New World": The Courts' Need to Appreciate the Evidentiary Differences Between Validity and Proficiency Studies*, 61 BROOK. L. REV. 1247, 1254-56 (1995) ("[I]n a proficiency study the object of the test is a particular analyst or laboratory."); *see also* Simon A. Cole, *Grandfathering Evidence: Fingerprint Admissibility Rulings from* Jennings *to* Llera Plaza *and Back Again*, 41 AM. CRIM. L. REV. 1189, 1212-13 (2004) (discussing the distinction between proficiency tests and validation studies).

53 *See infra* notes 79-81 and accompanying text.

54 Joseph L. Peterson & Penelope N. Markham, *Crime Laboratory Proficiency Testing Results, 1978–1991, I: Identification and Classification of Physical Evidence*, 40 J. FORENSIC SCI. 994, 994 (1995).

55 *Id.* at 994-95.

56 *See* David D. Dixon, Note, *The Admissibility of Electrophoretic Methods of Genetic Marker Bloodstain Typing Under the* Frye *Standard*, 11 OKLA. CITY U. L. REV. 773, 809-10 (1986) (noting the "proposal was resoundingly rejected by crime lab personnel"). It was, and remains, rare for results of proficiency tests to be published in an academic setting. *Id.* at 813 (explaining that the goal of achieving proficient experts is better accomplished by providing for confidential retraining and retesting of the examiner rather than a public pronouncement that he or she has erred).

1990s, when its operations were folded into the Collaborative Testing Services (CTS), a company that conducts quality assurance work through the regular administration of proficiency tests and other measures. In 1994, federal legislation called for an advisory board to "specify criteria for quality assurance and proficiency tests to be applied to the various types of DNA analysis used by forensic laboratories."[57] That legislation has expired.

Since the 1990s, CTS has conducted proficiency testing across the forensic disciplines. Demand has increased now that accrediting bodies require it, and CTS has become a leading provider of proficiency tests.[58] CTS makes available on its website the results of its proficiency tests, and each year in each discipline the reports show that errors are made.[59] However, CTS has stated that its reports should not be used to infer any "error rate" in any given discipline, because the "design of an error rate study would differ considerably from the design of a proficiency test."[60] In fact, an error rate study might well reveal higher levels of error, for CTS does not claim that its proficiency tests simulate realistic crime-scene samples, and nor can it control test taking conditions, therefore participants know that they are being tested, and may consult with others during the tests.[61]

For objective methods, accuracy and reliability can be assessed by "measuring the accuracy, reproducibility, and consistency of each of its individual steps."[62] Thus, a DNA test that relies on equipment to identify the genetic markers in a sample may not require any subjective judgment, and the quality of the method can be assured by seeing that the equipment uses

---

57 42 U.S.C. § 14131(a)(C)(3) (2012).

58 *See* Michael D. Risinger, *Cases Involving the Reliability of Handwriting Identification Expertise Since the Decision in* Daubert, 43 TULSA L. REV. 477, 484 (2007) ("The ASCLD-LAB standards require proficiency testing as a condition of accreditation, and the CTS tests are an approved provider (possibly the only one for Document Examination) of such tests. This has resulted in a large increase in the number of tests ordered and returned." (footnote omitted)).

59 *See Reports*, COLLABORATIVE TESTING SERVS., INC., http://www.ctsforensics.com/reports/default.aspx?F_CategoryId=19 [https://perma.cc/Q374-EF4B] (providing access to proficiency test summary reports for the forensic disciplines).

60 COLLABORATIVE TESTING SERVS., INC., CTS STATEMENT ON THE USE OF PROFICIENCY TESTING DATA FOR ERROR RATE DETERMINATIONS 3 (Mar. 30, 2010), http://www.ctsforensics.com/assets/news/CTSErrorRateStatement.pdf [https://perma.cc/RM5P-BVJT] ("[T]he results found in CTS' Summary Reports should not be used to determine forensic science discipline error rates.").

61 *See* Cole, *supra* note 25, at 1029 ("First, there are design flaws in the tests themselves. The tests were conducted by mail under unproctored, untimed conditions. It is not known whether the tests were completed by individual examiners or 'by committee.' Second, no metric exists for measuring the degree of difficulty of the latent print comparison." (footnote omitted)); Koehler, *supra* note 9, at 1091 ("Collaborative Testing Services (CTS) offers hundreds of laboratories the opportunity to participate in two fingerprint proficiency tests each year. Test participation is voluntary, examinees know that they are participating in a test, and it is not clear whether examinees work by themselves, in groups, or with assistance from supervisors." (footnotes omitted)).

62 PCAST Report, *supra* note 6, at 5.

scientifically sound design and is functioning properly and that human sources of error from confusing samples or the like are minimized by following standardized operational protocols.[63]

Many forensic methods involve a mix of both objective and subjective analysis. For example, while a simple DNA test may be largely objective, a DNA test in which the results may contain genetic markers from more than one individual may require some considerable "subjective choices" by the analysis.[64] When a technique involves subjective decisions, then part of the method is a "black box," based on the experience and judgment of the person. That subjective decisionmaking can be tested for its reliability and accuracy, but only by using a proficiency test.

How is proficiency measured? In many domains, such as sporting events and chess competitions, it is possible to identify consensual objective measures of good performance.[65] Where it is difficult to evaluate the quality of performance *in situ* because an outcome may be ambiguous or factors beyond the expert's performance may contribute to the outcome, it is possible to develop tests that simulate in-the-field conditions but can provide objective feedback on performance. For example, cytologists can be asked to judge whether cells from biopsies exhibit evidence of cancerous malformation using cells from samples where it has been determined through other means that cancer was or was not present. Or chemists and their drug-testing machinery can be given samples of materials to test for the presence of narcotics using samples of known chemical origins. Likewise, fingerprint and DNA analysts can be presented with samples from known persons to test whether the analysts properly match samples from the same persons. The consistency in a person's judgments can also be measured by including multiple samples from the same sources; it is not uncommon for analysts to provide difference judgments about two samples drawn from the same source.[66]

---

63 *Id.* at 7.

64 *Id.* at 8.

65 Or one may use multiple objective measures of good performance and look for convergence across measures to dispel debates about the best measure. For instance, various measures of patient care could be combined to identify physicians who perform at the highest levels.

66 *See, e.g.*, Daniel Kahneman et al., *Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making*, HARV. BUS. REV. (Oct. 2016), https://hbr.org/2016/10/noise [https://perma.cc/EEP8-UCYE] ("The prevalence of noise has been demonstrated in several studies. Academic researchers have repeatedly confirmed that professionals often contradict their own prior judgments when given the same data on different occasions. For instance, when software developers were asked on two separate days to estimate the completion time for a given task, the hours they projected differed by 71%, on average. When pathologists made two assessments of the severity of biopsy results, the correlation between their ratings was only .61 (out of a perfect 1.0), indicating that they made inconsistent diagnoses quite frequently.").

To be sure, in some domains, it can be more difficult to identify uncontroversial objective measures of expert performance. One might propose rates of reversal as a measure of judicial expertise, for example, but good arguments can be made against that and any other supposed objective measure of judicial expertise.[67] In many fields, consensus exists regarding the best methods for doing something and the best answers to questions, but proficiency testing can still be used when no such consensus exists.

If an expert contends that a method or body of knowledge exists that should be consulted to provide opinions or conclusions relevant to the case, then that method or body of knowledge should be capable of generating a proficiency test. If not, that absence begs the question of how the expert can contend that she has reached a trustworthy conclusion. That is not to say that disagreement among experts renders testimony inadmissible, for there may be disagreement on how to interpret data, the best method to use, or how best to use a method, but expert witnesses do need to employ a method that aims at providing right answers to questions that can have right and wrong answers. Otherwise, the expert seeks to offer nothing more than her own personal opinions or speculation on some matter.[68]

### 1. Results: False Positive, False Negative, and Inconclusive

The results reported in proficiency tests can take multiple forms. In many of the proficiency tests of relevance to expert witness testimony, the question is whether a putative expert can correctly categorize a test item. For instance, on a ballistics proficiency test, a firearms expert might be presented with a firearm and several bullet fragments and be asked to categorize each fragment as (a) having been fired from the gun, (b) not having been fired from the gun, or (c) incapable of categorization. True-positive responses identify those bullet fragments that were in fact fired from the gun, and true-negative responses correctly identify those fragments not fired from the gun. False-positive responses erroneously categorize a fragment as being fired from the gun, and false-negative responses erroneously categorize a fragment as not being fired

67 *See generally* Gregory Mitchell, *Evaluating Judges*, *in* THE PSYCHOLOGY OF JUDICIAL DECISION MAKING 221 (David Klein & Gregory Mitchell, eds., 2010) (discussing the many contested ways of judging judicial performance). In general, where evaluations of outcomes involve opinions over socially or politically contested matters, then it will be difficult to develop uncontroversial proficiency tests. However, few fields of expertise whose members attempt to testify in court offer opinions on matters that cannot be subjected to uncontroversial proficiency testing.

68 There may be some very limited areas (e.g., aesthetic judgments or valuations of rare items) in which an expert's judgment is authoritative simply because it is the expert's judgment, but usually the achievement of expert status alone is not a sufficient basis for an opinion. *See* Gen. Elec. Co. v. Joiner, 522 U.S. 136, 146 (1997) ("[N]othing in either *Daubert* or the Federal Rules of Evidence requires a district court to admit opinion evidence that is connected to existing data only by the *ipse dixit* of the expert.").

from the gun. Whether an incapable-of-categorization response is valid depends on whether the fragment presents adequate data for categorization, as judged by the field's standards. Such proficiency tests can thus yield rates for true positives, true negatives, false positives, false negatives, and inconclusives. High numbers of inconclusive responses indicate that either a method is not very discriminating (i.e., it has difficulty dealing with ambiguous data) or that the test-taker has assumed a conservative stance to avoid committing clear errors. In addition to data on accuracy, error rates and conservatism, proficiency tests can be used to examine adherence to specified procedures and a range of other skills apart from getting right or wrong answers. However, hit rates (i.e., percentages of true positives and true negatives) and error rates (i.e., percentages of false positives and false negatives) most directly address expert proficiency on the task of interest.

Within the forensic domain, false positive rates raise special concerns because such mistakes can divert attention from the true offender and lead to wrongful convictions.[69] False negatives should also be of substantial concern, however, because they can lead to erroneous acquittals and contribute as well to wrongful convictions.

Nor should inconclusive rates be ignored, for a high number of inconclusives for test items that other experts correctly categorize reveals a lower level of proficiency. Moreover, a bias towards calling potential exculpatory evidence merely inconclusive could result in failing to clear innocent persons (e.g., the bullet fragment recovered from a crime scene that was not fired by the suspect's gun may be erroneously labeled inconclusive). Likewise, a bias towards calling potential inculpatory evidence inconclusive could result in substantial evidence of guilt going undetected. As Professor Simon Cole points out, an analyst could get a "perfect" score by calling all test items inconclusive, unless inconclusive answers matter—what is needed

69 *See* JOHN MONAHAN, PREDICTING VIOLENT BEHAVIOR: AN ASSESSMENT OF CLINICAL TECHNIQUES 79 (1981) (comparing the validity of clinical studies purporting to predict violent behavior, including false positive rates); Simon A. Cole, *The Prevalence and Potential Causes of Wrongful Conviction by Fingerprint Evidence*, 37 GOLDEN GATE U. L. REV. 39, 57-60 (2006) (discussing fingerprint misattribution cases, ten of which resulted in wrongful convictions, and noting how "these misattribution cases are important for understanding the potential for wrongful conviction by fingerprint *because misattributions may be expected to cause wrongful convictions*"); Stephen D. Hart et al., *A Note on Portraying the Accuracy of Violence Predictions*, 17 LAW & HUM. BEHAV. 695, 697 (1993) (comparing the rates, and methods for measuring them, of false positives in studies predicting whether an individual will be violent). On variation in the ways false positives are reported, see Cole, *supra* note 25, at 1030 ("There are a number of different ways of reporting false positives. Often the false positive rate has been reported as the number of participants who committed at least one false positive divided by the total number of participants."); William C. Thompson, *Subjective Interpretation, Laboratory Error and the Value of Forensic DNA Evidence: Three Case Studies*, 96 GENETICA 153, 155 (1995) ("The false positive rate of a test is most usefully stated as the ratio of false positives to the sum of true positives and false positives." (citation omitted)).

is a signal detection analysis examining relative rates of discrimination (i.e., the level at which true positives and true negatives can be discerned).[70]

### 2. Nature of the Proficiency Study

Proficiency tests may be conducted with test-takers knowing that they are taking a test or "blind," with the laboratory or individual being tested not knowing that a proficiency test is being conducted.[71] Even if the test resembles everyday casework (as it should to be of any value), when the individuals know they are being tested they may do their work differently.

The lack of blind proficiency testing has been a subject of repeated criticism over the years in a range of disciplines, particularly with respect to the forensic fields that regularly supply expert witnesses in criminal cases.[72] One group of scholars noted: "no 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."[73] Even in the area of DNA testing, data from blind proficiency tests is lacking.[74] Jonathan Koehler points out that, despite recommendations

---

70 Cole, *supra* note 25, at 1031. This approach is commonly used in studies of eyewitness accuracy. *See* NAT'L RESEARCH COUNCIL OF THE NAT'L ACADS., IDENTIFYING THE CULPRIT: ASSESSING EYEWITNESS IDENTIFICATION 83-85 (2014) (providing an overview of the use of Receiver Operating Characteristics, a method from signal detection theory, to examine discriminability and response bias for binary classification decisions, as applied in eyewitness memory research).

71 An example of blind testing is found at the Transportation Safety Administration, which places weapons and explosives inside luggage to determine whether TSA's airport screeners detect these objects. PCAST Report, *supra* note 6, at 58.

72 *See* Adina Schwartz, *A Systematic Challenge to the Reliability and Admissibility of Firearms and Toolmark Identification*, 6 COLUM. SCI. & TECH. L. REV. 1, 27 (2005) (noting CTS proficiency tests "are likely to have understated day-to-day laboratory error rates because the testing was declared, rather than blind"). *But see* Steve Gutowski, *A Response To:* A Systematic Challenge to the Reliability and Admissibility of Firearms and Toolmark Identification, *a Recently Published Article by Adiana* [sic] *Schwartz*, FORENSIC BULL., Winter 2005, at 22, 24 ("[P]roficiency tests probably overestimate the error rate due to the irritation many examiners felt, at least in the past, of unnecessary testing which stopped people from getting on with their 'real work'. . . .").

73 Jennifer L. Mnookin et al., *The Need for a Research Culture in the Forensic Sciences*, 58 UCLA L. REV. 725, 745 (2011).

74 Jonathan J. Koehler et al., *The Random Match Probability (RMP) in DNA Evidence: Irrelevant and Prejudicial?*, 35 JURIMETRICS J. 201, 201 (1995) ("[Random match probabilities] 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 J. 439, 447 (1997) (noting that "[a]lmost none of the proficiency testing done to date or planned for the future is truly blind").

from the National Research Council, "there have been virtually no blind proficiency tests designed to estimate case-relevant DNA match error rates."[75]

In addition to the use of blind testing procedures, the tests must be designed to simulate the real-world circumstances that confront the laboratories and analysts. Just as an easy exam (i.e., a test that all students can pass regardless of their levels of course knowledge and effort) is not a good measure of student expertise, an easy proficiency test can be a misleading measure of expert proficiency. "Proficiency tests have several limitations: analysts know they are being tested (which may cause them to perform differently during proficiency tests than when performing casework); the tests involve relatively few samples; and the tests are typically designed to be relatively easy for a competent analyst to pass."[76] In hearings conducted by the National Commission on Forensic Science, the President of CTS stated "that he has been under commercial pressure to make proficiency tests easier."[77] This disclosure illustrates the concern that self-regulating industries will fail to impose on themselves a realistic and exacting blind proficiency testing system.

