1-1-2016

An Empirical Research Agenda for the Forensic Sciences

Jonathan J. Koehler
Assistant Professor of Law *Northwestern Pritzker School of Law*, jay.koehler@northwestern.edu

John B. Meixner Jr.
Assistant Professor of Law *University of Georgia School of Law*, John.Meixner@uga.edu

0091-4169/16/10601-0001
THE JOURNAL OF CRIMINAL LAW & CRIMINOLOGY Vol. 106, No. 1

Printed in U.S.A.

SYMPOSIUM: THE ROLE OF THE COURTS IN IMPROVING FORENSIC SCIENCE

AN EMPIRICAL RESEARCH AGENDA FOR THE FORENSIC SCIENCES†

JONATHAN J. KOEHLER* &

JOHN B. MEIXNER JR.**

After the National Academy of Sciences issued a stunning report in 2009 on the unscientific state of many forensic science subfields, forensic science has undergone internal and external scrutiny that it had managed to avoid for decades. Although some reform efforts are underway, forensic science writ large has yet to embrace and settle upon an empirical research

---

† This research was supported, in part, by a grant from the National Science Foundation to the first author (BCS-1048484), by the Searle Center at Northwestern Law School, and by the Northwestern Pritzker School of Law Faculty Research Program. Many of the ideas for empirical studies described in this article were offered in some form by participants at the Workshop on Cognitive Bias and Forensic Science held at the Northwestern University School of Law on September 23-24, 2010. We thank the following workshop participants for their contributions: Hal Arkes, Deborah Boehm-Davis, Joshua Correll, Shari Diamond, Itiel Dror, David L. Faigman, Thomas D. Gilovich, Lesley Hammer, Reid Hastie, Joshua Klayman, Glenn Langenburg, Craig R. M. McKenzie, Jennifer L. Mnookin, Emily Pronin, Michael J. Saks, Jay Siegel, William C. Thompson, and Paul Windschitl. We also thank Mark Weiss at the National Science Foundation for suggesting and championing this workshop. The statements, conclusions, and recommendations offered in this article do not necessarily reflect the views of the workshop participants, the National Science Foundation, the Searle Center, or Northwestern University.

* Jonathan J. Koehler, Ph.D., Beatrice Kuhn Professor of Law, Northwestern Pritzker School of Law. Correspondence for this chapter should be addressed to Jonathan J. Koehler, Northwestern University School of Law, 375 E. Chicago Avenue, Chicago, IL 60611. Contact: jay.koehler@northwestern.edu, 312-503-4469.

** John B. Meixner Jr., J.D., Ph.D., Associate, Schiff Hardin LLP.

agenda that addresses knowledge gaps pertaining to the reliability of its methods. Our paper addresses this problem by proposing a preliminary set of fourteen empirical studies for the forensic sciences. Following a brief discussion of the courtroom treatment of forensic science evidence, we sketch a series of studies that should be conducted to increase understanding of what forensic examiners are doing, how accurately they are doing it, and how cognitive bias may affect the work product. We also propose several studies that examine how the specific questions examiners are asked might affect the validity and persuasiveness of examiners' responses. We conclude by affirming the importance of developing a research culture within the forensic sciences that includes a commitment to conducting, participating in, and relying upon high quality empirical research.

Keywords: Empirical, Forensic science, Judicial decision making, Juries, Scientific evidence

TABLE OF CONTENTS

INTRODUCTION

John and Sally Sweek were brutally stabbed to death in Texas in 1987.[1] Steven Chaney was charged with their murder.[2] While Chaney knew the victims and there was evidence that Chaney owed the Sweeks approximately $500 for drugs, the key evidence against Chaney was a bite mark on John Sweek's arm.[3] Two forensic dentists testified that Chaney's teeth matched the bite mark.[4] One of the dentists said that there was just "1 to a million chance" that someone other than Chaney was the source of the bite mark.[5] Chaney was convicted and sentenced to life imprisonment. On October 12, 2015, after spending the previous twenty-eight years in prison, Steven Chaney's conviction was reversed after a court concluded that the bite mark testimony was junk science.[6] Even the dentist, who thought it was practically impossible that anyone other than Chaney was the biter, now believes that his own testimony was unfounded.[7]

---

[1] Chaney v. State, 775 S.W.2d 722, 723–24 (Tex. App. 1989).

[2] *Id.* at 724–25.

[3] *Id.*

[4] *Id.* at 725–26.

[5] Brandi Grissom, *Junk Science Cited in Bid to Clear Man in '89 Dallas Killing*, THE DALLAS MORNING NEWS (Oct. 11, 2015), http://www.dallasnews.com/news/crime/headlines/20151010-junk-science-cited-in-bid-to-clear-man-in-89-dallas-killing.ece.

[6] *Id.*

[7] *Id.* (discussing an affidavit filed in 2015 in which the dental expert who had offered the "1 to a million chance" claim in 1987 wrote, "[c]onclusions that a particular individual is the biter and their dentition is a match when you are dealing with an open population are now understood to be scientifically unsound.").

Although trial courts have routinely admitted bite mark evidence for decades,[8] Chaney was the twenty-sixth person since 2000 whose conviction was released or indictment dismissed based on discredited bite mark testimony.[9]

In February 2016, just a few months after Chaney's release, the Texas Forensic Science Commission called for an end to the use of bite mark evidence in criminal trials.[10] According to the *New York Times*, the Commission concluded that "the validity of the technique has not been scientifically established."[11] That is an understatement. According to a 2009 National Academy of Sciences (NAS) report, "[a]lthough the majority of forensic odontologists are satisfied that bite marks can demonstrate sufficient detail for positive identification, no scientific studies support this assessment, and no large population studies have been conducted."[12] Bite mark analysis has not fared any better in studies conducted since this NAS report appeared. A 2015 study showed that experienced, certified forensic odontologists often disagreed both about who was the source of a bite mark in a crime scene photograph and whether the marking in question was a bite mark at all.[13]

The research community has long known that the scientific basis for bite mark analysis is thin.[14] Why, then, do courts routinely admit this

---

[8] Paul C. Giannelli, *Bite Mark Analysis*, 43 CRIM. L. BULL. 930, 943–45 (2007).

[9] *Dallas District Attorney and Innocence Project Move to Reverse Conviction Based on False Bite Mark Testimony*, INNOCENCE PROJECT (Oct. 12, 2015), http://www.innocenceproject.org/news-events-exonerations/dallas-district-attorney-and-innocence-project-move-to-reverse-conviction-based-on-false-bite-mark-testimony.

[10] ERIK ECKHOLM, *Texas Panel Calls for an End to Criminal IDs via Bite Mark*, N.Y. TIMES (Feb. 12, 2016), http://www.nytimes.com/2016/02/13/us/texas-panel-calls-for-an-end-to-criminal-ids-via-bite-mark.html.

[11] *Id.*

[12] NAT'L RESEARCH COUNCIL, STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES: A PATH FORWARD 87, 176 (2009) [hereinafter NAS Report]; *see also* Iain A. Pretty & David Sweet, *The Scientific Basis of Human Bite Mark Analyses – A Critical Review*, 41 SCI. & JUST. 85, 86 (2001) ("Despite the continued acceptance of bitemark evidence in European, Oceanic, and North American Courts, the fundamental scientific basis for bitemark analysis has never been established.").

[13] Adam J. Freeman & Iain A. Pretty, *Construct Validity of Bitemark Assessments Using the ABFO Decision Tree*, AM. ACAD. OF FORENSIC SCI. ANN. MEETING (Feb. 19, 2015). The method and results of this as yet unpublished study are discussed in an amicus curiae brief in Richards v. Fox, No. S223651(Cal. 2015) (brief available at 2015 WL 5779457). According to the amicus curiae brief, thirty-nine examiners in the study reviewed injuries depicted in 100 crime scene photographs. In only four of the 100 cases did all examiners agree on whether an injury was a bite mark or not. In seventy-one of the 100 cases, less than 70% of the forensic odontologists agreed about whether the injury was or was not a bite mark.

[14] Pretty & Sweet, *supra* note 12, at 86 ("The fundamental scientific basis for bite mark

evidence[15] and permit testimony of the sort that led to the conviction of Steven Chaney? Is there something about bite mark evidence in particular that has fooled courts into treating it as reliable science? Or is the problem a more general one pertaining to the beliefs that people have about the reliability and accuracy of forensic science[16] evidence? In our view, the available evidence supports the latter conclusion. Although bite mark analysis is surely among the weakest of the forensic sciences,[17] it is not the only forensic science that lacks a sufficient scientific foundation to connect an evidentiary sample to its source.[18]

The idea that many forensic sciences lack a sufficient scientific foundation is not original with us, nor is this the takeaway point of our paper. Instead, we offer the untested nature of many of the forensic sciences as motivation for recommending a series of scientific studies that may provide guidance to legal decision makers about the reliability and validity of forensic science conclusions. In calling for additional research, our target audience is not so much those who have already made up their minds about the value of forensic science evidence as it is those who want and need to know what forensic methods can and cannot achieve in practice and how to evaluate the strength of forensic evidence, as promoted by unbiased, empirical data.

In September 2010, the National Science Foundation sponsored a two-day workshop on forensic science and cognitive bias at Northwestern Law School.[19] Many of the workshop participants were experimental psychologists with expertise in conducting studies that describe and

---

analysis has never been established."); *see also* Brief for Michael J. Saks, et al. as Amici Curiae Supporting Petitioner, *In re* William Richards on Habeas Corpus, No. S223651, 2015 WL 5779457 at *45 (Cal. 2015) ("The claims of forensic dentistry have for decades outrun empirical testing of those claims. Rather than confirming the field's claims, recent research . . . has confirmed that the foundations of bite mark identification are unsound.").

[15] *Id.* at 86.

[16] Forensic science is the application of science to legal matters. Forensic science identification techniques include DNA analysis, fingerprints, handwriting analysis, bite marks, hair and fiber analyses, ballistics, tool marks, etc. *See* RICHARD SAFERSTEIN, CRIMINALISTICS: AN INTRODUCTION TO FORENSIC SCIENCE 4 (10th ed. 2011).

[17] NAS Report, *supra* note 12, at 176 (referring to "the inherent weaknesses involved in bite mark comparison").

[18] *Id.* ("With the exception of nuclear DNA analysis, however, no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source.").