When states have required that their crime laboratories be accredited,[78] not much is required in the way of proficiency testing. The longtime leading U.S. forensic accreditation organization, ASCLD/LAB, required only that labs "participate annually in at least one external proficiency test" in each discipline and use approved test providers, if available.[79] But ASCLD/LAB did not require that the accredited labs make public the results of those proficiency tests, nor does the accrediting agency monitor proficiency testing

---

75 Koehler, *supra* note 25, at 1381 (emphasis omitted). The first NRC report on DNA testing noted "there is no substitute for rigorous external proficiency testing via blind trials. Such proficiency testing constitutes scientific confirmation that a laboratory's implementation of a method is valid not only in theory, but also in practice." NAT'L RESEARCH COUNCIL, DNA TECHNOLOGY IN FORENSIC SCIENCE 55 (1992). However, the subsequent 1996 NRC report used less emphatic language: "In open proficiency tests, the analyst knows that a test is being conducted. In blind proficiency tests, the analyst does not know that a test is being conducted. A blind test is therefore more likely to detect such errors as might occur in routine operations. However, the logistics of constructing fully blind proficiency tests [to ensure the laboratory will not suspect that it is being tested] are formidable." NAT'L RESEARCH COUNCIL, THE EVALUATION OF FORENSIC DNA EVIDENCE 24 (1996).

76 NAT'L COMM'N ON FORENSIC SCI., VIEWS OF THE COMMISSION, OPTIMIZING HUMAN PERFORMANCE IN CRIME LABORATORIES THROUGH TESTING AND FEEDBACK 4 (May 27, 2016), https://www.justice.gov/ncfs/file/864776/download [https://perma.cc/G6XV-P72E].

77 *Id.* at 4 n.10.

78 Some states only require that labs conducting DNA testing for the defense be accredited; all DNA labs given access to the CODIS databank must be accredited due to FBI rules for participation in the databank. ERIN MURPHY, INSIDE THE CELL 59 (Nation Books 2015); *see also Combined DNA Index System (CODIS)*, FBI, https://www.fbi.gov/services/laboratory/biometric-analysis/codis [https://perma.cc/2C72-MMUL] (discussing how CODIS enables "forensic laboratories to exchange and compare DNA profiles electronically").

79 ASCLD-LAB, LABORATORY ACCREDITATION BOARD 2005 MANUAL 38, https://www.scribd.com/document/86600881/ASCLD-LAB-Legacy-Manual-2005-Copy [https://perma.cc/H9MT-45UF]. Not all crime labs in the U.S. are accredited.

results. Now ASCLD/LAB has folded in its accreditation services with another organization, the ANSI-ASQ National Accreditation Board (ANAB), which similarly requires that labs conduct external proficiency testing and submit reports from approved test providers.[80]

This state of affairs is slowly beginning to change. While some in the forensics community have long maintained that it is not feasible in many settings to conduct blind proficiency testing, some laboratories have begun to do blind proficiency testing as a matter of routine.[81] Most notably, the Houston Forensic Science Center has begun to use blind proficiency testing for firearms and chemistry analysis, and it plans to extend that proficiency testing to DNA testing and latent print examination.[82] And despite concerns about the feasibility of blind proficiency testing for DNA analysis, experiments have shown that realistic and blind proficiency testing can be done even in this complex setting.[83] With some effort and thought, proficiency testing can be done for any question on which experts routinely opine. Those proficiency tests should be blind, as well as realistic and resembling the task for which the person seeks to testify as an expert in court.

A second question is what level of proficiency should be demanded. Professional organizations and laboratories may demand a high level of proficiency, and then respond to poor test results with additional training and testing.[84] To qualify as an expert in court, judges should insist at the very least that the person perform better than chance and better than a layperson would at the task. For some tasks, false positives may be of special concern, and for some tasks false negatives may be of concern, and for some tasks both may be

---

80 *Forensic Accreditation*, ANSI-ASQ NAT'L ACCREDITATION BD., https://www.anab.org/forensic-accreditation [https://perma.cc/E4YT-63J5]; *see also* ANAB, ACCREDITATION MANUAL FOR FORENSIC SERVICE PROVIDERS 24 (July 19, 2017), https://anab.qualtraxcloud.com/ShowDocument.aspx?ID=7183 [https://perma.cc/QZ7C-7RDL].

81 *See, e.g.*, Jonathan J. Koehler, *Proficiency Tests to Estimate Error Rates in the Forensic Sciences*, 12 LAW, PROB. & RISK 89, 94 (2013) ("Blind proficiency testing has been used in some forensic science areas, including the Department of Defence's forensic urine drug testing programme and the HIV testing programme." (citation omitted)).

82 Hous. Forensic Sci. Ctr., *In a National First, HFSC Begins Blind Testing in DNA, Latent Prints*, FORENSIC MAG. (Nov. 21, 2016), https://www.forensicmag.com/news/2016/11/national-first-hfsc-begins-blind-testing-dna-latent-prints [https://perma.cc/WPB8-WPC5].

83 *See* Joseph L. Peterson et al., *The Feasibility Of External Blind DNA Proficiency Testing. II. Experience With Actual Blind Tests*, 48 J. FORENSIC SCI. 1, 8 (2003) ("We have shown that external blind proficiency testing in forensic DNA laboratories is possible.").

84 For example, the originator of proficiency tests for clinical pathology laboratories has explained the goal as maintaining high standards and identifying ways to improve laboratory performance. *See* F. William Sunderman, *Twenty-five Years of Proficiency Testing for Clinical Laboratories*, 2 ANNALS OF CLINICAL LABORATORY SCIENCE 420, 422 (1972) ("When the results of analyses for the solutions of any given month fall outside the allowable range of values, the directors of the laboratories are encouraged to take an understanding and constructive approach in their efforts to ascertain the causes for the inaccuracies and to bring about correction.").

of great concern. As we will develop in the next Section, beyond meeting minimal levels of proficiency in order to be qualified as an expert, we counsel that, whatever a person's performance on proficiency tests, proficiency information should be discoverable and admissible at trial.

### C. *A Case Study: Fingerprint Examiner Proficiency*

The story of four decades of fingerprint proficiency testing illustrates the pitfalls of relying on commercial providers, largely unregulated, to assure proficiency. In this Section, we report the results of over twenty years of data on fingerprint proficiency testing that we have collected and analyzed. While data from the past few years were readily available online, data from several years ago had to be collected by contacting scholars and practitioners in the community. The results of this study uncovered widely varying error rates, with false positive rates ranging from 1% to 23% for 39 proficiency tests conducted from 1995 to 2016. Overall, the tests had an average 7% false positive rate and 7% false negative rate during the time period. These results suggest that far more should be done to carefully test fingerprint examiner proficiency and to test proficiency more broadly in other fields. After all, fingerprint analysis is a widely used form of expert evidence; it has been used for well over a hundred years, with examiners claiming in the past that their work was "infallible."[85] The White House PCAST report strongly emphasized proficiency testing is "essential" and should be "required."[86] Our findings support that conclusion.

Although examiners long claimed not to make mistakes, little was traditionally done to test whether that was true. From the 1980s through the mid-1990s, latent fingerprint proficiency testing was conducted only sporadically.[87] By the mid-1990s, however, with annual proficiency testing as a condition of lab accreditation becoming increasingly common, CTS began offering annual proficiency tests (and in the last decade, more than one round of testing per year).[88] The test results were closely watched by observers interested in rates of error in fingerprint testing, and labs were concerned with what the results might say about their work.[89] Yet without good information about what makes a latent fingerprint more or less challenging

85 PCAST Report, *supra* note 6, at 9.

86 *Id.* at 10.

87 *See* Cole, *supra* note 25, at 987 ("Latent print examiners have long claimed that fingerprint identification is 'infallible.'" (footnote omitted)); *see also* Koehler, *supra* note 9, at 1077 & n.2.

88 Cole, *supra* note 25, at 1029-33; Joseph L. Peterson & Penelope N. Markham, *Crime Laboratory Proficiency Testing Results, 1978–1991, II: Resolving Questions of Common Origin*, 40 J. FORENSIC SCI. 1009, 1010-12 (1995).

89 *Id.*

to match to a known print, it is not completely obvious to a provider like CTS how to calibrate the difficulty of its tests.

So, when a commercial proficiency test proves too challenging for the participants, the reaction in the industry may range from "shock to disbelief,"[90] paired with "consternation."[91] CTS administered such a test for latent fingerprint identification in 1995, with 22% of the participants making at least one error.[92] Some have suspected that CTS's fingerprint proficiency tests since 1995 have been "less difficult."[93] But the collected data do not clearly support that hypothesis. In fact, in the very next year, high error rates persisted. The report of results for 1996 noted that 38 of the 147 laboratories (19.89%) "correctly identified less than six of the [nine] latent prints," and recommended that those laboratories "review the experience levels of their examiners and provide additional training," as well as consider "[a]dditional internal proficiency testing."[94]

We collected the results of 39 CTS proficiency tests in the area of latent fingerprint comparison conducted from 1995 to 2016 (in some years there were as many as three tests).[95] Participants in these tests knew they were taking a test (i.e., blind proficiency testing was not used). Laboratories and participants received test packages in the mail from CTS, and there were no required time limits on the tests and no required external control to prevent collaboration or sharing of results.[96]

The tests provide a set of prints (ranging from seven to twelve prints) and often involve exemplars from more than one known person. This setup makes the test easier than if one were comparing unknown prints to similar prints,

---

90 David L. Grieve, *Possession of Truth*, 46 J. FORENSIC IDENTIFICATION 521, 524 (1996).

91 Collaborative Testing Servs., Inc., Forensic Testing Program, Latent Prints Examination, Report No. 9508 (1995) (on file with authors); Cole, *supra* note 52, at 1213.

92 Cole, *supra* note 52, at 1213.

93 Koehler, *supra* note 9, at 1093 n.53.

94 Collaberative Testing Servs., Inc., Forensic Testing Program, Latent Prints Examination, Report No. 9608 (1996) (on file with authors).

95 We obtained the proficiency test reports from 2010 through 2015 from CTS itself, and we are extremely grateful to Simon Cole for sharing prior reports from 1995 through 2009, which he was able to obtain only through subpoena. CTS makes its reports generally available online. *See Forensic Summary Reports*, COLLABORATIVE TESTING SERVS., INC., http://www.ctsforensics.com/reports/main.aspx [https://perma.cc/VKN6-MYLZ]. It specifically places its latent fingerprint reports online here: *Reports*, COLLABORATIVE TESTING SERVS., INC., http://www.ctsforensics.com/reports/default.aspx?F_CategoryId=21 [https://perma.cc/JK79-STSJ]. Professor Simon Cole has previously and carefully analyzed the earlier reports from 1995 through 2004 in *Grandfathering Evidence*, *supra* note 52, at 1213.

96 *See* Cole, *supra* note 25, at 1029-32, 1072-73 & tbl. ("The tests were conducted by mail under unproctored, untimed conditions."); Cole, *supra* note 52, at 1213 (examining results from 1995 through 2003); Lyn Haber & Ralph Norman Haber, *Error Rates for Human Fingerprint Examiners*, *in* AUTOMATIC FINGERPRINT RECOGNITION SYSTEMS 339 (Ratha and Bolle eds., 2003). *See generally* Peterson & Markham, *supra* note 54 (examining earlier results).

because even in their gross pattern type (i.e., whether it is a whorl pattern or a left loop or a right loop or an arch pattern) many of the prints from different people can be eliminated.[97] The tests also usually provide quite clear latent prints, of questionable resemblance to the often-distorted latent prints collected from real crime scenes.[98] The crime scenarios differ from year to year as well, as do the number of prints that are to be compared, and the number of unknown individuals to be compared to the sets of known prints from individuals identified for the purpose of the test. Complicating interpretation of the data is the fact that CTS's reporting of results has varied over time: only in more recent years has CTS reported inconclusive results for one or more prints, and for some testing cycles the reporting omits participants (sometimes for stated and sometimes for unstated reasons). There are also questions concerning how a result is treated if a response is left blank.

With these caveats in mind, Table 1 summarizes error rates in terms of the number of participants who made *at least one error* divided by the total number of participants.[99] Thus, these rates focus on how many individuals made an error and not on the total numbers of comparisons across individuals (we also do not measure numbers of errors per examiner). Inconclusive rates are the number of participants labeling at least one print inconclusive, or not of value or suitable for comparison, divided by the total number of participants.[100]

97 Cole, *supra* note 25, at 1030 ("Fourth, the number of 'elimination latents,' or latent prints that should not be attributed to any of the known prints provided, is relatively small. This may mitigate the difficulty of these tests.").

98 Lyn Haber & Ralph Norman Haber, *Scientific Validation of Fingerprint Evidence Under* Daubert, 7 LAW, PROB. & RISK 87, 95 (2008) ("Further, the prints used in proficiency tests do not reflect normal casework. They are predominately or entirely of value in contrast to casework, in which the majority of latent prints are of no value . . . . The results cannot be generalized to the examiner's performance on the job, or accuracy in court, because the difficulty of the test items is unknown, and the other parameters do not correspond to normal casework.").

99 Alternatively, one could examine mistaken comparisons divided by the total number of comparisons. CTS unfortunately does not always clearly report inconclusive results, making it difficult to know the total number of comparisons.

100 All of these rates are somewhat imprecise, partly because CTS itself reports its results in imprecise and sometimes inaccurate ways. *See* Cole, *supra* note 25, at 1072-73 (noting errors in CTS reporting). In some years, for example, CTS reports an initial number of false positives, but then notes that, in addition, some number of participants who made false negatives *also* made false positives. When that occurs, those numbers are added to the number of false positives. In some years (e.g., test numbers 2013-516 and 10-517-518), the numbers CTS provides do not add up correctly (it is not always clear when CTS excludes and includes a participant for some error calculations).

Table 1: CTS Fingerprint Proficiency Test Results, 1995–2016

| Year (Test #) | N of Prints | N of Test Takers | False Positive Rate | False Negative Rate (N) | Inconclusive Rate |
|---|---|---|---|---|---|
| 1995 (9508) | 7 | 156 | 22% (34) | 43% (67) | 3% (6) |
| 1996 (9608) | 11 | 184 | 8% (14) | N/A | N/A |
| 1997 (9708) | 11 | 204 | 10% (21) | 28% (58) | N/A |
| 1998 (9808) | 11 | 219 | 6% (14) | N/A | 35% (77) |
| 1999 (99-516) | 12 | 231 | 7% (16)[101] | N/A | 32% (75) |
| 2000 (00-516) | 10 | 278 | 4% (11) | N/A | N/A |
| 2001 (01-516) | 11 | 296 | 3% (8) | 18% (54) | N/A |
| 2001 (01-517) | 11 | 120 | 20% (24) | N/A | N/A |
| 2002 (02-516) | 11 | 303 | 4% (13) | N/A | 1% (2) |
| 2002 (02-517) | 10 | 146 | 3% (5) | 4% (6) | 1% (2) |
| 2002 (02-518) | 12 | 31 | 0% (0) | 3% (1) | 3% (1) |
| 2003 (03-516) | 10 | 336 | 1% (4) | 8% (26) | N/A |
| 2003 (03-517) | 12 | 188 | 1% (5) | 12% (22) | N/A |
| 2003 (03-518) | 9 | 28 | 7% (2) | 11% (3) | N/A |
| 2004 (04-516) | 12 | 206 | 4% (12) | 3% (7) | N/A |
| 2004 (04-517-518) | 15 | 259 | 6% (15) | 2% (5) | N/A |
| 2005 (05-516) | 16 | 327 | 1% (3) | 9% (28) | At least 1 |
| 2005 (05-517-518) | 16 | 250 | 5% (12) | 2% (6) | N/A |
| 2006 (06-516) | 15 | 333 | 23% (78) | 3% (11) | N/A |
| 2007 (07-516) | 15 | 351 | 4% (14) | 5% (18) | N/A |
| 2007 (07-517-518) | 15 | 315 | 4% (13) | 14% (45) | 6% (20) |
| 2008 (08-516) | 15 | 300 | 1% (3) | 5% (14) | N/A |
| 2008 (08-517-518) | 15 | 391 | 1% (5) | 2% (6) | 1% (2) |
| 2009 (09-516) | 16 | 321 | 11% (35) | N/A | N/A |
| 2009 (09-517 -518) | 16 | 419 | 1% (5) | 2% (8) | 1% (4) |
| 2010 (10-516) | 16 | 331 | 8% (26) | 2% (5) | N/A |
| 2010 (10-517-518) | 16 | 463 | 13% (60) | N/A | N/A |
| 2011 (11-516) | 15 | 335 | 9% (30) | N/A | 1% (3) |
| 2011 (11-517-518) | 16 | 478 | 4% (17) | 0 | .2% (1) |
| 2012 (12-515-516) | 16 | 350 | 2% (6) | 2% (6) | N/A |
| 2012 (12-517-518) | 12 | 555 | 3% (16) | 1% (8) | N/A |

101 This figure is based on Cole's count of the underlying results. *See id.* at 1074-75.

| | | | | | |
|---|---|---|---|---|---|
| 2013 (13-515-516) | 12 | 409[102] | 2% (8) | .2% (1) | 6% (24) |
| 2013 (13-517-518) | 15 | 469 | 3% (12) | 8% (38) | N/A |
| 2014 (14-515-516) | 12 | 424 | 4% (18) | 3% (12) | N/A |
| 2014 (14-517-518) | 12 | 587 | 11% (63) | 9% (53) | .3% (2) |
| 2015 (15-515/516) | 11 | 536 | 7% (39) | 11% (58) | .1% (1) |
| 2015 (15-517) | 11 | 509 | 4% (21) | 11% (57) | .1% (1) |
| 2015 (15-519) | 3 | 292 | 23% (36)[103] | N/A | N/A |
| 2016 (16-515-516) | 16 | 431 | 10% (41) | 3% (11) | N/A |

Note how few inconclusive results occurred in the most recent tests (i.e., where a participant for at least one of the prints reported being unable to reach a determination). Only .1% did so in a 2015 test, and, in recent years, the no identification notations often were due to blank responses as opposed to affirmative claims of inconclusiveness. In some earlier years, when the data were reported, there were high numbers of inconclusive results—and one could imagine far more if more of the prints provided were of the realistic crime-scene type that are often truly unsuitable for any comparison. Reporting on false negatives also varied, with sometimes quite high rates, and sometimes no information at all provided.

Comments from test-takers (which CTS, to its credit, publishes) cast doubt on any belief that these proficiency tests are more difficult than the ordinary crime scene comparison. Comments on the CTS's 1999 test include, for example, "Quite easy!" and "The test was a poor gauge of proficiency. All the comparisons were very easy," though another comment said, "Good, fair realistic test. Very similar to real case work."[104] For a 2015 test, one person commented, "This test was not as good as last year's test. Some of the photographs were not clear and did not appear to be in complete focus or either lost clarity upon mass duplication,"[105] and, in response to a 2014 test, another

102 CTS excluded 199 participants from its results for this test, noting that they were from a "single subscriber" and "the potential for such a large number relative to the total number of participants to skew the overall results," where the participants were said to be "inexperienced," and took the test under "proctored, structured conditions in a classroom setting." Collaborative Testing Servs., Inc., Forensic Testing Program, Latent Prints Examination, Test No. 13-515 and 13-516, at 3, https://www.ctsforensics.com/assets/news/3316_Web.pdf [https://perma.cc/ZB9E-QURL]. It appears from the responses that those 199 participants made far more errors.

103 This test used a different format designed to assess whether participants could determine whether latent prints were present on three pieces of evidence.

104 Collaborative Testing Servs., Inc., Forensic Testing Program, Latent Prints Examination, Test No. 99-516, at 46-47 (on file with authors).

105 Collaborative Testing Servs., Inc., Forensic Testing Program, Latent Print Examination, Test No. 15-515/516, at 25, https://www.ctsforensics.com/assets/news/3516_Web.pdf [https://perma.cc/B8QA-DPC6].

wrote, "This test seemed less difficult than previous tests taken."[106] Each test also has a different crime scenario, resulting in this comment on a test from 2013: "I didn't know the 'Glee' cast were criminals!!!"[107]

CTS itself also comments on the tests, to explain results and sometimes why participants were excluded. In 2012, CTS explained the exclusion of ten test-takers: "Ten participants made a note in their additional comments that latent print 5H appeared to be a footprint. This latent print was created using the palm of an individual for whom no inked prints were provided."[108] And in 2010, a particular latent print was excluded from the computation of test results because 29% of participants failed to identify it correctly.[109]

Our results reinforce the concerns of other scholars who have analyzed a subset of these CTS scores in the past and noted the persisting error rates, uneven reporting, and simplicity of the test designs. As Simon Cole has observed, since the 1995 test and through 2003, false positive rates dropped and ranged from 1% to 6%.[110] Michael Saks and Jonathan Koehler wrote in *Science* that these tests are "obviously imperfect indicators" but they nonetheless show that fingerprint examiners are not error-free.[111] Others, however, have criticized reliance on CTS results as measures of proficiency on grounds that many of the errors in attributing prints to individuals are likely "clerical errors," due to oversight.[112]

---

106 Collaborative Testing Servs., Inc., Forensic Testing Program, Latent Print Examination, Test No. 14-515/516, at 20, https://www.ctsforensics.com/assets/news/3416_Web.pdf [https://perma.cc/3P93-98RY].

107 Collaborative Testing Servs., Inc., Forensic Testing Program, Latent Print Examination, Test No. 13-517/518, at 21, https://www.ctsforensics.com/assets/news/3317_Web.pdf [https://perma.cc/A7CQ-TQXN]. The real import of this note is that it reveals that test-takers obviously know they are taking a test and may not approach the matter with the same seriousness as they do real criminal investigations. Thus, a lack of motivation may contribute to error rates. On the other hand, the lack of stress and pressure to make a comparison should facilitate accurate responding.

108 Collaborative Testing Servs., Inc., Forensic Testing Program, Latent Print Examination, Test No. 12-515 and 516, at 3, https://www.ctsforensics.com/assets/news/3216_Web.pdf [https://perma.cc/N73U-79YQ].

109 CTS explained that "the results for Item 5D were such that CTS did not consider that a consensus result had been obtained and therefore no inconsistencies were assigned to the reported results for this Item." Collaborative Testing Servs., Inc., Forensic Testing Program, Latent Print Examination, Test No. 10-516, at 3 (on file with authors). In other years, CTS reports a higher error rate for a particular item but notes lower error rates for other items. *See, e.g.*, Collaborative Testing Servs., Inc., Forensic Testing Program, Latent Print Examination, Test No. 06-516, at 3 (on file with authors).

110 Cole, *supra* note 52, at 1213.

111 Michael J. Saks & Jonathan J. Koehler, *The Coming Paradigm Shift in Forensic Identification Science*, 309 SCI. 892, 895 (Aug. 5, 2005).

112 Kasey Wertheim, Glenn Langenburg & André Moenssens, *A Report of Latent Print Examiner Accuracy During Comparison Training Exercises*, 56 J. FORENSIC IDENTIFICATION 55, 59 (2006); for discussion and further analysis of the data, see Cole, *supra* note 69, at 81 (noting that "removing 'clerical errors' eliminates around forty percent of the false positives and reduces the false positive error rate from 0.5 percent to 0.3 percent" and asking whether "right-finger-wrong-person" errors should correctly be considered to be "clerical").

On the central question whether these proficiency tests simulate real casework or not, CTS itself, as noted, disavows any comparison to realistic casework.[113] Yet one central problem is that we do not yet have accepted measures used to assess the difficulty of latent fingerprints. And as Simon Cole summarizes, "among other problems," such commercial proficiency tests are "taken under unproctored, untimed conditions (test items were mailed to laboratories and mailed back), and the difficulty of the tests, relative to the usual tasks performed by fingerprint examiners, remains unmeasured and unknown."[114]

Whether one sees proficiency testing simply as a form of training or as a measure of quality control, the value of the proficiency testing depends on its ability to simulate the task of interest. To our knowledge, no one has ever credibly claimed that CTS's tests present fingerprint examiners with a task that is *more difficult* than their real-world task. Nor has anyone shown that only inexperienced, poorly-performing examiners take the CTS tests. Indeed, results may be reported for an entire laboratory.[115] Absent evidence that the tests are invalid because they are too difficult or the participant sample is not representative of the ability distribution found among practicing examiners, it is fair to use the error rates revealed on the CTS tests as evidence of the lower bound of error rates within the discipline of fingerprint examination.

To the extent an individual laboratory or examiner fears being portrayed unfairly by these collective error rates, a simple solution exists: reveal the results of their own tests showing that they perform better. Until such individualized proof is provided, we are left with the CTS data showing that it is not uncommon for many fingerprint examiners in any given year to exhibit both false positives and false negatives in their fingerprint identifications on tests that surely do not present more difficult tasks than many real cases. Given the tremendous volume of work many fingerprint examiners conduct, an error rate above 1% means that *each examiner* in a single year may reach several, if not scores, of false positive and false negative conclusions. Perhaps laboratory verification procedures will catch many of these individual errors (and that is the goal of the ACE-V method[116]), but laboratories themselves are not foolproof.

The CTS data, despite its flaws, provides insight into the proficiency of latent fingerprint examiners, many of whom have likely given testimony in court. The question we now turn to is whether and how this proficiency information should be used by the judge and jury. Despite its apparent

---

113 *See* text accompanying notes 23 and 77 *supra*.

114 Cole, *supra* note 52, at 1213.

115 Cole, *supra* note 25, at 1029 ("The tests were conducted by mail under unproctored, untimed conditions. It is not known whether the tests were completed by individual examiners or 'by committee.'" (footnote omitted)).

116 *See infra* note 123.

relevance on the question of expert qualifications and for the probative value of a fingerprint identification or exclusion, judges are surprisingly reluctant to consider or allow use of proficiency data at trial.

## II. JUDICIAL ATTITUDES TOWARD PROFICIENCY DATA

We lack information about how well a range of scientific disciplines perform when experts do work used in litigation. Worse, what we do know suggests that we have been relying on expert evidence that is far less reliable than how it has been presented in court. This state of affairs is just now receiving substantial attention from policymakers and scholars. Most notably, the 2016 PCAST report described how "proficiency testing, where it had been conducted, showed instances of poor performance by specific examiners," in the area of forensic science.[117] However, the gaps in our knowledge have been apparent for some time. The failure of the judiciary to attend to the problem of proficiency can be seen in a pair of influential 2002 decisions in which Judge Louis Pollak of the Eastern District of Pennsylvania considered whether the epidermal structures that create fingerprints are unique and permanent identifiers of individuals, and whether identifications of individuals based on latent fingerprints recovered from crime scenes are sufficiently reliable to be admissible at trial.[118]

After hearing from experts on the biology of fingerprints and the method used by FBI agents to make fingerprint-based identifications,[119] Judge Pollak initially took judicial notice that the skin's friction ridges create fingerprints that are unique and permanent identifiers of individuals.[120] But Judge Pollak barred the FBI's fingerprint examiners from opining that a defendant's fingerprints did or did not match latent fingerprints found at the crime scene because their method of identification was not sufficiently reliable to pass

---

117 PCAST Report, *supra* note 6, at 4.

118 United States v. Llera-Plaza, 179 F. Supp. 2d 492, 494-95 (E.D. Pa. 2002), *vacated*, 188 F. Supp. 2d 549 (E.D. Pa. 2002) [hereinafter *Llera-Plaza I*]. Amendments to Federal Rule of Evidence 702 prompted by the Supreme Court's *Daubert*, *Frye*, and *Kumho Tire* rulings in the 1990s, which made clear that expert opinions must be based on reliable methods and principles and not simply methods that had become generally accepted in the relevant scientific community, prompted several criminal defendants to challenge the admissibility of fingerprint identifications. *See* FED. R. EVID. 702 advisory committee's note to 2000 amendment. The challenge before Judge Pollak became particularly noteworthy both because of the evidence presented to Judge Pollak and the opinions written by Judge Pollak.

119 The evidence Judge Pollak initially considered actually came from a hearing in another case. *See Llera-Plaza I*, 179 F. Supp. 2d at 494.

120 *See id.* at 502. The utility of fingerprints as a means of identifications depends on the assumption that fingerprints are permanent, unique identifiers of individuals. The more unique the identifying characteristic, the greater the probative value the characteristic will have on questions of identity.

muster under Federal Rule of Evidence 702.[121] In particular, Judge Pollak concluded that, although the FBI's method of latent fingerprint identification had become generally accepted,[122] the method had not been subjected to scientific testing or peer review, the method's error rate had not been established, and the ACE-V method then (and now) used in latent fingerprint identification is not a system of uniform scientific standards.[123]

Judge Pollak's initial decision took the criminal justice community by storm.[124] As another federal judge later put it, the decision "immediately provoked an uproar," and "huge pressure" was placed on Judge Pollak to reverse course.[125]

The federal government asked for leave to supplement the evidentiary record in the case to bolster the reliability of latent fingerprint identifications in hopes of convincing Judge Pollak to change his mind.[126] Judge Pollak agreed to reconsider his opinion, and held a new evidentiary hearing at which the government put on two witnesses who emphasized that the FBI's fingerprint

---

121 *Id.* at 515 ("While fingerprint examinations conducted under the general ACE V rubric are generally accepted as reliable by fingerprint examiners, this by itself cannot sustain the government's burden in making the case for the admissibility of fingerprint testimony under Federal Rule of Evidence 702."). Judge Pollak did plan to allow the government to put on evidence of prints recovered from the crime scene and evidence of the defendant's fingerprints, but he planned to prevent either prosecution or defense experts from opining that the crime scene prints matched or did not match the defendants' prints; he planned to leave the match/no-match evaluation to the jury. *See id.* at 517-18.

122 The FBI's fingerprint examiners, and other professional print examiners, employ what is known as the ACE-V method, which stands for Analysis, Comparison, Evaluation, and Verification. *See, e.g.*, Peter E. Peterson et al., *Latent Prints: A Perspective on the State of the Science*, 11 FORENSIC SCI. COMMS. (Oct. 2009), *https://archives.fbi.gov/archives/about-us/lab/forensic-science-communications/fsc/oct2009/review* [https://perma.cc/A4M6-HFCM] (describing the ACE-V methodology). At the time of Judge Pollak's decision, most examiners expressed their opinions in terms of an identification (i.e., match), elimination (i.e., no match), or insufficient information to identify or eliminate. *See Llera Plaza I*, 179 F. Supp. 2d. at 498-500.

123 *See id.* at 514-16.