[19] *See* JONATHAN J. KOEHLER & JOHN B. MEIXNER, WORKSHOP ON COGNITIVE BIAS AND FORENSIC SCIENCE, FINAL REPORT 8 (Aug. 7, 2013), www.law.northwestern.edu/faculty/conferences/workshops/cognitivebias/documents/NSFWorkshopReportFinal.pdf (explaining cognitive biases are "systematic distortions in thinking that occur when information passes through the subjective filters of human beliefs, attitudes, and experiences").

improve upon human judgment and decision making. A primary goal of that gathering was to identify the role that the natural limits and biases of human decision makers play in the forensic science process. In thinking about the type of empirical research in forensic science that would be most helpful, we drew liberally on the ideas and proposals offered at this workshop.

Our paper is organized as follows: Section I provides background information on the admission of forensic science evidence in court and recent developments that have spurred calls for reform. Section II proposes a series of scientific studies that consider what exactly examiners are doing, how accurately they are doing it, and how cognitive bias may affect an examiner's work. Section II also proposes several studies that examine how the particular questions that examiners are asked affect the scientific validity of their responses and how factfinders weigh these responses. Part III concludes with a call to make the empirical studies we propose here part of the forensic reform efforts that are under way.

I. HISTORY

In *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,[20] the U.S. Supreme Court introduced a new standard for the admissibility of scientific evidence.[21] Drawing on Federal Rule of Evidence 702, *Daubert* held that the "trial judge must ensure that any and all scientific testimony or evidence admitted is not only relevant, but reliable."[22] According to the court, the hallmark of scientific knowledge is its reliability,[23] and so it stands to reason that courts should require proof that the principles and methods that lie behind scientific evidence and testimony are demonstrably reliable. *Daubert* also provided general guidance for trial judges on how they might go about assessing the reliability of proffered scientific evidence. These so-called *Daubert* factors are nonexclusive and include a consideration of (a) the extent to which the underlying scientific theory has been tested, (b) the existence of peer-reviewed publications, (c) the "known or potential rate of error" of the method, (d) the existence of "standards controlling the technique's operation," and (e) general acceptance within the scientific

---

[20] 509 U.S. 579 (1993).

[21] *Id.* at 593–94.

[22] *Id.* at 589.

[23] When it referred to the "reliability" of scientific evidence or scientific knowledge, the Court had in mind something more akin to what scientists refer to as validity. That is, reliable knowledge, according to the Court, is knowledge that is valid and can be trusted. *See id.* at 590 n.9 (discussing "evidentiary reliability" and its relationship to validity).

community.[24]

However, as noted in the 2009 NAS Report, "[r]eview of reported judicial opinions reveals that, at least in criminal cases, forensic science evidence is not routinely scrutinized pursuant to the standard of reliability enunciated in *Daubert*."[25] Indeed, trial courts typically rely on the long history of admitting various types of forensic science evidence, and essentially give this type of evidence a pass when it comes to proof of reliability.[26]

At the same time, it is undeniably true that a paradigm shift is underway in the criminal justice system with respect to forensic science evidence.[27] Forensic science results no longer have the same aura of infallibility that they had as recently as a decade ago. Crime lab scandals, fraud, unsupported assumptions, high profile errors, and wrongful convictions that point to faulty forensic techniques and testimony at the trial level have all contributed to a national movement to investigate and reassess the value of different types of forensic science evidence.[28] As evidence of the paradigm shift, in 2013, the U.S. Government established the National Commission on Forensic Science (NCFS).[29] NCFS is charged

---

[24] *Id.* at 593–94.

[25] NAS Report, *supra* note 12, at 106. The same may not be true for civil cases. *See* D. Michael Risinger, *Navigating Expert Reliability: Are Criminal Standards of Certainty Being Left on the Dock*?, 64 ALB. L. REV. 99 (2000) (arguing that the heightened scrutiny of scientific evidence *Daubert* requires has continued to expand in civil, but not criminal, cases); *see also* NAS Report, *supra* note 12, at 98 ("[I]ronically, the appellate courts appear to be more willing to second-guess trial court judgments on the admissibility of purported scientific evidence in civil cases than in criminal cases.").

[26] Simon A. Cole, *Grandfathering Evidence: Fingerprint Admissibility Rulings from Jennings to Llera Plaza and Back Again*, 41 AM. CRIM. L. REV. 1189, 1216–19 (2004) (arguing that fingerprint evidence has never been scrutinized by trial courts using the *Daubert* factors because this type of evidence was too important to the criminal justice system to risk being ruled inadmissible).

[27] Michael J. Saks & Jonathan J. Koehler, *The Coming Paradigm Shift in Forensic Identification Science*, 309 SCI. 892, 895 (2005) (describing "a paradigm shift in the traditional forensic identification sciences in which untested assumptions and semi-informed guesswork are replaced by a sound scientific foundation and justifiable protocols"); M. Chris Fabricant & Tucker Carrington, *The Shifted Paradigm: Forensic Science's Overdue Evolution from Magic to Law*, 4 VA. J. CRIM. L. 1, 1 (2016) ("[T]he predicted paradigm shift has occurred.").

[28] NAS Report, *supra* note 12, at 44 ("In recent years, the integrity of crime laboratories increasingly has been called into question, with some highly publicized cases highlighting the sometimes lax standards of laboratories that have generated questionable or fraudulent evidence and that have lacked quality control measures that would have detected the questionable evidence.").

[29] *See generally* U.S. DEP'T OF JUSTICE, NATIONAL COMMISSION ON FORENSIC SCIENCE, https://www.justice.gov/ncfs (last visited May 23, 2016).

with making policy recommendations to the Department of Justice that will improve the "validity and reliability of the forensic sciences."[30] Though it is still too early to know how NCFS will change the forensic science landscape, early indications point to a focus on eliminating exaggerated, unproven, and unscientific claims that have been made about forensic science evidence.[31]

II. WHERE TO FROM HERE?

With or without assistance from trial courts, forensic science in the United States is undergoing a kind of internal and external scrutiny that it has avoided for decades. Reform efforts to date from the NCFS have focused largely on quality management,[32] laboratory accreditation,[33] ethics,[34] system upgrades,[35] curriculum development,[36] awareness of potential biases,[37] and providing more cautious scientific conclusions in

---

[30] U.S. DEP'T OF JUSTICE, NATIONAL COMMISSION ON FORENSIC SCIENCE, CHARTER (2015), https://www.justice.gov/ncfs/file/624216/download./file/624216/download.

[31] NAT'L COMM'N ON FORENSIC SCI., RECOMMENDATIONS TO THE ATTORNEY GENERAL REGARDING USE OF THE TERM "REASONABLE SCIENTIFIC CERTAINTY," FINAL DRAFT (Mar. 3, 2016), http://www.ascld.org/wp-content/uploads/2016/03/Final-Draft-Recs-Doc-on-The-Use-of-The-Term-Reasonable-Scientific-CertaiFalsepdf ("Forensic discipline conclusions are often testified to as being held 'to a reasonable degree of scientific certainty' or 'to a reasonable degree of [discipline] certainty.' These terms have no scientific meaning and may mislead factfinders about the level of objectivity involved in the analysis . . . . The Attorney General should direct all attorneys appearing on behalf of the Department of Justice (a) to forego use of these phrases when presenting forensic discipline testimony . . . ."); NAT'L COMM'N ON FORENSIC SCI., DRAFT POLICY RECOMMENDATION ON EXPERT TESTIMONY 2 (Oct. 12, 2014), https://www.justice.gov/sites/default/files/pages/attachments/2014/10/15/draft_on_expert_testimony.pdf ("Experts should not use misleading terms that suggest that the methodology or the expert is infallible when testifying.").

[32] NAT'L COMM'N ON FORENSIC SCI., RECOMMENDATION TO THE ATTORNEY GENERAL REGARDING TRANSPARENCY OF QUALITY MANAGEMENT SYSTEM DOCUMENTS (Mar. 22, 2016), https://www.justice.gov/ncfs/file/839706/download.

[33] NAT'L COMM'N ON FORENSIC SCI., VIEWS OF THE COMMISSION REGARDING CRITICAL STEPS TO ACCREDITATION (Mar. 22, 2016), https://www.justice.gov/ncfs/file/839701/download.

[34] NAT'L COMM'N ON FORENSIC SCI., RECOMMENDATION TO THE ATTORNEY GENERAL NATIONAL CODE OF PROFESSIONAL RESPONSIBILITY FOR FORENSIC SCIENCE AND FORENSIC MEDICINE SERVICE PROVIDERS (Mar. 22, 2016), https://www.justice.gov/ncfs/file/839711/download.

[35] NAT'L COMM'N ON FORENSIC SCI., DIRECTIVE RECOMMENDATION: AUTOMATED FINGERPRINT INFORMATION SYSTEMS (AFIS) INTEROPERABILITY (Aug. 11, 2015), https://www.justice.gov/ncfs/file/786576/download.

[36] NAT'L COMM'N ON FORENSIC SCI., FORENSIC SCIENCE CURRICULUM DEVELOPMENT (Dec. 8, 2015), https://www.justice.gov/ncfs/file/818206/download.

[37] NAT'L COMM'N ON FORENSIC SCI., ENSURING THAT FORENSIC ANALYSIS IS BASED UPON TASK-RELEVANT INFORMATION (Dec. 8, 2015), https://www.justice.gov/ncfs/file/

reports and testimony.[38] While these reform efforts are important, they have not focused on creating a body of knowledge about the various forensic sciences that legal decision makers need. What are the best forensic examiners doing? What is the probative value of forensic conclusions? What are the factors that affect the accuracy of forensic conclusions or the confidence with which examiners defend their conclusions? How, if at all, do the non-forensic contextual features of a case affect forensic conclusions? Are legal decision makers affected by the manner in which forensic scientists describe their findings? These types of questions are best addressed by empirical study.

Various types of empirical studies may be used to address these questions. Field studies, field experiments, experimental simulations, and laboratory experiments are several options. In a field study, the researcher makes systematic observations within the naturally occurring system under study.[39] Exploratory studies that examine how forensic examiners conduct their analyses could use this approach. A field experiment is similar to a field study, but here the researcher deliberately manipulates one or more variables and then measures its effect on one or more dependent variables.[40] For example, if one group of fingerprint examiners working on a case was told that Suspect #1 had not confessed, whereas another group of examiners was told instead that Suspect #2 had confessed, a field experiment could measure the effect of the confessions on the examiners' judgments. An experimental simulation involves constructing a setting that captures a naturally-occurring setting.[41] It is similar to a field experiment in that the researchers manipulate one or more variables. However, the researcher conducting a field experiment is also responsible for recreating a setting that appears naturally. In a laboratory experiment, the researcher does not try to recreate a naturally occurring setting. Instead, the researcher creates an artificial setting that permits a close examination of various measured variables and other potential causal variables.[42] The focus in a laboratory experiment is more on internal validity (i.e., our confidence that the manipulated variable caused a change in the measured variable) than external validity (i.e., our confidence that the study's result generalizes to other situations and people).