124 *See, e.g.*, Michael Higgins, *Fingerprint Evidence Put on Trial*, CHI. TRIB. (Feb. 25, 2002), http://articles.chicagotribune.com/2002-02-25/news/0202250139_1_district-judge-louis-pollak-public-defender-fingerprint [https://perma.cc/8UQD-59RZ] (highlighting that defense attorneys in Chicago and throughout the country regard Judge Pollak's decision as an invitation to challenge laboratory results and the reliability of fingerprint evidence); Andy Newman, *Judge Rules Fingerprints Cannot Be Called a Match*, N.Y. TIMES (Jan. 11, 2002), http://www.nytimes.com/2002/01/11/us/judge-rules-fingerprints-cannot-be-called-a-match.html [https://perma.cc/NXV7-JBJU] (underscoring the importance of Judge Pollak's decision because it was the first ruling that fingerprint evidence does not meet the standards for scientific testimony and likely could lead to challenges in other jurisdictions and to challenges to other forensic techniques, such as ballistics and handwriting analysis).

125 Jed S. Rakoff, Plenary Speech at National Institute of Standards and Technology Conference on Forensics: Are Judges Losing Confidence in Forensic Science? (Dec. 3, 2014).

126 Government's Motion for Reconsideration of the Court's Exclusion of Fingerprint Identification Evidence and for Permission to Present Limited Additional Evidence at ii, *Llera-Plaza I*, 179 F. Supp. 2d 492 (E.D. Pa. 2002) (No. 98-362).

examiners achieve high scores on proficiency tests.[127] According to the government's main witness, the FBI's fingerprint examiners had performed "spectacularly well" on proficiency tests conducted since 1995,[128] making errors on fewer than 1% of the test items.[129] The defendants, in response, contended that this low proficiency error rate was achieved not through the application of a reliable method but rather through the use of easy tests that failed to simulate the noise and distortions found in latent prints lifted from real crime scenes.[130] According to one of the defense's fingerprint experts, an esteemed examiner formerly employed by the Scotland Yard, if he gave his experts tests like those used by the FBI "they'd fall about laughing."[131]

After the hearing, Judge Pollak maintained that the FBI's proficiency tests were "less demanding than they should be,"[132] but he also concluded that "there is no evidence that the error rate of certified FBI fingerprint examiners is unacceptably high."[133] Comparing fingerprint methods in the UK to the FBI's methods led Judge Pollak to determine that there was "sufficient uniformity [in controlling standards] within the principal common law jurisdictions to satisfy *Daubert*."[134] Together, this new information led Judge Pollak to change his mind and conclude that latent fingerprint identifications are sufficiently reliable to permit examiners to opine on whether crime scene prints matched the prints of defendants.[135]

Beyond its practical significance, Judge Pollak's second decision is noteworthy for the emphasis it placed on evidence of proficiency testing to determine the general error rate for a forensic method and for its placing the burden *on opponents* of forensic evidence to demonstrate an unacceptably high error rate. So long as courts take the approach ultimately adopted by Judge Pollak, the forensic evidence professions have little incentive to engage in rigorous proficiency testing, yet defendants have no means of compelling rigorous proficiency testing. Indeed, many courts today deny defense requests for discovery of proficiency testing information pertaining to the forensic experts the government intends to call at trial.[136]

Almost fifteen years after Judge Pollak's landmark final decision on this matter—a decision that still carries weight when new challenges to

127 United States v. Llera-Plaza, 188 F. Supp. 2d 549, 553 (E.D. Pa. 2002) [hereinafter *Llera-Plaza II*].

128 *Id.* at 565.

129 *Id.* at 556.

130 *Id.* at 557.

131 *Id.* at 558.

132 *Id.* at 565.

133 *Id.* at 566.

134 *Id.* at 570.

135 *Id.* at 576.

136 *See infra* subsection B.3.d.

fingerprint identifications arise—we know little more than we did about the proficiency of fingerprint examiners and the various other forensic experts who testify every day in courts throughout the nation.[137] This unfortunate state of affairs becomes more pressing when we focus on a curiosity in Judge Pollak's reasoning. Judge Pollak altered his view that fingerprint examiners do not work under uniform quality control standards when he realized that key jurisdictions had done away with numerical requirements regarding the minimum number of points of similarity needed to declare a match between a defendant's print and a latent print recovered from the crime scene.[138] In other words, the fingerprint examiner community achieved uniformity in standards *by removing* any standards with respect to the minimum number of similarities needed to declare a match.[139] Generally, as the degree of objective control over application of a method goes down, the need for rigorous proficiency testing and evidence of an individual expert's proficiency should go up. Evidence of proficiency is the only assurance we can have that an individual's idiosyncratic applications of a general method tend to be reliable, and this evidence provides only weak protection against intentional and unintentional abuses of a loose-leash method in particular cases.

This Part turns to the ways in which courts assess proficiency when making decisions concerning how to handle scientific evidence. Often proficiency is litigated as one issue among many when challenging or defending scientific evidence, just as in the *Llera Plaza* case. Courts vary widely in their use of proficiency data—with different courts seeing the same data as evidence of both good or bad proficiency—and they use proficiency in different ways at different stages in litigation. The sections below discuss how courts have examined questions concerning proficiency in the context of deciding whether to: (a) consider proficiency information at all, (b) find evidence admissible, (c) consider proficiency information as relevant to weight, and (d) permit discovery on proficiency information. In the final section we summarize our proposed approach towards judicial regulation of proficiency data and expert evidence.

---

137 *See, e.g.*, Lyn Haber & Ralph Norman Haber, *Error Rates for Human Latent Fingerprint Examiners*, *in* AUTOMATIC FINGERPRINT RECOGNITION SYSTEMS 339, 339 (Nalini K. Ratha & Ruud Bolle eds., 2004) ("It is impossible to determine from existing data whether true error rates are miniscule or substantial."); William A. Tobin & William C. Thompson, *Evaluating and Challenging Forensic Identification Evidence*, CHAMPION, July 2006, at 12, 19-20 ("[P]roficiency testing in forensic science is frequently worthless as a true indicator of examiner proficiency."); *see also* Saks & Koehler, *supra* note 111.

138 *Llera Plaza II*, 188 F. Supp. 2d at 575-76.

139 *Id.* at 575 (noting that the English fingerprint identification system, "stripped of any required minimum number" of similarities, "corresponds almost exactly with the ACE-V procedures followed by the FBI").

### A. *Disregarding Proficiency*

Many judicial opinions discussing the qualifications of experts presume proficiency from credentials and experience, as developed in Part I. Thus, courts allow doctors to testify as experts offering opinions on the causes of a condition without any evidence of their accuracy and reliability at differentiating true from false causes. As the Ninth Circuit recently put it: "Medicine partakes of art as well as science, and there is nothing wrong with a doctor relying on extensive clinical experience when making a differential diagnosis."[140] Such cases often emphasize the qualifications of the relevant professionals with no discussion of the expert's track record using differential diagnosis. Likewise, psychiatrists are given wide latitude to testify so long as they have training or experience and profess familiarity with relevant diagnostic criteria and treatment standards.[141] Similarly, courts emphasize the credentials and experience of forensic experts when qualifying them to testify. For example, the First Circuit in *United States v. Vargas* admitted testimony by a fingerprint examiner, noting that the examiner had "considerable" qualifications, having completed "two FBI courses," and other training courses, having worked in the field for twenty years, and having been found qualified "in over one-hundred previous cases."[142] In such rulings, having found the expert qualified, the court goes no further in its analysis.

Given their deference to trial judges' decisions on expert evidence issues, appellate courts rarely discuss the role of proficiency in the admissibility of expert evidence. This silence applies even in the domain of fingerprint expertise, where proficiency data is available, as discussed in the last Part. Although all of the federal circuits but the Second have considered the question whether fingerprint identifications should be admissible, only about half of those courts discuss or mention proficiency as either supportive of or cutting against admissibility.[143] Judicial disregard of proficiency information

---

140 Messick v. Novartis Pharm. Corp., 747 F.3d 1193, 1198 (9th Cir. 2014); *see also* Bitler v. A.O. Smith Corp., 391 F.3d 1114, 1123-24 (10th Cir. 2004) (collecting cases regarding admissibility of differential diagnosis by doctors). For criticism, see, e.g., Joe G. Hollingsworth & Eric G. Lasker, *The Case Against Differential Diagnosis:* Daubert, *Medical Causation Testimony, and the Scientific Method*, 37 J. HEALTH L. 85, 98 (2004) (claiming that doctors' clinical causation opinions based on differential diagnosis are scientifically unreliable and fail to satisfy *Daubert*'s requirements).

141 *See, e.g.*, Skidmore v. Precision Printing & Packaging, Inc., 188 F.3d 606, 617-18 (5th Cir. 1999) (noting that an expert witness must employ the level of "intellectual rigor" in the courtroom that characterizes the practice of an expert in his or her field (quoting Kumho Tire Co. v. Carmichael, 526 U.S. 137, 152 (1999))).

142 471 F.3d 255, 262 (1st Cir. 2006).

143 *See id.* (discussing the background and experience of the expert but not proficiency); *see also* United States v. Straker, 800 F.3d 570, 631 (D.C. Cir. 2015) (discussing the failure of the expert witness "to articulate the rate of human error," but not proficiency); United States v. Watkins, 450 Fed. Appx. 511, 516 (6th Cir. 2011) (noting that the error rate is just one of several factors to be

reflects the view that qualifying an expert, as discussed in Part I, is largely a formality, primarily based on credentials and experience. While *Daubert* and Rule 702 tighten the standards for the reliability of the methods and data the expert uses, the standard for qualification of an expert, as currently interpreted by judges, remains traditional and not empirically informed in many courts.

### B. *Admissibility and Proficiency*

Those courts that have considered proficiency tend to do so not as part of the threshold question whether to qualify an expert, but rather as part of *Daubert* and Rule 702 reliability analysis. One of the factors the Supreme Court mentioned in *Daubert* for assessing reliability was the "known or potential rate of error" associated with the expert's technique.[144] Therefore, the focus in these cases has been on the general error rate associated with a technique, rather than an individual lab or expert's proficiency.[145] Some courts have found proficiency evidence supportive of admissibility, while others have found proficiency evidence (sometimes exactly the same evidence) as cutting against admissibility. Other courts have expressed concerns that proficiency testing itself is inadequate and not informative, while yet other courts have found proficiency data irrelevant to the question of admissibility. We have argued proficiency should be considered at the threshold qualification stage. While proficiency of a particular expert is related to the broader question regarding the reliability of the method used, carefully considering proficiency

---

considered and that the examiner testified regarding "the system of proficiency testing within her lab"); United States v. John, 597 F.3d 263, 275 (5th Cir. 2010) (noting briefly that the evidence "has been routinely subject to peer review" and that "the error rate is low"); United States v. Baines, 573 F.3d 979, 989-92 (10th Cir. 2009) (discussing adequacy of proficiency testing); United States v. Spotted Elk, 548 F.3d 641, 663 (8th Cir. 2008) (discussing expert testimony but not discussing proficiency); United States v. Abreu, 406 F.3d 1304, 1307 (11th Cir. 2005) (discussing expert testimony on fingerprint evidence but not proficiency); United States v. George, 363 F.3d 666, 672-73 (7th Cir. 2004) (finding fingerprint testimony admissible and noting that "the FBI annually tests its fingerprint examiners with sets of prints whose sources are known to the testers, but unknown to the test-takers"); United States v. Janis, 387 F.3d 682, 690 (8th Cir. 2004) (discussing fingerprint expert testimony but not proficiency); United States v. Mitchell, 365 F.3d 215, 242 (3rd Cir. 2004) (raising concerns regarding proficiency); United States v. Crisp, 324 F.3d 261, 268-69 (4th Cir. 2003) (discussing proficiency testing as evidence in support of admissibility); United States v. Sherwood, 98 F.3d 402, 408 (9th Cir. 1996) (discussing the *Daubert* standard for expert testimony but not proficiency). The Seventh Circuit affirmed, in *George*, 363 F.3d at 673, its prior ruling in *Havvard*, which did not consider CTS proficiency test data proffered on appeal because the data were not part of the district court record. United States v. Havvard, 260 F.3d 597, 600-01 (7th Cir. 2001).

144 *See* Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 594 (1993); *see also* FED. R. EVID. 702 advisory committee's note to 2000 amendments.

145 Ronald G. Nichols, *Defending the Scientific Foundations of the Firearms and Tool Mark Identification Discipline: Responding to Recent Challenges*, 52 J. FORENSIC SCI. 586, 592 (2007) ("[P]roficiency tests can offer to the court a reliable practical indicator of how often the profession, using accepted procedures, practices, and controls, makes a false identification.").

at the threshold qualification stage can avoid the more complex inquiry into the *Daubert* and Rule 702 factors. As the sections that follow describe, courts have been quite confused in rulings discussing proficiency.

### 1. Use of Proficiency Data to Exclude Evidence Entirely

Very few courts have concluded that proficiency testing uncovers such troubling error-rates that a proffered expert should not be allowed to testify. But one area in which courts agree that the risk of error is too great is that of polygraph-based opinions. Although some states permit lie detector test results admissible if the parties stipulate to their admissibility, most states and the federal courts exclude opinions based on the results of a polygraph test. For instance, in a Michigan Court of Appeals ruling in 1974, the judges continued the ban on polygraph evidence, noting the expert in the case testified how "even at the highest level of proficiency, polygraph examinations have a ten percent error factor."[146] In more recent years, courts have taken note of proficiency concerns about handwriting analysis or document examination after studies alarmingly suggested that experts were no better than laypeople at detecting forged handwriting.[147] One federal judge who excluded testimony from a forensic document examiner in 2001 stated that, "[i]n addition to there being a lack of empirical evidence on the proficiency of document examiners, there has been little empirical testing done on the basic theories upon which the field is based."[148] The judge went on to note that, "[a]s to some tasks, there is a high rate of error and forensic document examiners may not be any better at analyzing handwriting than laypersons."[149] In reference to proficiency tests administered by CTS in which the task is "to compare written letters in the natural hand of the writers with known exemplars of several suspects," the judge noted that a "test involving hand printing produced only 13% correct answers."[150] Still more remarkable, "[i]n a test asking examiners to identify the author of a forgery, the examiners were wrong 100% of the time."[151] In another study, "laypersons were given the same material as experts were given in the 1987 proficiency study. The true positive accuracy rate of laypersons was the same as that of handwriting examiners; both groups were correct 52% of the time."[152] A judge considering whether to qualify another handwriting analyst

146 People v. Levelston, 221 N.W.2d 235, 236 (Mich. Ct. App. 1974).

147 *See generally* Risinger, *supra* note 58.

148 United States v. Saelee, 162 F. Supp. 2d 1097, 1102 (D. Alaska 2001).

149 *Id.* at 1103.

150 *Id.*

151 *Id.*

152 *Id.* (footnote omitted). For an analysis of five sets of proficiency tests conducted by the Forensic Science Foundation, see DAVID L. FAIGMAN ET AL., 4 MOD. SCI. EVID. § 33:21 (2016–2017 ed.), which notes "If we assume that inconclusive examinations do not wind up as testimony in court . . . and remain

noted that, in addition to lacking credentials such as having authored authoritative texts or completing training for certification, there is "no evidence that he is routinely subjected to proficiency tests or that his work is regularly reviewed by at least one qualified document examiner."[153]

In a child pornography case in which an expert for the government sought to opine on whether images in the possession of the defendant represented real or virtual children, the judge asked whether the expert's proficiency at this authentication task had been tested by supervisors: "[The analyst] responded, 'I've never been tested.'"[154] The judge concluded that, "[a]bsent this type of proficiency testing, neither this Court nor the jury can assess the reliability" of the specific person's work, a particular concern "where, as here, the field as a whole has no known error rates, making it impossible to guess how often [the analyst] is likely to be right or wrong."[155]

Outside of polygraph examiners, document examiners, and the occasional fingerprint examiner,[156] it is rare for judges to examine proficiency data, much less exclude an expert for a lack of demonstrated proficiency. Indeed, state courts have interpreted Judge Pollack's opinion in the *Llera-Plaza* case as signaling good proficiency on the part of fingerprint examiners.[157] Thus, not only do most judges disregard proficiency at the threshold when deciding

---

as generous as possible within the bounds of reason, then the most we can conclude is this: Document examiners were correct 57% of the time and incorrect 43% of the time.". *But see* Oliver Galbraith, Craig S. Galbraith & Nanette Galbraith, *The "Principle of the Drunkard's Search" as a Proxy for Scientific Analysis: The Misuse of Handwriting Test Data in a Law Journal Article*, 1 INT'L J. FORENSIC DOCUMENT EXAMINATION 7 (1995); *see also* Moshe Kam et al., *Signature Authentication by Forensic Document Examiners*, 46 J. FORENSIC. SCI. 884, 887 (2001) (finding that expert document examiners made fewer errors than laypeople on tests); Moshe Kam & Erwei Lin, *Writer Identification Using Hand-Printed and Non-Hand-Printed Questioned Documents*, 48 J. FORENSIC SCI. 1391 (2003) (same); Jodi Sita et al., *Forensic Handwriting Examiners' Expertise for Signature Comparison*, 47 J. FORENSIC SCI. 1, 4 (2002) (finding that examiners detected forgeries in 55% of the test cases while laypeople did so in 57% of the test cases).

153 Dracz v. Am. Gen. Life Ins. Co., 426 F. Supp. 2d 1373, 1378-79 (M.D. Ga. 2006).

154 United States v. Frabizio, 445 F. Supp. 2d 152, 165 (D. Mass. 2006).

155 *Id.*

156 In the case of *State v. Rose*, the judge excluded a fingerprint examiner from the FBI after he claimed "no error rate" for fingerprinting and "100 percent certainty." State v. Rose, No. K06-0545, at 24-25 (Cir. Ct. Baltimore Co. Oct. 19, 2007) ("Mr. Meagher 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 [the fingerprint technique] ACE-V." (footnote omitted)). That Maryland ruling was then vacated when state prosecutors dropped the charges and federal prosecutors refiled the case in federal court—and the federal judge found the fingerprint evidence admissible. United States v. Rose, 672 F. Supp. 2d. 723, 726 (D. Md. 2009).

157 *See, e.g.*, State v. Escobido-Ortiz, 126 P.3d 402, 411 (Haw. Ct. App. 2005) (noting that fingerprint "[a]nalysts are also consistently subjected to testing and proficiency requirements") (citing *Llera Plaza II*, 188 F. Supp. 2d 549, 566-71 (E.D. Pa. 2002) and United States v. Havvard, 260 F.3d 597, 599 (11th Cir. 2011)); Barber v. State, 952 So.2d 393, 420-22 (Ala. Crim. App. 2005) (citing *Llera Plaza II* repeatedly to show that fingerprint analysis has "strong general acceptance, not only in the expert community, but in the courts as well").

whether to qualify an expert, but they do not engage with the proficiency data at the admissibility or reliability stage of the analysis either.

### 2. Concerns with General Inadequacy of Proficiency Testing

Some courts, while not ultimately excluding the expert evidence, have found that proficiency testing is relevant and that it raises important—though not dispositive—concerns regarding admissibility. Such rulings at least make proficiency data salient, while not considering it as a question of expert qualification as we have recommended. In *United States v. Mitchell,* the Third Circuit highlighted the lack of *any* proficiency testing on the part of the individual examiners, warning that "prosecutors would be well-advised to elicit testimony about their experts' personal proficiency, rather than relying on the discipline's good general reputation among lay jurors"—but ultimately deeming the evidence admissible.[158] In the area of DNA testing, several courts have discussed the recommendation of the initial National Research Council report on DNA evidence, in 1992, that recommended blind proficiency testing.[159] One federal judge explained that the lack of blind testing was "troubling" but nevertheless admitted the evidence.[160] Following suit, another federal judge added: "Absent evidence demonstrating that the particular quality control procedures followed by the FBI laboratory violated a statute, regulation or a generally accepted industry requirement, these issues impact the weight of the evidence rather than its admissibility."[161]

Other judges have raised concerns that proficiency tests are available, but appear not to be informative. Thus, Judge Pollack in the second *Llera-Plaza* decision did express concern that the FBI's proficiency tests were "less demanding than they should be" and therefore "can be of little assistance in providing the test makers with a discriminating measure of the relative competence of the test takers."[162] Similarly, when ruling on handwriting evidence, a federal judge noted that proficiency testing was "not meaningful" when "all of [the expert's] peers *always* passed."[163] Another judge reasoned, in response to a challenge to handwriting testimony based on error rates from CTS proficiency tests, that "this data lacks a control group and other hallmarks

158 365 F.3d 215, 242 (3d Cir. 2004).

159 *See supra* note 75 and accompanying text.

160 United States v. Bonds, 12 F.3d 540, 560 (6th Cir. 1993).

161 United States v. Lowe, 954 F. Supp. 401, 420 (D. Mass. 1996); *see also* Commonwealth v. Teixeira, 662 N.E.2d 726, 729 (Mass. App. Ct. 1996), *review denied,* 664 N.E.2d 1197 (Mass. 1996) ("Weaknesses in the laboratory's proficiency testing went to the weight to be ascribed to the evidence of match, not to its admissibility.").

162 *Llera Plaza II,* 188 F. Supp. 2d 549, 565 (E.D. Pa. 2002)

163 United States v. Lewis, 220 F. Supp. 2d 548, 554 (S.D. W. Va. 2002).

of scientific rigor."[164] In considering a challenge to a firearms expert, Judge Nancy Gertner wrote that, "[b]ecause of the subjective nature of the matching analysis, a firearms examiner must be qualified through training, experience, and/or proficiency testing to provide expert testimony."[165] In response to evidence that none of 255 test-takers gave an incorrect response on a CTS proficiency test for bullet cartridge case examination, the judge noted: "One could read these results to mean that the technique is foolproof, but the results might instead indicate that the test was somewhat elementary."[166]

### 3. Judicial Acceptance of Proficiency

Occasionally courts cite proficiency data from a field of expertise or from an individual expert to *support* the admissibility of expert evidence, particularly with respect to latent fingerprint examiners. Such rulings raise the concern that judges do not inquire carefully into whether the proficiency testing is realistic or into the proficiency test results of the examiner seeking to testify.

#### a. *General Proficiency*

Courts have cited proficiency testing within a field to support the admissibility of drug testing,[167] DNA evidence,[168] boot-print matches,[169] firearms examination,[170] and document examinations.[171] To judges, field-wide data can apparently make up for the lack of individualized data: an expert on footwear impressions was permitted to testify despite a lack of evidence on this expert's "proficiency in obtaining impression evidence, how frequently he utilizes his training in the area of impression evidence, how often he is proficiency tested, etc." because the government "established that

164 United States v. Hidalgo, 229 F. Supp. 2d 961, 964 n.7 (D. Ariz. 2002).

165 United States v. Monteiro, 407 F. Supp. 2d 351, 355 (D. Mass. 2006).

166 *Id.* at 367.

167 *See, e.g.*, United States v. Diaz, No. CR 05-0167, 2006 WL 3512032, at *10-11 (N.D. Cal. Dec. 6, 2006) (reasoning that a crime lab's identification of cocaine was admissible because the results of blind proficiency tests in the lab showed that the error rate was "exceedingly low").

168 *See, e.g.*, United States v. Peters, No. CR 91-395-SC, 1995 U.S. Dist. LEXIS 20950, at *58 (D.N.M. Sept. 7, 1995) (noting that external proficiency testing would be "the ideal" but finding internal FBI proficiency testing supportive of admissibility); *see also* Morris v. Presley, No. 1:05cv-330-LG-RHW, 2008 U.S. Dist. LEXIS 84767, at *6 (S.D. Miss. Aug. 8, 2008) (finding Reliagene (a company that conducts DNA testing) proficiency testing supportive of admissibility).

169 *See, e.g.*, United States v. Turner, 287 Fed. App'x. 426, 434 (6th Cir. 2008) (reasoning that yearly proficiency testing by an independent agency supported admissibility of shoeprint analysis).

170 *See, e.g.*, United States v. Otero, 849 F. Supp. 2d 425, 433-34 (D.N.J. 2012) (finding evidence by firearms examiners to be admissible where proficiency testing indicated a low error rate).

171 *See, e.g.*, United States v. Mooney, 315 F.3d 54, 62 (1st Cir. 2002) (holding that testimony by a handwriting expert was admissible where the expert testified that he submitted to proficiency tests regularly); United States v. Jones, 107 F.3d 1147, 1160-61 (6th Cir. 1997) (holding that expert handwriting analysis was admissible where the expert had been subject to training and testing).

examiners of footwear and other impression evidence are routinely tested to ensure proficiency in the field."[172]

Spurred by Judge Pollak's decision in the *Llera-Plaza* case, federal courts have most commonly discussed proficiency data with respect to the field of fingerprint examinations.[173] The various Courts of Appeals to have considered the question of fingerprint admissibility—the First, Third, Fourth, Fifth, Sixth, Seventh, Eighth, Ninth, Tenth, Eleventh and D.C. Circuits—have all found fingerprint examinations to be sufficiently reliable to be admissible under *Daubert*.[174] Few of these courts have engaged in close analysis of the proficiency data, however.

The Tenth Circuit, in its decision in *United States v. Baines*, did examine the proficiency question more closely, stating that FBI fingerprint analysts "have undergone demanding training culminating in proficiency examinations, followed by further proficiency examinations at regular intervals during their careers."[175] The judges concluded that, "[a]lthough these proficiency examinations have been criticized on several grounds, most notably that they do not accurately represent conditions encountered in the field, we see no basis in this record for totally disregarding these proficiency tests."[176] Subsequent courts have endorsed the Tenth Circuit's conclusion.[177]

In *United States v. Mitchell,* the defendant presented at a *Daubert* hearing proficiency tests showing that fingerprint examiners make both false negatives and false positives.[178] The Third Circuit Court of Appeals panel stated that this proficiency data "is troubling, but we view it as evidence

---

172 United States v. Allen, 207 F. Supp. 2d 856, 864, 866 (N.D. Ind. 2002). The judge reasoned that because "the processes for obtaining footwear impression evidence and fingerprint identification evidence are similar," the court could rely on the holdings of *Daubert* and *Kumho Tire* to conclude that the process is sufficiently reliable. *Id.* at 867.

173 Of eight cases in the federal Courts of Appeals located through a Westlaw search that included the term "proficiency" in the same paragraph as "Daubert," one case discusses boot print analysis, *see Turner*, 287 Fed. Appx. 426, *supra* note 170, one discusses handwriting analysis, *see Jones*, 107 F.3d 1147, *supra* note 172, and the remaining six discuss fingerprints.

174 *See* United States v. Straker, 800 F.3d 570, 630-32 (D.C. Cir. 2015); United States v. Watkins, 450 Fed. Appx. 511, 515-16 (6th Cir. 2011); United States v. John, 597 F.3d 263, 274-75 (5th Cir. 2010); United States v. Baines, 573 F.3d 979, 989-92 (10th Cir. 2009); United States v. Spotted Elk, 548 F.3d 641, 663 (8th Cir. 2008); United States v. Vargas, 471 F.3d 255, 265-66 (1st Cir. 2006); United States v. Abreu, 406 F.3d 1304, 1307 (11th Cir. 2005); United States v. Mitchell, 365 F.3d 215, 244-46 (3rd Cir. 2004); United States v. George, 363 F.3d 666, 672-73 (7th Cir. 2004); United States v. Crisp, 324 F.3d 261, 268-69 (4th Cir. 2003); United States v. Sherwood, 98 F.3d 402, 408 (9th Cir. 1996) .

175 573 F.3d at 990.

176 *Id.*

177 *E.g.*, United States v. Love, No. 10cr2418-MMM, 2011 WL 2173644, at *3 (S.D. Cal. 2011) ("The FBI also conducts proficiency examinations of its examiners, which—even if taken under conditions that 'do not accurately represent [those] encountered in the field'—are of some value in assessing the reliability of individual examiners." (quoting *Baines*, 573 F.3d at 990)).

178 *Mitchell*, 365 F.3d at 239-40.

relating only to the competency of those practitioners, leaving undisturbed the government's evidence about the near-absence of false positive identifications" for the field as a whole.[179]

b. *Individual and Lab Proficiency*

When individualized proficiency data of the kind the Third Circuit sought exists, courts are mixed in their use of this data. Some courts highlight this data in support of admissibility, for example citing how a hospital passed proficiency tests in its analysis of blood to assess alcohol concentration of drivers,[180] or how a DNA laboratory regularly gave proficiency tests to its analysts and compared their work to that in other laboratories.[181] One court noted that "[s]ince its inception, Reliagene has undergone proficiency testing by an outside agency approved by the American Society of Crime Lab Directors—the lab, every analyst, every technician undergoes such testing twice a year, and ReliaGene's results have always proved correct."[182] A First Circuit ruling highlighted how an examiner, conducting footwear comparisons, is "subject to annual proficiency testing by an outside agency."[183]

Other courts disregard individualized proficiency data. One court held, for example, that a lack of proficiency testing was not an obstacle to admissibility of DNA tests, even though the laboratory had lost its accreditation at the time it conducted the relevant DNA testing.[184] Another court treated individualized proficiency information as irrelevant to the question whether the expert evidence was reliable:

> 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. Since Rule 702's reliability requirement focuses on the validity of the test rather than the proficiency of the tester, the absence of a laboratory error rate will rarely be dispositive if the rest of the evidence establishes that the test has been properly validated.[185]

Endorsing the same view, another court rejected a defendant's attempt to exclude evidence by pointing to errors found in proficiency testing and characterized the defendant as trying to "challenge the proficiency of the tester," when Rule 702 and *Daubert* are concerned only with "the reliability of

---

179 *Id.* at 240.

180 Barna v. Comm'r of Pub. Safety, 508 N.W.2d 220, 222 (Minn. Ct. App. 1993).

181 Johnson v. Runnels, No. C 02-5537 CW (PR), 2006 WL 823060, at *6 (N.D. Cal. Mar. 29, 2006).

182 Morris v. Presley, No. 1:05cv330-LG-RHW, 2008 WL 4186932, at *3 (S.D. Miss. Aug. 8, 2008).

183 United States v. Mahone, 453 F.3d 68, 70 (1st Cir. 2006).

184 J.H.H. v. State, 897 So.2d 419, 425-26 (Ala. Crim. App. 2004).

185 United States v. Shea, 957 F. Supp. 331, 340 (D.N.H. 1997). This opinion has been influential. *See* United States v. Ewell, 252 F. Supp. 2d 104, 114 (D.N.J. 2003) (quoting *Shea*, 957 F. Supp. at 340).

the test."[186] In the view of these courts, individual- or lab-level proficiency data may be relevant to weight but not admissibility.

c. *Weight and Proficiency*

Factfinders should have a full opportunity to assess the performance of an expert and not just the expert's credentials and experience. Proficiency data can give the factfinder a superior and empirically-informed picture of that expert's performance. While proficiency information should be considered in the qualification and *Daubert* reviews, if an expert is permitted to take the stand, proficiency data should be admitted because it is relevant to the weight that should be given to the expert's opinions. Judges have largely correctly treated proficiency data as relevant to weight. Unfortunately, a number of courts treat proficiency data as going only to the weight of the evidence rather than admissibility. For example, in a case in which a DNA analyst had failed a proficiency test, the court stated that this "was relevant to the weight that her opinion should carry generally," but not to the "general issue of the admissibility of DNA comparisons."[187] The court went on to explain that "failure had no more relevance to the admissibility of the PCR technique any more than a particular fingerprint expert's poor eyesight has on the general admissibility of fingerprint comparison evidence."[188] This failure "was relevant to impeach the credibility of [the expert] personally, not DNA evidence generally."[189] In that case, the prosecution only disclosed the proficiency results during trial; the defense objected to the late disclosure, but the appellate court ultimately concluded that the judge's instructions to the jury cured the failure to disclose.[190] To avoid problems of the kind encountered by the defense in this case, discovery of proficiency data should be permitted, as we develop more in the next Section.