818196/download.

[38] NAT'L COMM'N ON FORENSIC SCI., *supra* note 31.

[39] PHILIP J. RUNKEL & JOSEPH E. MCGRATH, RESEARCH ON HUMAN BEHAVIOR: A SYSTEMIC GUIDE TO METHOD 1, 90 (1972).

[40] *Id.* at 94–95.

[41] *Id.* at 96.

[42] *Id.*

All of these types of studies have scientific merit and all of them can be used to expand our knowledge about the forensic sciences. In the sections below, we briefly describe fourteen scientific studies that could and should be conducted in support of this goal.

DESCRIPTIVE STUDIES: EXAMINER METHODS

Most of the studies we suggest in this paper seek *knowledge and clarity* in the forensic sciences, rather than *improvement*. Of course, improvement is a likely byproduct of scientific study, and there is nothing wrong with collecting data that address prescriptive questions about how forensic science examiners should conduct themselves in light of the practical constraints they face. But in light of the reality that forensic science currently plays and will likely continue to play a pivotal role in our criminal justice system, the research priority at this juncture should be on providing consumers of forensic science (i.e., courts and juries) with the information they need to evaluate this evidence. Perhaps the most useful descriptive account for purposes of the legal system would focus on identifying the strengths and weaknesses of various types of forensic evidence—e.g., how often do examiners make errors? Are certain types of errors more frequent than others? Are there particular contexts in which errors are more common? Such information is largely unavailable, yet it is critically important for a judge who deliberates over the admissibility of forensic evidence or a factfinder who thinks about how much weight to give this evidence once admitted. However, we cannot expect to understand these more complex questions without a basic grasp of how forensic examiners do their work. Surprisingly, there is little formal standardization of methods within the forensic communities—even the best of the non-DNA forensic sciences can be considered as much art as science.[43] In the three studies described below, we seek to grasp—at a very basic level—the methods forensic examiners use in making their determinations.

Study 1: What do examiners generally look for in making comparisons?

While many forensic disciplines have guidelines that examiners follow when making comparisons between samples, those guidelines rarely have a rigorous, detailed structure. For example, while bite mark examiners typically "compar[e] the pattern size, and shapes of the suspect's teeth with

[43] DONALD E. SHELTON, FORENSIC SCIENCE IN COURT: CHALLENGES IN THE TWENTY-FIRST CENTURY 50 (2011) ("[F]ingerprint comparison, without scientific standards of point similarity, is perhaps more art than science.").

the unknown bite mark using transparent comparison overlays,"[44] there is no standardized method for how to make the comparison[45] or requirement as to what tools should be used to make the comparison.[46] Fingerprint comparison—perhaps the most established forensic discipline—likewise offers little more than a general framework for examiners to follow. Under the ACE-V (Analysis, Comparison, Evaluation, and Verification) method, examiners compare a latent fingerprint (a questioned sample) with a reference sample (a known sample) by gathering relevant data from the two fingerprints, such as the pattern of ridges or orientation of loops in the fingerprints.[47] Most descriptions of the method do not provide specific, detailed rules for examiners to follow when making comparisons or source judgments.[48] Should one assess certain features first before moving to others? How detailed should the initial analysis be before making comparisons between samples? How much distinction between the samples can there be without eliminating the possibility of a match? The literature

---

[44] David Sweet et al., *Computer-Based Production of Bite Mark Comparison Overlays*, 45 J. FORENSIC SCI. 1050, 1050 (1998).

[45] Giannelli, *supra* note 8, at 936–37 ("Although the expert's conclusions are based on objective data, the opinion is essentially a subjective one. There is no accepted minimum number of points of identity required for a positive identification. The experts who have testified in reported bite mark cases have used a low of eight points of comparison to a high of fifty-two points. Like fingerprint and firearms identifications . . . , the conclusions are based on the examiner's experience and expertise.").

[46] *See* Iain Pretty, *Unresolved Issues in Bitemark Analysis*, *in* BITEMARK EVIDENCE 547, 553–54 (Robert B. J. Dorion, ed., 2005) ("An essential component of the determination of the validity of bitemark analysis is that the techniques used in the physical comparison between suspect dentition and physical injury have been assessed and found valid. One of the fundamental problems with this task is the wide variety of techniques that have been described in the literature. Techniques using confocal, reflex and scanning electron microscopes; complex computer systems; typing of oral bacteria; special light sources; fingerprint dusting powder; and overlays have all been reported.").

[47] *See, e.g.*, IGOR PACHECO ET AL., MIAMI-DADE POLICE DEP'T FORENSIC SERVS. BUREAU, MIAMI-DADE RESEARCH STUDY FOR THE RELIABILITY OF THE ACE-V PROCESS: ACCURACY & PRECISION IN LATENT FINGERPRINT EXAMINATIONS 14–15 (2014).

[48] *See, e.g.*, *id.* (describing ACE-V at only a high level of generality); Glenn Langenburg, *A Performance Study of the ACE-V Process: A Pilot Study to Measure the Accuracy, Precision, Reproducibility, Repeatability, and Biasability of Conclusions Resulting from the ACE-V Process*, 59 J. FORENSIC IDENTIFICATION 219, 226 (2009) (describing flexibility for examiners in determining which comparisons are relevant and how to document those comparisons); Philip J. Kellman et al., *Forensic Comparison and Matching of Fingerprints: Using Quantitative Image Measures for Estimating Error Rates Through Understanding and Predicting Difficulty*, 9 PLOS ONE 1, 3 (2014) ("[T]here is no formalized process for any of [the ACE-V] steps. There is no method or metric for specification of which features should be used for comparison, nor any general measure for what counts as sufficient information to make a decision. Examiners rely on their experience and training rather than formal methods or quantified rubrics at each step of the process.").

does not appear to provide uniform guidance to examiners on these, and a host of other, important questions.

We suggest beginning with a series of basic, observational studies. Several methods could provide insight. We recommend that a researcher obtain a reasonably random sample of twenty or so examiners and either (1) ask them to walk through the specific methods they use when making a comparison, or (2) observe them making comparisons[49] while "thinking out loud" and explaining their specific procedures.[50] Both methods could be employed together, though there would be some risk that interviewing the examiner prior to observing her methods might have some influence on the methods themselves.[51] Researchers should try to identify as many specific aspects of examiners' procedure as possible, but they might specifically focus on: (1) the types and numbers of features that examiners track (e.g., ridges and bifurcation minutia in fingerprint analyses, size and mold characteristics in shoe print analyses), (2) the order in which examiners conduct their analyses, (3) the weight examiners place on various features they analyze, and (4) the criteria that examiners use when drawing conclusions. We offer no hypotheses, but anticipate that these studies will reveal great variability across—and even within—the various forensic science subfields.[52]

Study 2: How much variability is there in examiner methods?

Once a baseline of methods in a particular forensic discipline is established, a natural follow-up question is, "do those methods differ between examiners within a single laboratory, between laboratories within a region, or between regions?" This question is important because, as discussed above, many forensic sciences appear to lack standard protocols

---

[49] For the sake of ensuring appropriate power, researchers will want to consider the extent and specificity of the questions they plan to ask examiners when deciding how many samples each examiner should analyze under observation.

[50] For a classic discussion of the value of think aloud protocols, see K. Anders Ericsson & Herbert A. Simon, *Verbal Reports as Data*, 87 PSYCHOL. REV. 215 (1980).

[51] Alternatively, employing both methods could provide some unique insight into the differences, if any, in how examiners *think* they conduct their analyses and how they actually conduct them.

[52] We also recognize that examiners may not be able to explicitly explain certain parts of their analyses, as there are likely implicit aspects of processing that are relevant to their decision making. *See* Kellman et al., *supra* note 48, at 2 ("It would be a mistake . . . to infer that the processes of pattern comparison and the determinants of difficulty are . . . fully available for conscious report or explicit description. As in many other complex tasks in which learning has led to generative pattern recognition . . . and accurate classification, much of the relevant processing is likely to be at least partly implicit.").

that are broadly followed by practitioners.[53] Anecdotal evidence suggests that laboratories and regions vary widely in terms of the methods they use and the ways that common methods are deployed.[54] Reliable data on this matter are sorely needed. If the methods and practices of examiners in a common forensic subfield vary, then studies conducted in one region or laboratory may not tell us much about what examiners elsewhere are doing. At a more basic level, it would be alarming if examiner methods differ greatly between laboratories, as this could limit any broad conclusions that could be made about quality and validity of the examiners' conclusions.[55] This research is a natural follow-up to Study 1. The basic observational approach also seems appropriate here, though perhaps over a larger, more diverse set of examiners. Follow-up analysis will likely depend on the extent of variability found in examiner methods.

Study 3: Do the most effective examiners employ unique methods?

Once we know more about the different methods employed by forensic scientists, it is appropriate to try to determine which methods produce the "best" outcomes. One way to accomplish this goal would be to identify the

---

[53] NAS Report, *supra* note 12, at 6 ("Often there are no standard protocols governing forensic practice in a given discipline. And, even when protocols are in place (e.g., SWG standards), they often are vague and not enforced in any meaningful way.").

[54] Scott Bader, *Peak Height: DNA*, *in* WILEY ENCYCLOPEDIA OF FORENSIC SCIENCE 2007, 2008 (2009) (In the DNA context, "[t]here is some debate about the threshold value above which a peak can be declared as a 'real' peak that represents a piece of DNA, as opposed to chance occurrence of noise of sufficient intensity to appear as a peak. Different laboratories use different thresholds, decisions that may be based on objective principle, scientific validation, or rule-of-thumb experience."); NAT'L INSTIT. OF STANDARDS & TECH., LATENT PRINT EXAMINATION AND HUMAN FACTORS: IMPROVING THE PRACTICE THROUGH A SYSTEMS APPROACH 7 (2012) (In the fingerprint context, "[t]he thresholds for these decisions can vary among examiners and among forensic service providers. Some examiners state that they report identification if they find a particular number of relatively rare concurring features, for instance, eight or twelve. Others do not use any fixed numerical standard. Some examiners discount seemingly different details as long as there are enough similarities between the two prints. Other examiners practice the one-dissimilarity rule, excluding a print if a single dissimilarity not attributable to perceptible distortion exists. If the examiner decides that the degree of similarity falls short of satisfying the standard, the examiner can report an inconclusive outcome. If the conclusion is that the degree of similarity satisfies the standard, the examiner reports an identification."); NAS Report, *supra* note 12, at 139 (In the fingerprint context, "[i]n the United States, the threshold for making a source identification is deliberately kept subjective, so that the examiner can take into account both the quantity and quality of comparable details. As a result, the outcome of a friction ridge analysis is not necessarily repeatable from examiner to examiner.").

[55] NAS Report, *supra* note 12, at 16 ("The fragmented nature of the enterprise raises the worrisome prospect that the quality of evidence presented in court, and its interpretation, can vary unpredictably according to jurisdiction.").

most successful examiners, and then study the ways in which their methods differ from those of other examiners. Unfortunately, this is easier said than done. Forensic examiners do not come with "batting averages," and it is rarely possible to truly know whether an examiner did or did not make the correct call in a case due to the lack of ground truth. Therefore, the first step in this process would be to identify one or more ways to spot the most effective examiners.[56]

One possible method could involve proficiency tests in which ground truth is available. The methods used by top scorers on a proficiency test that includes challenging, realistic samples could be compared to those used by others. A second, less direct approach would be to develop and rely on a reputation index within the forensic community. The forensic science community could be polled, in essence, to determine who among them is believed to be part of the most skilled or accurate subgroup. Once this elite subgroup is identified, the observation would proceed as in Studies 1 and 2, with a focus on identifying differences in method between more skilled and less skilled examiners. Obviously, the second method is simpler to administer than the first. But the first method would likely yield more reliable results and has other ancillary benefits, such as providing evidence about which types of samples are more or less likely to be analyzed correctly by forensic scientists of varying ability levels.

DESCRIPTIVE STUDIES: EFFECTS OF DIFFERENCES IN SAMPLE AND METHODOLOGY ON ACCURACY

Beyond the baseline studies described above that identify *what* examiners are doing, the most legally relevant questions regarding forensic science involve *how well* examiners are doing. These questions cut straight to the heart of the *Daubert* Court's emphasis on the validity[57] of an expert's methods and the likelihood that those methods will produce accurate data for the trier of fact. Although the courts have largely given a pass to the forensic sciences on matters pertaining to error rate,[58] fingerprint scholars have conducted some research on error rates.[59] Although these error rate

---

[56] *See, e.g.*, Jason M. Tangen et al., *Identifying Fingerprint Expertise*, 22 PSYCHOL. SCI. 995 (2011); Matthew B. Thompson et al., *Expertise in Fingerprint Identification*, 58 J. FORENSIC SCI. 1519 (2013); Kellman et al., *supra* note 48, at 3, 10–11.

[57] *See Daubert*, 509 U.S. 579, 590 n.9 (1993); *supra* note 23.

[58] Nancy Gertner, *Commentary on the Need for a Research Culture in the Forensic Sciences*, 58 UCLA L. REV. 789, 790 (2011).

[59] *See, e.g.*, Thomas A. Busey & John R. Vanderkolk, *Behavioral and Electrophysiological Evidence for Configural Processing in Fingerprint Experts*, 45 VISION RES. 431, 436 (2005); Bradford T. Ulery et al., *Accuracy and Reliability of Forensic Latent Fingerprint Decisions*, 108 PROC. NAT'L ACAD. SCI. U.S. 7733, 7737 (2011); PACHECO ET

studies are seriously flawed,[60] the fact that such tests were undertaken in one forensic science domain suggests movement in the direction of science and data, and away from the unsupported opinions and beliefs of forensic scientists.[61] While we support a comprehensive program of blind proficiency testing for the forensic sciences, administered by disinterested parties using realistic samples,[62] we propose five additional studies below (Studies 4–8) that examine the effects of *sample difficulty* on errors, the *distribution* of errors, and the *value* of particular types of examination.

Study 4: Does the difficulty of the sample affect accuracy?

Generally, a forensic examiner makes comparisons between one or more questioned samples (often recovered from a crime scene) and one or more known samples. Questioned samples are highly variable in quality. A shoe may only have marked a portion of a surface, a bite may not have been complete, or a fingerprint may have smeared. One implication of this variability is that some comparisons an examiner is asked to make are more difficult than others. Relatedly, some forensic sources may leave markings that are not particularly distinctive, and this may also make it more difficult for examiners to narrow the set of potential donors of a questioned sample. Other factors may influence difficulty as well. Whether this variation affects the accuracy of an examiner's conclusions is a relevant question for a judge or factfinder because the difficulty of a decision could affect the weight a decision maker gives to the decision. In the fingerprint domain, some studies have already examined this question with somewhat mixed results.[63] We propose similar research in other disciplines.

AL., *supra* note 47, at 2; Kellman et al., *supra* note 48; Tangen et al., *supra* note 56.

[60] Jonathan J. Koehler, Forensics or Fauxrensics? *Testing for Accuracy in the Forensic Sciences*, 49 ARIZ. ST. L.J.. (forthcoming January 2018) (on file with author), http://ssrn.com/abstract=2773255 (criticizing recent fingerprint error rate studies for using volunteer participants, a non-blind test format, and for being conducted by researchers who may have a career stake in demonstrating very low rates of error).

[61] One important reason rigorous testing is needed across the forensic sciences is that many people are naively optimistic about the risk of forensic science error. *See* Jonathan Koehler, *Intuitive Error Rate Estimates for the Forensic Sciences*, 57 JURIMETRICS J. 7 (forthcoming Winter 2017), http://ssrn.com/abstract=2817443 (providing evidence from an online study that shows mock jurors estimate the false positive error rate for various forensic sciences to be on the order of 1 in 1,000,000).

[62] Jonathan J. Koehler, *Proficiency Tests to Estimate Error Rates in the Forensic Sciences*, 12 LAW, PROBABILITY & RISK 89, 93–94 (2013).

[63] *Compare* Kellman et al., *supra* note 48, at 7 (finding correlation between average accuracy and self-reported difficulty of comparison) *and* Matthew B. Thompson et al., *Humans Matching Fingerprints: Sequence and Size*, 54 PROC. OF THE HUM. FACTORS & ERGONOMICS SOC'Y ANN. MEETING 478, 480–81 (2010) (amount of visible area in a target

Following Kellman's work, a study could begin by generating a set of questioned samples of varying difficulties. Kellman generated a varied set of questioned fingerprint samples by instructing participants to touch surfaces in different ways.[64] Those samples were then recovered and evaluated for difficulty, but only by the examiner-participants who were being tested.[65] Alternatively, the samples could be graded for difficulty in advance by independent "graders," and investigators could then provide examiners with samples that vary in terms of difficulty level.[66] If difficulty served as a within-subjects independent variable, researchers might be able to determine the point at which difficulty begins to affect examiner accuracy and by how much.

Study 5: Applying signal detection theory: Can examiners' decision thresholds be shifted?

Many forensic domains center around binary classifications in which an examiner decides whether a questioned and known sample share a common source. Holding aside "inconclusives," each match / no match decision falls into one of four categories: (1) true positive (i.e., a correct match decision), (2) true negative (a correct nonmatch decision), (3) false positive (an incorrect match decision), or (4) false negative (an incorrect nonmatch decision). Two measures are commonly used to describe performance in such a task: (1) *d prime* (the extent to which an observer correctly identifies a distinctive stimulus), and (2) *Beta* (the threshold for making a decision as to whether the stimulus is present).[67] Even if two examiners have the same discriminability to detect matches and nonmatches, one examiner might have a strict criterion before calling a match, whereas a second examiner might have a modest criterion. The beta value for the individual examiner represents this threshold. The beta value is particularly important in forensic decision making for at least two reasons: (1) in legal contexts where a normative decision has been made to

print is positively correlated with classification accuracy), *with* PACHECO ET AL., *supra* note 47, at 57 (finding effects of difficulty on the number of inconclusive determinations made by examiners, but no effects on accuracy when a determination is made).

[64] Kellman et al., *supra* note 48, at 6.

[65] *Id.* at 10.

[66] *See generally* Drew P. Pulsifer et al., *An Objective Fingerprint Quality-Grading System*, 231 FORENSIC SCI. INT'L 204 (2013) (attempting to provide an objective system or grading of difficulty).

[67] David A. Balota & Elizabeth J. March, *Cognitive Psychology: An Overview*, *in* COGNITIVE PSYCHOLOGY: KEY READINGS 1, 3–4 (David A. Balota & Elizabeth J. March eds., 2004); Hal R. Arkes & Barbara A. Mellers, *Do Juries Meet Our Expectations?*, 26 LAW & HUM. BEHAV., 625, 628–30 (2002).

place a heavy burden on one party (e.g., the beyond-a-reasonable-doubt standard for a criminal conviction), we may want to ensure that a similarly high threshold is used in assessing potentially very powerful evidence that could negate that burden; and (2) we want to ensure consistency in examiners such that there are not some who trend toward a high true positive rate at the risk of a high false positive rate or a high true negative rate at the risk of a high false negative rate.

We suggest a study with two goals in mind: (1) determining whether and when there is variability among examiners in their beta thresholds,[68] and (2) determining whether those thresholds can be modified to push examiners toward consistency (or toward a threshold that is set from a policy perspective). Much like Study 4, the simplest study would involve presenting forensic examiners with a set of samples and asking them to make match / no match decisions. In order to have sufficient statistical power to identify significant but small differences in beta thresholds, we recommend a larger sample of both examiners and forensic stimuli in this study than in others. For each examiner, the set of stimuli must be extensive enough to generate outcomes in each of the four categories described above. After gathering data, a researcher could determine the extent of differences in beta thresholds among examiners. If there are differences, a follow-up study could attempt to push outlier examiners toward a more moderate threshold by either (1) instructing them to change their thresholds, or (2) providing trial by trial outcome feedback to outlier examiners about the conclusions reached by other examiners who have more desirable beta values.

Study 6: Does examiner confidence correlate with accuracy?