Courts have also ruled that the failure to properly conduct regular proficiency testing goes to the weight of the evidence. This issue has most frequently been litigated in cases involving DNA testing, and a large number

186 United States v. Wrensford, Criminal Action No. 2013-0003, 2014 WL 1224657, at *11 (D.V.I. Mar. 25, 2014) (quoting *Ewell*, 252 F. Supp. at 114).

187 People v. Tillet, 108 Cal. Reptr. 2d 76, 91 (Cal. Ct. App. 2001) (emphasis omitted). The court stated: "The issue of [the expert's] failing a proficiency test was relevant to the weight that her opinion should carry generally; *it had no relevance to the general issue of the admissibility of DNA comparisons by the PCR method,* which is the purpose of a *Kelly* hearing." *Id.*; *see also* United States v. Yagman, No. CR 06-227(A)-SVW, 2007 WL 4409618, at *6 (C.D. Cal. May 22, 2007) (ruling in the context of studies of proficiency in handwriting, but not proficiency data concerning the examiner testifying in the case).

188 *Tillet*, 108 Cal. Reptr. 2d at 91.

189 *Id.* at 92.

190 *Id.* at 93.

of courts have ruled that it is an issue of weight, often in cases in which DNA labs did not routinely proficiency test their analysts.[191]

While some courts do allow questions about proficiency on cross-examination of forensic experts, courts have often found that the denial of proficiency data or of questioning about proficiency are harmless errors because the topics only affect the credibility of a witness. For example, a Virginia court noted proficiency information "would not have affected the admissibility of the DNA evidence, but rather, would have only affected the weight the fact finder accorded the DNA evidence."[192] In that case, however, the analyst had passed prior proficiency tests.[193] Had the proficiency information been negative, we are skeptical the proficiency of an expert would be discounted by jurors. Such data might have a strong impact on outcomes; more work should be done to assess the question of how laypeople evaluate proficiency data.[194]

Courts should be reluctant to qualify experts for whom no proficiency data is available. If that person is proficient and qualified, however, proficiency information should go to weight, meaning that proficiency data should be admissible and presented to the jury. While there may be difficult questions regarding what level of proficiency should be demanded in particular fields, one solution to that challenge is to simply let the jury determine how an expert's proficiency level, particularly as compared to that of any competing experts, affects the credibility of the expert's opinions.

d. *Discovery of Proficiency Data*

Before a party may inquire into the proficiency of an expert, wise counsel will first possess proficiency data. Without advance discovery and access to the information needed to impeach untruthful answers, it could be perilous for a lawyer to inquire into an expert's proficiency. As the Advisory Committee's Note to Federal Rule of Criminal Procedure 16 emphasizes, "it is difficult to test

191 *See* United States v. Beasley, 102 F.3d 1440, 1448 (8th Cir. 1996) (ruling that the alleged lack of "frequent external proficiency testing" and "double blind external tests to check results and to show that proper procedures are being followed" will "go to the weight of the DNA evidence"); State v. Tankersley, 956 P.2d 486, 492-93 (Ariz. 1998) (discussing the question of a lab's lack of "current proficiency testing" where the lab had not participated in proficiency tests for two years as a matter that goes to weight), *abrogated by* State v. Machado, 246 P.3d 632 (Ariz. 2011) on other grounds; Commonwealth v. Teixeira, 662 N.E.2d 726, 729 (Mass. App. Ct. 1996) ("Weaknesses in the laboratory's proficiency testing went to the weight to be ascribed to the evidence of match, not to its admissibility."); State v. Cauthron, 846 P.2d 502, 512 (Wash. 1993) (noting that cross-examination addressed errors made in a state proficiency study).

192 Keen v. Commonwealth, 485 S.E.2d 659, 663 (Va. Ct. App. 1997).

193 *Id.* at 662.

194 A work in progress by the Authors presents the results of an experiment asking laypersons to evaluate proficiency data in the context of a hypothetical criminal case.

expert testimony at trial without advance notice and preparation."[195] However, Rule 16 does not provide for disclosure of credentials or proficiency data of experts, but rather just the reports and summaries of conclusions the experts reached.[196] Under the Federal Civil Rules, parties must disclose an expert witness's qualifications, including publications, compensation to be paid, and lists of cases in which the witness testified, but not necessarily evidence of proficiency.[197] Nonetheless, judges have discretion to order broader expert discovery under both criminal and civil procedure rules.

The rarity with which discovery rulings are the subject of appeal or published opinions makes the survey of discovery practices an imperfect enterprise. But one still finds cases in which the courts conclude that proficiency data is not material to a criminal defendant's defense or of sufficient relevance to compel discovery in civil cases.[198] In criminal cases, discovery about forensics more generally is quite limited. Defense requests for proficiency data may be part of omnibus requests for information about the underlying bench notes and reports underlying the forensic conclusions, laboratory protocols and procedures, and information about the expert's training and experience.[199] Courts that are reluctant to grant broad discovery regarding other aspects of the expert's work and the laboratory's practices are likely to be similarly reluctant to grant discovery of materials regarding proficiency. Recently, the National Commission on Forensic Science encouraged greater pre-trial discovery with regard to forensic evidence, but it may be some time before these recommendations gain traction.[200]

Other courts have ruled that proficiency evidence is discoverable and should be shared by the parties. The Maryland Court of Appeals, for example, highlighted how the analyst's "record in proficiency tests," was "relevant to the

---

195 FED. R. CRIM. P. 16 advisory committee's note to 1974 amendment.

196 FED. R. CRIM. P. 16(A)(1)(D)–(G).

197 FED. R. CIV. P. 26(2)(B).

198 *See e.g.*, Samatar v. Clarridge, No. 2:04-CV-1108, 2006 WL 355684, at *1 (S.D. Ohio Feb. 16, 2006) (explaining that petitioner's request for discovery of proficiency tests was denied).

199 *See, e.g.*, Notice of Motion and Motion to Exclude Fingerprint Identification Testimony and Request for *Daubert* Hearing, And Motion for Discovery, United States v. Ablett, No. CR 09 0749 RS, 2012 WL 8499482, at *2 (N.D. Cal Jan. 17, 2012) (holding that the Court could not effectively perform its evidentiary gatekeeping function under *Daubert* and Rules 104(a) and 702 because the government "fail[ed] to provide the defendant with the actual points of identification being relied upon by the expert, her laboratory protocol for performing the evaluation, her proficiency tests, and other materials necessary to the fair presentation of this motion.").

200 NAT'L COMM'N ON FORENSIC SCI., RECOMMENDATIONS TO THE ATTORNEY GENERAL REGARDING PRETRIAL DISCOVERY (Jan. 16, 2016), http://www.ascld.org/wp-content/uploads/2016/03/Initial-Draft-Recommendations-on-Pretrial-Discovery.pdf [https://perma.cc/RF5N-92FU]. We note that the recommendations direct federal prosecutors to provide pre-trial discovery on a number of subjects, including the witness's qualifications—but this document did not specifically highlight proficiency as a relevant aspect of the witness's qualifications.

weight the fact-finder might give the test results."[201] Courts have specifically held that discovery on the issue of proficiency is important with respect to black-box experts: "[T]he subjectivity of firearms toolmark identification methodology places a great degree of emphasis on the individual's training and proficiency . . ."[202] The Court of Appeals of South Carolina held that a trial court abused its discretion by not ordering the prosecutor to turn over proficiency tests of DNA experts, including as potential impeachment material—but the state Supreme Court reversed the lower appellate court, finding "that the nondisclosure of proficiency test results was not material."[203] The court reasoned "the test results could only have reduced the probabilities. Even if the 'lab error rate' resulted in a 90% reduction . . . those numbers would still be staggering: 1 in 450 million Caucasians and 1 in 37 million African–Americans."[204] The court added that the analyst "acknowledged during his testimony that errors are made in every lab, and that those errors affect the validity of the probability determination."[205]

A party may separately seek discovery of an entire laboratory's proficiency. Courts appear to be even more reluctant to order discovery of this broader data. For example, a Virginia appellate court ruled that a lower court could refuse further discovery of proficiency information after the government provided a memo stating the results of three proficiency tests by the analyst in question but not others sought by the defendant.[206]

We view such rulings denying discovery on questions of proficiency as misguided: because many individuals within a laboratory other than the testifying expert may participate in the testing connected to a particular case, lab-wide proficiency information would shed light on how expert the laboratory as a whole is at the task that generated evidence against the defendant. Generally, courts should permit discovery on proficiency and allow the factfinder to consider individual-level proficiency as well, just as the factfinder may consider an expert's credentials and experience. Proficiency can allow the factfinder to far better assess what weight to place on an expert's testimony. Jurors should have information about the actual performance of an expert and not just their professional pedigree.

---

201 Cole v. State, 835 A.2d 600, 610 (Md. 2003).

202 United States v. Willock, 696 F. Supp. 2d 536, 578 (D. Md. 2010).

203 State v. Proctor, 358 S.C. 417, 480 (S.C. 2004) (discussing State v. Proctor, 347 S.C. 587, 603 (S.C. Ct. App. 2001)).

204 *Id.*

205 *Id.*

206 Hodges v. Commonwealth, 492 S.E.2d 846, 851 (Va. Ct. App. 1997).

e. *Rethinking Proficiency and Judicial Gatekeeping*

As described in this Part, many judges have seen little value in proficiency data, and proficiency data is not directly addressed in discovery rules. The rules for the qualification of experts focus on credentials and experience rather than actual performance and proficiency. When proficiency is addressed, largely in criminal cases, it is used in general ways to justify admissibility, even when the limited proficiency testing data available suggests real problems with the testimony. While case-by-case adjudication may not effectively set up a comprehensive framework to regulate quality of scientific evidence, we suggest that judges could use individual expert proficiency information to make better rulings on expert evidence.

Our argument is that before courts even reach the question of whether an expert used sufficient data and a reliable method to reach a reliable opinion in this case, the court should investigate more critically whether the putative expert truly has expertise in the domain in which the witness is offered as an expert. First, the expertise inquiry is much simpler to conduct than the reliability inquiry. A judge can look for empirical evidence that the person is proficient at a particular task or has esoteric knowledge on some point of relevance. If not, then no further inquiry is needed. Second, this inquiry is particularly important for the sub-group of experts who employ subjective methods to reach their opinions. Any person who claims to be an expert in some domain, whether that expertise supposedly comes from using a subjective or objective method, should be able to demonstrate that expertise through performance-based proficiency testing. Third, accompanying such a threshold inquiry into proficiency, judges should routinely require discovery on the proficiency of experts. Fourth, judges could take proficiency into account, and if a proffered expert is lacking, the judge could retain a more proficient expert. Federal Rule of Evidence 706 is a tool that is "virtually unquestioned" and yet it is lamented by scholars that it is so infrequently used—it allows the judge to exercise discretion to appoint an expert of its choosing—and appointing another expert of greater proficiency would be a sound exercise of discretion.[207] The Court in *Daubert* highlighted how Rule

207 FED. R. EVID. 706 advisory committee's notes to 1972 proposed rules. A pre-*Daubert* survey of federal judges sponsored by the Federal Judicial Center found "uneasiness with court-appointed experts" due to a system in which judges value "adversarial presentation of evidence." Joe S. Cecil & Thomas E. Willging, *Court-Appointed Experts*, *in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 525, 530 (Fed. Jud. Ctr. ed., 1994). A more recent survey of state judges in states that do not follow *Daubert* found many judges did report appointing experts and were willing to do so. Stephanie Domitrovich, Mara L. Merlino, and James T. Richardson, *State Trial Judge Use of Court Appointed Experts: Survey Results and Comparisons*, 50 JURIMETRICS 371, 371 (2010). Many of those judges stated they believed court-appointed experts should be more common and 37% stated a preference for court-appointed experts. *Id.* at 388.

706 provides an additional tool to procure expert witnesses.[208] By reorienting expert qualification around proficiency, Rule 706 could also become a far more useful tool for judges.

i. Proficiency, Qualification, and Admissibility

The proficiency of a proffered expert should be central to the threshold question asking whether a person should be qualified as an expert, as we have proposed, but that proficiency information can additionally inform the second stage of the analysis, focusing on the multi-factored *Daubert* and Rule 702 examination of the validity and reliability of the method. For a subjective or "black box" method, the process is not transparent and the only way to assess its accuracy and reliability is to have proficiency data. Thus, *Daubert* analysis can be informed by the proficiency and error rates of the particular expert, and not just of the field as a whole. Where the expert cannot point to an objective, validated method for reaching a conclusion, proficiency testing is the only means to discern how reliable the expert's method really is.

A more difficult question is what minimal proficiency judges should demand in order for evidence to be admissible. No one threshold may be imposed. We have already stated that any expert should at minimum be able to show that she can perform better than chance and better than a layperson, in order to claim expertise in a subject. Any evidence that analysts had repeatedly done poorly on proficiency tests, by the standards of their discipline, would be important information for a judge to have. What to do with that information may depend on what use the evidence is being put to, and more work would have to be done to assess what thresholds are appropriate for particular disciplines, including as to false positive, false negative, and inconclusive results on proficiency tests. However, courts should be wary of proficiency results derived from tests that have extremely low error rates in the aggregate: either such tests are very easy or the task at hand is very easy—raising questions about whether expertise is even needed to perform the task. An analyst may be quite proficient at DNA testing generally, but a case involving a difficult and more specific question of interpreting a possible DNA mixture may implicate a different type of proficiency. Judges must be sensitive to these differences.

ii. Rethinking Proficiency in the Courtroom

Assuming judges typically find expert evidence admissible, as they currently do, they should permit the jury to hear detailed information about

208 Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 595 (1993) ("Rule 706 allows the court at its discretion to procure the assistance of an expert of its own choosing.").

proficiency. Proficiency matters when an expert takes the stand. Indeed, the most common rulings on the subject involve judges holding that proficiency goes to weight and not to admissibility. If so, judges should be open to discovery on the question of proficiency, should allow questioning regarding proficiency, and should permit other experts to testify regarding the lack of proficiency in a forensic field.[209]

There are constitutional implications to the failure to provide access to proficiency information in pre-trial discovery and at trial in criminal matters. While the Supreme Court has held over the past decade that the Sixth Amendment right to confront adverse witnesses includes the right to question adverse expert witnesses, we submit that having sound proficiency information can make the right to question expert witnesses far more meaningful.[210] A jury should hear detailed information about the proficiency of a testifying expert. Without such information, jurors may have little way to assess the accuracy of the evidence being presented to them. Indeed, there is evidence that jurors assume that the accuracy of forensic science evidence is very high and that the risks of error are vanishingly low.[211] There is also evidence, from a study that we have just conducted, that jurors understand and take into account information about an expert's proficiency when determining the weight of the expert's opinion.[212]

If proficiency testing is not conducted, the other side should be permitted to benefit from an adverse inference concerning the expert evidence—and point out to the jury that the proficiency of the expert is unknown. The result may encourage careful collection of proficiency data in the first instance.