A question related to whether forensic comparison difficulty affects accuracy is whether examiners' confidence in their conclusions correlates with their accuracy. Substantial research in psychology indicates a weak or even nonexistent relationship between accuracy and confidence in a variety of tasks (including eyewitness testimony).[69] However, some research in the fingerprint domain has found the correlation between confidence and accuracy to be at least as high as that between difficulty and accuracy.[70] This link is important in forensic science for several reasons. First, if

---

[68] *See* Ulery et al., *supra* note 59, at 7737 (making some progress on this front, though it did not include extensive discussion of the nature of examiner variability in beta threshold).

[69] *See, e.g.*, Sigfried Ludwig Sporer et al., *Choosing, Confidence, and Accuracy: A Meta-analysis of the Confidence-accuracy Relation in Eyewitness Identification Studies*, 118 PSYCHOL. BULL. 315, 315 (1995).

[70] Kellman et al., *supra* note 48, at 7.

examiner confidence is a reliable indicator of accuracy, examiners may be able to assist judges and factfinders by providing confidence estimates for their conclusions about the similarity between samples. Second, because jurors rely, in part, on experts' confidence when assessing the credibility of their testimony,[71] it would be helpful for jurors to have a more objective indicator of experts' confidence rather than try to infer it from body language or other peripheral cues.

We propose a study that uses methods similar to those proposed in Study 4 (which examines the difficulty-accuracy relationship) to examine the confidence-accuracy relationship. Confidence ratings could be made on a multipoint Likert-type scale,[72] and the relationship with accuracy should be measured separately for match / no match conclusions, as discussed above with regard to signal detection theory. In domains where pilot testing indicates that agreement among examiners is likely to be high, the study will need to include a relatively larger number of stimuli or examiners. By way of a sports analogy, if we wish to test whether some outside influence (e.g., noise) affects the ability of professional basketball players to make layups or some similarly easy shot, we would need to test many players shooting many layups to obtain a sufficient quantity of missed layups in the analysis.

Study 7: Does the use of a computer database affect match report accuracy?

Over the past several decades, some forensic disciplines have begun to develop databases of known forensic samples for future comparison with questioned samples. The most well-known database is the Automated Fingerprint Identification System (AFIS), a database that includes millions of fingerprints.[73] As the forensic subfields continue to develop databases

---

[71] Robert J. Cramer et al., *Expert Witness Confidence and Juror Personality: Their Impact on Credibility and Persuasion in the Courtroom*, 37 J. AM. ACAD. PSYCHIATRY L. 63, 68–69 (2009).

[72] A Likert-type scale is a response scale that is commonly used on surveys and questionnaires. A typical scale item provides study participants with a statement (e.g., "I like chocolate ice cream") and asks participants to indicate their degree of agreement or disagreement with the statement using a numbered scale, which often ranges from one (strongly disagree) to seven (strongly agree). The goal is to allow the researcher to measure the intensity of a participant's feelings about a particular matter. For further detail, see Michael S. Matell & Jacob Jacoby, *Is There an Optimal Number of Alternatives for Likert Scale Items?*, 31 EDUC. & PSYCHOL. MEASUREMENT 657 (1971).

[73] *See* NAT'L INST. OF JUST., FINGERPRINTS: AN OVERVIEW n.1 (July 1, 2016), http://www.nij.gov/topics/forensics/evidence/impression/pages/fingerprints.aspx#note1 (showing that other forensic disciplines maintain databases as well).

that store prints and markings, a question arises as to whether the risk of examiner error increases as the databases become larger. As others have observed, searches through such a large database may uncover prints from different fingers that nonetheless appear similar to the questioned (latent) sample:

> Ironically, the practical importance of understanding when and why fingerprint comparison errors occur is likely to increase as technology advances. It is common for a latent print to be submitted to an AFIS . . . database, where automated routines return a number of most likely potential matches. Error rates (especially of the false-positive type) may increase as databases get larger (currently some databases include tens of millions of prints). The reason for this is that as a database grows, an AFIS searching that database is increasingly likely to find close non-matches.[74]

The notions that use of a database entails special error considerations,[75] and that the risk of error may increase as the size of the database increases, are empirical questions.[76]

One method for studying this issue could be similar to that of Study 4, replacing the sample difficulty independent variable in that study with a use of computerized database (i.e., yes, no) independent variable. In the fingerprint context, the stimulus set could include an even split of (1) known samples selected from AFIS based on similarity to a corresponding questioned print, and (2) known samples that are similar to both one another and the corresponding questioned print, but which are not selected from any database. If results indicate that the use of a database does affect accuracy at the individual-examiner level, a follow-up study could examine the effect of the size of the database. For example, the stimulus set could include an even split of (1) known samples selected from the entire AFIS database based on similarity to a corresponding questioned print, and (2) known samples selected from only half (or some other subset) of AFIS. One might expect that for larger databases, the similarity of the selected known samples to the questioned sample would also increase, thereby increasing

---

[74] Kellman et al., *supra* note 48, at 2; *see also* Itiel E. Dror & Jennifer L. Mnookin, *The Use of Technology in Human Expert Domains: Challenges and Risks Arising from the Use of Automated Fingerprint Identification Systems in Forensic Science*, 9 LAW, PROBABILITY & RISK 47, 57 (2010).

[75] *See* Itiel E. Dror et al., *The Impact of Human-technology, Cooperation and Distributed Cognition in Forensic Science: Biasing Effects of AFIS Contextual Information on Human Experts*, 57 J. FORENSIC SCI. 343, 343–44 (2012) (showing there is evidence that the judgments of fingerprint examiners who use databases are affected by the position of the matching print in the AFIS list).

[76] *See generally id.* at 343 (recently examining other potential issues with AFIS, notably the biasing impact of the order in which AFIS returns potential matches, but did not examine the variables suggested here).

the chance of false positive errors.[77]

The policy implications of this study may be difficult to navigate. Even if the use of large databases increases the risk of false positive error in some domains (because the database has returned close non-matches), it may decrease the risk of false negative errors (because in a large database, the best match is more likely to be a true match due to the greater inclusiveness of the database). If true, then policy makers will need to assess whether the decreased risk of one type of error associated with large databases provides adequate compensation for the increased risk of the other type of error.

Study 8: How many points of similarity should examiners use?

Comparison between known and questioned forensic samples typically involves examination of noteworthy class and individuating features in the samples.[78] However, none of the traditional forensic sciences provide clear guidance to examiners about the number of matching features that must be examined prior to reaching a conclusion about who or what made the print or marking in question.[79] One might assume that increasing the signal by increasing the number of minutia points observed would decrease the risk of false positive errors, but it would likely also increase the risk of false negative errors. Further, there may be a point where increases in the number of signals provide little or no gain. Ultimately, these are empirical questions.

We propose a study in which the number of points of similarity used in a questioned sample analysis is manipulated as a within-subjects independent variable. The methods would largely mirror those of Study 4. Forensic examiners would be assigned a number of sample pairs and would make match / no match decisions on each. However, the number of points of similarity to be used—which an examiner would typically determine on a case-by-case basis—would be assigned randomly on each trial. This variable could then be examined for effects, if any, on error rates (both false

---

[77] Computer searches generally provide a rank-ordered set of candidate matches from the database (e.g., the closest twenty prints). The examiner then makes a series of pairwise comparisons between the questioned print and the known prints until the examiner is satisfied that there is or is not a match. We are suggesting that with more prints in the database, the non-matching candidates that the computer search generates may be more similar to the questioned print than the non-matching prints that would be generated if the database were smaller. This increased similarity between the questioned print and the non-matching computer generated prints may increase the risk of a false positive error.

[78] *See supra* notes 44–48.

[79] NAS Report, *supra* note 12, at 141.

positive and false negatives).[80]

DESCRIPTIVE STUDIES: EFFECTS OF BIASING INFORMATION AND METHODS ON ACCURACY

Since the early 2000s, there has been a growing awareness of the role cognitive bias may play in forensic science.[81] Much of the relevant research has focused on whether and how examiners may be influenced by information that is unrelated to the actual process of making a scientific assessment.[82] Some forensic scientists appear to be influenced by extraneous information, particularly when the information is itself highly probative of a material issue. For example, when fingerprint examiners are aware that a suspect has confessed or that other examiners have reported a match, they appear to be more likely to report a match.[83] But what about

---

[80] *See* Cedric Neumann et al., *Computation of Likelihood Ratios in Fingerprint Identification for Configurations of Any Number of Minutiae*, 52 J. FORENSIC SCI. 524 (2007) (examining questions along these lines in the fingerprint domain; this study found evidence for increased discriminability up to twelve points of similarity, though there were decreasing returns when approaching that number); *see also* NAS Report, *supra* note 12, at 61–63.

[81] *See* D. Michael Risinger et al., *The* Daubert/Kumho *Implications of Observer Effects in Forensic Science: Hidden Problems of Expectation and Suggestion*, 90 CALIF. L. REV. 1 (2002); *see also* William C. Thompson, *Interpretation: Observer Effects*, *in* WILEY ENCYCLOPEDIA OF FORENSIC SCIENCE 1575 (2009).

[82] *See generally* Gary Edmond et al., *Contextual Bias and Cross-Contamination in the Forensic Sciences: The Corrosive Implications for Investigations, Plea Bargains, Trials and Appeals Examination Casework*, 14 LAW, PROBABILITY & RISK 1, 1 (2015) (explaining "that lawyers and courts have not recognized how contextual bias and cognitive processes may distort and undermine the probative value of expert evidence"); Bryan Found & John Ganas, *The Management of Domain Irrelevant Context Information in Forensic Handwriting Examination Casework*, 53 SCI. & JUST. 154 (2013) (describing a procedure to reduce the risk that potentially biasing, domain irrelevant information, reaches a handwriting examiner); Sherry Nakhaeizadeh et al., *The Power Of Contextual Effects In Forensic Anthropology: A Study of Biasability In The Visual Interpretations of Trauma Analysis on Skeletal Remains*, 59 J. FORENSIC SCI. 1177 (2014); Nikola K. P. Osborne et al., *Does Contextual Information Bias Bitemark Comparisons?*, 54 SCI. & JUST. 267, 272 (2014) (reporting the results of a study showing that "bitemark comparisons – whether they are made by people with or without dental experience – are susceptible to contextual influences"); Mark Page et al., *Context Effects And Observer Bias: Implications for Forensic Odontology*, 57 J. FORENSIC SCI. 108 (2012).

[83] *See, e.g.*, Itiel E. Dror et al., *Contextual Information Renders Experts Vulnerable to Making Erroneous Identifications*, 156 FORENSIC SCI. INT'L 74 (2006); Saul M. Kassin et al., *The Forensic Confirmation Bias: Problems, Perspectives, and Proposed Solutions*, 2 J. APPLIED RES. MEMORY & COGNITION 42, 50 (2013) ("judges and juries need to know that forensic science conclusions that appear to corroborate a confession or eyewitness identification may, in fact, have been influenced by these previously collected forms of evidence"); Glenn Langenburg et al., *Testing for Potential Contextual Bias Effects During the Verification Stage of the ACE-V Methodology When Conducting Fingerprint*

subtler forms of contextual bias that provide less probative information? Do the questions that examiners are asked to answer, or the order in which they're asked to answer those questions affect forensic judgment? We suggest four additional studies (Studies 9–12) below related to this issue.

Study 9: Does biasing information interact with the questions examiners are asked to answer?

Psychologists have long known that the way questions are asked can exert large effects on the answers that people provide in legal and non-legal contexts.[84] Less clear is whether the way questions are asked of trained professional forensic examiners affect the conclusions that they reach. Sometimes forensic examiners are asked by investigators or attorneys to identify the source of evidentiary material. Whose DNA is it? Which carpet did that fiber come from? Other times forensic examiners are asked whether the DNA *could have come from* the suspect, or whether that carpet fiber is consistent with the carpet in the victim's car. And still other times forensic examiners simply identify similarities or dissimilarities between two forensic samples without offering conclusions or opinions about whether the items could or did come from a common source. In this study, we seek to examine *the interaction* between these various question types and the presence or absence of contextually biasing information. In other words, are any of these questions more or less likely to lead to an erroneous response when combined with extraneous contextual information?

When an examiner is asked to *identify the source* of a questioned sample, it is arguably reasonable for the examiner to take into account any and all information he or she may know about the case. Indeed, from a Bayesian standpoint,[85] both forensic and non-forensic information are

Comparisons, 54 J. FORENSIC SCI. 571 (2009).

[84] *See* ELIZABETH F. LOFTUS, EYEWITNESS TESTIMONY 96 (1979) (describing, among other things, an experiment that showed that people judged the speed of cars that were involved in a videotaped accident differently depending on whether the cars were said to have smashed, collided, bumped or contacted).

[85] DAVID H. KAYE ET AL., THE NEW WIGMORE: EXPERT EVIDENCE, § 14.3.1 pp. 639–42 (2d ed. 2011) (Bayes' Theorem is a mathematical tool that tells decision makers how they should update their probabilistic beliefs about a hypothesis in the face of new evidence. Suppose a decision maker wanted to assess the probability that a suspect is the source of a questioned fingerprint in light of evidence from a forensic examiner that the suspect's print matches the questioned print. In this situation, Bayes counsels the decision maker to (1) identify a prior probability that the suspect is the source of the print (i.e., prior to the introduction of the evidence of the forensic match), (2) identify the strength of the evidence associated with the reported match, and (3) combine the prior probability with the evidence strength to form a "posterior probability" that the suspect is the source of the questioned print.).

required to answer the question.[86] However, if the question is restricted to *identifying similarities or dissimilarities* between the questioned and known samples, non-forensic case information is less relevant and potentially biasing. A simple version of such a restricted question might be, "how similar are the two fingerprints to one another?" More formally, examiners might be asked to estimate a likelihood ratio that corresponds to the evidentiary strength of their observations (provided that sufficient data for such estimates exist). If this were done, then the trier of fact could combine the forensic evidence with the non-forensic case evidence to form judgments about the source and guilt likelihoods for themselves. In this manner, forensic judgments would retain their independence and avoid being double counted.[87] We suggest a basic experiment to test whether the specific questions posed to examiners interacts with potentially biasing contextual information on the answers they provide.

We recommend a 2 x 2 between-subjects design, crossing question type (item similarity vs. source) and biasing information (present vs. absent).[88] A sample of forensic examiners would be asked to make comparisons between pairs of forensic samples. Half of the examiners would be asked a source question (e.g., "was the questioned sample derived from the same source as the known sample?"), and half would be asked a more restricted question that does not invite the use of contextually biasing information (e.g., "on a scale of 1–7, how similar are the two samples?"). Completing the full cross, half of the examiners would be presented with biasing information, while half would complete the comparisons without

---

[86] *Id.*; *see also* Jonathan J. Koehler & Daniel Shaviro, *Veridical Verdicts: Increasing Verdict Accuracy Through the Use of Overtly Probabilistic Evidence Methods*, 75 CORNELL L. REV. 247, 255–56 (1990) (providing a detailed account of how Bayes theorem may be used to combine different items of evidence in a legal context).

[87] *But see* Jonathan J. Koehler, *The Influence of Prior Beliefs on Scientific Judgments of Evidence Quality*, 56 ORG. BEHAV. & HUM. DECISION PROCESSES 28, 47–48 (1993) (arguing that double counting may actually be normatively appropriate within the Bayesian framework).

[88] A 2 x 2 between-subjects design is one in which there are two manipulated independent variables (in this case, question type and biasing information) and two levels of each of those independent variables (e.g., the feature is absent or present). In this manner, each study participant would receive exactly one of four possible stimuli. To put this experimental design in a more intuitive context, an experimenter interested in studying the impact of the color (red and green) and size (small or large) of a tip jar on people's willingness to tip in coffee shops might assign customers to view either a small red jar, a small green jar, a large red jar, or a large green jar. If there were 100 people in the study, it would be common for each person to be assigned, at random, to view one of the four tip jars such that, in the end, about twenty-five people viewed each of the four jars.

biasing information.[89]

Study 10: Does the presence of multiple samples or the order in which samples are examined bias conclusions?

Examiners likely use a broad range of sample-present and sample-absent features when comparing questioned and known samples. Reliance on a broad set of features could, however, introduce potential biases. For example, examiners' awareness of specific features in questioned or known samples could affect the way they allocate their subsequent attention, their visual searches, and even their thresholds for determining a match.[90] Here, we focus on two particular potential effects: (1) effects of the mere availability of a known sample that is available to be directly and repeatedly compared with a questioned sample, and (2) the order in which the two samples are examined.[91]

We suggest a between-subjects experiment with four groups. Each group would contain an equal number of forensic examiners who would conduct analyses on a set of questioned samples. The first group of examiners would begin by examining a questioned sample for information. Once completed, these examiners would examine a known sample, and then draw conclusions. The second group would initially examine only the known sample for information, before examining the questioned sample and drawing conclusions. A third group would rely on the examination

---

[89] One might also consider varying the type of biasing information presented: for example, experimenters might indicate to the examiner the race or gender of the suspect, or tell the examiner that incriminating evidence was found in the suspect's possession.

[90] *See* Itiel E. Dror et al., *Cognitive Issues in Fingerprint Analysis: Inter- and Intra-Expert Consistency and the Effect of a "Target" Comparison*, 208 FORENSIC SCI. INT'L 10 (2011); *see also* Itiel E. Dror & Simon A. Cole, *The Vision In "Blind" Justice: Expert Perception, Judgment, and Visual Cognition in Forensic Pattern Recognition*, 17 PSYCHONOMIC BULL. & REV. 161, 162 (2010).

[91] *See, e.g.*, Gary L. Wells et al., *Eyewitness Identification: Bayesian Information Gain, Base-Rate Effect-Equivalency Curves, and Reasonable Suspicion*, 39 LAW & HUM. BEHAV. 99, 99–100 (2015) (extensively documenting the ways in which the structure of a lineup can affect decision outcomes; relatedly, while forensic comparisons are typically pairwise (between a single known sample and a single questioned sample), such an approach may introduce bias if the examiner, for whatever reason, has an *a priori* belief that the suspect is guilty or is likely to be the source of the unknown); *see also* Larry S. Miller, *Procedural Bias in Forensic Science Examinations of Human Hair*, 11 LAW & HUM. BEHAV. 157 (1987) (hinting that the lineup approach could reduce examiner errors, though the sample size in that study is small and the participant population was relatively inexperienced). Study 10 (or a follow-up study) could be expanded to examine the effects of the inclusion of multiple known samples to compare with a single questioned sample, much the way an eyewitness seeks to compare a single mental image of an unknown individual viewed at a crime scene with multiple knowns in the context of a suspect lineup at the police station.

results of the first group for the questioned sample features, and the results of the second group for the known sample features. A fourth group would have access to both samples for the duration of each examination. Various dependent variables could be examined including the rates of accuracy and inconclusives in each group.

Study 11: Are examiners affected by knowledge of a forthcoming review?

Most forensic disciplines do not routinely include a "verifier" to check the work of an examiner. Although verification imposes extra costs, it would seem to be an important part of a purportedly scientific enterprise. Fingerprint analysis appears to be the only discipline that routinely employs a verifier.[92] Though there is some evidence that verification is an effective method for catching false positive errors,[93] examiners may behave differently and reach different conclusions if they know or suspect that a review of their work is forthcoming. Potential effects could be either positive or negative. On the one hand, examiners' overall accuracy rates might increase if they are motivated to be more thorough and careful when assessing the information in the samples, knowing the potential for reversal. On the other hand, examiners might exhibit a sort of social loafing,[94] knowing that another examiner will be there to correct any errors. Forensic researcher Glenn Langenburg has noted a similar problem related to verification that he refers to as a "bias loop," in which examiners who know that their work will be checked are affected by that knowledge, as are the verifiers who know that they are merely verifying the work of another examiner who has presumably studied the matter carefully.[95] Further, knowledge of a forthcoming review might not affect false positive and false negative rates equally. If examiners believe that there is greater professional or societal harm associated with committing a false positive error relative to a false negative error (for example, because of the potential harm to an innocent victim; a result that the justice system is designed to

---

[92] *See, e.g.*, Langenburg, *supra* note 48, at 219–20 (stating that verification is the norm in fingerprint analysis: The "V" in the ACE-V fingerprint procedure stands for "verification"); PACHECO ET AL., *supra* note 47, at 7–8; *see also* NAS Report, *supra* note 12, at 64 (finding through an internal survey, 69% of fingerprint units reported having some system for verifying results).

[93] *See, e.g.*, PACHECO ET AL., *supra* note 47, at 20; Ulery et al., *supra* note 59, at 7737–38.

[94] *See, e.g.*, Steven J. Karau & Kipling D. Williams, *Social Loafing: A Meta-Analytic Review and Theoretical Integration*, 65 J. PERSONALITY & SOC. PSYCHOL. 681, 681 (1993) (describing social loafing as "the reduction in motivation and effort when individuals work collectively compared with when they work individually or coactively").