---

209 Some of the cases deeming the evidence of proficiency sufficient note that no contrary expert testimony was offered. *See, e.g.*, United States. v. Foster, 300 F. Supp. 2d 375, 377 (D. Md. 2004) (stating that "[n]o contrary expert testimony was offered").

210 *See* Brandon L. Garrett, *Constitutional Regulation of Forensic Evidence*, 73 WASH. & LEE L. REV. 1147, 1150 (2016) (discussing those Sixth Amendment rulings).

211 See Jonathan J. Koehler, *Intuitive Error Rate Estimates for the Forensic Sciences*, 57 JURIMETRICS J. (forthcoming 2018) ("[C]ourts should take seriously the possibility that jurors will overweight various types of forensic science evidence because they mistakenly believe that the risk of error is infinitesimal."); Joel D. Lieberman, Courtney A. Carrell, Terance D. Meithe & Daniel A. Krauss, *Gold Versus Platinum: Do Jurors Recognize the Superiority and Limitations of DNA Evidence Compared to Other Types of Forensic Evidence?*, 14 PSYCHOL., PUB. POL'Y., & L. 27, 27 (2008) (finding that DNA evidence strongly influences verdict decisions, whether the evidence is incriminating or exculpatory); *see also* Brandon L. Garrett & Gregory Mitchell, *Forensics and Fallibility: Comparing the Views of Lawyers and Judges*, 119 W. VA. L. REV. 621, 636 (2016) ("[O]ur results suggest that most jurors will enter the courtroom with a default view that both DNA and fingerprinting evidence are at least somewhat reliable, if not nearly infallible.").

212 Garrett & Mitchell, *supra* note 211.

## III. REGULATING EXPERT EVIDENCE

In addition to the role of judges as gatekeepers for the quality of expert evidence in litigation, a regulatory approach towards the production and use of expert evidence could improve proficiency across a range of disciplines and ensure that there is sound proficiency data on which judges and factfinders could rely. Scholars have increasingly described administrative law approaches towards law enforcement and prosecution generally, where individual case-specific criminal procedure rules do not adequately address systemic issues.[213] In the area of scientific evidence, the scientific community itself recommends a regulatory model. The National Academy of Sciences in its 2009 recommendations, the White House PCAST report, and other scientific groups and policymakers have all called for federal involvement in the evaluation and oversight of forensic evidence.[214] Scientific research and laboratory research is more commonly regulated, and not just when the evidence used is produced for litigation. The benefits of regulation flow to judges, who may not be in a position to assure sound quality control in lab settings in the first instance, while administrative agencies can directly oversee quality. Existing federal regulation of proficiency for clinical laboratories provides a model for the regulation of forensic laboratories, and helpful guidance can also be found in other nations that regulate the proficiency of forensic analysts.

### A. *Federal Regulation of Proficiency*

In most areas, expert evidence lacks the type of comprehensive regulation that we recommend in this Article, but routine proficiency testing is both feasible and required by regulation in certain fields. In one area in particular, there has been substantial federal regulation of proficiency. Unlike the forensics context, in which laboratories self-regulate through voluntary accreditation and there is no federal agency tasked with developing standard proficiency tests, in the medical context, the federal Centers for Disease

213 *See, e.g.*, Andrew Manuel Crespo, *Systemic Facts: Toward Institutional Awareness in Criminal Courts*, 129 HARV. L. REV. 2049, 2105-06 (2016) (urging courts to broaden the audience that can evaluate expert proficiency); Brandon L. Garrett, *Aggregation in Criminal Law*, 95 CALIF. L. REV. 383, 383 (2007) (advocating for aggregation in criminal proceedings as a way to "create a more efficient, accurate, and fair criminal justice system").

214 NAS Report, *supra* note 21, at 19 (recommending that Congress establish the National Institute of Forensic Science to enforce standards in the forensic science field); PCAST Report, *supra* note 6, at 14-15.

Control and Prevention and other agencies at the state level provide medical laboratories with proficiency tests.[215]

This improved process is due, in large part, to the enactment of the Clinical Laboratory Improvement Act (CLIA) of 1967[216] and subsequent strengthening of this act. In the mid-1980s, journalists wrote about misdiagnosed cancer and lax standards at labs conducting cytology tests of Pap smears.[217] There were concerns with labs failing to identify Pap smears as abnormal, resulting in "large numbers of false negative results, contributing to unnecessary suffering and even death in women who did not receive prompt treatment for cervical cancer."[218] The resulting concern with laboratory proficiency led to the enactment in 1988 of a law strengthening the CLIA and extending the CLIA to all clinical laboratories, whether public or not (so long as they participate in the federal Medicare program or ship items in interstate commerce). "Regular proficiency testing was vital evidence of a laboratory's competence."[219] All medical laboratories must either be CLIA-exempt or have a certificate of registration and compliance and accreditation under the relevant federal regulations.[220] That law required that proficiency testing reflect "to the extent practicable . . . normal working conditions," to make tests realistic.[221] In litigation courts have enforced that requirement.[222] The law also permits the agency to conduct "announced and unannounced on-site proficiency testing of such individuals."[223]

Today, proficiency testing must be done across a wide range of clinical laboratory specialties, including not only cytology but also specialties ranging from bacteriology, to routine chemistry, toxicology, and virology.[224] The proficiency tests must be conducted with "the laboratory's regular patient

---

215 *See Laboratory Registry*, CTRS. FOR MEDICARE & MEDICAID SERVS., https://www.cms.gov/Regulations-and-Guidance/Legislation/CLIA/Laboratory_Registry.html [https://perma.cc/N45Q-PMAM].

216 42 U.S.C. § 263a (2012) (setting certification standards for all clinical laboratories, defined as "a facility for the biological, microbiological, serological, chemical, immuno-hematological, hematological, biophysical, cytological, pathological, or other examination of materials derived from the human body for the purpose of providing information for the diagnosis, prevention, or treatment of any disease or impairment of, or the assessment of the health of, human beings").

217 *See, e.g.*, Walt Bogdanich, *Lax Laboratories: The Pap Test Misses Much Cervical Cancer Through Labs' Errors*, WALL STREET J., Nov. 2, 1987, at A1 (noting that the Pap-screening industry "often ignores what few laws exist to protect women from slipshod testing").

218 Consumer Fed'n of Am. v. United States Dep't of Health and Human Servs., 83 F.3d 1497, 1500 (D.C. Cir. 1996).

219 H.R. REP. NO. 100-899, at 11 (1988), *reprinted in* U.S.S.C.A.N. at 3831; S. REP. NO. 100-561, at 3-4 (1988).

220 42 C.F.R. § 493.5(c) (2016).

221 42 U.S.C. § 263a(f)(4)(B)(iv) (2012).

222 *Consumer Fed'n of Am.*, 83 F.3d at 1500 (noting that CLIA requires proficiency testing to reflect "normal working conditions").

223 42 U.S.C. § 263a(f)(4)(B)(iv) (2012).

224 42 C.F.R. § 493.911-937 (2016).

workload" and use "the laboratory's routine methods."[225] Labs may not communicate with other labs about tests, and they must carefully document each step in the testing process.[226] Any organization that seeks to prepare proficiency tests for labs must have its proficiency test evaluated and approved in advance by the Department of Health and Human Services (HHS).[227] The proficiency tests, however, are not blind. They may be conducted by a supervisor in a lab, and the regulations impose detailed rules for the design and scoring of particular types of proficiency tests.[228] In the area of cytology, individuals who do not receive scores of at least 90 percent must be retested, and if they fail a second test they must receive remedial training and have all of their case work reexamined; if they fail a third test, they may not resume work absent remedial training and retesting.[229]

The CLIA and accompanying regulations do much more to monitor quality at the labs beyond the proficiency testing. They also impose detailed requirements on qualifications of personnel, and on casework in labs, including that potential false negatives (i.e., results finding no anomalies) must be routinely retested to double check the results.[230] All labs must permit random samples to be validated through inspections, and the federal agency can monitor and supervise on-site any labs not found to be fully compliant.[231]

One early challenge to proficiency rules under the CLIA, by public interest groups Consumer Federation of America ("Consumer Federation") and Public Citizen, argued that the regulations adopted by the Department of Health and Human Services, among other things, did not regulate proficiency of cytology labs stringently enough.[232] The regulations tested cytologists by asking them to examine five slides per hour, and not the typical 12.5 slides per hour, and that the test could include a "much higher proportion of abnormal slides than would occur in the average work day."[233] The agency explained only that its

225 42 C.F.R. § 493.801(b)(1) (2016).

226 42 C.F.R. § 493.801(b) (2016).

227 42 C.F.R. § 493.901 (2016).

228 *See, e.g.*, 42 C.F.R. § 493.945 (2016) (discussing the requirements for proficiency testing for gynecologic examinations).

229 42 C.F.R. § 493.855 (2016).

230 *See* 42 USC § 263a(f)(4)(B)(iii)–(iv) (2012) (requiring various rescreening protocols—including "random rescreening of cytology specimens determined to be in the benign category"—and "periodic confirmation and evaluation of the proficiency of individuals involved in screening or interpreting cytological preparations").

231 *See* 42 C.F.R. § 493.61(b)(4) (2016) (requiring accredited laboratories to "[p]ermit random sample validation and complaint inspections"); 42 C.F.R. § 493.563(a) (2016) ("[A] CMS agent may conduct an inspection of an accredited laboratory . . . in response to a substantial allegation of noncompliance."). *See also* 42 C.F.R § 493.1274(c) (2016).

232 Consumer Fed'n of Am. v. Dep't of Health & Human Servs., 83 F.3d 1497 (D.C. Cir. 1996).

233 *Id.* at 1500.

program was based on one used in the state of Maryland.[234] The D.C. Circuit ruled the agency must "articulate a convincing rationale" or engage in rulemaking to create a new proficiency testing protocol.[235] The agency subsequently withdrew the proposed rule, supplemented the record, sought input from the Centers for Disease Control and Prevention, and ultimately concluded that a supervised, time-limited proficiency test should be conducted with slightly different conditions than regular working conditions.[236] This type of review under the Administrative Procedure Act, provides a far more sensible approach to adopting, reviewing, and regulating proficiency than purely voluntary accreditation.[237]

---

234 *Id.* at 1506 (quoting the agency as stating that its testing rate and scoring system was "modell[ed] . . . after that in use in the State of Maryland").

235 *Id.* at 1507 ("[W]e remand to the agency to articulate a convincing rationale for its protocol or to continue the rulemaking process it has already commenced for issuing a new one.").

236 *See* CLIA Program; Cytology Proficiency Testing, 65 Fed. Reg. 14,510-02 (Mar. 17, 2000) (announcing the withdrawal of the "proposed rule on cytology proficiency testing" at issue in *Consumer Federation of America*, reaffirming the belief that said "regulations are appropriate," and "supplying a supplementary statement that further explains the rationale" to that effect).

237 The CLIA itself, as opposed to the ensuing regulations, has not been the subject of much litigation. Federal district courts held that the CLIA did not create a private cause of action for individuals to sue laboratories that do not comply with its provisions. *See, e.g.*, Whitehead v. Edmondson, No. 1:97CV29-S-D, 1998 U.S. Dist. LEXIS 23347, at *5 (N.D. Miss. Mar. 24, 1998) ("[T]he court finds that . . . CLIA does not provide a private right of action to individuals . . . . "); Jewell v. Pinson, No. 255661, 2005 Mich. App. LEXIS 2152, at *18 (Mich. Ct. App. Sept. 1, 2005) ("[W]e are persuaded that the language of the CLIA does not create a private cause of action."); Wood v. Schuen, 760 N.E.2d 651, 659 (Ind. Ct. App. 2001) ("[B]ecause a laboratory director does not incur personal liability for a private negligence action based upon alleged CLIA violations, the trial court did not err in granting summary judgment in favor of Schuen."). For a ruling denying a motion to dismiss on a federal False Claims Act case premised on CLIA violations, see United States *ex rel.* Porter v. HCA Health Servs. of Okla., Inc., No. 3:09-CV-0992, 2011 U.S. Dist. LEXIS 115853, at *15 (N.D. Tex. Sept. 30, 2011). For cases using a CLIA violation as part of a negligence or malpractice case, see McClung v. Lab. Corp. of Am. Holdings, No. 2:06-0336, 2007 U.S. Dist. LEXIS 102611, at *26 (S.D. W. Va. Aug. 1, 2007) and Wilkerson v. Temple Univ. Health Sys., No. 5114, 2005 Phila. Ct. Com. Pl. LEXIS 407, at *11-12 (Phila. Ct. Com. Pl. 2005). One also sees CLIA issues litigated in cases by former employees that allege they were fired in retaliation for whistleblowing regarding CLIA violations that if reported can result in revocation of the laboratories' certification, or litigation by employees fired for violating CLIA rules. *See, e.g.*, Zeigler v. Univ. of Miss. Med. Ctr., 877 F. Supp. 2d 454, 464 (S.D. Miss. 2012) ("Plaintiffs respond that the issue is . . . whether plaintiffs' reporting 'of what may have been criminally illegal conduct' was a motivating factor in the decision to terminate their employment."); Falk v. Phillips, No. 4:06CV00506, 2007 U.S. Dist. LEXIS 93883, at *10-11 (E.D. Ark. Dec. 20, 2007) ("[The plaintiff] alleges defendants terminated him in retaliation for speaking with the surveyor from CLIA."). *See also* Roberts v. St. Agnes Hosp., No. GJH-13-3475, 2015 U.S. Dist. LEXIS 82400, at *2, *22 (D. Md. June 25, 2015) (adjudicating a race-discrimination and retaliation case regarding the termination of a hospital lab technician whose job performance "jeopardized the Hospital's [CLIA] accreditation"). In one case, a state laboratory inspector was fired after federal inspectors uncovered a lack of recordkeeping. Wynn v. State Civ. Serv. Comm'n (Dep't of Health), No. 475 C.D. 2013, 2013 Pa. Commw. Unpub. LEXIS 785, at *1-4 (Pa. Commw. Ct. Oct. 25, 2013). In addition, laboratories have challenged the suspension of their certificates for violating the CLIA, including for checking their

### B. *Regulating the Quality of Proficiency Testing*

A comprehensive regulatory scheme like the CLIA can best ensure a consistent and monitored framework for proficiency testing across laboratories and disciplines. In short, we need something like the CLIA across the wide range of disciplines that produce expert evidence for litigation. Improved judicial review of proferred experts, as we recommended in Part II, might also incentivize such regulation.

Proficiency testing cannot assure reliability of a particular expert or of a field of expertise if it is not conducted in a manner designed to mimic real-world conditions and carefully test the accuracy of the expert's work. Proficiency testing should be independent and blind. The tests should be calibrated at realistic levels of difficulty by a national scientific body. The American Bar Association's Resolution on Forensic Science states that all crime labs should be required to "conduct proficiency testing using blind tests prepared internally or externally and submitted as normal casework evidence or by re-examination by another examiner on completed casework."[238] More recently, the White House's PCAST report recommended that "proficiency testing needs to be improved by making it more rigorous, by incorporating it systematically within the flow of casework, and by disclosing tests for evaluation by the scientific community."[239] Unfortunately, in few laboratories or fields in which evidence is produced for litigation are such recommendations adopted.

Scientific research is being done to develop measures of the difficulty of latent fingerprints, and when this research is completed, then proficiency tests could be set at objectively-defined difficulty levels (which would better permit use of signal detection theory to assess how well the experts can discriminate between signals and noise).[240] Such an approach should be taken across forensic disciplines.