[95] Langenburg, *supra* note 48, at 242.

avoid), they might alter their usual thresholds for making match decisions, lowering their false positive error rate but increasing their false negative error rate.

Last, knowledge of forthcoming review might interact with the presence of contextually biasing information: examiners might be more aware of potential bias when a second examiner, who may have no knowledge of the biasing information, will be checking their work.[96]

We recommend a 2 x 2 between-subjects design similar to the one proposed in Study 9, crossing knowledge of a forthcoming review (present vs. absent) and biasing information (present vs. absent). A sample of forensic examiners would be asked to compare a set of forensic sample pairs. Half of the examiners would be told that a second examiner will review their work and make a separate determination, while half would be told that they are the only examiners reviewing each sample. Completing the full cross, half of the examiners would be presented with biasing information as described in Study 9.

Study 12: Can examiners be debiased?

As research demonstrating the pernicious effects of contextually biasing information on forensic examiners continues to appear, attention has begun to shift toward ways to reduce these biases.[97] The psychological literature suggests that even where the introduction of contextually biasing information cannot be avoided, there may be ways to mitigate the biasing effect on the forensic examiner herself. For example, providing examiners with a general education on the way bias can influence judgments might be useful and has been suggested as an important tool.[98] Relatedly, research in a number of domains suggests that requiring decision makers to consider various alternative hypotheses and explanations might also be a useful way to reduce overconfidence and debias judgment.[99] Importing this idea into

---

[96] *See* Roger Koppl, *How to Improve Forensic Science*, 20 EUR. J. L. & ECON. 255, 256 (2005) (proposing model in which forensic examiners compete with each other to improve performance); Philip E. Tetlock, *Accountability and the Perseverance of First Impressions*, 46 SOC. PSYCHOL. Q. 285 (1983).

[97] *See* Itiel E. Dror, *Practical Solutions to Cognitive and Human Factor Challenges in Forensic Science*, 4 FORENSIC SCI. POL'Y & MGMT. 1, 5–6 (2013); Kassin et al., *supra* note 83, at 49–50; Elizabeth J. Reese, *Techniques for Mitigating Cognitive Biases in Fingerprint Identification*, 59 UCLA L. REV. 1252, 1280–88 (2012); Dan E. Krane et al., *Sequential Unmasking: A Means of Minimizing Observer Effects in Forensic DNA Interpretation*, 53 J. FORENSIC SCI. 1006 (2008).

[98] Dror, *supra* note 97, at 5.

[99] *See, e.g.*, Edward R. Hirt & Keith D. Markman, *Multiple Explanation: A Consider-an-Alternative Strategy for Debiasing Judgments*, 69 J. PERSONALITY & SOC. PSYCHOL. 1069

the forensic arena, it may be that requiring examiners to consider specific alternative explanations to their beliefs about whether a pair of samples does or does not share a common source will lead to less biased judgments. We suggest a study to measure the effects of these two debiasing strategies.

This study could simply compare five groups of examiners who are asked to make match / no match decisions on a set of questioned forensic samples in a given domain in the presence of contextually biasing information. Group 1 would not be exposed to any sort of debiasing procedure. Group 2 would receive generalized training on the dangers of cognitive bias in decision making prior to evaluating the sample pairs. Group 3 would be required to explain what her hypothesis is (both at a broad match / no match level and at a narrower minutiae-point level) and identify potential alternative explanations for each sample pair. Group 4 would go through both debiasing procedures (education plus identify alternative explanations). Group 5, a control group, would not be exposed to contextually biasing information and would not go through any debiasing procedure.

DESCRIPTIVE STUDIES: HOW EXAMINERS REPORT THEIR RESULTS AND HOW JUDICIAL ACTORS INTERPRET THEM

Our discussion so far has focused on the examiners' methods in collecting data and drawing conclusions regarding forensic samples—what could be called the *input* process for generating forensic data. In order for those data to impact the legal process, they must be *output* to a third party, such as a judge acting as an evidentiary gatekeeper or a jury acting as a factfinder. Thus, the current problems with forensic science cannot be solved simply by addressing issues with inputs on the front end (e.g., taking steps to increase consistency or reduce bias). Instead, we must also consider what can be done on the back end to ensure that judges, jurors and others give forensic science its proper weight. Empirical questions arise as to how this back end output process does and should occur. We propose two preliminary studies (Studies 13 and 14) below.

(1995); Lutz Kaufmann et al., *Debiasing the Supplier Selection Decision: A Taxonomy and Conceptualization*, 40 INT'L J. PHYSICAL DISTRIBUTION & LOGISTICS MGMT. 792 (2010); Derek J. Koehler, *Explanation, Imagination, and Confidence in Judgment*, 110 PSYCHOL. BULL. 499, 500 (1991); Charles Lord et al., *Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence*, 37 J. PERSONALITY& SOC. PSYCHOL. 2098 (1979); Charles G. Lord et al., *Considering the Opposite: A Corrective Strategy for Social Judgment*, 47 J. PERSONALITY & SOC. PSYCHOL. 1231, 1233 (1984). *But see* Hal R. Arkes, *Impediments to Accurate Clinical Judgment and Possible Ways to Minimize Their Impact*, 49 J. CONSULTING & CLINICAL PSYCHOL. 323, 326 (1981) (arguing against the use of generalized education regarding bias as a debiasing method).

Study 13: How do forensic examiners actually testify in court?

Forensic scientists have been criticized for exaggerating the strength of the evidence they report.[100] For example, forensic scientists have often testified that a person or an object is the source of an evidentiary item "to a reasonable degree of scientific certainty," or words to that effect.[101] But as a draft report from the National Commission on Forensic Science recently pointed out, this common phrase has "no scientific meaning and may mislead factfinders about the level of objectivity involved in the analysis, its scientific reliability and limitations, and the ability of the analysis to reach a conclusion."[102] This rebuke raises the question of how often such misleading terminology is actually used and, more generally, what types of language examiners use to convey their findings. To our knowledge, the issue has not been studied in any systematic way. How often do examiners make source statements? How do they phrase those source statements? How do they explain their level of certainty or the possibility of an error? How do they explain potential inconclusive decisions? How much variability is there on this matter across jurisdictions, laboratories, and individual examiners working in the same laboratory? A systematic examination of these questions would provide a helpful starting point for any type of reform in this area.

We suggest beginning with an archival study of a random sample of trial transcripts that seeks to classify the various types of testimony provided by forensic scientists. One classification could be as simple as whether the testimony did or did not include exaggerated, misleading, or false scientific claims. Other classifications could consider whether the testimony made direct claims about the source of a sample (as opposed to claims about the similarity between samples), or whether the testimony included "weighting guides" for the factfinder regarding the examiner's confidence or the strength of similarity between samples. Such transcripts will not be easy to obtain, as they are typically proprietary and not readily

---

[100] *See, e.g.*, Alex Biedermann et al., *The Subjectivist Interpretation of Probability and the Problem of Individualisation in Forensic Science*, 53 SCI. & JUST. 192 (2013); Christophe Champod, *Fingerprint Examination: Towards More Transparency*, 7 LAW PROBABILITY & RISK 111 (2008); Jonathan J. Koehler, *Error and Exaggeration in the Presentation of DNA Evidence*, 34 JURIMETRICS J. 21, 22 (1993); C. Neumann et al., *Quantifying the Weight of Evidence from a Forensic Fingerprint Comparison: A New Paradigm*, 175 J. ROYAL STAT. SOC'Y 371 (2012); Mark Page et al., *Uniqueness in the Forensic Identification Sciences—Fact or Fiction?*, 206 FORENSIC SCI. INT'L 12 (2011); Michael J. Saks & Jonathan J. Koehler, *The Individualization Fallacy in Forensic Science Evidence*, 61 VAND. L. REV. 199, 200 (2008).

[101] NAT'L INSTITUTE OF STANDARDS AND TECHNOLOGY, *supra* note 54, at 118–20.

[102] NAT'L COMM'N ON FORENSIC SCI., *supra* note 31.

available from online databases.[103] Even if a non-random sample of transcripts could be cobbled together, it would still be useful to review the language used by testifying forensic scientists under direct and cross examination as a way to generate testable research hypotheses about expert testimony.

Study 14: How should examiners present evidence in court?

After learning more about how forensic scientists in the various domains present their evidence in court, we should turn our attention to the back-end process of how *consumers* of forensic science evidence respond to that evidence. Over the past thirty years, psychologists have conducted many controlled experiments that examine how people process forensic science evidence.[104] Much of this research suggests that people may not weigh forensic science evidence appropriately,[105] or that they may be influenced by the way in which the forensic science statistics are presented.[106] At this point, the field should focus on developing

---

[103] In light of the importance of trial transcripts for examining all aspects of expert testimony at trial, we hope that some researchers or agency will take steps to create a database of forensic trial transcripts.

[104] David H. Kaye & Jonathan J. Koehler, *Can Jurors Understand Probabilistic Evidence?*, 154 J. ROYAL STAT. SOC'Y 75 (1991) (an early review of the probabilistic studies); Jonathan J. Koehler, *On Conveying the Probative Value of DNA Evidence: Frequencies, Likelihood Ratios, and Error Rates*, 67 U. COLO. L. REV. 859 (1996); Jonathan J. Koehler, *If* the *Shoe Fits, They Might* Acquit: The Value of Forensic Science Testimony, 8 J. EMPIRICAL LEGAL STUD. 21 (2011); Samuel Lindsey et al., *Communicating Statistical DNA Evidence*, 43 JURIMETRICS J. 147 (2003).

[105] Dale A. Nance & Scott B. Morris, *Juror Understanding of DNA Evidence: An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Small Random-Match Probability*, 34 J. LEGAL STUD. 395, 418–36 (2005) (reporting that people undervalue DNA match testimony); Jason Schklar & Shari Diamond, *Juror Reactions to DNA Evidence: Errors and Expectancies*, 23 LAW & HUM. BEHAV. 159, 176 (1999) (reporting that DNA study participants "misaggregated the probabilistic evidence with their prior probability of guilt estimates"); Nicholas Scurich, *The Differential Effect of Numeracy and Anecdotes on the Perceived Fallibility of Forensic Science*, 22 PSYCHIATRY, PSYCHOL. & L. 616, 616 (2015) (reporting that "innumerate" participants based their valuations of DNA evidence on anecdotal information rather than scientifically derived error rate information); William C. Thompson & Eryn J. Newman, *Lay Understanding of Forensic Statistics: Evaluation of Random Match Probabilities, Likelihood Ratios, and Verbal Equivalents*, 39 LAW & HUM. BEHAV. 332, 343 (2015) (reporting that study participants undervalued shoeprint evidence relative to Bayesian norms); William C. Thompson et al., *Do Jurors Give Appropriate Weight to Forensic Identification Evidence?*, 10 J. EMPIRICAL LEGAL STUD. 359, 359 (2013) (reporting that DNA study participants made judgments "consistent with Bayesian expectations, although people overvalued the DNA evidence when the probability of a false report of a match was high relative to the random match probability").

[106] Jonathan J. Koehler, *When Are People Persuaded by DNA Match Statistics?*, 25 LAW

presentation protocols that are designed to maximize jurors' understanding of the probative value of the evidence, reduce the risk of probabilistic fallacies (such as the source probability error[107] or the prosecutor's fallacy[108]), and move jurors' relevant beliefs in normatively appropriate amounts.

A study that addresses these issues will necessarily have rather low ecological validity[109] because actual cases rarely translate directly into a normative scenario. An actual criminal case will typically include so many different considerations (e.g., eyewitness testimony, evidence that impeaches that testimony, evidence about motives and opportunity, alibi evidence, etc.) that it would be impossible to determine the exact amount by which evidence of a forensic science "match" should influence a juror's judgment about a defendant's guilt or innocence.

Still, a number of researchers have constructed artificial legal scenarios that permit comparison of jurors' judgments with Bayesian norms. For example, Professor William Thompson and his colleagues have offered a useful paradigm for eliciting probabilistic estimates from mock jurors in simple cases involving forensic science evidence.[110] In this Bayesian paradigm, the relevant information is constrained in ways that allow the researcher to determine whether mock jurors are undervaluing or overvaluing forensic evidence.[111] Informed by the findings of Study 13 regarding the different ways that forensic scientists present their findings, future studies could use Professor Thompson's approach to pit each of a

---

& HUM. BEHAV. 493, 508–10 (2001); Jonathan J. Koehler, *The Psychology of Numbers in the Courtroom: How to Make DNA-Match Statistics Seem Impressive or Insufficient*, 74 S. CAL. L. REV. 1275, 1277 (2001); Jonathan J. Koehler & Laura Macchi, *Thinking About Low-Probability Events: An Exemplar-Cuing Theory*, 15 PSYCHOL. SCI. 540 (2004) (finding that people were less persuaded by low probability DNA evidence when it was presented in an example-friendly way than when it was not).

[107] Koehler, *supra* note 100, at 22 (identifying the source probability as the error that occurs when one equates the random match probability (RMP)—the probability that a randomly selected person will match by coincidence—with the probability that a matching defendant is not the source of the forensic evidence); Thompson & Newman, *supra* note 105, at 335.

[108] William C. Thompson & Edward L. Schumann, *Interpretation of Statistical Evidence in Criminal Trials*, 11 LAW & HUM. BEHAV. 167, 171 (1987) (identifying the prosecutor's fallacy as the error that occurs when one equates the RMP with the probability that a matching defendant is not guilty).

[109] The ecological validity of a study refers to how well the experimental setting mimics real world settings of interest. *See* Jonathan J. Koehler & John B. Meixner, *Jury Simulation Goals*, *in* THE PSYCHOLOGY OF JURIES: CURRENT KNOWLEDGE AND A RESEARCH AGENDA FOR THE FUTURE (Margaret Bull Kovera, ed.) (forthcoming).

[110] Thompson & Newman, *supra* note 105; Thompson et al., *supra* note 105, at 360–62.

[111] Thompson & Newman, *supra* note 105, at 347.

series of different approaches against one another to determine which ones most reliably produce Bayesian-appropriate responses.

CONCLUSION

The movement to reform the forensic sciences is well under way. The National Commission on Forensic Science will likely take steps to reinforce the recommendations in the 2009 NAS report to eliminate some obviously unscientific forensic science practices. For example, we are approaching the end of exaggerated 100% certainty and 0% error rate claims. Testimony about having individualized a marking to its one and only source in the world to the exclusion of all others will likely also disappear. Some of the weaker subfields, such as hair microscopy and bite mark analysis seem destined to join comparative bullet lead analysis, voiceprint identification, and arson "indicators" in the forensic science trash heap.[112] There may be a push to take testimony only from certified forensic examiners who work in accredited laboratories. Procedural changes that reduce the risk of cognitive bias, such as the use of sequential unmasking and blind verifiers, may take hold. Similarly, there are efforts under way to increase the independence of laboratories from law enforcement to help reduce prosecutorial bias that some claim infects the forensic sciences.[113]

Such reforms are a good start, but they are not enough. The most important reform, in our view, is one that would imbue the entire forensic science enterprise with a research culture. This idea was developed most thoroughly in a *UCLA Law Review* article that was co-authored by a broad and diverse group of people who have written widely about the forensic sciences.[114] Adoption of a research culture entails a commitment to conducting, participating in, and relying upon high quality empirical research. Research is needed to address such fundamental issues as what

---

[112] Michael J. Saks & Ashley M. Votruba, "*. . . and the Courts Have Been Utterly Ineffective*," 54 JUDGES' J. 28 (2015) ("In recent years, a number of forensic science technique have been found to be so lacking in validity that they have been laid to rest . . . ."); *see also* ECKHOLM, *supra* note 10.

[113] ROGER KOPPL, REASON FOUND., CSI FOR REAL: HOW TO IMPROVE FORENSICS SCIENCE 6 (2007), http://reason.org/files/d834fab5860d5cf4b3949fecf86d3328.pdf ("About 80 percent of all U.S. crime labs are within law enforcement agencies, and approximately 90 percent of the accredited ones are organized under police agencies."); Simon A. Cole, *Response: Forensic Science Reform: Out of the Laboratory and into the Crime Scene*, 91 TEX. L. REV. 123, 130 (2013) ("Laboratory independence has long been perhaps the chief proposed reform among those American scholars who have been engaged in work calling for forensic reform.").

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

the various forensic methods can achieve, how reliably they can achieve them, and under what conditions.[115] Adoption of a research culture would also entail transparency, an "ongoing critical perspective," and a willingness to change and adapt that is different from how the forensic sciences have traditionally operated.[116]

Here we note that a call for empirical studies in the forensic sciences to assess their respective degrees of reliability[117] should not be confused with a claim that the forensic science methods used today are unreliable. Unproven is not the same as unreliable. As attorney and Harvard doctoral candidate Nathan J. Robinson recently wrote, "the problem with forensic science is not that it is wrong, but that it is hard to know when it is right."[118] The problem Robinson points out is compounded by the fact that people apparently believe, quite strongly and with little justification, that forensic science is hardly ever wrong. As is true in all areas of scientific evidence, the burden of demonstrating threshold reliability—and providing decision makers with scientific information about error and accuracy rates—rests with the evidentiary proponent. We hope that some of the studies we propose will help address this burden.[119]

---

[115] *Id.* at 740 (explaining that a research culture is one "in which the question of the relationship between research-based knowledge and laboratory practices is both foregrounded and central. We mean a culture in which the following questions are primary: What do we know? How do we know that? How sure are we about that? We mean a culture in which these questions are answered by reference to data, to published studies, and to publicly accessible materials, rather than primarily by reference to experience or craft knowledge, or simply assumed to be true because they have long been assumed to be true.").

[116] *Id.* at 743–44. The thoughtful historian of science Professor Simon Cole offers a somewhat different perspective on moving toward a scientific culture in forensic science. He says that the culture we should want has more to do with carefulness, documentation, and honesty than it does with traditional scientific values like testing hypotheses or adopting a skeptical mindset. *See* Simon A. Cole, *Acculturating Forensic Science: What Is "Scientific Culture," and How Can Forensic Science Adopt It?*, 38 FORDHAM URB. L.J. 435, 457 (2010) (In evidence collection, "the main concern is that we want people who are careful, meticulous, and honest.").

[117] PRESIDENT'S COUNCIL OF ADVISORS ON SCIENCE AND TECHNOLOGY, REPORT TO THE PRESIDENT, FORENSIC SCIENCE IN CRIMINAL COURTS: ENSURING SCIENTIFIC VALIDITY OF FEATURE-COMPARISON METHODS (September 2016) (calling for rigorous studies to assess the foundational validity and accuracy of many forensic sciences).

[118] Wesley Vernon et al., *Should We Trust Forensic Science?*, BOSTON REVIEW (Feb. 18, 2016), http://bostonreview.net/books-ideas/vernon-nirenberg-respond-robinson-forensic-pseudoscience; *see also* Nathan J. Robinson, *Forensic Pseudoscience*, BOSTON REVIEW (Nov. 16, 2015), http://bostonreview.net/books-ideas/nathan-robinson-forensic-pseudoscience-criminal-justice.

[119] The goal of much of the descriptive research we propose is not necessarily to improve the practice of forensic science, though this is a potential side benefit. A more central goal is to provide a set of empirical findings that will inform consumers of forensic

Finally, we join former U.S. District Court Judge Nancy Gertner in calling for the judiciary to join the forensic reform party by *requiring* greater participation from the forensic science community.[120] It is not enough for trial judges to hold occasional *Daubert* hearings to assess the reliability of proffered forensic science evidence if those judges continue to rely on the unsupported claims of forensic science supporters rather than the results of high quality empirical research conducted by disinterested scientists. As Judge Gertner opined in *United States v. Green*,[121] in the context of toolmark evidence, "[t]he more courts admit this type of toolmark evidence without requiring documentation, proficiency testing, or evidence of reliability, the more sloppy practices will endure; we should require more."[122] We agree, and hope that we have provided some constructive suggestions as to what more could and should be done by way of scientific testing.

science information about the value and limits of forensic science evidence.

[120] Nancy Gertner, *Commentary on the Need for a Research Culture in the Forensic Sciences*, 58 UCLA L. REV. 789, 790 (2011) ("Until courts address the deficiencies in the forensic sciences—until courts do what *Daubert* . . . requires that they do—there will be no meaningful change here.").

[121] 405 F. Supp. 2d 104 (D. Mass. 2005).

[122] *Green*, 405 F. Supp. 2d at 109; Judge Nancy Gertner (Ret), *Opinions I Should Have Written*, 110 NW. U. L. REV. 423, 437 (2016) (More recently, Judge Gertner stated that she wished she had excluded the ballistics testimony in *Green* altogether, rather than merely limiting it).