Conducting blind proficiency testing can pose practical challenges given the current workflow of most crime laboratories and their close relationships with law enforcement. Sample evidence must be designed so that it can be realistically incorporated into the casework in a laboratory and so that the experts do not realize that it is a test. If lab analysts frequently contact law enforcement to obtain further information about cases, then they might realize that the sample case is a test. To minimize cognitive bias, however,

---

answers with another laboratory's equipment, or for operating a laboratory and not obtaining CLIA certification. *E.g.* Wade Pediatrics v. Dep't of Health & Human Servs., 567 F.3d 1202, 1203 (10th Cir. 2009) (denying a petition by a testing facility whose CLIA certificate was suspended for one year); Anghel v. Daines, 927 N.Y.S. 2d 710, 715-16 (N.Y. App. Div. 2011) (discussing petitioner's negligence in failing to obtain a CLIA certificate).

238 *See Resolution 111B*, A.B.A. SEC. OF CRIM. JUST. REP. 8 (2004).

239 PCAST Report, *supra* note 6, at 10.

240 *Id.*

such contacts should be eliminated (i.e., analysts should be blind to police beliefs about sources to avoid contamination of the analysis). A lab that conducts blinded analysis can far more readily engage in blind proficiency testing. Only a few U.S. labs, like the Houston Forensic Science Center,[241] have implemented routine blind forensic testing.

A government agency could take on a role in regulating proficiency testing, just as under the CLIA. Of course, scholars for years have called for greater independence of crime laboratories from law enforcement.[242] Few crime laboratories are independent, although more now have scientific oversight. Whether independent or not, they can be regulated by an independent scientific entity. The National Academy of Sciences and many scientists and academics have called for the creation of a National Institute of Forensic Science (NIFS), along the lines of the National Institute of Health (NIH) to comprehensively regulate forensic science.[243] This was one of many proposals in the lengthy 2009 NAS Report, but the committee called NIFS "the greatest hope for success in [reform]," and noted that all of the "remaining recommendations . . . are crucially tied to the creation of NIFS."[244] The proposal to create a NIFS was applauded by many academics and some in the forensic science community.[245] The NIFS proposal was opposed by law enforcement and some in the forensic science community, and it has not been adopted, although legislation continues to be introduced that would create such an agency or otherwise fund federal regulatory efforts in the area.[246] Members of Congress soon focused on the National Institute

241 *See supra* note 82 and accompanying text.

242 *See, e.g.*, M. A. Thompson, *Bias and Quality Control in Forensic Science: A Cause for Concern*, 19 J. FORENSIC SCI. 504, 512 (1974) (proposing that laboratories be controlled by the judicial branch, rather than by the police); *see also* Paul C. Giannelli, *The Abuse of Scientific Evidence in Criminal Cases: The Need for Independent Crime Laboratories*, 4 VA. J. SOC. POL'Y & L. 439, 478 (1997) ("Independent crime labs should be established as part of an augmented Medical Examiner system.").

243 NAS Report, *supra* note 25, at 19-20.

244 *Id.* at 20.

245 *See, e.g.*, Paul C. Giannelli, Daubert *and Forensic Science: The Pitfalls of Law Enforcement Control of Scientific Research*, 2011 U. ILL. L. REV. 53, 53-54 (investigating the dangers of allowing law enforcement agencies to conduct forensic investigations and arguing for an independent forensic science academy); *see also* Quintin Chatman, *How Scientific Is Forensic Science?*, CHAMPION, Aug. 2009, at 36, 37-38 (discussing existing problems with forensic investigation and arguing that the NIFS could help rectify them); *National Research Council's Publication "Strengthening Forensic Science in the United States: A Path Forward": Hearing Before the Subcomm. on Crime, Terrorism, and Homeland Security*, 111th Cong. 32 (2009) [hereinafter Hearing May 2009] (statement of Peter Neufeld, Co-Director, The Innocence Project) (arguing for a strong centralized authority in the forensic sciences to control quality and encourage best practices).

246 *See, e.g.* Forensic Science and Standards Act of 2016, S. 3259, 114th Cong. § 4 (2016), https://www.congress.gov/bill/114th-congress/senate-bill/3259/text/is?format=txt [https://perma.cc/9U4N-M2FX] (proposing the foundation of a National Forensic Science Research Initiative to "improve, expand, and coordinate Federal research in the forensic sciences"); CONG. RESEARCH SERV., S. 2022: FORENSIC SCIENCE AND STANDARDS (Dec. 8, 2014), https://www.govtrack.us/congress/bills/113/s2022/summary

of Standards and Technology (NIST),[247] and NIST has convened a group of committees composed of scientists, lawyers, and judges, to consider improvements to forensic disciplines and standards.[248] The Department of Justice created a National Commission on Forensic Science in 2013 to review standards for forensics, but it was disbanded in early 2017.[249] An entire federal agency need not be created to regulate and improve proficiency testing. Instead, federal grants could be conditioned on compliance with proficiency rules, such as with CLIA. A statute could alternatively require that as a condition for use of federal forensic databanks any laboratory must meet minimum proficiency standards. An existing agency, such as NIST, could certify compliance with regulations.

Unless and until such a comprehensive framework exists, individual labs can adopt proficiency testing and encourage the use of such testing as a matter of best practices, but such a model would not ensure consistent and high-quality proficiency testing.[250] While not perfect, in part because it does not insist on blind testing, CLIA contains comprehensive regulation of quality control at clinical laboratories. It requires setting out in advance the range of tasks at laboratories and their complexity, with testing designed to assess proficiency at each task.[251] The consequences of failed proficiency tests are also set out in advance: retesting, reexamination of the person's casework, and remedial training, and if that does not

---

[https://perma.cc/WQ6R-DS3D] (proposing quality control standards for forensic science); Criminal Justice and Forensic Science Reform Act of 2011, S. 132, 112th Cong. § 101 (2011) (proposing the establishment of an office and board of forensic science). Crime lab directors made strong statements in support of the NIFS proposal. Hearing May 2009, *supra* note 245, at 16 (statement of Peter M. Marone) (advocating for increased funding for the forensic sciences). *But see id.* at 25-26 (statement of Dean Gialamas, President, and Beth Greene, President-Elect, American Society of Crime Laboratory Directors).

247 *Strengthening Forensic Science in the United States: The Role of the National Institute of Standards and Technology: Hearing Before the Subcomm. on Technology and Innovation of the H. Comm. on Science and Technology*, 111th Cong. 3-4 (2009) (introducing the problems plaguing forensic science and how the establishment of the NIFS could help); *see also* D. Michael Risinger, *The NAS/NRC Report on Forensic Science: A Path Forward Fraught with Pitfalls*, 2010 UTAH L. REV. 225, 238 (discussing the extensive hearings on forensic science conducted before Congress).

248 *Forensic Science Standards Effort Takes Shape as NIST Appoints Scientific Area Committees Members*, NAT'L INST. STANDARDS & TECH. (Sept. 3, 2014), http://www.nist.gov/forensics/sac-members-announcement.cfm [https://perma.cc/C2N7-V5W9] (discussing the various appointments made by the NIST).

249 *Department of Justice and National Institute of Standards and Technology Announce Launch of National Commission on Forensic Science*, NAT'L INST. STANDARDS & TECH. (Feb. 15, 2013), https://www.nist.gov/news-events/news/2013/02/department-justice-and-national-institute-standards-and-technology-announce [https://perma.cc/8XTK-3DW9].

250 The PCAST Report asks that the FBI conduct routine blind proficiency testing in its regular casework and that the FBI assist other labs in doing so. PCAST Report, *supra* note 6, at 17. Whether the FBI will take on that role remains to be seen.

251 42 C.F.R. § 493.17 (2016). The entire proficiency testing framework is informed by experts on a CLIA advisory committee, tasked with engaging in ongoing review of testing and standards, including proficiency testing standards. 42 C.F.R. § 493.2001 (2016).

work, then the person cannot do casework until further training produces acceptable results.[252] A similar framework should be adopted for all laboratory work. Case-by-case adjudication cannot create such a framework.

### C. *International Approaches*

Many countries do not have the resources for large crime laboratories like those operating in the United States, and yet they have established far more rigorous systems for evaluating proficiency in forensics. In Germany, an organization called GEDNAP conducts proficiency testing of DNA laboratories.[253] The group was independent of the laboratories themselves, and was founded by the German Society for Legal Medicine.[254] It is run with central involvement of research scientists; the tests are designed by a laboratory at the University of Munster. The program has expanded to include over 220 laboratories from 38 countries, with two tests per year, permitting an international framework for quality control.[255] The testing is not blind for the laboratories; the participants know that they are being tested. However, GEDNAP does review the samples blind by anonymizing the test submissions.[256]

In the U.K., laboratories are required by the United Kingdom Accreditation Service to "define the level and frequency of participation" in proficiency testing and each laboratory must "be prepared to justify their policy and approach" in appropriate proficiency testing.[257] A lab plan for proficiency testing must be "regularly reviewed" in response to changes in the lab.[258] These

252 *See supra* note 229 and accompanying text.

253 *See GEDNAP Stain Commission*, GERMAN DNA PROFILING, http://www.gednap.org/ [https://perma.cc/UD9L-6P22] (providing information regarding the history and standards set by GEDNAP).

254 *Id.*

255 *Id.*

256 *Id.* (discussing the anonymity of the tests performed by the GEDNAP). For a detailed description of GEDNAP, see generally Steven Rand et al., *The GEDNAP (German DNA Profiling Group) Blind Trial Concept*, 116 INT. J. LEGAL MED. 199 (2002) (discussing the procedures and standards GEDNAP uses when performing DNA analysis). Research has been done on the results and the testing provided a set of important lessons identifying common problem areas in laboratories, including "human carelessness" in transcription errors, as well as problems in interpretation of DNA mixtures. S. Rand et al., *The GEDNAP Blind Trial Concept Part II. Trends and Developments*, 118 INT. J. LEGAL MED. 83, 85 (2004) (discussing the developments in GEDNAP's blind trials that help minimize errors).

257 U.K. ACCREDITATION SERV., UKAS POLICY ON PARTICIPATION IN PROFICIENCY TESTING 4.3–4.5 (Nov. 2013), https://www.ukas.com/download/publications/Technical%20Policy%20Statements/TPS%2047%20-%20Edition%203%20-%20November%202016.pdf [https://perma.cc/P4QY-44ZG] (requiring all UK laboratories to participate in proficiency testing).

258 *Id.* at 4.3 (requiring labs to have a plan for participating in proficiency testing that is regularly reviewed).

requirements are not overly detailed. However, these requirements are more rigorous than those in the U.S., for example, adopted by the leading U.S. organization for accrediting forensic labs, the ASCLD/LAB.

In Ontario, Canada, the Centre of Forensic Sciences (CFS) is a branch within the Ministry of Community Safety and Correctional Services.[259] Routine proficiency tests as required by accreditation have been supplemented by a program of blind proficiency testing managed by the CFS Quality Assurance unit. Police and fire investigation agencies submit dummy cases that resemble actual cases that are used to test the lab analysts. Following the reporting of results, feedback is provided to analysts and supervisors.[260] Each of these models suggests far more can be done in the U.S. to regulate proficiency and assure the quality of work performed by forensic laboratories and their analysts.

## CONCLUSION

In a decision eventually reversed by the Tenth Circuit, a district judge threw out evidence of a dog's detection of narcotics in the defendant's luggage because, among other reasons, the dog's handler could present no evidence of the dog's proficiency at detecting drugs:

> Maintenance of a dog's reliability depends on progressive training and daily documentation of the dog's activities. [Defendant's expert] emphasized that thorough and complete daily documentation is "extremely important" and is mandatory for proper on-going training. The handler must design the dog's training schedule based on perceived problems in the field, and it is very important to document false alerts in the field because they are warning signs to the handler that there may be problems with the dog. It is insufficient for the handler to rely on other law enforcement officers to keep detailed records on the dog because they will not be designing or conducting the training sessions. If potential problem areas are unknown, training has little value because it is not tailored to address the dog's deficiencies and the training exercises will have low task difficulty.[261]

A handler must be "constantly vigilant to make sure that the dog does not pick up false cues."[262]

---

259 *Centre of Forensic Sciences*, MINISTRY OF COMMUNITY SAFETY & CORRECTIONAL SERVS., http://www.mcscs.jus.gov.on.ca/english/centre_forensic/CFS_intro.html [https://perma.cc/H6M8-YHZ8] (providing information about the Centre of Forensic Sciences).

260 Email from Jonathan Newman, Deputy Director, Center of Forensic Sciences, to Brandon Garrett (Oct. 20, 2016) (on file with author).

261 United States v. Kennedy, 955 F. Supp 1331, 1335 (D.N.M. 1996).

262 *Id.*

What is true for dogs is true all the more for human experts, who are susceptible to all sorts of misleading cues when retained to assist a party in a case.[263] No technique is immune from error, and for any technique proficiency is relevant. To continue with the dog analogy, just as a dog's pedigree does not ensure top performance at the Westminster Kennel Club Dog Show, an expert's pedigree does not ensure expert performance in court. Judges have been focused on pedigree when they should be looking at empirical evidence of proficiency. As the White House PCAST Report put it: "[N]either experience, nor judgment, nor good professional practices (such as certification programs and accreditation programs, standardized protocols, proficiency testing, and codes of ethics) can substitute for actual evidence of foundational validity and reliability."[264] This is not to say that credentials and experience do not matter; they may be correlated with performance. Or they may not be. One cannot know whether an individual's performance is up to par without evaluating it.

A wide range of disciplines involve some degree of subjective analysis, and for these forms of expertise, performance data is particularly critical. Where an expert's conclusions are formed through steps that are not objectively set out, but rather require some degree of subjective judgment, the expert is a "black box." For these experts, the only way to know how reliable their method is involves proficiency testing.[265] If jurors do not hear that individual experts have known error rates, they may assume the experts are infallible. No expert is infallible.

Before qualifying a person as an expert, judges should open up the black box: judges should directly pose the question whether the person can perform the task accurately and reliably. If the expert cannot demonstrate proficiency, then the expert should not be qualified. Once the expert demonstrates proficiency adequate to be deemed an expert, then the proficiency data can be used to inform the factfinder's assessment of the weight to be given to the expert's testimony. Within a judicial regime that requires proof of proficiency, proficiency testing will follow. But the courts should go further and insist on rigorous proficiency testing. Only by demanding data from realistic blind proficiency testing will courts ensure that parties and their experts come forward with the data needed to ensure that an expert truly is an expert. In mandating this information, judges will greatly simplify the question of expert admissibility, avoiding the more complex methodological inquiries called for by Rule 702 and *Daubert*.

---

263 *See generally* Itiel E. Dror, *A Hierarchy of Expert Performance*, 5 J. APPLIED RES. IN MEMORY & COGNITION 121 (2016) (discussing the susceptibility of experts to bias and using certain criteria to build a hierarchy of expert performance that assesses the reliability of their conclusions).

264 PCAST Report, *supra* note 6, at 6.

265 *Id.* at 5-6 (discussing the subjectivity of certain forensic tests and comparisons conducted by experts).

If courts continue to be indifferent to the problem of proficiency, then comprehensive regulation of forensic proficiency, along the lines of the regulations for clinical laboratories, can ensure that forensic laboratories produce the data. Once the data exists, courts will be far more likely to admit the data as relevant to weight. A consistent regulatory approach would also greatly advance our understanding of the reliability of expert evidence.

Federal Rule of Evidence 702 should be interpreted to require proof that a putative expert is proficient at applying proper tools to proper data to produce correct answers. Experts with unknown proficiency should no more be admitted than evidence of unknown provenance. The judiciary should make empirical evidence of proficiency the touchstone for expert qualification.